KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Politechnika Wrocławska

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

ROZWÓJ POTENCJAŁU I OFERTY DYDAKTYCZNEJ POLITECHNIKI WROCŁAWSKIEJ

Wrocław University of Technology

Business Information Systems

Jacek W. Mercik

BUSINESS STATISTICS
Six Lectures on Statistics

Wrocław 2011

Projekt współfinansowany ze środków Unii Europejskiej w ramach
Europejskiego Funduszu Społecznego

Wrocław University of Technology

Business Information Systems

Jacek W. Mercik

BUSINESS STATISTICS
Six Lectures on Statistics

Wrocław 2011

Reviewer: Tadeusz Galanc

ISBN 978-83-62098-33-0

Published by PRINTPAP Łódź, www.printpap.pl

Introduction

1. Average (mean) values, variability and the distribution of elements in a sample

2. Random variables

3. The normal distribution

4. Distributions of statistics.

5. Confidence intervals

6. Parametric tests

Introduction

General description

Year of Study	: 1
Number of Credits	: 3 credits
Duration in Weeks	: 8 weeks
Contact Hours Per Week	: Lecture and Computer Laboratory (2 hours)
Pre-requisite Course(s)	: Probability theory

Course Aims

The course aims to:

1. provide a quantitative foundation in statistical analysis for business;
2. equip students with knowledge in various statistical techniques applicable to business problems;
3. enable students to interpret analytical and statistical results; and
4. give students an overall appreciation of the role and benefit of computers in statistical analysis.

Course Objectives

Upon completion of this course, the students should be able to:

1. understand the techniques of selecting, collecting and organizing data;
2. understand and apply the techniques of summarizing, analyzing and presenting statistical data in a business environment;
3. understand statistical measures and inference and apply them to business problems;
4. interpret quantitative and statistical analysis.

Course Outline

1. <u>Average (mean) values, variability and the distribution of elements in a sample n of data</u>

    2.1 Frequency distributions

    2.2 Cumulative frequency distributions

    2.3 Graphical means of presentation

    2.4 Exploratory data analysis

2. <u>Measures of central tendency and dispersion</u>

    3.1 Mean

    3.2 Median

    3.3 Mode

    3.4 Range

    3.5 Variance and standard deviation

    3.6 Coefficient of variation

3. <u>Discrete and continuous probability distributions</u>

    5.1 Binomial probability distributions

    5.2 Poisson probability distributions

    5.3 Normal probability distribution

    5.4 Normal approximation of the binomial and Poisson distributions

4. Estimation

6.1 Distribution of sample means

6.2 Estimators

6.3 Confidence intervals for means

6.4 Confidence intervals for proportions

5. Tests of hypotheses

7.1 Establishing hypotheses

7.2 Hypotheses regarding one or two means

7.3 Hypotheses regarding one or two proportions

Teaching Approach

The course will be taught via lectures and computer laboratories. Students will be introduced to realistic problems and various quantitative techniques used. Equal emphasis will be placed on the calculation of quantitative measures and the interpretation of results. The computer laboratories will be used to further develop and explore quantitative methods for problem solving.

Assessment

Final Examination	100%

Resources

Principal Reading:
Aczel, Sounderpandian, Complete Business Statistics, 7th edition, McGraw Hill, 2009.

Supplementary Reading:

1. Levine, Statistics for Managers Using Excel and Student CD Package, 5th edition, 2008, Prentice Hall. (ISBN-10: 0136149901)
2. Aczel, Complete Business Statistics, 6th edition, McGraw Hill, 2006.

Computer packages:

1) MS Excel

2) Open Office: http://download.openoffice.org/

3) Statistical macros:

http://sourceforge.net/projects/ooomacros/files/OOo%20Statistics/

4) SPSS: the version for students can be obtained from room 317, build. B-1.

The lectures presented in this script are intended to enrich students' statistical knowledge. Still, it is advisable to begin studying the problem from Aczel's book and then to return to the appropriate lecture from the script.

The lecture notes presented here are designed to go with the educational materials for the statistics laboratory. In this regard, the notes for these two classes constitute the basic material for the laboratory, but it is still advisable to return to Aczel's book. Combined reading of all these materials should result in a good knowledge of statistics.

Lecture I: Average (mean) values, variability and the distribution of elements in a sample

The basic question when analyzing a sample is the question regarding what values could be used to characterise the entire sample in a synthetic way. For example, we know that a tram should arrive at a certain stop at 15 minute intervals and our requirement for regularity should not be incompatible (it seems) with the actual observations of the arrival rate. For example, we observed the following times between arrivals (in minutes): 13, 17, 16, 16, 14, and 14. By saying that the tram runs every 15 minutes, we mean that this is the arithmetic average (mean) of these observations, i.e.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ ,}$$

where x_i denotes the *i-th* observation, and $\bar{x}$ denotes the arithmetic mean of these observations.

Let's characterize the properties of such a sample:

1) $n\bar{x} = \sum_{i=1}^{n} x_i$, that is, knowing the average value and the number of observations in a sample we can determine the sum of the observed values.

In the example of the trams, we have 6 observations; the sum is 13 +17 +16 +16 +14 +14 = 90. The average value (also called the expected value for the sample) is $\bar{x} = 15$. Thus, 6 times 15 is equal to 90. This result is obvious, but how often do you use it without thinking about its origin? For example, if the average return on one share is PLN 10, you expect to get a PLN 1000 profit from 100 shares. The key to the correctness of this reasoning is the notion of average value, which, as we show next is not, however, always a unique concept.

2) $x_{\min} \le \overline{x} \le x_{\max}$. This is another characteristic of the average of the sample, which seems pretty obvious. *min (13, 17, 16, 16, 14, 14) = 13*, *max (13, 17, 16, 16, 14, 14) = 17.* So, the average value must satisfy $13 < \overline{x} < 17$, but none of the observed values are equal to it. Therefore, let us remember: the mean value may never be observed!

3) $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, i.e. the average value is “central" with respect to all observations and this "centring" is the best possible.

It is easy to show that the functional $\sum_{i=1}^{n}(x_i - a)^2$ attains its minimum value for $a^* = \overline{x}$:

$$\frac{\partial\left(\sum_{i=1}^{n}\left(x_i - a^*\right)^2\right)}{\partial a} = 0$$

$$\frac{\partial\left(\sum_{i=1}^{n}\left(x_i^2 - 2x_i a^* + a^{*2}\right)\right)}{\partial a} = 0$$

$$-2\sum_{i=1}^{n} x_i + 2na^* = 0$$

$$a^* = \frac{1}{n}\sum_{i+1}^{n} x_i = \overline{x}$$

4) The mean of a representative sample is a good approximation of the average in the population.

Note that this property is often used to justify the following practical reasoning: because we have a representative sample with a given average value, then we believe that the average in the population is similar. Underlying this reasoning is the so called “law of large numbers”, which ensures that we do not commit an error by using such an argument. Here, we refer the inquisitive reader to the recommended literature.

5) The arithmetic mean is sensitive to extreme values.

Suppose in our example of the trams that one of these trams came almost immediately after the previous tram had left the stop (the phenomenon of tram convoys, which sometimes occurs to the annoyance of passengers,). So our data will look like this:

0, 17, 16, 16, 14, 14,

the mean value for this sample is $\bar{x} = 12,83$, a decrease of 15%. This is a very radical change!

From the above properties, we could imagine that one way to estimate the expected value in the population would be to use:

$$m_1^* = \frac{x_{\min} + x_{\max}}{2},$$

where m_1^* denotes the estimated expected value for the population. We can see immediately that this expression is very sensitive to the extreme values (minimum and maximum). Hence, such estimates are very untrustworthy. One can show that by using the following formula we obtain a significantly better estimate:

$$m_2^* = \frac{x_{\min-1} + x_{\max-1}}{2},$$

where m_2^* denotes the estimated expected value in the population, $x_{\min-1}$ and $x_{\max-1}$, denote the 2nd smallest and 2nd largest observations, respectively, i.e. the new extreme values after the minimum and maximum values have been deleted.

In the example of the trams, the expected value for the sample was $\bar{x} = 15$. Deletion of the extremes means that the sample looks as follows: *16, 16, 14, 14*. Its expected value (and thus the estimate of the expected value in the population) is still 15, although we have achieved this result with less effort. Unfortunately, we do not always get such good results. Hence, the recommendation is to use such an estimate only when we want a fast, easy way to obtain results. Obviously, the larger the sample, the better the estimate obtained.

Let us return to our considerations related to the expected value (average) from the sample. Does it always work? Consider the following example:
Example.

The work of three workers was observed for eight hours:
Worker A: took 2 minutes to make a single component, worker B - 3 min, and worker C - 6 min. What is the average time needed to make a single item?
Calculating the arithmetic mean, we obtain $\bar{x} = \frac{2+3+6}{3} = 3\frac{2}{3}$. Thus using property #1 of the arithmetic mean, we obtain that in one shift (480 min.) a "statistical" workman should produce $480 : \frac{11}{3} = 130,91$ items, and three workers respectively 392.73 elements. However, how many components were produced in reality?
Worker # 1 produced 240 items (480 / 2 = 240).
Worker # 2 produced 160 items (480 / 3 = 160).
Worker # 3 produced 80 items (480 / 6 = 80).
The three workers thus produced 480 elements and not, as obtained (using the arithmetic mean) 392.73 elements. We should infer from this that the arithmetic mean is not suitable for estimating the mean of a set of rates (such as labour productivity in units per hour, effort in minutes per task, speed in kilometres per hour, etc.). Here we use a different measure, namely the harmonic mean:

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Let’s conduct our estimates on the production of these three workers again, but now using the harmonic mean,

$$\overline{x}_H = \frac{3}{\frac{1}{2}+\frac{1}{3}+\frac{1}{6}} = 3\,.$$

Thus, a worker produces an average of 160 elements per shift (480 / 3 = 160) and three workers 480 components, respectively. Note that this result is in line with their actual production.

Example.

The population of a town in three consecutive periods is 5,000, 6500, and 7450 people, respectively. What is the average relative increase in population?

We calculate the relative population increase year to year:

$$x_1 = \frac{6500}{5000} = 1{,}3 \qquad x_2 = \frac{7450}{6500} = 1{,}147$$

Thus the average population growth is:

$$\overline{x} = \frac{1{,}3 + 1{,}147}{2} = 1{,}2230$$

Let's see whether this result is correct? Assume that in both years the population grew at the average rate of 1.2230, i.e. from 5000 to 6115 in the first year and from 6115 to 1.223 times 6115=7479 in the second year. We see that even taking into account the effect of approximation, this result does not give the actual size of the population, which is 7,450 people.

Note that if instead of the average (arithmetic) rate of population growth, we use the so called geometric mean:

$$\overline{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}\ ,$$

this result is almost exact.

We obtain the geometric mean $\bar{x}_G = \sqrt{1,3 \cdot 1,47} = 1,2206$. Assume that the population grew at the geometric average rate of 1.2206 per year, i.e. in the first year from 5000 to 6103 and in the second year from 6103 to 1.223 times 6103 = 7450. This is the correct result.
We see here that the geometric mean is suitable, for example, to study the rate of population growth.

At the end of our discussion of the expected values for a sample, let's ask whether there may be two different samples with the same mean value. The obvious answer is that this is possible. For example, consider the following two trivial two-element samples: *A = (2, -2)* and *B = (1000000, -1000000)*. The average value for both samples is the same and equals 0.

Variability and Measures of variability.

One of the basic research questions regarding a sample of data is a study of its variability. Variability in data is so important that it can even be stated in general that if there is no variation in the surveyed population, it usually cannot be stated with certainty that we know this population on the basis of observed data.
Of course, if we have a large number of identical observations, we are almost sure (statisticians say that "with 95% certainty" or greater) that the test population is made up of such observations, regardless of whether our knowledge is stored in the form of numbers (measurable characteristics) or as a categorisation. For example, if we pick a crop on a dark night and we only get strawberries, in a way we can say, almost with certainty (i.e. at least 95% certainty), that this is a field of strawberries. Especially if we have already have 1000 strawberries! However, intuitively we feel that our confidence is significantly reduced if, for example, we have only ten strawberries (because it might be that one farmer planted a bed of strawberries on the edge of a corn field, or field of cauliflowers), not to mention

that when we have 10 strawberries we might find, for example, a cauliflower or an ear of maize. Let us not forget that no prior knowledge is given to us in this case (as usual in practice), i.e. we cannot see the field! We must recognize this field before harvesting the fruit.

Let's try to think about what characterises the variability of a sample. Here are examples of three data sets:

1) (apple, apple, ... , apple) (in other words, all apples)

2) (apple, apple, ..., apple, pear) (apples and a pear), and

3) (apple, pear, pear, apple, ..., pear) (a mixture of apples and pears).

Of course, these examples can also be presented in the form of sets of numbers, as is usually done in textbooks on statistics:

1) (1, 1, ..., 1),

2) (1, 1, ..., 1, 0),

3) (1, 0, 0, 1, ..., 0).

If we think of variability, it's pretty intuitive that we can say that the first sample is a set of constant (fixed) values, while the second set is much less volatile than the third set. Therefore, any way of measuring the variability of a set should be consistent with our intuition and "able to" distinguish between these sets in terms of volatility.

The most basic, easy to calculate, measure of the variability of any set of numbers is the range, R, defined as follows:

$$R = x_{\max} - x_{\min} \ ,$$

where $x_{\max}$ denotes the maximum value in a set of numbers, and $x_{\min}$ denotes the minimum value in this set.

Note that, in terms of variation, this measure of variability sufficiently distinguishes the first of the three sets above from the others, but is not sensitive enough to distinguish the second set from the third. The range for the first set is 0 (and this is true for all invariant sets) and 1-0=1 for the second and third sets (in general, $R > 0$ for any set in which there is at least one element that differs from the others).

Thus the range is a measure of variability that is capable of detecting samples in which there is no variation, but cannot often distinguish between the intensity of variation when all observations are within a given range. The information that we obtain from the use of the range of the sample as a measure of variability is purely the knowledge of the difference between the extremes. Even so, we should start any study the variability from this measure.

A measure which allows for a more meaningful description of the intensity of variability is the sample variance, *s*^{2}. It is defined as follows:

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

where n is the number of observations (sample size), x_i the *i-th* element of this sample and $\bar{x}$ the arithmetic average of these numbers: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Equivalently, we can determine the sample variance using:

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

Again, note that the sample variance is zero when all the observations are of equal value and that for any set of numbers in which at least one element is different from the others, the sample variance is positive. So, the sample variance can distinguish the first set of the three considered above from the other two, as does the range. Moreover, the sample variance is capable of distinguishing the intensity of variability, which cannot be known just from the range of a sample. Consider the following example of two sets *A = (0, 0, 0, 1)*, and *B = (1, 0,*

1, 0). You will notice that the set *B* is characterized by a greater variability than the set *A*.
The range for both sets is 1, so the range does not distinguish between these two sets in terms of variability.

The variance of set *A* is 0.25. The corresponding variance for the set *B* is 0.33 (we used the definition with n-1 in the denominator to calculate these – it is so called unbiased variance). Thus the set B is more volatile than the set A (according to our intuitive understanding of variability).
One of the problems with using the sample variance as a measure of variability is the unit in which it is measured. It can be seen from the definition of the sample variance that it is the average of the squared distance from the arithmetic mean. Hence, if the units of our measurements are e.g. miles, years, etc., then the units of the sample variance are the square of these units i.e. miles^2 years^2 , respectively). This distorts our understanding of the ratio between such measures. Therefore, in practice, we most frequently use the sample standard deviation, which is the square root of the variance:

$$\sigma = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \quad .$$

The sample standard deviation (also sometimes denoted as *s*) has the same characteristics as the variance, but it is much easier to interpret the results. For example, for the above sets *A* and *B*, we obtain, respectively: $\sigma_A = 0{,}5$ and $\sigma_B = 0{,}58$. The set *B* is therefore characterized by greater variability and the resulting ratio between these standard deviations probably adequately reflects the difference between the variability of both sets.
Consider another example, which will allow us even greater precision in measuring variability. Define the following sets: *C = (0, 0, 0, 1)* and *D = (99, 99, 99, 100).* These samples are characterized by the same variance (one element differs

from the rest by 1; $\sigma_C = \sigma_D = 0{,}5$), but we clearly see that the "consequences" of this variability are much smaller for the set *D* than the "consequences" for the set *C*: a difference of one is not so important if the point of reference is 99 or 100 when compared to the same difference if the reference point is 0 or 1. Therefore, another measure of variability takes into account this aspect. This is the coefficient of variation, *V*, (usually given as a percentage):

$$V = \frac{\sigma}{\overline{x}} \ .$$

For the sets *C* and *D*, it is respectively: $V_C = 200\%$ and $V_D = 0{,}5\%$. This is the correct measure of the variability of these samples.

Now consider an experiment involving the analysis of 9 samples of size 12 (see Table 1)

Table 1 Descriptive parameters for some experiments.

	Data in individual experiments								
	I	II	III	IV	V	VI	VII	VIII	IX
	1	2	0.99	0.98	0.97	0.96	0.92	0.88	0.75
	1	1	1.01	1.02	1.03	1.04	1.08	1.12	1.25
	1	1	0.99	0.98	0.97	0.96	0.92	0.88	0.75
	1	1	1.01	1.02	1.03	1.04	1.08	1.12	1.25
	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1
	1	1	0.99	0.98	0.97	0.96	0.92	0.88	0.75
	1	1	1.01	1.02	1.03	1.04	1.08	1.12	1.25
	1	1	0.99	0.98	0.97	0.96	0.92	0.88	0.75
	1	1	1.01	1.02	1.03	1.04	1.08	1.12	1.25
Descriptive parameters									
Minimum	1	1	0.99	0.98	0.97	0.96	0.92	0.88	0.75
Maximum	1	2	1.01	1.02	1.03	1.04	1.08	1.12	1.25
Range	0	1	0.02	0.04	0.06	0.08	0.16	0.24	0.5
Mean value	1	1.083333	1	1	1	1	1	1	1
Sample variance	0	0.083333	7.27E-05	0.000291	0.000655	0.001164	0.004655	0.010473	0.045455
Standard deviation	0	0.288675	0.008528	0.017056	0.025584	0.034112	0.068224	0.102336	0.213201
Variability coefficient	0.00%	26.65%	0.85%	1.71%	2.56%	3.41%	6.82%	10.23%	21.32%

In the first experiment, we observed the value 1 on each of 12 occasions. What can we say about the phenomenon, which has just been observed in the form of twelve ones? Actually, nothing meaningful, except the assumption that this is a constant value. However, without additional information, we cannot assess the certainty of this statement. Of course, if we had some additional knowledge, our situation would be quite different. For example, suppose we know that these observations represent the fact that a sample of twelve randomly drawn school pupils was composed entirely of girls (a frequently used coding scheme: 1 represents a girl, a boy could be represented by 2). General knowledge tells us that the proportions of boys and girls among pupils should be more or less equal. Hence the probability that we get twelve ones in a row is 1 in 2^{12}, or approximately 0.000244. Because this is an extremely unlikely event, our conclusion (and this is the domain of statistical inference) in this example should read: we chose pupils from a girls' school. This conclusion is almost certain. However, the importance of additional information should be stressed (the proportions of boys and girls in the population of children), without which this inference is not possible.

In the second experiment, we observed 11 ones and one two. In the above example, the convention would mean that we have found 11 girls and one boy among the 12 randomly chosen school. Could we still say that we are almost certainly in a girls’ school? Using statistical inference, we could probably continue to argue for this case (although with less certainty), but, let us stress again that we cannot use any additional information to resolve this (assuming, of course, that the boy was there by chance, visiting his sister, or the "2" recorded is the result of an error in the coding - because after all, this happens from time to time). Note that such consideration is highly specific in that the observation of just one male student means that it is not a school that is 100% for girls. Another thing, on the basis of these data, can we conclude that the population of students consists of 10% boys and 90% girls? Clearly the sample size for making such a statement is too small!

Experiment 2 differs from experiment 1 in that variability is observed in experiment 2 and, in an intuitive way, we can conclude that the observations in experiment II are more volatile than in experiment I. But is the observed variability so large that it already allows us to accept the hypothesis that there are two genders of pupils? Or put another way, is the cause of the observed variation the "presence" of two genders of students? This is a problem that we will return to in the analysis of volatility.

In experiments III - IX, we observe more and more volatility, so we can track how the values of the measures of variability change . It's not surprising that these values increase with the volatility of the data. Some surprise may be caused by the fact that the introduction of just one observation different from the others (experiment II) can create such a strong distortion that only the clear introduction of variability in experiment IX is comparable, even though the coefficient of variation *V* in experiment IX, is approximately 5 percentage points less than the value of this coefficient in experiment II.

In practice, it is considered that data for which the coefficient of variation *V* is less than 10% do not lend themselves to the search for the causal factors of such variability. This is a very important piece of information because, according to it, data sets for which the coefficient of variation is less than 10% are considered to be quasi-constant, and therefore not useful in giving the relevant information. Once again, this is to point out that data without volatility are undesirable, as they do not furnish additional information.

Consider another measure, the coefficient of variation *V*. It is not dependent on the order of the observations (this also applies to other measures of volatility, i.e. the sample range, variance and standard deviation). Therefore, its value can be inferred from the characteristics of the ordered data. In addition, the coefficient of variation reacts more strongly to deviations among the data in a small sample than in a large sample. Consider the data in the following example:

	Data strings of different lengths			
	4	8	12	16
	1	1	1	1
	0	0	0	0
	1	1	1	1
	0	0	0	0
		1	1	1
		0	0	0
		1	1	1
		0	0	0
			1	1
			0	0
			1	1
			0	0
				1
				0
				1
				0
V=	115,47%	106,90%	104,45%	103,28%

The above-mentioned measure of variability will be the basis for further discussion in which we analyze the causes of the variability observed.

Let us return to the problem of analysing the average of the sample. Recall the example regarding the existence of two samples with the same mean value. Consider the samples: *A = (2, -2)* and *B = (1000000, -1000000).* The average value is the same for both samples and equal to 0. We already know that what differentiates between these samples is their variability: the standard deviation for sample A is $\sigma_A = 2,8284$, while for sample B the standard deviation is

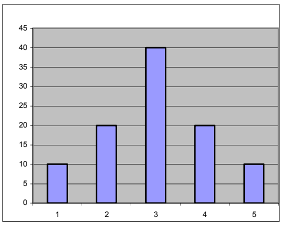
$\sigma_B = 1414214$, i.e. 500,000 times larger. Maybe samples that have the same mean value and the same standard deviation (equal variability) are identical? Unfortunately, the answer is to this question is also no. We will show this in the following example.

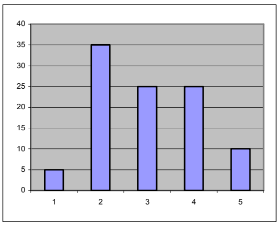
Example.

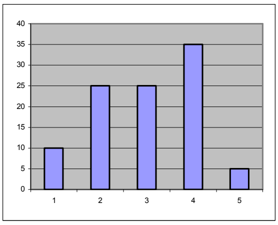
Consider a study of the hourly wage rates in three different companies, For simplicity, assume that they employ the same number of employees: 100 people.

	Percentage of employees		
Hourly wages	Comp. I	Comp. II	Comp. III
10-20	10	5	10
20-30	20	35	25
30-40	40	25	25
40-50	20	25	35
50-60	10	10	5
Total:	100	100	100
Average:	35	35	35
Variance:	120	120	120

So we have three 100-element samples, which have the same average value (35) and the same variability (120). But these are different samples. The diversity of these samples can be seen even better when we draw their histograms.







Thus, in addition to the average (expected) value and the variability, we should also consider the symmetry of the distribution of a sample. The histogram for company I (left chart) is symmetric. The histogram for company II (middle chart) is right skewed. The histogram for company III (right chart) is left skewed. It remains for us to find a way of determining the type of asymmetry (skewness) and "distinguishing" it from symmetry.

Positional characteristics.

It has not yet been mentioned but we should always order a sample. This involves ranking items according to their values. This can be done in two ways: from the largest (maximum) to the smallest (minimum) element, or vice versa. Note that such an arrangement itself provides a wealth of information about the sample: the observed value of the maximum, minimum, range (including the position in which the average value is found), etc.

The primary way of assessing the distribution of the elements in the ordered sample between the minimum and maximum value is to provide the so-called percentiles, i.e. according to position in the ordered series, which "divides" the ordering according to the given proportion. So, the 50-th percentile of the sample (also called the median) is the value below which 50% of the ordered sample lies (the median, depending on the exact definition of percentile, is defined as one of the following two divisions: 50% of observations are below or above the median). For example, consider the following definition of the 100 λ^{th} -

percentile: the $\left[n\lambda\right]+1$*-th* observation in the ordered list, where *n* is the sample size and *[a]* denotes the integer part of *a*, we find that the median (denoted *Me*), for the tram example (after ordering the arrival times of the trams are 13, 14, 14, 16 , 16, 17) is the value of the item in position $\left[6\cdot\frac{1}{2}\right]+1=4$, i.e. *Me* = 16

Note that the question regarding the top 20% of the waiting times for the tram on the basis of the sample, is the question regarding the 80$^{\text{th}}$ percentile (80% of the observations are to the left and 20% of the observations are to the right of it), which we find from the value of the element located in position $\left[6\cdot 0,8\right]+1=5$. This value is 16, or in other words 20% of the observations concerning the arrival times of trams are not less than 16 minutes.

Another characterisation of the sample is given by the modal value, which is the value which appears most often in the sample. Note that the modal value is not necessarily unique: in the sample of the arrival times of trams there are two observations of both 14 and 16. Depending on the definition used, it is assumed that in this case there is no modal value for the sample in question, or we take the smallest of the possible modal values (for example, in calculations using Excel).

Therefore, let us return to the example of the structure of the hourly wages of the three companies. We have found that all these samples have the same average value (35) and the same variance (120). What differs, however, are the median values: *Me* (I) = 35, *Me* (II) = 34, and *Me* (III) = 36. We also see that the corresponding modal values are (usually denoted as *Mo*): 35, 27.5 and 42.5, respectively.

Knowing the median, modal and average values enables us to resolve the problem regarding the symmetry of the distribution of the sample. Hence,

- For symmetrical distributions: $\bar{x}=Me=Mo$,
- For right skewed distributions: $\bar{x}>Me>Mo$, and

- For left skewed distributions $\overline{x} < Me < Mo$.

We obtain the following relevant indicators (measures) of asymmetry:

- Index of skewness: $\overline{x} - Mo$,
- Standardized skewness ratio: $A_S = \dfrac{\overline{x} - Mo}{s}$ and
- Coefficient of asymmetry $A_{as} = \dfrac{m_3}{s^2} = \dfrac{\frac{1}{n}\sum\left(x_i - \overline{x}\right)^3}{s^2}$.

At the end of our discussion on the structure of a sample, we present the so called "3 sigma" rule. This says that for sufficiently large samples (generally assumed to be greater than 30 observations) over 90% of the surveyed population (not only the observations!) is in the range $\left[\overline{x} - 3s,\ \overline{x} + 3s\right]$. The "three sigma" rule comes directly from Chebyshev's[1] inequality (for its own sake, please read the recommended literature) and has a very practical significance. For example, by analyzing the arrival times of trams, we find that the average value is $\overline{x} = 15$ and the standard deviation (sigma) $s = 1,55$. Assuming for simplicity that the test sample is sufficiently large, we find from the "three sigma" rule that 90% of the average waiting times are found in the range $\left[10,35;\ 19,65\right]$. This result applies to the whole population, and therefore to the arrival times of trams other than those in our sample. Colloquially speaking, 90% of the time we do not have to wait for a tram for more than 20 minutes.

[1] One can find a brief description of the life and the work of Chebyshev at http://en.wikipedia.org/wiki/Pafnuty_Chebyshev

Lecture II. Random variables.

I hope you have all seen the beginning of a match when the referee tosses a coin to decide which team will start on which half of the pitch. We assume that the coin is an honest and fair judge. By the fairness of the coin, we mean that its two sides have the same chance of falling in a single toss. Let's try to present the situation in a more formal manner. Now, let P (.) denote the probability of the event given in brackets. We thus assume that:

$$P(\text{head})=P(\text{tail})= ½$$

Because there are many such scenarios that follow a similar rule, it will be more convenient to talk about success and failure, and the context will decide what we call a success and what we call a failure. So, if throwing a head is a success (in a football match this could mean playing with the sun behind you) our assumption is that:

$$P(\text{failure})=P(\text{success})= ½$$

Note that if the chance of failure is two times greater than the chance of success, then the above statement would look different:

$$P(\text{success})=1/3,\ P(\text{failure})=2/3,$$

and this clearly does not describe a coin toss (where we expect failure and success with an equal chance).

If now we code failure and success using numerical values, we obtain a mathematical expression, which we call a random variable, usually denoted by a capital letter at the end of the alphabet, such as X. Encoding success and failure in the following manner: 1 if a success occurred, 0 if a failure occurred, we obtain the so called coin pattern:

$$P(X=1)=P(X=0)=\frac{1}{2},$$

or more generally:

$$P(X=1)=p;\;\; P(X=0)=1-p,\;\; \boldsymbol{for}\ 0<p<1$$

Obviously, coding the values of success and failure using the numbers 0 and 1 is totally arbitrary, although, as we will show later, convenient. But we could equally well attribute a value of 1,000,000 to success and -3.14 to failure. In general, such a scheme, therefore, can be defined as follows:

$$P(X=x_1)=p;\;\; P(X=x_2)=1-p,\;\; \boldsymbol{for}\ 0<p<1,$$

where x_1 and x_2, are the numerical values (generally these are real numbers) representing success and failure, respectively.

It should be noted that the random variable described above is called a random variable with a two-point distribution. The description given here is only intuitive and it is suggested that the reader look in their notes or any textbook of probability and recall the formal definition of a random variable.
Thus, if an event takes place according to such a two point distribution (sometimes we say according to the coin scheme, although it is not in general a fair coin), we say that the event has a two-point distribution. Note that if a student asked about the chances of passing the statistics examination, an answer given on the basis of a coin toss would give the correct answer according to the fair coin scheme. This clearly does not apply to students of Business Information Systems, since we believe that their probability of success in the statistics exam is significantly higher. We shall return to the concept of significance later.

Let us return to the two-point distribution:

$$P(X=1)=p;\;\; P(X=0)=1-p,\;\; \boldsymbol{for}\ 0<p<1.$$

The convenience of such a definition lies, among other things, in the fact that if we define Y as the sum of 0-1 random variables defined in such a way, then Y is simply the number of successes. For example, if we are dealing with n coin tosses, $Y = \sum_{i=1}^{n} X_i$ will be the sum of the resulting ones and zeros (i.e. just the ones) and hence we count the number of successes (if one denotes a success). For example, throwing a coin 10 times, if tails is coded as 1 (success), the situation in which we never throw tails ($Y = 0$) might occur and other extreme situation occurs when we throw only tails ($Y = 10$). Of course, we correctly think that a mixture of heads and tails is rather more likely than such extreme results, although they are not impossible. So if *k* is the number of successes in n trials, then *k* takes values from 0 to n, which is written as $k = 0,1,...,n$.

It is clear that the sum of random variables is itself a random variable, i.e. Y is a random variable. To complete the definition of Y we need to give the probability of obtaining each possible value of *k*, i.e. we must answer the following question: what is the distribution of this random variable?

Before we give the formula for the distribution defining Y as a random variable, let us turn our attention to another aspect of tossing a coin. Any reasonable reader, based on his or her life experience, will agree that it is not possible, for example, if we know what we have obtained in the first five throws, to state what will fall in the sixth throw: heads or tails (success or failure)? We know that the results of successive throws are independent: every single coin toss is carried out spontaneously and the results already obtained have no impact on future results. In probability theory, such events are called independent. Again, we ask the reader to refresh their knowledge on this subject.

Returning to the variable Y, we see that we are dealing here with the sum of *n* independent random variables each from the same 0-1 distribution. The distribution of the random variable Y is called the Bernoulli distribution (or binomial distribution) and often denoted B(n, p, k) . Its definition is as follows:

$$P(Y=k)=\binom{n}{k}p^k(1-p)^{n-k}\,,\quad 0<p<1,\quad k=0,1,\dots,n\,.$$

In other words, we have answered the question regarding what is the probability of obtaining *k* successes in *n* independent identical trials.

Note that this provides a general pattern for the definition of random variables. You must specify the values that a random variable can take and the likelihood of each of these values. So, generally a distribution of this type, where the random variable *X* can take *m* values $x_1, x_{2,\dots}\, x_m$ (sometimes we say discrete values x_i) each with probability p_i (*m* may be finite or infinite):

$$P(X=x_i)=p_i\,,\quad 0\le p_i\le 1 \text{ and } \sum_{i=1}^{m}p_i\,,\quad i=1,2,\dots$$

Each time we run into some random phenomenon leading to such a variable, in general we can define the random variable using this pattern . Fortunately, such distributions tend to be an example of one of a few standard distributions. In addition to the two-point and Bernoulli distributions (also known as the binomial distribution) defined above, we give a few others below. They describe the majority of distributions occurring in the economic analysis of random patterns.

The negative binomial distribution is useful for determining the probability of the number of trials carried out until the desired number of successes is achieved in a sequence of Bernoulli trials[2]. It counts the number of trials X needed to achieve s successes with p being the probability of success on each trial.

[2] One can find a brief description of the life and work of Jacob Bernoulli at http://en.wikipedia.org/wiki/Jacob_Bernoulli

$$P(X=k)=\binom{k-1}{s-1}p^s(1-p)^{(k-s)}$$

Example:

Suppose that the probability of a manufacturing process producing a defective item is 0.001 , which means that one item per thousand manufactured is defective (life experience tells us that for various reasons, ideal production is not possible). Suppose further that the quality of any one item is independent of the quality of any other item produced. If a quality control officer selects items at random from the production line, what is the probability that the first defective item is the 10th item selected.

Here k = 10, s = 1, and p = 0.001.

For instance: $P(X=10)=\binom{10-1}{1-1}0.001^1(1-0.001)^{(10-1)}=0.000991$.

Thus,

$$P(X\leq 10)=0.008955,$$

which is a relatively small number. This means that we are dealing with an unlikely event, so we either belong to a "lucky" group, which encounters an unlikely event, or our idea of the level of deficiencies does not reflect reality, since the already the tenth item is defective. Common sense dictates that it is likely that the proportion of defectives is higher than the declared one item produced in 1000.

Geometric distribution.

Within the context of a binomial experiment, in which the outcome of each of *n* independent trials can be classified as a success or a failure, a geometric random variable counts the number of trials until the first success

$$P(k) = pq^{k-1} \text{ for } k=1,2,3,\ldots .,$$

where q=1-p

Example.

From data from the Department of Computer Science and Management, Wroclaw University of Technology, it is known that 83% of students come from an urban area. A company interested in developing rural regions has established a scholarship for candidates from rural areas. Suppose that students are chosen at random in sequence. What is the probability that the first student selected is from a rural region? What number of selections should be expected before finding a student who comes from a rural area? Let's calculate the corresponding probabilities:

- Picking a student from a rural area as the first person

$$P(1) = 0.17 \cdot 0.83^{1-1} = 0.17,$$

- Needing two selections to find such a person

$$P(2) = 0.17 \cdot 0{,}83^{2-1} = 0.1411$$

- Needing three selections before finding a student from a rural area

$$P(3) = 0.17 \cdot 0{,}83^{3-1} = 0.1171$$

- The probabilities for four, five, six, seven and eight selections before finding a rural candidate are 0.0972, 0.0806, 0.0669, 0.0555 and 0.0461 respectively.

It is not a coincidence that we carried out these eight calculations. In statistics, it is assumed that we generally compare the probability of extreme events with a set value. Frequently this value is 0.05. Therefore, we must reckon with the fact that we will need to make as many as minimum seven selections or at least the chances of such a course of events is (as statisticians say) significant.

The hypergeometric probability distribution.

The hypergeometric probability distribution is useful for determining the probability of a number of occurrences when sampling without replacement. It counts the number of successes (*k*) in *n* selections, without replacement, from a population of *N* elements, *s* of which are successes and (*N*-*s*) of which are failures.

$$P(k) = \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}}$$

Example (Aczel, 2006):

Suppose that automobiles arrive at a dealership in lots of 10 and that for time and resource considerations, only 5 out of each 10 are inspected for safety. The 5 cars are randomly chosen from the 10 on the lot. If 2 out of the 10 cars on the lot are below standards for safety, what is the probability that at least 1 out of the 5 cars to be inspected will be found to not meet safety standards?

$$P(1) = \frac{\binom{2}{1}\binom{(10-2)}{(5-1)}}{\binom{10}{5}} = \frac{\binom{2}{1}\binom{8}{4}}{\binom{10}{5}} = \frac{\frac{2!}{1!1!}\frac{8!}{4!4!}}{\frac{10!}{5!5!}} = \frac{5}{9} = 0.556$$

$$P(2) = \frac{\binom{2}{1}\binom{(10-2)}{(5-2)}}{\binom{10}{5}} = \frac{\binom{2}{1}\binom{8}{3}}{\binom{10}{5}} = \frac{\frac{2!}{1!1!}\frac{8!}{3!5!}}{\frac{10!}{5!5!}} = \frac{2}{9} = 0.222$$

i.e. the required probability is $P(1)+P(2)=0.556+0.222=0.778$. It is questionable from our (i.e. the car manufacturer's) point of view that such a control system is acceptable, since there is a significantly high likelihood of not finding either defective car.

Poisson Scheme (distribution).

From a formal point of view (the so called Poisson theorem), consider the sequence of random variables $\{X_n\}$, where each *X_i* has a Bernoulli distribution *B(n, p, k*) with the following property: $n \cdot p = const \ \ \boldsymbol{for} \ n \to \infty$. Let $X = \lim X_n$. X is a random variable taking values $k = 0,1,...$. We say that X has a Poisson distribution. To complete the definition of this random variable it is necessary to determine the values taken ($k = 0,1,...$) and their probabilities. The Poisson probability distribution is defined as follows:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$

where

- e is the base of the natural logarithm (e = 2.71828...)

- k is the number of occurrences of an event - the probability of which is given by the function above

- k! is the factorial of k

- λ is a positive real number, equal to the expected number of occurrences of an event that occur during a given interval. For instance, if the event occurs on average 5 times per minute and you are interested in the probability of the event occurring k times in a 10 minute interval, you would use a Poisson distribution with λ = 10×5 = 50 as your model.

Required conditions for the Poisson distribution to be applied:

The probability that an event occurs in a short interval of time or space is proportional to the size of the interval.

The probability that two events will occur in a very small interval is close to zero.

The probability that a given number of events will occur in a given interval is independent of where the interval begins.

The probability of a given number of events occurring over a given interval is independent of the number of events that occurred prior to that interval.

The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume. For example, the number of customers waiting for service in front of a supermarket cash register can be modelled as a random variable with a Poisson distribution with intensity equal to λ.

The random variables examined so far have one thing in common: the realizations of these random variables are natural numbers. Variables of this type are called discrete variables. But clearly we can see that some phenomena cannot be described as natural numbers. For example, all measurable phenomena are realized as positive real numbers (measured in centimetres, seconds, etc.) or even negative real numbers (such as profit and loss). The essential characteristic of real numbers is that they are everywhere dense. This means that between any two real numbers one can always put another real number. Thus there is a problem with the definition of a random variable given above, according to which it is necessary to give the set of values a random variable can take and the likelihood of occurrence for each of these values.

Thus, formally, a discrete variable satisfies the following:

- Has a countable number of possible values,
- Has discrete (discontinuous) jumps between consecutive values
- Has a measurable probability for each possible value.

Commonly met discrete variables are counts (i.e. number of children, number of occurrences of an event).

In contrast to discrete random variables, continuous random variables have different characteristics:

- Such variables are measured and not counted (e.g. height, weight, speed, etc.)
- Take an uncountable (thus infinite) number of possible values,
- The possible values change in a continuous manner,
- No probability is ascribed to any single value
- Probability can only be ascribed to ranges of values.

Continuous random variables are defined by a function *f(x),* called the density function. It has the following characteristics:

1) $f(x) \geq 0$,

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Thus, the probability that some random variable X takes a value between a_1 and a_2 ($a_1 < a_2$) corresponds to the integral of the function *f(x)* over the interval [a_1, a_2], that is

$$P(a_1 < X < a_2) = \int_{a_1}^{a_2} f(x)dx$$

From the properties of integral calculus, we are sure that the value of this expression is non-negative and less than or equal to 1.

Now we consider a random variable with a uniform distribution on the interval [a, b] (Fig. 1).

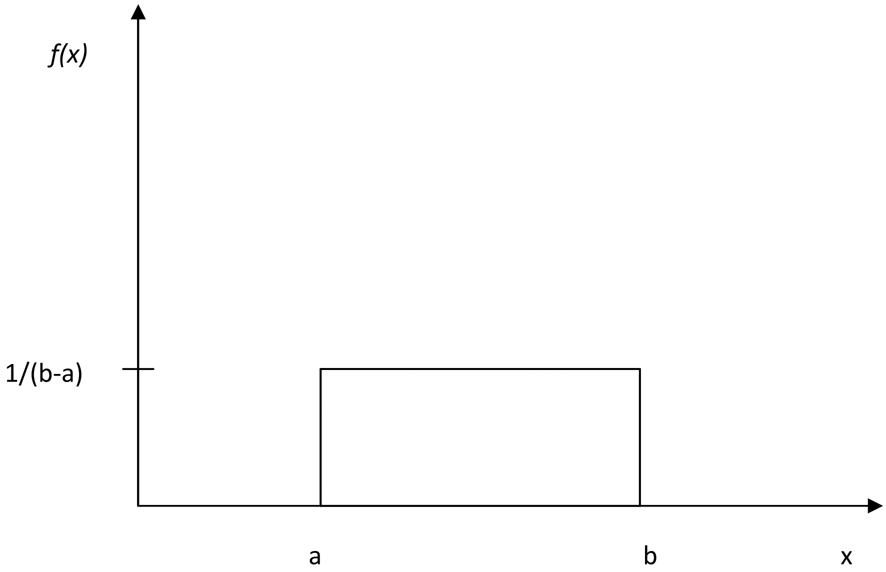


Figure 1 Uniform probability density function on the interval [a, b].

The density function, *f(x),* of the uniform distribution on the interval [a, b], has the following analytical form:

$$f_{[a,b]}(x) = \begin{cases} \dfrac{1}{b-a} & \boldsymbol{for} \ \ x \in [a,b] \\ 0 & \boldsymbol{for} \ \ x \notin [a,b] \end{cases}$$

It is easy to show that this is a density function, i.e., has the properties:

1) $f_{[a,b]}(x) > 0$, because $a < b$, and

2) $\int_a^b f_{[a,b]}(x)dx = \frac{1}{b-a}\int_a^b dx = \frac{1}{b-a}(b-a) = 1$.

Note that if we have two intervals [c, d] and [e, f] such that both are in the range [a, b] and have the same length, then the probability that such a random variable X takes a value from the interval [c, d] is the same as the probability of it taking a value in the interval [e, f]. This is called monotony or uniformity.
Using these two properties of density functions, we can define virtually any number of continuous distributions (as there are any number of functions with these properties). Fortunately, as with discrete distributions, the number of continuous distributions that are useful for modelling economic and social phenomena is not that big. Below we describe some of them.

One-sided exponential distribution.

Suppose the number of occurrences of an event in an interval has a Poisson distribution. The time between two occurrences has an exponential distribution. Its density function is as follows:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \textbf{\textit{for}} \;\; \lambda, x > 0 \\ 0 & \textbf{\textit{for}} \;\; x \le 0 \end{cases}$$

Example

The time for which a particular machine operates before breaking down (time between breakdowns) is known to have an one sided exponential distribution with parameter λ = 2 (where the units of time are hours). What is the probability that the machine will work continuously for at least one hour? We require, therefore, the probability that a variable with a one-sided exponential distribution with parameter λ = 2, takes a value greater than 1, when time is measured in hours.

Thus,

$$P(X \geq 1) = 2\int_{1}^{+\infty} e^{-2x}dx = e^{-2} = 0.1353 .$$

This is not a very high probability. We would say that we have a much greater chance that the machine breaks down within an hour (the probability of this, since it the complement to the phenomenon studied, is 1 - 0.1353 = 0.8647). The key here is the role of the parameter λ. If the parameter value λ was ½, then $P(X \geq 1) = 0.6065$. What, then, does this parameter measure?

Before answering the question about the role played by the parameter λ in the exponential distribution, we return to the concept of the expected value of a sample. The expected value is a numerical characteristic of the entire sample, the value which we would expect “on average”. We also remember that the variance of a sample (or a similar measure) is characteristic of the variability observed in the sample. Given these two values, we can estimate the range of values that might appear in a sample, or in the population as a whole. A probability distribution describes a random phenomenon. The expected value and variance of such a distribution are numeric measures of the population expected value (mean) and population variance. Moreover, if a sample is representative, the expected value and variance of the samples are approximations (estimators) of the expected value and variance in the population. The issue of the parameters of a distribution will return in a moment.

The most important of the continuous distributions is the normal distribution, also called the Gaussian distribution. We will devote a separate lecture to this distribution.

The gamma distribution.

Another important continuous probability distribution is the gamma distribution. Its density function (PDF) is defined as follows:

$$f(x;r,c) = x^{r-1}\frac{e^{-\frac{x}{c}}}{c^{r}\Gamma(r)} \quad \boldsymbol{for}\ \ x \ge 0 \ \ and \ \ r,c > 0$$

Alternatively, the gamma distribution can be parameterized in terms of a shape parameter α = r and an inverse scale parameter β = 1/c, called the rate parameter:

$$g(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \quad \boldsymbol{for}\ \ x \ge 0$$

If *n* is a positive integer, then $\Gamma(n) = (n-1)!$.

Both parameterizations are common, because either can be more convenient depending on the situation.

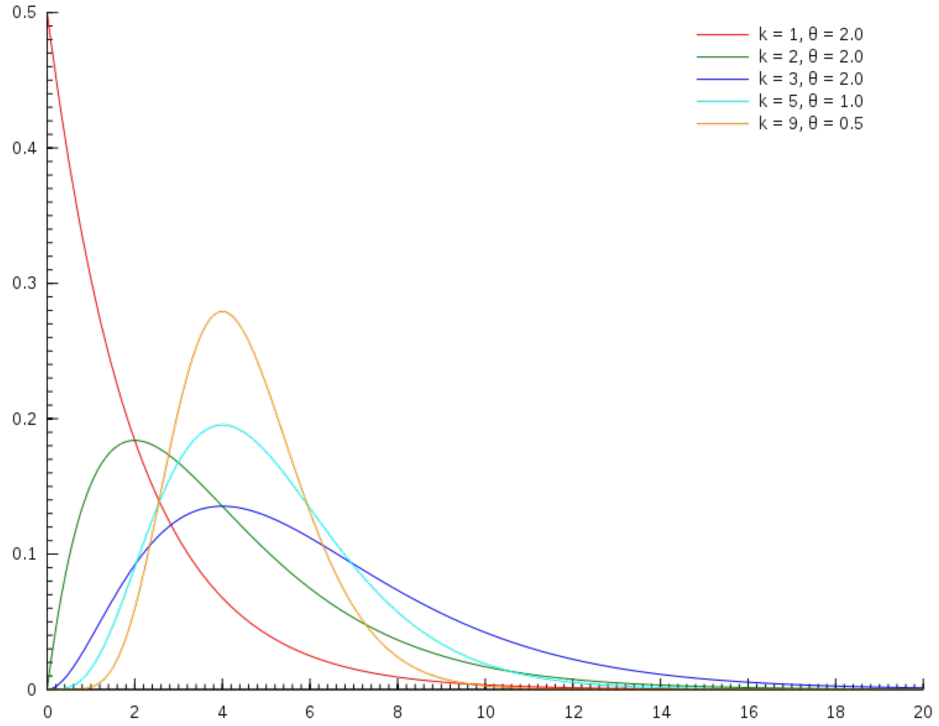


Fig. 2. Illustration of the Gamma PDF for various parameter values *r = k* and $c = \theta$.

It can be shown that the one-sided exponential distribution is a special case of the gamma distribution with parameters .

$$r = 1, \;\; c = \frac{1}{\lambda}$$

Note that the gamma distribution is ideal for modelling the time to carry out tasks, such as the estimation of the duration of activities on the basis of expert assessments as in the PERT method, where the expected value is estimated as $\overline{x} = \dfrac{x_{\min} + 4x_{sp} + x_{\max}}{6}$ and the variance as $s^2 = \left(\dfrac{x_{\max} - x_{\min}}{6}\right)^2$, where *x_{min}, x_{sp}* and *x_{max}* are respectively the optimistic time, most likely time and pessimistic time of a given activity's duration. These estimates are the result of the application of the gamma distribution to describe the duration time of an activity.

Now, we return to the concept of the parameters of a distribution. From a formal point of view, the expected value of a random variable is called the first moment of this variable.

The ordinary moment of order k (*k = 1, 2, ...)* of a random variable is the expected value of the *k*-th power of this variable.

$$m_k = E\left(X^k\right) = \int_{-\infty}^{\infty} x^k dF(x) = \begin{cases} \sum_i x_i^k p_i & \textbf{\textit{(1)}} \\ \int_{-\infty}^{\infty} x^k f(x)dx & \textbf{\textit{(2)}} \end{cases}$$

where: *X* - random variable, *E(X)* - expected value of the random variable *X*, p - the probability function, f - a density function.

Patterns (1) and (2) should be used for a random variable with probability distributions of discrete and continuous type, respectively.

For k = 1, we obtain the formula for the expected value, so the expected value can be treated as the ordinary first moment m_1.

Similarly, the variance can be presented as a special case of a so-called central moment of a random variable. The central moment of order k (k = 1, 2, ...) of the random variable X is the expected value of the function $[X - E(X)]^k$, i.e.:

$$\mu_k = E[X - E(X)]^k = \begin{cases} \sum_i [x_i - E(X)]^k p_i & (1) \\ \int\limits_{-\infty}^{\infty} [x - E(X)]^k f(x)dx & (2) \end{cases}$$

where: X - random variable, *E(X)* - expected value of the random variable *X*, p - the probability function, f - a density function.

Patterns (1) and (2) should be used for a random variable with probability distributions of discrete and continuous type, respectively.

The case k = 2 corresponds to the formula for the variance and, therefore, the second central moment. The third central moment is also common and allows you to measure the asymmetry of a distribution. The fourth central moment is useful in calculating the kurtosis.

Central moments of a random variable can be expressed by means of ordinary moments. The following expression of the second order central moment (variance) is particularly useful: $\mu_2 = m_2 - m_1^2$ The reader is encouraged to check this equality. In fact, ordinary and central moments convey all the subtleties of a probability distribution and allow you to distinguish between distributions that are very slightly different.

In practical applications of an economic and social nature, it can be assumed that the moments of any order and type always exist. But this is not a general trait: there are probability distributions, for which there are no moments. In particular, note that this may lead to distributions for which there is no expected value. One example of such a distribution is the Cauchy distribution (with density function

$f(x)=1/(\pi(1+x^2))$. It is worth raising another aspect of the distribution of moments. Is it true that knowledge of the moments of all orders uniquely specifies a random variable? The general answer is negative, although in the context of our discussion: the practical issues of a socio-economic nature, knowledge of the first four moments is sufficient to specify a random variable. Note that there is a theorem stating that if there is a moment of order *k*, all the moments of order less than *k* exist. Using the "descriptive statistics" menu, Excel can find the expected value, dispersion, and excess kurtosis for a sample which can be used as approximations for the respective values in the population. Calculation of these values requires knowledge of the first four moments. Please refer to the appropriate definitions from the recommended reading.

At the end of this section, for the convenience of the reader we state the most commonly used moments for the aforementioned discrete and continuous distributions.

distribution	Mass probability function or density	Expected value (moment of first order)	Variance (central moment of second order)
Two-point	$P(X=x_1)=p;\ P(X=x_2)=$ $1-p,\ \boldsymbol{for}\ 0<p<1$	$x_1p+x_2(1-p)$	$p(1-p)(x_1-x_2)^2$
Binomial (Bernoulli's)	$P(Y=k)=\binom{n}{k}p^k(1-p)^{n-k}$, $0<p<1,\quad k=0,1,...,n$	np	$np(1-p)$
The negative binomial	$P(X=k)=\binom{k-1}{s-1}p^s(1-p)^{(k-s)}$	$s\dfrac{p}{1-p}$	$s\dfrac{p}{(1-p)^2}$

Geometric	$P(k)=pq^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergeometric	$P(k)=\frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}}$	$\frac{ns}{N}$	$\frac{ns(N-n)(N-s)}{N^2(N-1)}$
Poisson	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ
Uniform [a, b]	$f_{[a,b]}(x)=\begin{cases}\frac{1}{b-a} & \boldsymbol{for}\ \ x\in[a,b]\\ 0 & \boldsymbol{for}\ \ x\notin[a,b]\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	$f(x)=\begin{cases}\lambda e^{-\lambda x} & \boldsymbol{for}\ \ \lambda,x>0\\ 0 & \boldsymbol{for}\ \ x\le 0\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$g(x;\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ $\boldsymbol{for}\ \ x\ge 0$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$

At the end, we introduce yet another particularly useful feature of the moments of random variables. Any random variable *Y* with expected value *m* and standard deviation *σ* may be standardized using the following operation

$$Y=\frac{X-m}{\sigma}$$

This operation transforms random value Y into a random value with expected value 0 and standard deviation 1.

Lecture III. The normal distribution.

The normal distribution is an extremely important probability distribution in many areas. It is also called the Gaussian distribution, particularly in physics and engineering. In fact, this is a family of infinitely many distributions, defined by two parameters: the average (responsible for the location of the distribution) and standard deviation (scale). The standard normal distribution is a normal distribution with mean zero and standard deviation of one. Since the graph of the normal distribution density function resembles a bell, it is often called the bell curve. As already mentioned, the normal distribution belongs to the class of continuous distributions.

History

- the normal distribution was first presented by de Moivre in an article in 1773 (reprinted in the second edition of “The Doctrine of Chance”, 1783) in the context of approximating certain binomial distributions for large n. This result were further developed by Laplace in his book “The Analytical Theory of Probability” (1812) and is now called the de Moivre-Laplace assertion.
- Laplace used the normal distribution in the analysis of errors in experiments. The important method of least squares, used in probability theory, was introduced by Legendre in 1805. Gauss, who claimed that he had used this method since 1794, made great advances in 1809 by assuming errors had a normal distribution.
- The name of the curve comes from the bell curve of Joufrett, who coined the term “surface of a bell” in 1872 for a two-dimensional normal distribution with independent components. The name of the normal distribution was introduced simultaneously by Charles S. Peirce, Francis Galton and Wilhelm Lexis around the year 1875. This terminology is not the best, because it suggests that the vast majority of things have a normal distribution, while recent studies of economic and social phenomena indicate that most phenomena have a rather different distribution to the normal distribution (only about 20% of data sets show the characteristics of normality).

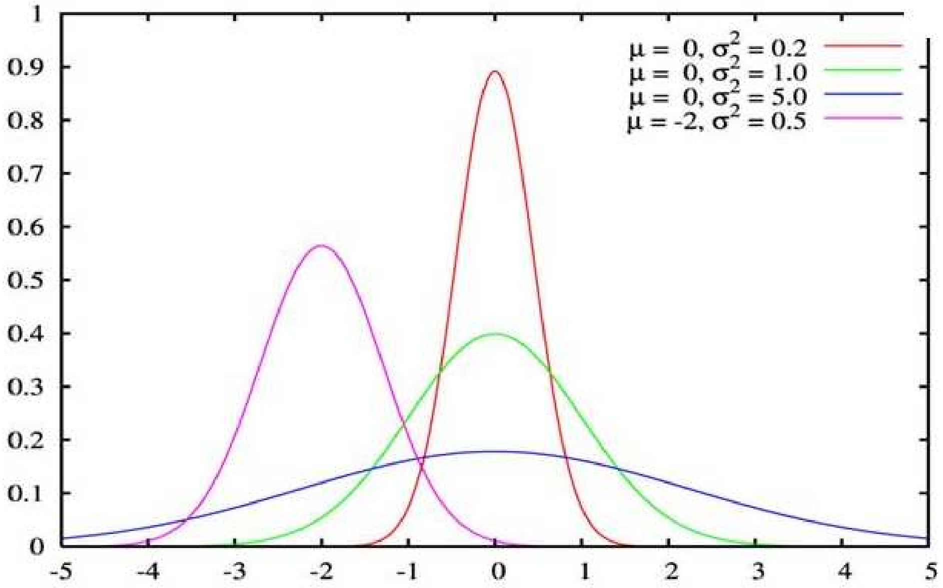
Parameters of a normal distribution:

The expected value	μ
Variance	σ^2
Modal value (mode)	μ
Median	μ
Skewness	0
Kurtosis	0

The density function of the normal distribution with mean μ and standard deviation σ (equivalently: variance σ^2) is an example of a Gaussian function. Typically, the normality of the random variable X is denoted $X \sim N(\mu,\sigma)$ (or sometimes $X \sim N(m,\sigma)$,(when just the values of the parameters are given be careful to note whether the second parameter is the standard deviation or the variance), which means that the random variable X has the following density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

Its shape is shown below.



All normal density functions are symmetric about the mean value of the distribution. About 68% of the area under the graph of the density function is within one standard deviation of from the average, about 95.5% within two standard deviations and about 99.7% within three standard deviations (the three sigma rule). Note that the density function for the normal distribution is greatest around its expected value.
The inflection point is within one standard deviation of the mean.

Cumulative Distribution Function.

We have defined random variables by using the probability mass function (for discrete random variables) or the density function (for continuous random variables.) The cumulative distribution function (cdf) can be used to define both types of random variable.
By definition, the cumulative distribution function (often shortened to distribution function) of a random variable X is a function of the real variable $x \in R$:

$$F_X(x) = P(X < x)$$

A cumulative distribution function F(x) has the following characteristics:

1) $\lim_{x \to -\infty} F(x) = 0,\ \lim_{x \to \infty} F(x) = 1$

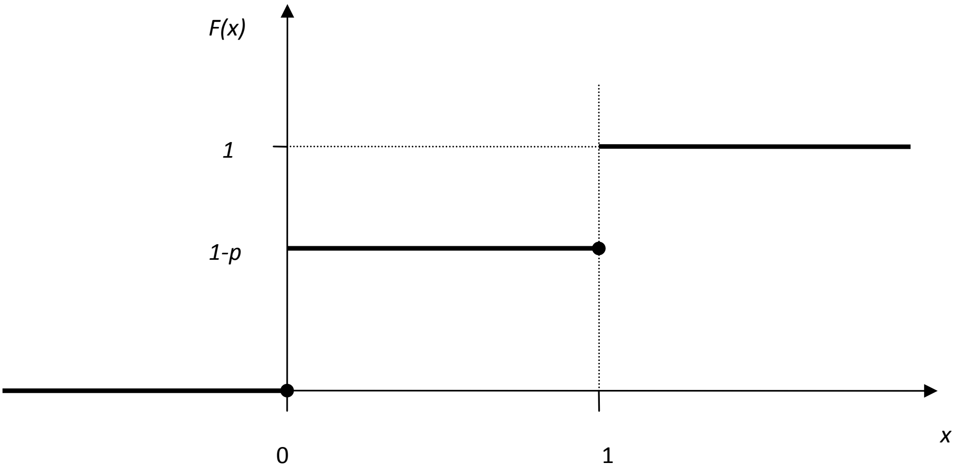
2) The function F(x) is a non-decreasing function, i.e. if $x_1 < x_2$, then $F(x_1) \le F(x_2)$, and

3) F(x) is a left-continuous function (here we refer the reader to the lectures on mathematical analysis)[3].

[3] This property is closely related to the strict inequality used in the definition of the cumulative distribution function. It is worth noting that left-handed continuity turns into right-hand continuity if we change this strict inequality into a non-strict one,, i.e. we use the following definition of the cumulative distribution function $F_X(x) = P(X \le x)$
Such a definition of the distribution function can be found in many textbooks on statistics.

Example.
Let us determine the cumulative distribution function of the two point 0-1 distribution.

$P(X=1)=p;\ \ P(X=0)=1-p,\ \ \boldsymbol{for}\ 0<p<1$. This is illustrated by the following figure.

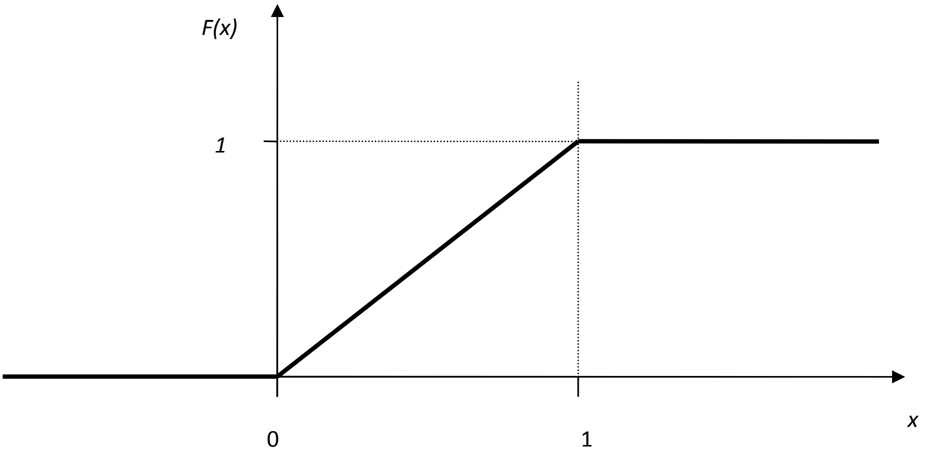


Note that this distribution function has two points of discontinuity, namely the points x = 0 and x = 1.

Example.
Let us determine the cumulative distribution function of the uniform distribution on the interval [0, 1]. It is defined using the following density function

$$f_{[a,b]}(x)=\begin{cases}\dfrac{1}{b-a} & \boldsymbol{for}\ \ x\in[a,b]\\ 0 & \boldsymbol{for}\ \ x\notin[a,b]\end{cases}$$

Therefore, its cumulative distribution function is as follows:

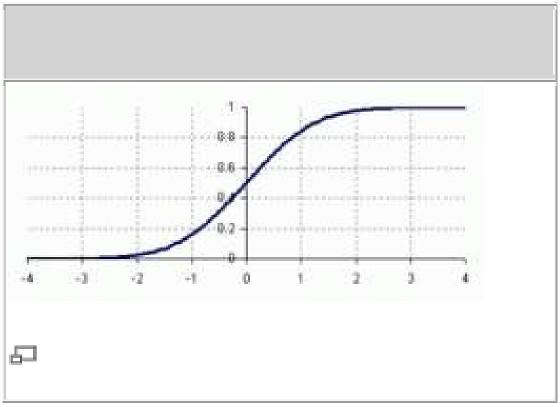


Note that the cumulative distribution function of this distribution has no points of discontinuity. In general, we can say that if a cumulative distribution function is continuous, then we are dealing with a continuous probability distribution.
It is also worth noting that any function which has properties 1) - 3) listed above is the cumulative distribution function of a probability distribution. Thus, if it turns out, for example, that a phenomenon studied by us has its own unique probability distribution, by using a cumulative distribution function we can create the appropriate probability distribution.

The cumulative distribution function of the normal distribution is defined as the probability that the variable X takes a value less than x and in terms of the density function is expressed by the formula

$$F_{N(\mu,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

To obtain the formula for the cumulative distribution function of the standard normal distribution (traditionally denoted by Φ), we simply substitute μ = 0 and σ = 1 into this general formula.



1. If X ~ N (μ, σ^2) , where a and b are real numbers, then aX + b = Y ~ N (aμ + b, (aσ)^2). Thus, a linear transformation of a random variable with a normal distribution gives another random variable with a normal distribution.
2. If X_1 ~ N (μ_1, σ_1^2), X_2 ~ N (μ_2, σ_2^2), and X_1 and X_2 are independent, then

$X_1 + X_2$ ~ N ($\mu_1 + \mu_2$, $\sigma_1^2 + \sigma_2^2$), i.e. the sum of independent normal random variables also has a normal distribution.
3. If X_1, ..., X_n are independent random variables with a standard normal distribution, then $X_1^2 + ... + X_n^2$ has a chi-squared distribution with n degrees of freedom. We will return to this distribution in the future.

Standardizing random variables with a normal distribution

The consequence of property 1 is that all random variables with a normal distribution can be transformed into a random variable with the standard normal distribution.

If X is normally distributed with mean μ and variance σ^2, then:

$$Z = \frac{X - \mu}{\sigma}$$

Z is a random variable with standard normal distribution N(0, 1).

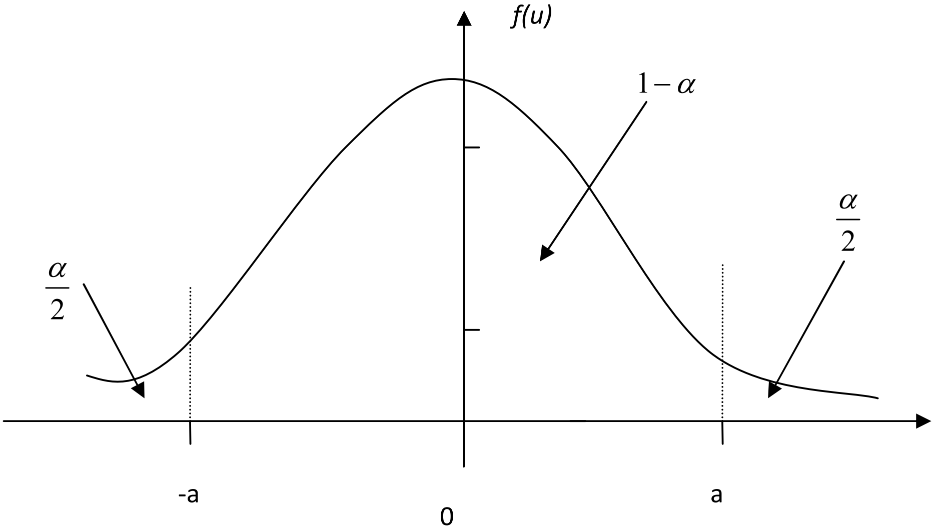
Conversely, if Z is a random variable with standard normal distribution and

$$X = \sigma Z + \mu ,$$

then X is a normally distributed random variable with mean μ and variance σ^2.

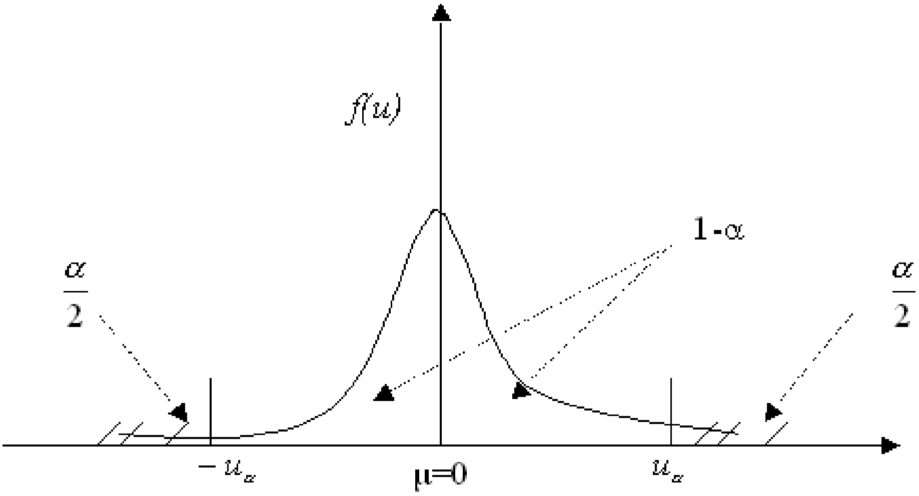
The distribution function of the standard normal distribution has been tabularised and other normal distributions are simple transformations of the standard distribution. In this way, we can use normal distribution tables to determine probabilities corresponding to any normal distribution.

Example.
Let us find the value of a (a> 0), for which a random variable X with normal distribution $X \sim N(0,\sigma)$ takes a value in the interval [-a, a] with probability $1-\alpha,\ 0<\alpha<1$. We are looking, therefore, for the area under the curve of the density function between *–a* and *a* to be equal to 1-α, see below.



Standardization transforms the normal random variable X into a random variable with expected value 0 and standard deviation 1 (still normal). So,

,

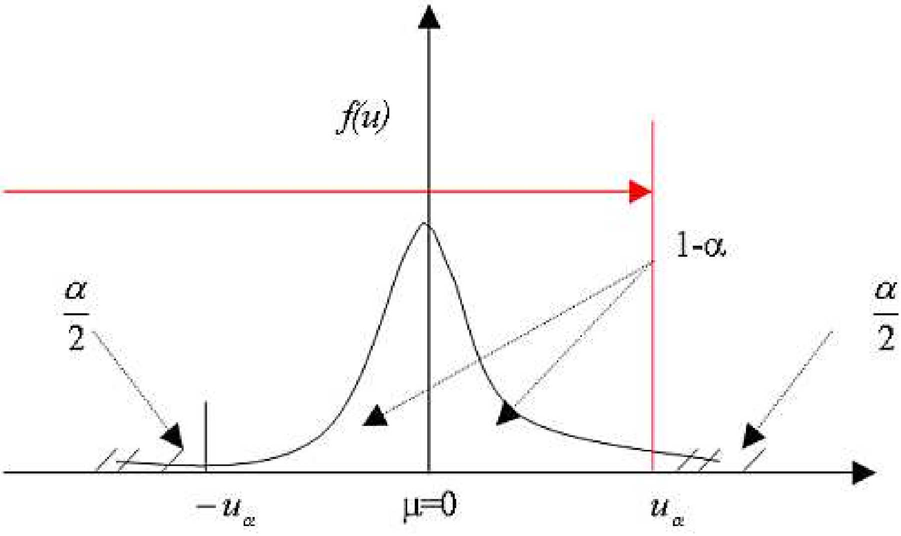


$$P\left(-u_\alpha < \overbrace{\boxed{\frac{X-\mu}{\sigma}}}^{\text{N (0, 1)}} < u_\alpha\right) = 1-\alpha$$

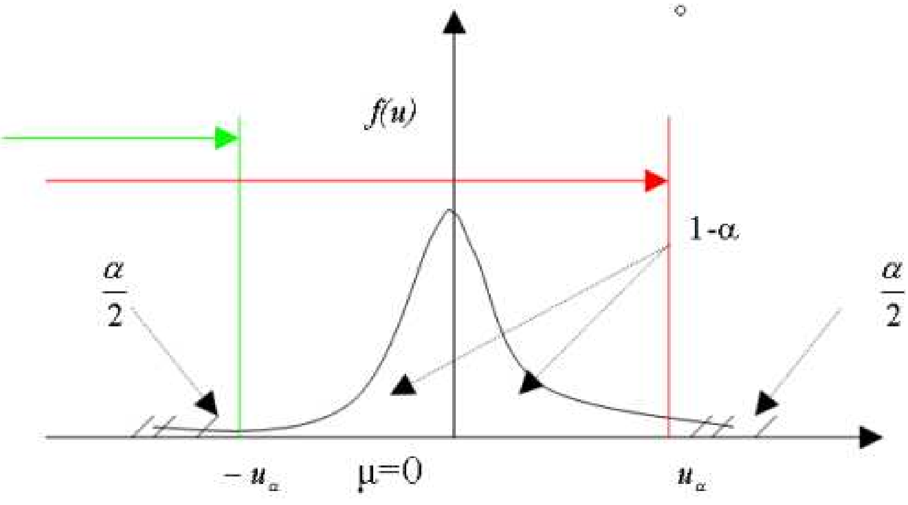
Where: $u_\alpha = \frac{a-\mu}{\sigma}$

Let us calculate: $P\left(-u_\alpha < Y = \frac{X-\mu}{\sigma} < u_\alpha\right) = 1-\alpha$

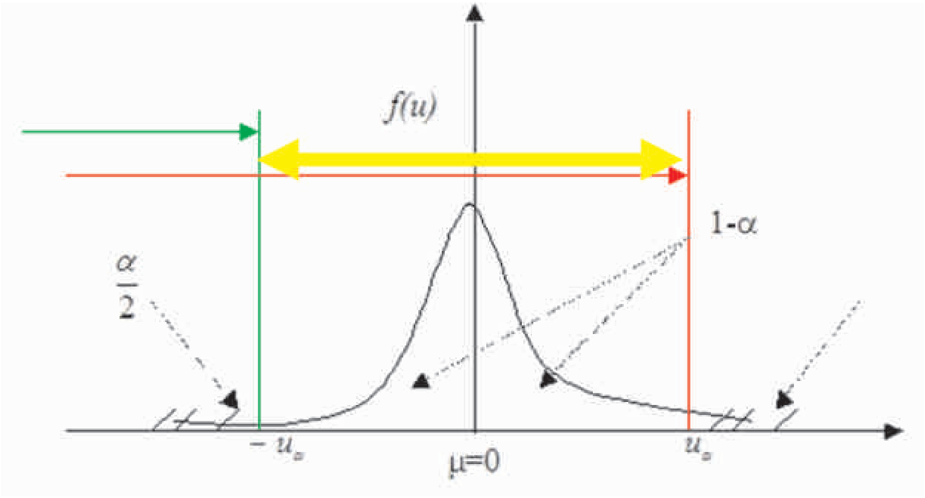
Step I: $P\left(Y < u_\alpha\right) = 1 - \frac{\alpha}{2}$



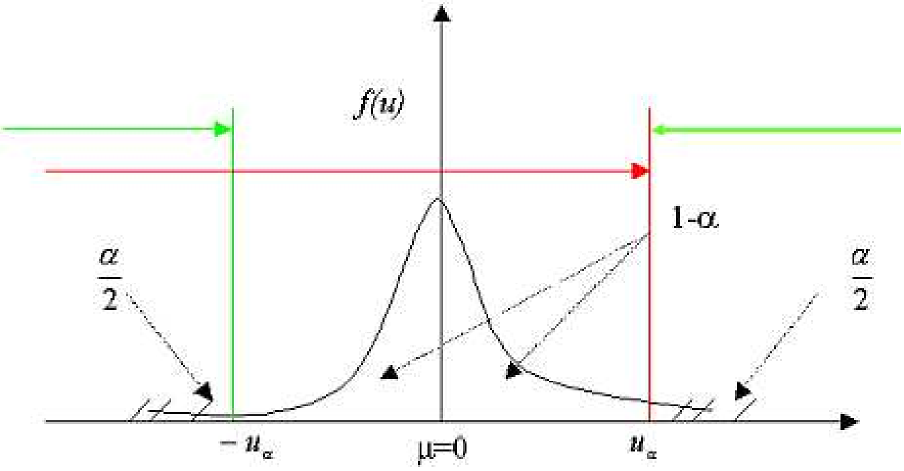
Step II: $P\left(Y < -u_\alpha\right) = \frac{\alpha}{2}$



Step III: $P(-u_{\alpha} < Y < u_{\alpha}) = P(Y < u_{\alpha}) - P(Y < -u_{\alpha}) = F_Y(u_{\alpha}) - F_Y(-u_{\alpha})$



Let us now use the symmetry of the normal distribution:



$$P(-u_\alpha < Y < u_\alpha) = P(Y < u_\alpha) - P(Y < -u_\alpha) = P(Y < u_\alpha) - P(Y > u_\alpha) =$$
$$= P(Y < u_\alpha) - [1 - P(Y < u_\alpha)] = 2P(Y < u_\alpha) - 1$$

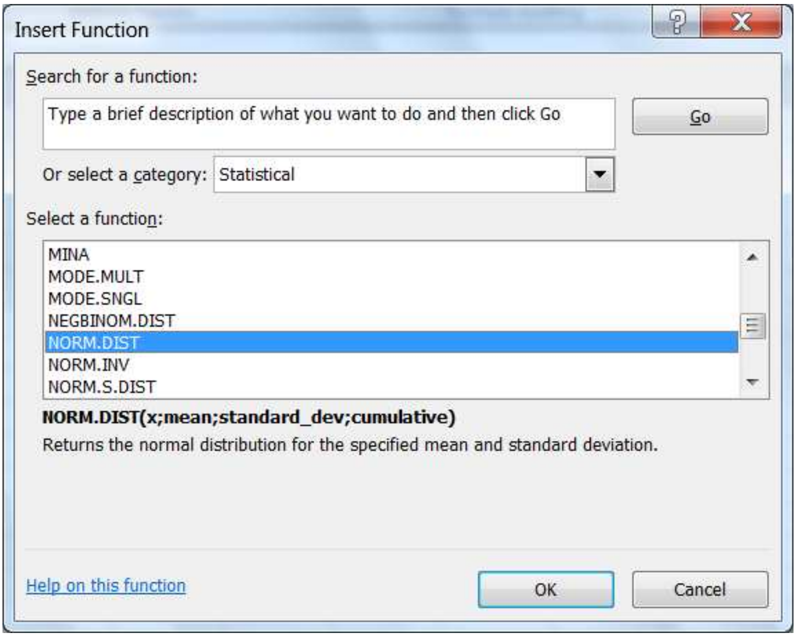
Therefore,

$$P(-u_\alpha < Y < u_\alpha) = 2P(Y < u_\alpha) - 1$$

As we know, the expression $P(Y < u_\alpha)$ is the value of the cumulative distribution function at u_α. If a random variable Y has the standardized normal distribution, its cumulative distribution function is denoted $\Phi(x)$. Thus:

$$P(-u_\alpha < Y < u_\alpha) = 2P(Y < u_\alpha) - 1 = 2\Phi(u_\alpha) - 1$$

The values of the distribution function for the N (0, 1) distribution are tabularised. They can also be read using a spreadsheet.
For u_α equal to 1.36, from the Excel spreadsheet



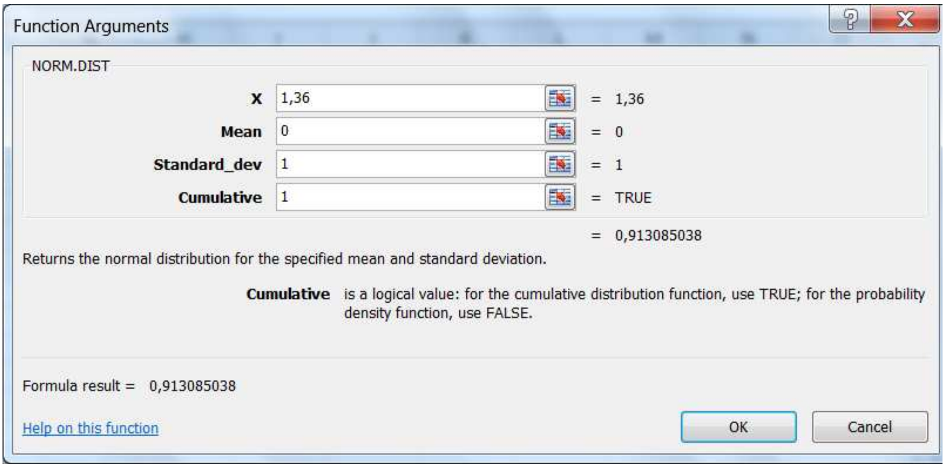
Search for a function: Type a brief description of what you want to do and then click Go

Or select a category: Statistical

Select a function: NORM.DIST

NORM.DIST(x;mean;standard_dev;cumulative)

Returns the normal distribution for the specified mean and standard deviation.



NORM.DIST

X 1,36 = 1,36

Mean 0 = 0

Standard_dev 1 = 1

Cumulative 1 = TRUE

= 0,913085038

Returns the normal distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability density function, use FALSE.

Formula result = 0,913085038

Example

- Let X~N(5, 4).
- Calculate P(-3<X<20)

First, we standardize:

P(-3 < $X_{N(5,4)}$ < 20) = $P\left(\frac{-3-5}{4} < \frac{X-5}{4} < \frac{20-5}{4}\right)$ = P (-2 < $Y_{N(0,1)}$ < 3,75)

Finally, we write this as the difference between standardized cumulative functions, Φ (3.75) - Φ (-2) :

Φ(3.75) - Φ(-2) = Φ(3.75) -[1- Φ(2)] = Φ(3.75) + Φ(2) - 1

= 0.9999 + 0.977 – 1 ≈ 0.977.

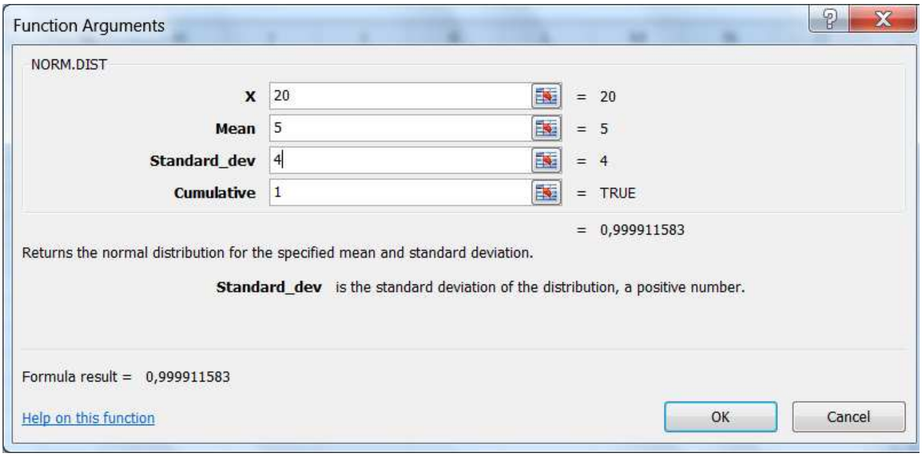
Note we find Φ (2) from the table as follows:

x	0	0.01	0.02	0.03	0.04
1.9	0.971	0.972	0.973	0.973	0.974
2.0	0.977	0.978	0.978	0.979	0.979
2.1	0.982	0.983	0.983	0.983	0.984
2.2	0.986	0.986	0.987	0.987	0.987
2.3	0.989	0.990	0.990	0.990	0.990

2.4	0.992	0.992	0.992	0.992	0.993
2.5	0.994	0.994	0.994	0.994	0.994
2.6	0.995	0.995	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998
3.0	0.999	0.999	0.999	0.999	0.999

The same calculation is carried out in the Excel spreadsheet as follows:

Let X~N(5, 4). Calculate P(-3<X<20)



X: 20 = 20

Mean: 5 = 5

Standard_dev: 4 = 4

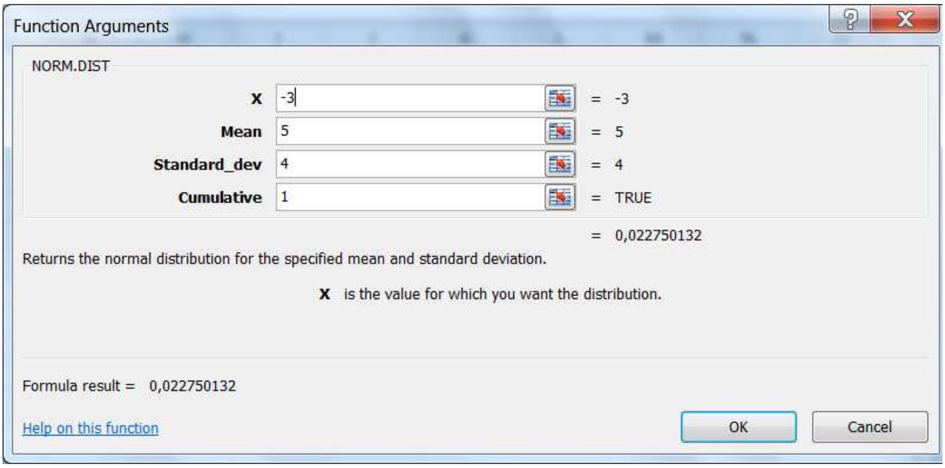
Cumulative: 1 = TRUE

= 0,999911583

Returns the normal distribution for the specified mean and standard deviation.

Standard_dev is the standard deviation of the distribution, a positive number.

Formula result = 0,999911583



NORM.DIST

X -3 = -3

Mean 5 = 5

Standard_dev 4 = 4

Cumulative 1 = TRUE

= 0,022750132

Returns the normal distribution for the specified mean and standard deviation.

X is the value for which you want the distribution.

Formula result = 0,022750132

Help on this function

Final remarks.

The first comment concerns the notation used for the normal distribution. $X \sim N(m, \sigma)$ is widely accepted to mean that the variable X is normally distributed with average µ and standard deviation σ and. However, in the English speaking world, the variance is often given i.e. $X \sim N (m, \sigma^2)$ is commonly used notation.

The second remark concerns the construction of tables for the standard normal distribution. In some tables 0.5 is substracted from the distribution function. Since this often leads to misunderstandings, it is worth checking what tables we are dealing with by finding the value of the distribution function at 0. If this value is 0 then we are dealing with the modified table (and therefore have to add 0.5 to obtain the value of the distribution function). If the value of the function at zero is 0.5, then we're dealing with a standard table. You can also read the value of the distribution function using Excel. Also note that some tables give P(Z>x), i.e. 1-Φ(x). In this case the values at the bottom of the table will be close to 0, and not close to 1 as in the standard table.

If you want to calculate the density function for the normal distribution, you should input 0 (the Boolean representation of false) into the appropriate box (Cumulative).

Note 3. The asymptotic properties of the normal distribution are particularly important from a practical point of view. In the literature, this phenomenon is often referred to as a “central limit theorem”.

Fourth Note. Let us return to the previously encountered concepts of the standardized skewness and kurtosis. If observations come from a normal distribution, these two measures will almost certainly be contained in the interval [-2, 2]. This means that given the values of the standardized kurtosis and skewness for a sample fall into the above range, we can presume (in statistics this is called hypothesizing) that the values observed come from an approximately normal distribution.

Lecture IV: Distributions of statistics.

Basic concepts:

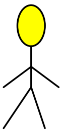
General population: population of elements from which we draw and whose characteristics will be investigated. Typically, a characteristic is determined as a random variable X, Y, Z, W, ...

A population may be finite (with a small or large number of elements) or infinite.

Example. The height of athletes.

Suppose we want to estimate the average height of athletes. Of course, our findings can be based only on a selected (in one way or another) group to examine, because investigating all athletes is non-feasible. Imagine that we were able to measure the height of five athletes and obtained the following results

Mr A	Miss B	Mr C	Miss D	Mr E
192 cm	165 cm	183 cm	170 cm	175 cm



Statistical Athlete:

X

Mean	177
Standard Error	4,785394
Median	175
Standard Deviation	10,70047
Sample Variance	114,5
Kurtosis	-0,84205
Skewness	0,5142

The observed values x_1, x_2, x_3, x_4 and x_5 are realizations of the same variable: X – height of statistical athlete

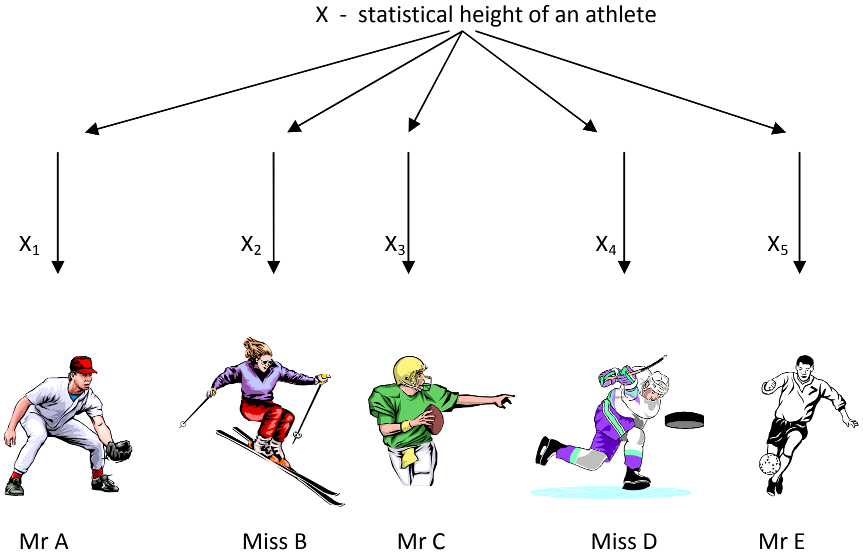
Descriptive statistics in Excel

Suppose the population has a distribution determined by the cumulative distribution function F(x), i.e. the feature *X* in this population has a distribution with distribution function F(x). This distribution is called the theoretical distribution function.

A random sample is given by a vector of random variables $(X_1, X_2, ..., X_n)$, each of which has the same probability distribution. In this example, the X_i can be interpreted as the heights our athletes before we measured them.

A random sample is said to be simple if $(X_1, X_2, ..., X_n)$ is a vector of independent random variables with the same distribution. Continuing our example, we can say that the height of the ice hockey player has no bearing on the height of the football player. This is the practical interpretation of independence. Unless stated otherwise, the samples considered here will be simple .

A realization of a simple sample is the set of the observed values of the vector$(X_1, X_2, ..., X_n)$, denoted by $x_1, x_2, ..., x_n$.



The sample distribution gives a statistical picture of the distribution of the characteristic X in the population as a whole.

The term statistics is also used in a narrower sense: a statistic is an arbitrary function of *n* random variables. Statistics are therefore random variables, and as

such have a probability distribution, exact or approximate.
For example, having a random sample $(X_1, X_2, ..., X_n)$, we can create the "mean" statistic (always labelled as $\overline{X}$) as follows:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$\overline{X}$ is simply the arithmetic mean of the *n* random variables from the sample and a sum of random variables is itself a random variable. Thus, $\overline{X}$ also has a probability distribution, which in turn has moments, such as the expected value, variance, standard deviation, etc.

Of course, there is the following relationship between the statistic $\overline{X}$ for the sample and the corresponding population parameters:

The average for the sample: $\overline{x}$ (number!). Realization of the statistic $\overline{X}$	The average for the population: an unknown or known parameter called the expected value of the statistic $\overline{X}$

Note that we have already met one example of such a statistic. A binomial distribution can be expressed as the sum of *n* independent random variables, each of which has a two-point 0-1 distribution. If one denotes a success, then the number of successes in a sample of size *n* has a binomial distribution. Each trial has the same probability of success, and the result does not depend on the results of other trials.
We may seek rules for various statistics (functions of random variables), in order to describe their probability distribution. For example, what is the distribution of the mean value of a sample?

Suppose the population mean and variance are *m* and σ^2. Let's calculate the expected value and variance of the sample mean:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X) = \frac{1}{n}n \cdot m = m$$

$$D^2\left(\overline{X}\right) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}D^2\left(\sum_{i=1}^{n} X_i\right) \underset{\substack{\uparrow \\ \text{Independency}}}{=} \frac{1}{n^2}\sum_{i=1}^{n} D^2(X_i) = \frac{1}{n^2}n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

Therefore,

- If $(X_1, X_2, ..., X_n)$ have (the same) normal distribution *N(m,σ)* and describe a simple sample (i.e. they are independent), then $\overline{X}$ is also normally distributed and $\overline{X} \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

- If $(X_1, X_2, ..., X_n)$ have the same distribution with average *m* and standard deviation σ, the sample is simple (independency!) and the sample is large (i.e. n> 30), then based on the so-called central limit theorem (the “sum of a large number of independent random variables has approximately a normal distribution”), *X* has approximately the same distribution as given above. We draw the reader's attention to the fact that this does not require knowledge of the distribution of the characteristic *X*!

Distributions of statistics play a central role in statistical inference (posing and test hypotheses), and therefore we will now review some useful statistics (functions of random variables.)

The χ^2 statistic[4]

If $X_1, \ldots, X_k$ are independent, standard normal random variables, then the sum of their squares

$$\chi^2(k) = \sum_{i=1}^{k} X_i^2$$

is distributed according to the chi-square distribution with *k* degrees of freedom.

Tables for this distribution — usually in its cumulative form — are widely available and the function is included in many spreadsheets and all statistical packages.

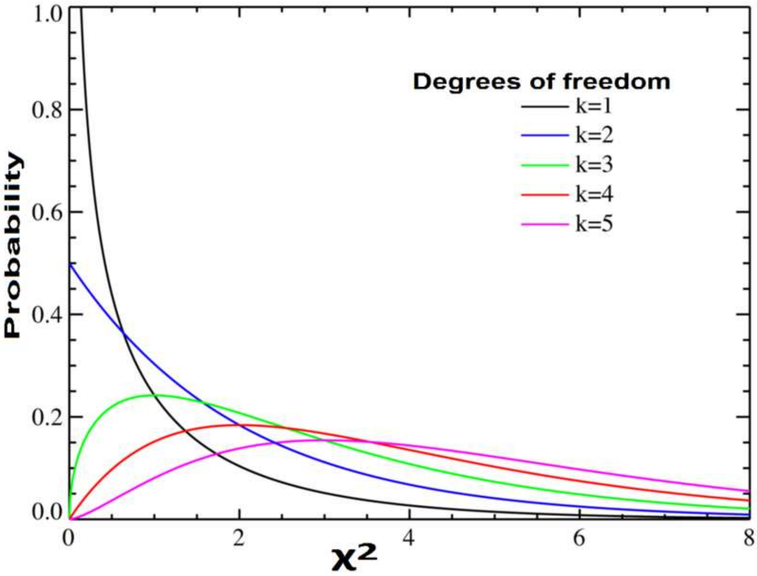


Fig. Density function of the χ^2 distribution.

The χ^2 distribution is tabularised for different *n* (called the number of degrees of freedom) , u$_0$ and for $\sigma = 1$. One may find the probability: $P\left(\chi^2 \geq u_0\right)$ from

[4] Those readers who are interested in the origin of the names used in mathematics are encouraged to have a look at the following page http://jeff560.tripod.com/c.html.

such a table. For different values of the standard deviation of the terms in the sum above, i.e. $\sigma \neq 1$, the formula $P\left(\chi^2 \geq u_0\right)$ should be replaced by

$$P\left(\chi^2 \geq z_0\sigma^2\right) = P\left(\frac{\chi^2}{\sigma^2} \geq z_0\right).$$

The t-Student (Gosset[5]) **statistic.**

Statistic:

$$t = \frac{\bar{X} - m}{S}\sqrt{n-1}$$

is called the t-student statistic with *n-1* degrees of freedom.

Generally, the t-statistic with n-1 degrees of freedom is any expression of the type

$$\frac{X \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right)}{\sqrt{\frac{Y}{n-1}} \sim \chi^2 \; \textbf{\textit{with (n-1) df}}}$$

Theorem: The sequence $\{F_n(t)\}$ of distribution functions of random variables with the t-Student distribution with *n-1* degrees of freedom satisfies the equation:

$$\lim_{n\to\infty} F(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{t^2}{2}} dt$$

So, the t-distribution converges to the standard normal distribution. In practice, when the number of degrees of freedom exceeds 30, then the difference between the standard normal distribution and the Student t-distribution is

[5] The derivation of the t-distribution was first published in 1908 by William Sealy Gosset, while he worked at the Guinness brewery in Dublin. He was not allowed to publish under his own name, so the paper was written under the pseudonym Student. We encourage the reader to become familiar with his biography http://en.wikipedia.org/wiki/William_Sealy_Gosset

insignificant. This explains why in most textbooks the table for the t-Student distribution stops at 30 degrees of freedom. This is evident from the graph of the density of the t-Student distribution.

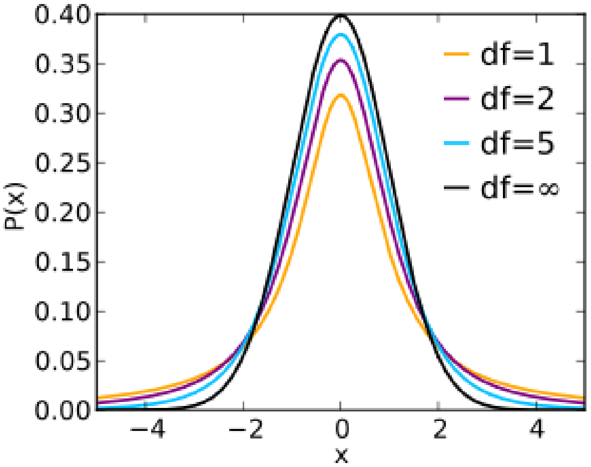


Fig. Density function of the t-Student distribution.

The U statistic.

Let $X_1, X_2, \ldots X_{n_1}, Y_1, \ldots, Y_{n_2}$ be two sequences of independent random variables from the $N(m, \sigma)$ distribution:

$$\overline{X} = \frac{1}{n_1}\sum_{k=1}^{n_1} X_k \qquad \overline{Y} = \frac{1}{n_2}\sum_{k=1}^{n_2} X_k$$

$$S_1^2 = \frac{1}{n_1}\sum_{k=1}^{n_1}\left(X_k - \overline{X}\right)^2 \qquad S_2^2 = \frac{1}{n_2}\sum_{k=1}^{n_2}\left(Y_k - \overline{Y}\right)^2$$

The mean statistics have the following distributions:

$$\overline{X} \sim N(m, \frac{\sigma}{\sqrt{n_1}}) \qquad \overline{Y} \sim N(m, \frac{\sigma}{\sqrt{n_2}})$$

Thus:

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sim N(0,1)$$

and

$$W = \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \sim \chi^2 \boldsymbol{with} \quad n_1 + n_2 - 2 \ \boldsymbol{df}$$

The U statistic is defined as:

$$U = \frac{Z}{\sqrt{\frac{W}{n_1 + n_2 - 2}}}$$

U has a t-Student distribution with $n_1 + n_2 - 2 \ \boldsymbol{df}$

$$U = \frac{\frac{\overline{X} - \overline{Y}}{\sigma} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}}{\frac{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}}}{\sqrt{n_1 + n_2 - 2}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)}$$

Note: it is assumed that the standard deviations of the *X*'s and the *Y*'s are equal. Since the t-distribution is asymptotically normal, we conclude from this that the U-statistic is asymptotically normal. This means that for large samples (n> 30), the distribution of the U-statistic is approximately normal

The S Statistic (sample standard deviation).

It can be shown that the statistic S has a gamma distribution with parameters

$p = \frac{1}{2}(n-1)$ and $b = \frac{n}{2\sigma^2}$.

The Z Statistic.

The Z statistic is defined as $Z = nS^2$. One can show that it has a gamma distribution with parameters $p = \frac{1}{2}(n-1)$ and $b = \frac{1}{2\sigma^2}$, which is equivalent to the χ^2 distribution with n-1 degrees of freedom.

The F-Snedecor[6] Statistic.

Let $X_1, X_2, ... X_{n_1}, Y_1, ..., Y_{n_2}$ be sequences of independent random variables with the $N(m, \sigma)$ distribution.

A random variable with an F-distribution arises as the ratio of two chi-squared variables:

$$F = \frac{S_1^2}{S_2^2}$$

with *(d_1=n_1-1, d_2=n_2-1)* degrees of freedom.

Note: normally it is assumed that the numerator is larger than the denominator (due to the form of standard tables for the F distribution).

[6] To find out more about the author see: http://en.wikipedia.org/wiki/George_W._Snedecor

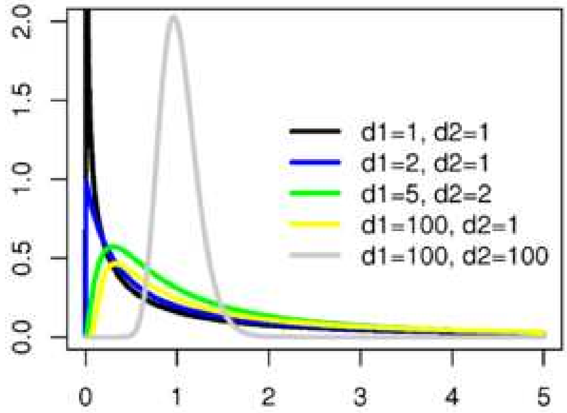


Fig. Density function of the F-Snedecor random variable.

If the characteristic X has a normal probability distribution, the distribution of selected statistics are given in the following table:

Distribution	Definition	Expected value	Variance
$\overline{X}$	$\frac{1}{n}\sum_{i=1}^{n} X_i$	*0* if $X \sim N(0,\sigma)$	$\frac{\sigma}{n}$
χ^2	$\sum_{i=1}^{k} X_i^2$	*k* if $X \sim N(0,1)$	*2k*
t-Student	$\frac{\overline{X}-m}{S}\sqrt{n-1}$	*0* if $X \sim N(m,\sigma)$	$\frac{n-1}{n-3}$ for *n*>3
U	$\frac{\overline{X}-\overline{Y}}{\sqrt{n_1S_1^2+n_2S_2^2}}\sqrt{\frac{n_1n_2}{n_1+n_2}(n_1+n_2-2)}$	*0* if	$\frac{n_1+n_2}{n_1n_2}$

		$X \sim N(m,\sigma)$	
S	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2}$	$\frac{n-1}{n}\sigma^2$ if $X \sim N(m,\sigma)$	$\frac{2(n-1)\sigma^4}{n^2}$
Z	nS^2	$(n-1)\sigma^2$ if $X \sim N(m,\sigma)$	$2(n-1)\sigma^4$

F-Snedecor	$\frac{S_1^2}{S_2^2}$	$\frac{n_2 - 1}{n_2 - 3}$
		for $n_2 > 3$
		if $X \sim N(0,1)$
			$\frac{2(n_2-1)^2(n_1+n_2-4)}{(n_1-1)(n_2-3)^2(n_2-5)}$
			for $n_2 > 5$

Lecture V. Confidence intervals.

The fact that it is not possible, based on a sample, to perfectly estimate an unknown parameter underlies the concept of confidence intervals. i.e. we cannot give one numerical value which corresponds to a given parameter, because by obtaining another sample from the population the estimated value of that parameter will almost certainly be different from the previously calculated value. Consider the following example.
When we analyzed the binomial distribution, we found that the probability of 4 successes in a sample of 10 elements with p = 0.5 is the same as the probability of six successes in such a sample, i.e. $P(X=4)=P(X=6)=0.205078$. Now imagine that we toss a coin 10 times and each result is written as 1 if you throw heads, and 0 if tails are thrown. Define the random variable *X* to be the sum of ten individual variables, each describing a single coin toss. The probability that heads is thrown 4 times in 10 throws (call it sample I) is the same as the probability that heads are thrown 6 times in 10 throws (sample II). Thus, each of these two situations are equally likely. However, calculating the mean for sample I we obtain $\overline{x}=0.4$, while the mean for sample *II* is $\overline{x}=0.6$. Note that both sample means can be used to estimate the expected value in the population! Both values are equally likely, and intuitively these two situations are quite possible in reality. So, what can we say in relation to the population mean. It is reasonably certain that the population mean is in the range [0.4, 0.6] and such an estimate is good enough, i.e. we wish to define an interval which includes the population mean with a high probability.

Suppose the distribution of some trait X in a population is dependent on some parameter Q (e.g. the population average and/or variance). It is assumed that the form of the density function $f(x,Q)$ or probability function $P(X=x_k)=p_k(Q)$ is known.

Let E be a vector of observations: $\left(X_1, X_2, ..., X_n\right)$. E is an n - dimensional random variable dependent on the parameter Q.

Let $\underline{U}(E), \quad \overline{U}(E)$ be functions of the random variable E such that $\underline{U}(E) \leq \overline{U}(E)$.

Let α be a real number $0 < \alpha < 1$.

If the following holds:

$$P(\underline{U}(E) \quad \leq \quad Q \quad \leq \quad \overline{U}(E)) \geq 1 - \alpha$$

(for continuous random variables = 1 - α), then the interval $\left\langle \underline{U}(E), \overline{U}(E) \right\rangle$ is called a confidence interval for Q, and $1 - \alpha$ the confidence level.

Theoretically, you can build a confidence interval for each parameter of the trait's distribution, but in practice they are used to construct confidence intervals for the population average (Q = m) and variance (Q = Var(X)). Below we show the corresponding confidence intervals for both parameters.

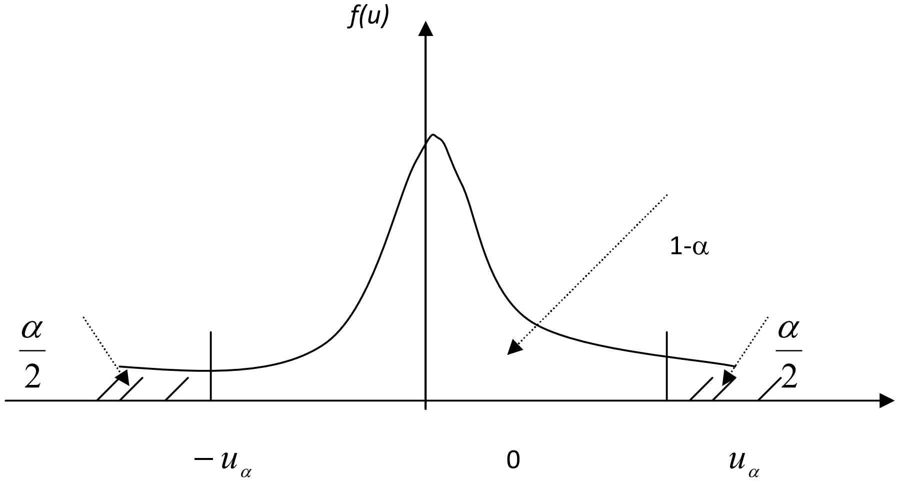
Confidence interval for the population mean (Q=m)

Model I.

Suppose a trait in some population has a N(m, σ) distribution. Let's assume that - *m* - is unknown, σ is known, and the sample is small (*n* < 30).

The estimate of the population average is the sample mean statistic $\overline{X}$. Given these conditions, $\overline{X}$ has a $N(m, \frac{\sigma}{\sqrt{n}})$ distribution.

Thus, the standardized random variable $U = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$ has a N(0, 1) distribution.



Hence, the random variable U satisfies:

$$P\left(-u_\alpha < U < u_\alpha\right) = 1-\alpha$$

For a given α, the value u_α can be read from the tables of the standardized normal distribution:

$$\Phi(u_\alpha) = 1-\frac{\alpha}{2}$$

Since U has a *N(0, 1)* distribution and we have $U = \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}$:

$$P\left(-u_\alpha < \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < u_\alpha\right) = 1-\alpha \ ,$$

$$P\left(\overline{X} - u_{\alpha}\frac{\sigma}{\sqrt{n}} < m < \overline{X} + u_{\alpha}\frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha \ .$$

Thus, for example for α = 0.05, the confidence interval is as follows (from the table u_{α} = 1.96):

$$\left\langle \overline{X} - \frac{1,96\sigma}{\sqrt{n}}, \overline{X} + \frac{1,96\sigma}{\sqrt{n}} \right\rangle .$$

This means that in 95 cases out of 100, the estimated „m" is in this range. In other words, the error in estimation is not greater than $\frac{1,96\sigma}{\sqrt{n}}$ in 95% of cases.

Example.

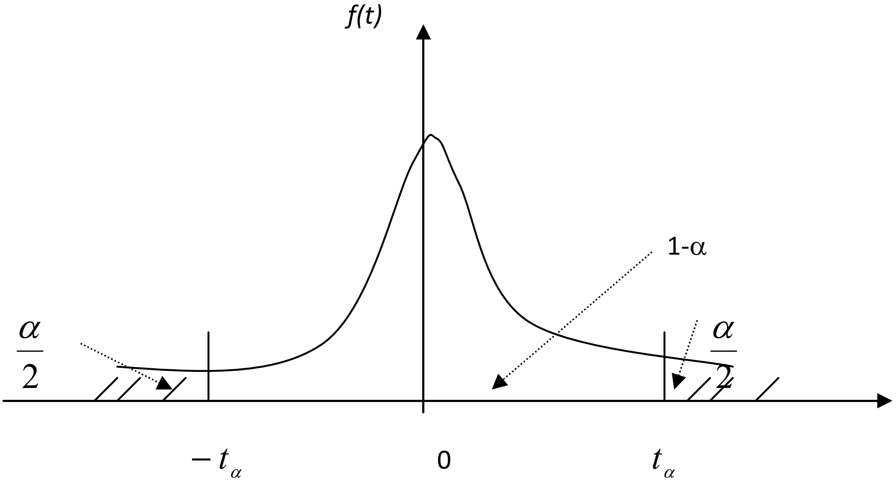
The waiting time for a tram has been studied and the following values obtained (in minutes): 12, 15, 14, 13, 15. Suppose that the waiting time for a tram has a normal distribution with unknown mean value (m) and a known standard deviation $\sigma = 2$.

Construction of the confidence interval for the mean involves finding the appropriate values in tables and some simple calculations . In our example, we find that $\overline{x} = 13.8$. Assuming that the confidence level is 0.95 (i.e. $\alpha = 0.05$) , we read the value $u_{\alpha} = 1.96$ from the table. Substituting these values into the formula, we obtain $\left\langle \overline{x} - 1.753; \overline{x} + 1.753 \right\rangle = \left\langle 12.0469; 15.5530 \right\rangle$. Thus with a probability of not less than $1 - \alpha = 0.95$, the desired population average lies in this interval.

Model II.

The trait X has a *N(m, σ)* distribution in some population, where neither *m* nor σ are known. To build a confidence interval for m, we will use the t-statistic with n-1 degrees of freedom:

$$t = \frac{\overline{X} - m}{S}\sqrt{n-1}$$



The value t_{α} is read from the tables for the Student distribution with n-1 - degrees of freedom:

$$P\left(-t_{\alpha} < \quad t < t_{\alpha}\right) = 1 - \alpha$$

$$P\left(-t_{\alpha} < \frac{\overline{X} - m}{\frac{S}{\sqrt{n-1}}} < t_{\alpha}\right) = 1 - \alpha \text{ ,}$$

$$P\left(\overline{X}-t_{\alpha}\frac{S}{\sqrt{n-1}}<m<\overline{X}+t_{\alpha}\frac{S}{\sqrt{n-1}}\right)=1-\alpha\ .$$

Thus, for example for α = 0.05, the confidence interval is as follows (t_{α} is read from the table for the Student distribution);
e.g. for n = 26, this value is 2.056):

$$\left\langle\overline{X}-\frac{2,056S}{5};\overline{X}+\frac{2,056S}{5}\right\rangle.$$

This means that in 95 cases out of 100, „m" lies in this range. In other words, the error in estimation is not greater than $\frac{2.056S}{5}$ in 95% of cases.

Note: this length of this interval is variable, since it depends on the value of S.

Example. Suppose that in the previous example, on the average waiting time for a tram, there was no information on the standard deviation in the population. Again, we calculate the average value $\overline{x}$, which is 13.8. We calculate the standard deviation for the sample $s=1.3038$: and read the value $t_{\alpha}=2.7764$ for α= 0.05 from the tables. We thus calculate the confidence interval to be:

$\langle\overline{x}-1.618931187;\overline{x}+1.618931187\rangle=\langle12.18107;15.41893\rangle$. Thus with a probability of not less than $1-\alpha=0.95$, the desired population mean lies within this confidence interval

Model III.

For large samples (n> 30), the central limit theorem states that

$\bar{X} \to N(m, \frac{\sigma}{\sqrt{n}})$ for $n \to \infty$, while the law of large numbers states that $S \to \sigma$.

Therefore, by substituting the population standard deviation α into model I with the sample standard deviation *s*, we get:

$$P\left(\bar{X} - u_\alpha \frac{s}{\sqrt{n}} < m < \bar{X} + u_\alpha \frac{s}{\sqrt{n}}\right) \cong 1 - \alpha .$$

Example. Suppose that in the analysis of the example on the average waiting time for a tram there was no information on the standard deviation in the population, but we managed to gather much more data: *n* = 34. From these data we computed $\bar{x} = 13.9411$, $s = 1.1791$. Hence, the required confidence interval is $\langle \bar{x} - 0.41142; \bar{x} + 0.41142 \rangle = \langle 13.5297; 14.3525 \rangle$. Thus with a probability of not less than $1 - \alpha = 0.95$, the desired population mean lies in this confidence interval. Note that, in accordance with our intuition, as a consequence of the large number of observations the length of this interval has decreased significantly, i.e. estimates become more accurate.

Model IV (confidence interval for a proportion).

If we examine a population according to the presence or absence of a certain characteristic (e.g. quality control - products classed as good and bad, non-smokers and smokers, etc.), it can be described by a two-point distribution:

$$P(X = 1) = p, \quad P(X = 0) = 1 - p ,$$

where the random variable X takes the value 1 if the feature is present and 0 if not present.

Thus, if a feature is observed m times in an n-element sample, an approximation of p is given by $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{m}{n} = \widehat{p}$.

We find that for 0.05 <p<0.95 and n> 100:

$$P\left\{\frac{m}{n} - u_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} < p < \frac{m}{n} + u_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\right\} \approx 1-\alpha$$

Estimation of the percentage of smokers among students. In an 1800-element sample, the number of smokers m = 600. For 1 - α = 0.95, $u_\alpha = 1{,}96$. Thus,

$\frac{m}{n} = \frac{600}{1800} = 0{,}333$, $\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} \cong 0{,}011$. Thus the 95% confidence level for the fraction of students smoking is:

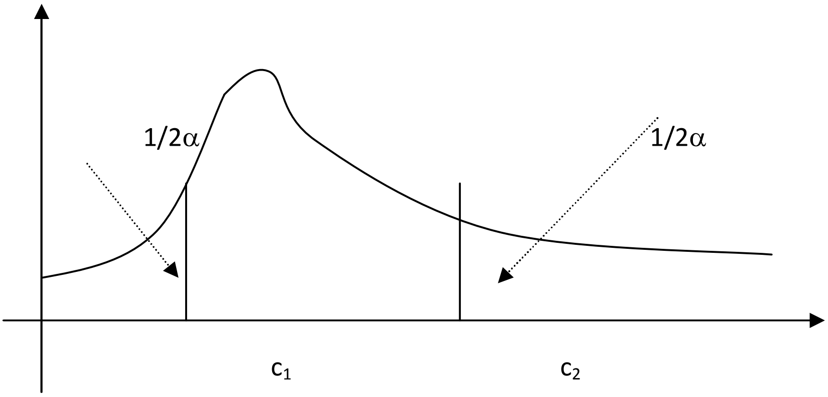
32,19% < *p* < 34,40%.

Confidence interval for the standard deviation - variance (Q = σ).

The second most commonly used population parameter is a measure of its volatility. Depending on ones requirements, this is measured using the variance or standard deviation. In general, deriving a confidence interval for the variance or standard deviation is not easy. However, if you assume that the feature *X* has a normal distribution, or at least approximately normal, we can use the χ^2 and Z distributions, already known to us .

Model V:

Suppose that the population mean m and population standard deviation σ are not known and the sample is large (n > 30). The statistic $Z = \frac{nS^2}{\sigma^2}$ has a χ^2 distribution with n-1 degrees of freedom:



$$P\left(c_1 < \frac{nS^2}{\sigma^2} < c_2 \right) = 1 - \alpha$$

where: $P\left\{ \chi^2 < c_1 \right\} = \frac{1}{2}\alpha \quad P\left\{ \chi^2 \geq c_2 \right\} = \frac{1}{2}\alpha$.

Note: Most tables give values of the form $P\left\{\chi^2 \geq a\right\}$ and therefore the derivation of the value of *c₁* will require some transformations.

We obtain the following confidence interval for the variance:

$$P\left(\frac{nS^2}{c_2} < \sigma^2 < \frac{nS^2}{c_1}\right) = 1 - \alpha .$$

Example. Consider again the data used in model III.

We calculate the sample variance: s=1.3903. Thus the observed value of the non-standardized χ^2 statistic (the numerator of the equation) is given by 1.3903 * 34 = 47.2727.

We now calculate the values of c_1 and c_2: c_1= 19.0466 and c_2=50.7251. Hence, the required confidence interval for the variance is:

$$\left\langle \frac{47.2727}{50.725}; \frac{47.2727}{19.0466} \right\rangle \equiv \left\langle 0.9319; 2.4819 \right\rangle$$

Model VI.

Assume that the trait of interest has a normal or close to normal distribution and the sample is large, i.e. n> 30. Interval estimation of the standard deviation is obtained from the following formula

$$P\left\{\frac{s}{1+\frac{u_\alpha}{\sqrt{2n}}} < \sigma < \frac{s}{1-\frac{u_\alpha}{\sqrt{2n}}}\right\} \approx 1 - \alpha$$

where: u_α is read from tables for the standard normal distribution, N(0, 1).

Example. Again, we use the data from the example for model III. In this case, *n* = 34, $s = 1.1791$. The value $u_{\alpha} = 1.96$ (from the table for the N(0, 1) distribution). Thus we obtain the following confidence interval for the population standard deviation

[0.9527; 1.5467].

Note that the confidence interval obtained in the previous example can also be used to obtain an interval estimate for the standard deviation. This is done by taking the square root of both ends of the interval. In this case, we obtain the following confidence interval: [0.96535, 1.5754]. As you can see, the interval estimate obtained is similar to the interval given above.

Determination of the sample size required for interval estimates of the average with a given precision and confidence level.

As already mentioned, the use of confidence intervals to estimate the unknown parameters of a population is due to constraints related to the information available about the population: we only observe a sample rather than the entire population. It is assumed that, as well as specifying a level of confidence, we also set the precision of estimates and thus meet expectations associated with obtaining an accurate estimate. Looking for more precision is not desirable in this context. Admittedly, the narrower the range of the interval, the better an estimate is. However, it must be remembered that the "price" we pay for this accuracy is an increased sample size. For various reasons, we use small samples, e.g. because it is not possible to obtain a larger sample, or the cost of collecting a greater amount of data is too large. In this case, we want to make sure that the sample size is not greater than required from the assumed level of confidence and accuracy of the estimate.

Note that in each case we give the required length of the interval, as the

difference between the right and left end of the confidence interval. This knowledge allows us to determine the necessary sample size, so that one can give interval estimates with a predetermined confidence level and precision.
We would like, therefore, to determine the appropriate sample size to obtain a confidence interval of a given length (accuracy) with the chosen confidence level *1* - α*.* Let 2*d* be the length of the interval.

Model VII

Using the formula for the confidence interval for the population mean in model I, we find that the length of the interval is $2d = 2\frac{u_\alpha \sigma}{\sqrt{n}}$. Hence, the required sample size is: $n = \frac{u_\alpha^2 \sigma^2}{d^2}$.

For example, using model I, it is assumed that $\sigma = 2$. The value $u_\alpha = 1.96$ is read from the table for the standard normal distribution. We can therefore answer the question of whether estimating the average waiting time for a tram, with an accuracy of, e.g. 30 seconds, is possible using the sample we have, i.e. of 5 elements. Calculating *n* for *d = 0.5*, we find that $n = \frac{1.96^2 \cdot 2^2}{0.5^2} = 61.47$. Thus with a confidence level of 95%, it is not possible for us to estimate the population mean with an accuracy of 1 minute. For this purpose, we need a minimum sample size of 62 observations.

Model VIII

Similarly, calculating the length of the interval from model II, we proceed as follows: $2d = 2\frac{t_\alpha s}{\sqrt{n-1}}$. Thus: $n = \frac{t_\alpha^2 s^2}{d^2} + 1$. However, given that we do not know the population standard deviation, we need an estimate. Hence, we take an initial sample (s is calculated from this sample) of size n_0. If $n > n_0$, then the remaining $n - n_0$ elements required must be sampled.

Returning to the data from the example for model *II.* We calculate the average value $\overline{x}$, which is 13.8. Then, we calculate the sample standard deviation: $s = 1.3038$ and read from the table the value of $t_\alpha = 2.7764$ for $\alpha = 0.05$. As before, we ask whether it is possible to estimate the average duration of the waiting for a tram to the nearest *30* seconds (*d = 0.5*)? Calculating the required sample size, we obtain *n* = 53.41. Hence, it is necessary to draw 49 more observations (54 - 5), in order to estimate the average waiting time for a tram to the nearest 30 seconds.

Model IX:

Under model III, the determination of the required sample size for a given level of accuracy *d* is the same as in model II. Let's use the data from the example for model III. We managed to collect much more data: $n_0 = 34$. First, the necessary statistics are computed, $\overline{x} = 13.9411$ and $s = 1.1791$. Assuming that the confidence level is 0.95 (i.e. $\alpha = 0.05$), we read the value $u_\alpha = 1.96$ from the table. The required value of *n* is thus $n = \frac{u_\alpha^2 s^2}{d^2} + 1$. Again, as before, we ask whether it is possible to estimate the average duration of the waiting for a tram to the nearest *d* = 30 seconds (*d* = 0.5)? We calculate the value of *n* = 22.36. Thus, in this case it is possible to estimate with a predetermined precision (*d* = 30 seconds) for a given confidence level ($\alpha = 0.05$), because we have a sample of 34 items, and the required sample size is only 23.

Model X:

In assessing the sample size required to estimate a population proportion, one can use the following two options:

a) if we know the magnitude of p, the required sample size is $n = \frac{u_\alpha^2 p(1-p)}{d^2}$:
b) If we do not know the order of magnitude of p, we use the inequality $p(1-p) \leq \frac{1}{4}$. Thus: $n = \frac{u_\alpha^2}{4d^2}$.

Let's look at this using the data from the example for model IV. We estimated the percentage of smokers among students. We observed m = 600 smokers in the sample of 1800. For

1-α= 0.95 $u_\alpha = 1.96$. Hence, the estimate of the proportion of all students who smoke is $\frac{m}{n} = \frac{600}{1800} = 0,333$ i.e. 33.3%. Is it possible to estimate this proportion to the nearest 5%?

If we know (e.g. from studies by other authors) the expected value of the percentage of students smoking, for example, 30%, the desired sample size is *n* = 322.69. So the sample size of 1800 students used is sufficient for this purpose.

If you do not know the expected value of the percentage of students smoking, then use the second formula and we obtain *n* = 384.16. The resulting sample size, although it is greater than the one previously calculated, still indicates that the sample of 1800 students is sufficient for our purposes.

In both situations, to achieve the required precision for the estimate of the population proportion (5%) with a 95% confidence level it is always possible to use a 385-element sample. Collecting data for 1800 students is clearly redundant!

Lecture 6. Parametric tests.

Any claim regarding the unknown distribution of a characteristic is called a statistical hypothesis.

A hypothesis which specifies only the numerical values of unknown parameters of the distribution of a characteristic is called a parametric hypothesis. In order to verify such hypotheses, we use parametric tests.

A hypothesis regarding other features of the distribution of a character (including its parameters) is called a non-parametric hypothesis. To verify such hypotheses, we use non-parametric tests.

Hypothesis testing is necessary to decide whether a given hypothesis can be considered to be true or false. Hence, we specify an initial hypothesis (called H_0 – the null hypothesis) and formulate an alternative hypothesis (called H_1), which will be considered to be true if it is concluded that the hypothesis H_0 is not true.

If only one hypothesis is formulated and the object of a statistical test is to check whether this hypothesis is true or not and we do not check other hypotheses, such a test is called a test of significance.

The following is an algorithm for testing a parametric hypothesis (significance testing):

1. We formulate the hypothesis H_0: $(Q = Q_0)$.
2. We set the significance level α.
3. Next, we observe an n - element simple sample.
4. Calculate the realization *u* of the relevant U statistic (significance test).

5. We look for the critical value u_0 of the statistic U for the selected $\neq$ satisfying the following inequality:

$$P\left(\left|U\right| \geq u_0\right) \leq \alpha$$

6. If $\left|u\right| \geq u_0$, then we reject the hypothesis H_0,

   If $\left|u\right| < u_0$, there is no reason to reject the hypothesis H_0.

Note: For small samples we use the exact distribution of the statistic U; for large samples the limiting distributions of these statistics is used.

Example.

Let *X* be a trait with a normal distribution $N(m,1)$ in the population of interest, where m is unknown. We believe that the unknown average value is 0, i.e., we test the hypothesis H_0: m = 0 against the alternative hypothesis H_1: m $\neq$ 0.

The only way to test our hypothesis is to compare it with a sample from the general population. Hence, a random sample of 10 items was chosen and the following results obtained:

-0.30023

-1.27768

0.244257

1.276474

1.19835

1.733133

-2.18359

-0.23418

1.095023

-1.0867

We need to assess whether the probability of obtaining such specific values, as in our 10 -element sample, is large enough to be able to claim that this does not conflict with the hypothesis H_0. In normal circumstances, it is usually assumed that if the probability of obtaining more extreme results under H_0 than those observed is less than 0.05 (the significance level), then we reject H_0.

We calculate the sample mean and standard deviation and obtain $\overline{x} = 0.04649$ and *s* = 1.2924. Thus, the difference between the average value from the null hypothesis and the average value calculated from the sample is 0.04649. Is this difference so significant (important) to believe that we must reject the null hypothesis? To assess this fact the t-statistic is used.

$$t = \frac{\overline{X} - m}{S}\sqrt{n-1}$$

where:

$\overline{x} = 0.04649$,

s=1.2924,

n = 10,

m=0.

The realisation of the t-statistic 9 degrees of freedom is equal to 0.114. The probability that the absolute value of the realisation of the test statistic is greater than this value (i.e. that the mean of such a sample is further away from 0 than 0.04649) under H_0 is 0.912. We interpret this as meaning that if the null hypothesis is true, then the chances (measure of the likelihood) of obtaining such a result amounts to 0.912. Quite a lot, if we remember that the upper limit is set to 1.

Technically, the value of this probability is called the p-value Our intuition is that a *p*-value close to 1 does not allow the rejection of the null hypothesis. But what should be the lowest p-value for which we accept the null hypothesis? After all, rare events happen. For example, throwing a coin 10 times, we expect an equal number of heads and tails. However, what imbalance between these results would compromise our certainty about the reliability of the coin? 7:3? 8:2? But 10:0 could happen! It is necessary to compromise between certainty and presumption. In practice, it is assumed that this lowest p-value is 0.05 and this is called the level of significance α = 0.05. This means that we consciously accept the possibility of rejecting a true null hypothesis H_0 after a rare event, given that this could happen. In our example, in order to get the p-value for this test given there are 8 heads, we should calculate the probability of either 8,9 or 10 heads or 8,9 or 10 tails – i.e. results that are at least as far away from the expected 5 heads and 5 tails, hence we do not reject if 8 heads (or tails) appear p=2*(1+10+45)/1024≈0.11, but reject if 9 heads (or tails) appear p=2*11/1024≈0.02. Thus, if the number of heads (or tails) was 8, we would not reject H_0 (that the coin is fair). If there were 9 or 10 heads (or tails), we would reject H_0. To reiterate: the level of significance is a compromise accepted by practitioners. The significance level can be changed: in fact, the more we are afraid of rejecting H_0, the lower the significance level should be. It is assumed that this level should be in the range [0.02, 0.1].

At the end of this section, considerations reaffirm that the similarity to a confidence level of 1-α is not incidental. The same symbol α is used, and very often the same value of 0.05. The confidence interval for a parameter can also be used to test hypotheses relating to the value of this parameter. If, for example, the null hypothesis is that the average value in the population $m = 0$, and the confidence interval for the population mean based on a sample has a negative left endpoint and positive right endpoint, then this result states that it is quite possible that the population average is zero (i.e. we do not reject H_0) , because *0* is inside the confidence interval. If both endpoints of the confidence interval for the population mean have the same sign, it means that *0* is not in it, and we reject H_0 at a significance level of α = 0.05 (i.e. it is not credible that the population mean is zero at this significance level).

For the record, let us also recall that if we accept the null hypothesis, it only means that there are no grounds for its rejection and this is to be interpreted as treating H_0 as true as long as we cannot find strong evidence contradicting it. Rejection of the null hypothesis is unconditional. However, in both cases, we may be wrong and e.g. the probability of rejecting H_0 when is true is given by α – the level of significance.

Let us now divide the test sample into two parts (two samples) by including the even-numbered observations in the first sample and odd-numbered observations in the second sample. Such a method can be used to test the homogeneity of the original sample. If it is homogeneous, the parameters (e.g. mean value and standard deviation) in the two new samples should not significantly differ. Thus we have the following two samples:

X	Y
-0.30023	-1.27768
0.244257	1.276474
1.19835	1.733133
-2.18359	-0.23418
1.095023	-1.0867

Let us examine whether both samples come from populations with the same population mean (which should be the case as they come from the same population), i.e. $m_X = m_Y$. We will use the statistic

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)}$$

We calculate the appropriate statistics:

	Variable 1	*Variable 2*
Mean	0.010762	0.0822094
Variance	1.888100953	1.866891821
Observations	5	5

The realisation of the U statistic is -0.08244548. The critical value for the t-Student distribution with 8 degrees of freedom found in the tables is $t_\alpha = 2.306$. Thus, any realisation of U contained in the interval [-2.306, 2.306] is not inconsistent with the null hypothesis stated above. There are, therefore, no grounds for its rejection.

Depending on the context, the alternative hypothesis (to the null hypothesis H_0: $(Q = Q_0)$) can have one of the following forms:

$$H_1 : (Q \neq Q_0),$$

$$H_1 : (Q > Q_0),$$

$$H_1 : (Q < Q_0).$$

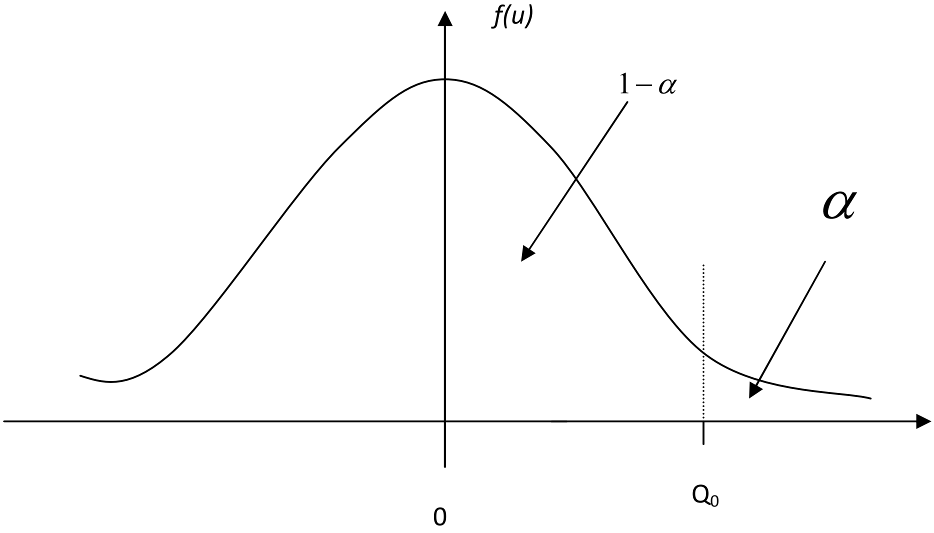
The first case is called a two-sided alternative and the corresponding parametric test is a two-sided test. The other two cases are called right-sided and left-sided alternatives, respectively. The examples considered above were two-sided tests. For example, examining the earnings of men and women we may formulate the following hypotheses:

a) two-sided, for example, the earnings of men and women in the same positions differ significantly. This hypothesis is two-sided, because we did not assume a particular direction for the difference.

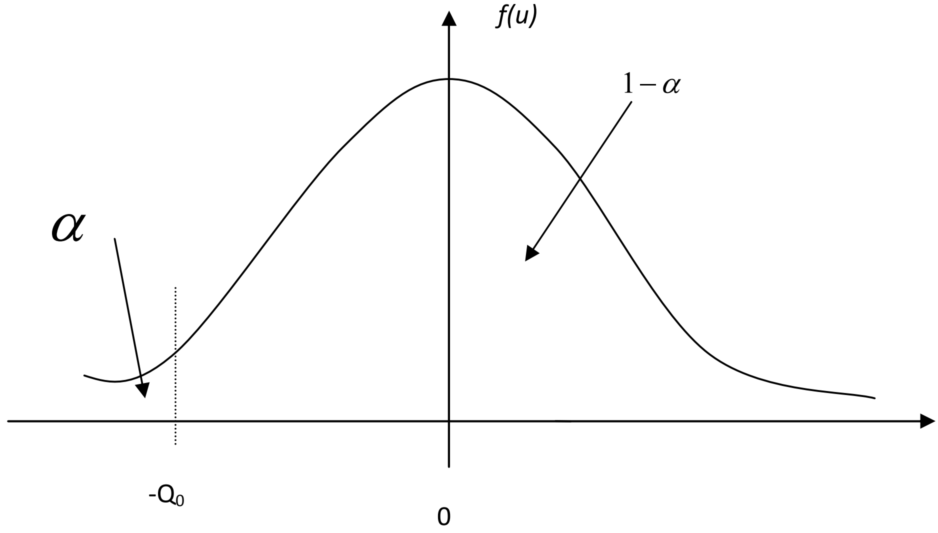
b) One-sided (directional), for example, men earn significantly more than women in the same job. This hypothesis implies the direction of difference, thus it is one-sided.

The two types of one-sided tests using t-statistics are illustrated below:

$H_1 : (Q > Q_0)$



$$H_1 : (Q < -Q_0)$$



Returning to the example concerned with the samples *X* and *Y*, now suppose that we think that the sample *Y* comes from a population with a mean value greater than the average in the population represented by the sample *X*. So we formulate the null hypothesis $m_X = m_Y$ against the alternative hypothesis $m_X < m_Y$. The realisation of the U statistic is t =- 0.08244548. The critical value for the t-Student distribution with 8 degrees of freedom (one-sided test) found in the tables is $t_\alpha = \mathbf{1.8595}$. So, therefore, there is no reason to reject the null hypothesis.

To finish, the data are used to illustrate another test: a comparison of variance with two small samples. We will use the F-Snedecor test:

$$F = \frac{S_1^2}{S_2^2}$$

with (*n_1-1=4, n_2-1=4*) degrees of freedom.

We get:

	Variable 1	*Variable 2*
Mean	0.010762	0.0822094
Variance	1.888100953	1.866891821
Observations	5	5
Df	4	4
F	1.011360665	
P(F<=f) one-tail	0.495763859	
F Critical one-tail	6.388232909	

We see that the assumption of the equality of variances is reasonable: you cannot reject the hypothesis of the equality of the variances in the populations from which these samples come. The realisation of the F statistic is 1.01136 and is much smaller than the critical value 6.3882.

Example.

A factory produces product A. Characteristic X has an $N(m,\sigma)$ distribution with an unknown value of σ. Technical requirements mean that the standard deviation of this characteristic should be 4.

To verify the quality of production, a simple sample was taken with *n* = 20 elements. The standard deviation for this sample is *s* = 4.4. Does production satisfy the technical requirements?

We therefore hypothesize $H_0(\sigma = 4)$.

We calculate the realisation of the statistic $\frac{Z}{\sigma^2}$, which has a χ^2 distribution with n-1 degrees of freedom:

$$\frac{Z}{\sigma^2} = \frac{20 \cdot (4.4)^2}{4^2} \cong 24.2$$

Let's assume α=0.01. We look for the value of z_0 which satisfies:

$$P\left(\frac{Z}{\sigma^2} > z_0\right) = 0.01$$

From the table for the χ^2 distribution, we found that for 19 degrees of freedom z_0 = 36.19.

Thus, assuming that the hypothesis H_0 is true, the probability $P\left(\frac{Z}{\sigma^2} > 24.2\right)$ is greater than α=0.01.

Conclusion: there is no reason to reject the hypothesis that the technical requirements are satisfied. Note that even a change in the level of significance from α = 0.01 to α = 0.05 does not affect this result: the critical value is then z_0 = 30.1435, and is still greater than 24.2.

Example.

Pieces of art from a given workshop are either good or flawed. The probability *p* of a piece being flawed is not known. A sample of 30 pieces was taken and 4 pieces were found to be defective. The workshop declares that the proportion of flawed items is 10%. Is this declaration realistic?

We put forward the hypothesis $H_0(p = 0,1)$. The alternative is right-sided, i.e. (p>0.1).

Denote:

$$X_k = \begin{cases} 1 & k - th\ item\ is\ good \\ 0 & k - th\ item\ is\ flawed \end{cases} \quad , \quad X = \sum_{k=1}^{30} X_k \quad .$$

If the hypothesis is true, then:

$$P(X = r) = \binom{30}{r} \cdot 0,1^r \cdot 0,9^{30-r}$$

We observe x = 4 . The p-value is given by:

$$P(X \geq 4) = 1 - P(X < 4) = 1 - \sum_{r=0}^{3} \binom{30}{r} \cdot 0.1^{r} \cdot 0.9^{30-r} = 1 - 11.134 \cdot 0.9^{27} = 0.3527$$

Conclusion: there is no reason to reject the hypothesis that the proportion of items that are flawed is 10%.

At the end of this lecture, let us examine an example of the series test used to test the quality of mass-produced products, which can be regarded as a non-parametric test, i.e. a test used to examine the compatibility of the empirical distribution with a theoretical distribution. Such tests are called nonparametric tests or tests of compliance.

We are examining the quality of manufactured components. 50 successive items are taken from a production line and classified as good (D) or defective (Z):

Length of series	6	1	8	1	2	1	13	2	13	1	1	50
Classified elements	DDDDDD	Z	DDDDDDDD	Z	DD	Z	DDDDDDDDDDDDD	ZZ	DDDDDDDDDDDDD	Z	D	
Number of series	1	2	3	4	5	6	7	8	9	10	11	11

Hypothesis H_0: good and bad components are produced in a random way.

The alternative hypothesis H_1: the technological process leads to subsequences in which defects are more likely than usual . Acceptance of the alternative hypothesis means that there may be a systematic trend in the process, which may enable the identification of ways to improve the quality of production.

For the data presented here, we use a test called the series test, whose test statistic has the following form:

$$Z = \frac{U - \mu_U}{\sigma_U},$$

which for n> 20 has an approximately normal distribution.

where: U - the number of series, n_1 - number of bad elements, n_2 - the number of good elements

- Mean value: $\mu_U = \frac{2n_1 n_2}{n} + 1$,
- Standard deviation $\sigma_U = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}}$

In our example: n = 50, n_1 = 6, n_2 = 44, U = 11 :

$$Z = \frac{U - \left(\frac{2n_1 n_2}{n} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}}} = -0,39$$

$Z_{\alpha=0,05} = -1,645$ (a left-sided test).

Conclusion: there is no reason to reject the hypothesis of homogeneity of production - there are no systematic changes in the likelihood of errors.