Course: Mathematics & Statistics

Field of study: Management and Production Engineering

Type of instruction and number of hours: lecture 15 h, laboratory 30 h

Number of ECTS credits: 4

Learning outcomes:

Knowledge:

- Student knows the concept of statistical series and types of statistical series.
- Student knows the definition and properties of probability.
- Student knows the basic probability distributions of discrete and continuous random variables and their characteristics.

Skills:

- Student can create and interpret frequency tables and determine descriptive statistics.
- Student is able to determine the expected value and variance for different distributions of a random variable.
- Student is able to evaluate linear correlation of variables and construct linear regression equations.
- Student is able to calculate the probability using the classical definition of probability and the distribution of a random variable.

Social competences:

- Student is aware of the importance of using statistical methods in business management.
- Student is aware of the advantages as well as the limitations of statistical tools for their practical use.

Evaluation methods of learning outcomes:
written test, activity during classes – solving tasks

List of course topics:

Lecture:

1. Descriptive statistics. Creating frequency tables and calculation of dispersion and position measures.
2. Classical definition of probability. Elements of combinatorics. Independence of events. Conditional probability, total probability, Bayes' formula.
3. The probability distribution of a random variable. Discrete and continuous distributions.
4. Interdependence of variables. Discrete two-dimensional random variable, correlation coefficient, linear regression model.
5. Confidence interval estimation.
6. Statistical hypotheses testing: parametric tests, non-parametric tests.

Laboratory:

1. Descriptive statistics. Creating frequency tables and calculation of dispersion and position measures.
2. Classical definition of probability. Probability properties. Elements of combinatorics. Independence of events. Conditional probability, total probability, Bayes' formula.

3. Linear correlation and regression.
4. Rank correlation.
5. The probability distribution of a random variable. Discrete and continuous distributions.
6. Confidence interval estimation.
7. Statistical hypotheses testing: parametric tests, non-parametric tests.

Bibliography

Basic literature

[1] Mendenhall W. M., Sincich T. L., *Statistics for engineering and the sciences*. 6th ed., CRC Press, Taylor & Francis Group, a Chapman & Hall Book, Boca Raton, 2016.

[2] Montgomery D. C., Runger G. C., *Applied statistics and probability for engineers*. 6th ed., international student version, John Wiley & Sons, Singapore, 2014.

Complementary literature

[3] Pritchett G. D. [et al.], *Fundamentals of quantitative business methods: business tools and cases in mathematics, descriptive statistics and probability*. 3 ed., McGraw-Hill Higher Education, New York, 1999.

Websites

[4] Virtual Laboratories in Probability and Statistics: http://www.math.uah.edu/stat/

The concept of statistics

Statistics is a field of science that deals with collecting, presenting and analysing data in order to discover regularities in mass phenomena and to support and improve the quality of decision-making. Statistics can be broadly divided into two areas: descriptive statistics and inferential statistics.

Descriptive statistics deals with methods of describing statistical data collected during a statistical survey – population parameters are counted directly from a data set available to a researcher. The purpose of using the methods of descriptive statistics is to summarise a set of data and draw some basic conclusions and generalisations about a set.
Descriptive statistics is usually used as the first and basic step in analysing collected data.

There are several techniques of descriptive statistics available:
1) tabular description;
2) graphical presentation of results;
3) determination of distribution measures:
    a) measures of position (e.g. quantile), including measures of central tendency, e.g. arithmetic mean, geometric mean, harmonic mean, quadratic mean, median, mode,
    b) measures of variation, e.g. standard deviation, variance, spread, quarterly spread, mean absolute deviation, quarterly deviation, coefficient of variation,
    c) measures of the shape of the distribution, including measures of asymmetry (e.g. coefficient of skewness, coefficient of asymmetry, third central moment) and measures of concentration (e.g. kurtosis).

Inferential statistics deals with the problems of generalising the results of a random sample to the entire population and estimating the errors resulting from such generalisation. This distinguishes statistical inference from the tools of descriptive statistics, which serve only to provide a basic description of the properties of a single sample.

There are two divisions of statistical inference (two groups of methods for generalising results):
1) estimation – estimating the values of unknown parameters of a distribution,
2) statistical hypothesis testing – checking the validity of assumptions about a distribution.

Descriptive statistics

There are several techniques of descriptive statistics available:
1) tabular description;
2) graphical presentation of results;
3) determination of distribution measures:
    a) measures of position (e.g. quantile), including measures of central tendency, e.g. arithmetic mean, geometric mean, harmonic mean, quadratic mean, median, quartiles, mode,
    b) measures of variation, e.g. standard deviation, variance, range, interquartile range, mean absolute deviation, quartile deviation, coefficient of variation,
    c) measures of the shape of the distribution, including measures of asymmetry (e.g. coefficient of skewness, third central moment) and measures of concentration (e.g. kurtosis).

In probability and statistics, mean (also known as ‘expected value’) refers to one measure of the central tendency.

For a data set, the terms 'arithmetic mean', 'mathematical expectation', and sometimes 'average' are used synonymously to refer to a central value of a discrete set of numbers; specifically, the sum of values (x_i) divided by the number of values (n):

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

Median is the number separating the higher half of a data sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from the lowest to the highest value and picking the middle one (e.g. the median of {1, 2, 6, 8, 14} is 6). If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of two middle values (the median of {2, 6, 8, 9} is (6+8)/2 = 7).

The quartiles of a ranked set of data values are three points that divide said data set into four equal groups, each group comprising of a quarter of the data. The first quartile (Q1, lower quartile) is defined as the middle number between the smallest number and the median of a data set. The second quartile (Q2) is the median of data. The third quartile (Q3, upper quartile) is the middle value between the median and the highest value of a data set.

Mode is the value that appears most often in a set of data.

The range R of a set of data is the difference between the largest and smallest values:

$$R = \max_i x_i - \min_i x_i$$

Variability may be represented numerically with the calculation of variance and standard deviation. Standard deviation (also represented by the Greek letter sigma σ or s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A standard deviation that is close to 0 indicates that the data points tend to be very close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.
The formulas for variance and standard deviation depend on whether we are dealing with the entire population or with a sample intended to represent a larger population:

- variance of a population:

$$s^2 = \frac{1}{n}\sum_{i=0}^{n}(x_i - \bar{x})^2$$

- standard deviation of a population:

$$s = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

- variance of a sample:

$$s^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2$$

- standard deviation of a sample:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

Assuming that there is a normal distribution, data values are distributed equally around the mean, decreasing as one moves away from that value, and measured in terms of standard deviations. The standard deviation (square root of the variance) gives insight into the spread of the data through the use of what is known as the '68-95-99.7' rule, also called the 'three sigma' rule. This rule is:

- approximately 68 percent of values will lie within one standard deviation of the mean,
- approximately 95 percent of values will be within two standard deviations of the mean,
- approximately 99.7 percent of values will be within three standard deviations of the mean.

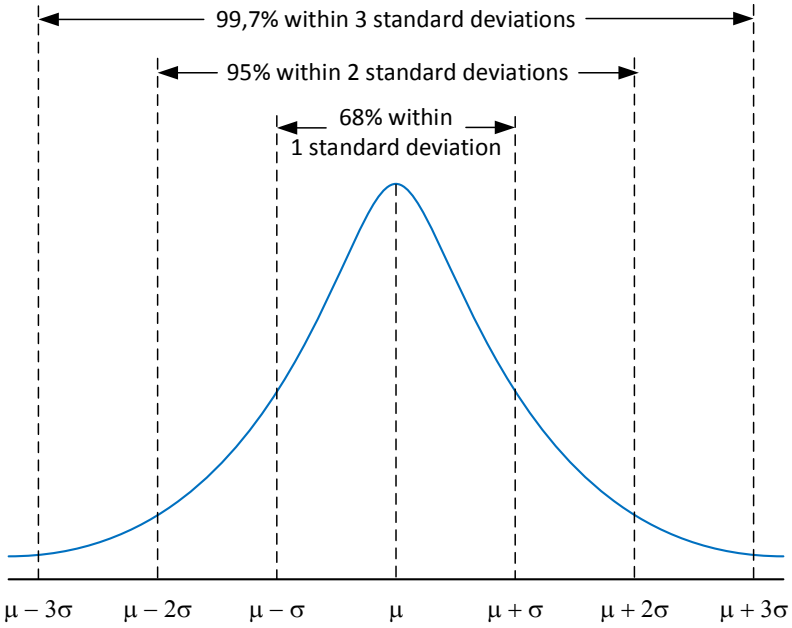


Figure 1. '68-95-99.7' rule: 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.

Coefficient of variation (CV), also known as 'relative standard deviation', is a standardised measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

To interpret the coefficient of variation, the most common assumptions are:

- CV < 25 % – low variation of the feature (low variability),
- 25% ≤ CV < 45% – average variation of the feature (average variability),
- 45% ≤ CV < 100% – high variation of the trait (high variability),
- CV ≥ 100% – very high variation of the feature (very high variability).

Other ranges are also accepted:

- CV < 20 % – low variability,
- 20% ≤ CV < 40% – average variability,
- 40% ≤ CV < 100% – high variability,
- CV ≥ 100% – very high variability.

The taken ranges usually depend on the type of analysed feature and the purpose of analysis.

Interquartile range (IQR), also called 'midspread' or 'middle fifty', is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles:

$$IQR = Q_3 - Q_1$$

Another statistical tool used to measure spread or, in other words, to measure dispersion, is the quartile deviation (QD):

$$QD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}$$

Interquartile range and quartile deviation are not as sensitive to outliers as the range and standard deviation.
Skewness is a statistical measure used to describe the shape of distribution. The coefficient of skewness is a measure of asymmetry in the distribution. Positive skew indicates a longer tail to the right, while negative skew indicates a longer tail to the left. For perfectly symmetric distribution, like normal distribution, the coefficient of skewness is equal to zero.

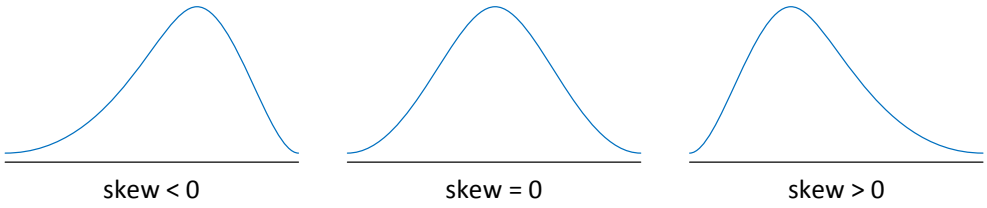


Figure 2.

Kurtosis is another statistical measure used to describe the shape of distribution. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution. Distributions with low kurtosis exhibit tail data that are generally less extreme than the tails of normal distribution.

There are three categories of kurtosis that can be displayed by a set of data. The first category of kurtosis is mesokurtic distribution. This distribution has a kurtosis statistic similar to that of the normal distribution and equal to 0, meaning the extreme value that is characteristic of the distribution is similar to that of a normal distribution.

The second category is leptokurtic distribution for kurtosis higher than 0. This distribution is characterised with long tails (outliers). The outliers stretch the horizontal axis of the histogram graph, making the bulk of the data appear in a narrow vertical range. Thus leptokurtic distributions are sometimes presented as 'concentrated towards the mean'.

The third type of distribution is platykurtic distribution. Extreme values of platykurtic distributions are less than that of the normal distribution.

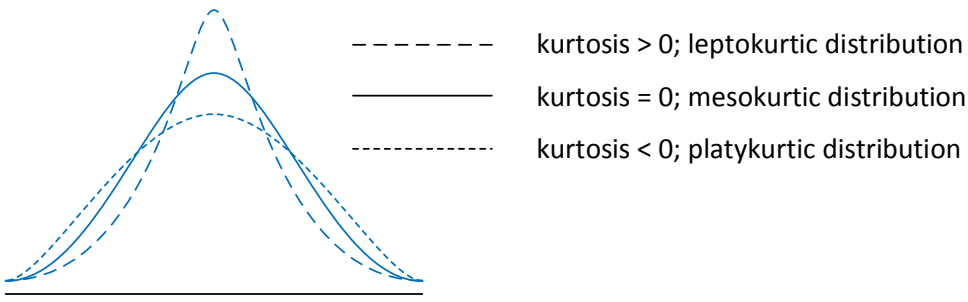


Figure 3.

Graphically depicting groups of numerical data through their quartiles is possible using a box plot method. Box plots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms 'box-and-whisker plot' and 'box-and-whisker diagram'. Outliers, i.e. data points that differ significantly from other observations, may be plotted as individual points in the diagram. An outlier may be due to variability in the measurement or it may indicate experimental error. Usually outliers are defined as observations outside the range:

$$[Q_1 - 1.5(Q_3 - Q_1); Q_3 + 1.5(Q_3 - Q_1)]$$

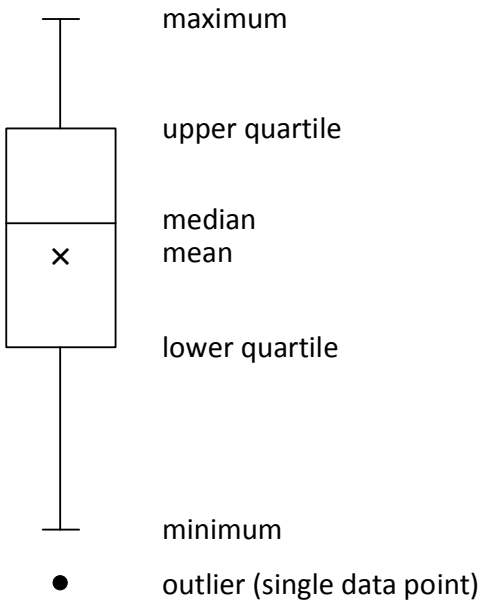


Figure 4. Box plot (box-and-whisker diagram)

Random variable

Random variable is a function defined on the space of elementary events Ω, with values in the real numbers R.
A random variable associated with an experiment is a variable that, as a result of the experiment, takes on a numerical value that depends on chance (cannot be determined before the experiment).

Random variables are denoted by uppercase letters, e.g. X, Y, Z, their values (also called realisations) – by corresponding lowercase letters x, y, z, often with subscripts, e.g. x_1, x_2, x_i. We denote the sets of values of the random variables X, Y, Z by W_X, W_Y, W_Z, respectively.

The following types of random variables can be listed:
1) discrete random variables,
2) continuous random variables.

The set of possible values of a discrete random variable is finite or infinite, but countable. They usually take the values of natural numbers.

The set of possible values of a continuous random variable is recalculable. A continuous random variable is a variable that can take any value from some numerical interval, finite or infinite.

Probability distribution of a random variable

The specific realisations of a random variable (the values taken by the random variable) are random events – so their probabilities can be determined. The function that assigns particular probabilities to particular realisations of a random variable X is the probability distribution function of the variable X.
For a discrete variable, this can be presented as follows:

$$P(X = x_i) = p_i \qquad (i = 1, 2, \ldots, n)$$

whereby:

$$0 \leq p_i \leq 1$$

and

$$\sum_{i=1}^{n} p_i = 1$$

Thus, a random variable X can be described by a sequence of pairs of numbers (x_1, p_1), (x_1, p_1),..., (x_n, p_n), where the first element denotes the possible value of the random variable and the second element denotes the probability of its realisation.

For a continuous random variable, the equivalent of distribution function $P(X = x_i)$ is the density function f(x) satisfying the conditions:

$$f(x) \geq 0$$

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

The distribution of any random variable is characterised by at least two parameters: the expected value of the random variable and the variance.

Example task – descriptive statistics

Task content
The table contains the data of the production volume and the number of man-hours within one month in 10 large factories of the shoe industry.

Factory number (i)	1	2	3	4	5	6	7	8	9	10
Number of man-hours [in thousands] (X)	100	127	150	168	184	182	185	240	248	256
Production volume [in thousands] (Y)	25	26	30	26	31	35	27	41	38	51

For both variables X and Y determine: mean (average), median, first quartile (lower quartile), third quartile (upper quartile), mode, range, variance, standard deviation, coefficient of variation, interquartile range and quartile deviation. Draw the box-and-whisker diagrams. Interpret the results.

Solution

Descriptive statistics can be calculated using formulas given above in the section *Descriptive statistics*. The easiest way is to use appropriate functions in Excel. The data sheet and formulas for each statistic are shown in the figure.

	A	B	C	D	E	F	G
1	Factory number (i)	Number of man-hours [in thousands] (X)	Production volume [in thousands] (Y)		Statistics	Number of man-hours [in thousands] (X)	Production volume [in thousands] (Y)
2	1	100	25		mean	=AVERAGE(B2:B11)	
3	2	127	26		median	=MEDIAN(B2:B11)	
4	3	150	30		lower quartile Q1	=QUARTILE.INC(B2:B11;1)	
5	4	168	26		upper quartile Q3	=QUARTILE.INC(B2:B11;3)	
6	5	184	31		mode	=MODE.SNGL(B2:B11)	
7	6	182	35		range	=MAX(B2:B11)-MIN(B2:B11)	
8	7	185	27		variance	=VAR.S(B2:B11)	
9	8	240	41		standard deviation	=STDEV.S(B2:B11)	
10	9	248	38		coefficient of variation	=F9/F2	
11	10	256	51		interquartile range	=F5-F4	
12					quartile deviation	=F11/2	

Figure 5.

The use of each function is similar. For example, to determine the mean, insert the statistical function 'mean' and then select the data series.

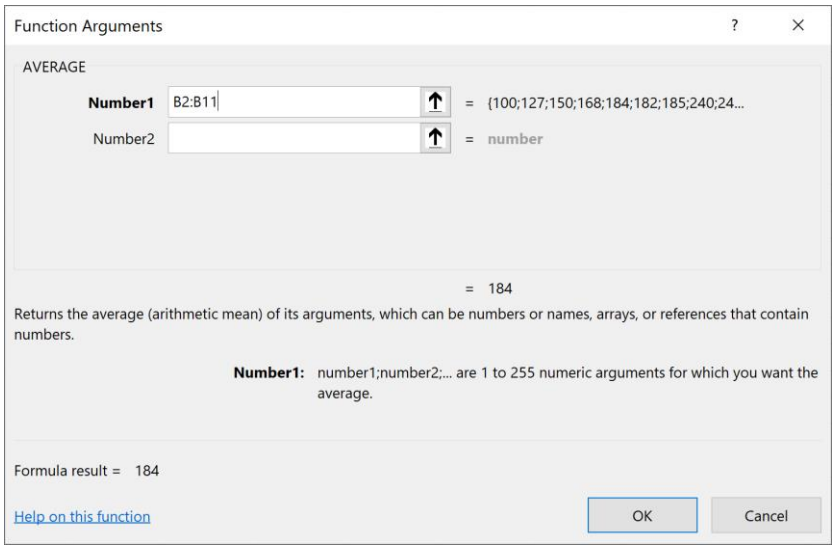


Figure 6.

The figure below shows the results obtained.

	A	B	C	D	E	F	G
1	Factory number (i)	Number of man-hours [in thousands] (X)	Production volume [in thousands] (Y)		Statistics	Number of man-hours [in thousands] (X)	Production volume [in thousands] (Y)
2	1	100	25		mean	184	33
3	2	127	26		median	183	30,5
4	3	150	30		lower quartile Q1	154,5	26,25
5	4	168	26		upper quartile Q3	226,25	37,25
6	5	184	31		mode	#N/A	26
7	6	182	35		range	156	26
8	7	185	27		variance	2682	69,78
9	8	240	41		standard deviation	51,79	8,35
10	9	248	38		coefficient of variation	28%	25%
11	10	256	51		interquartile range	71,75	11
12					quartile deviation	35,88	5,50

Figure 7.

Some statistics (including skewness and kurtosis) can be calculated using MS Excel Data Analysis. After preparing a worksheet and entering the data, choose *Data* → *Data Analysis* → *Descriptive Statistics* (if Data Analysis is not available, we have to activate it: *File* → *Options* → *Add-ins* → *Manage Excel Add-ins* → *Go* → *Analysis ToolPak*). The window with entered data of Descriptive Statistics is shown in the figure below.

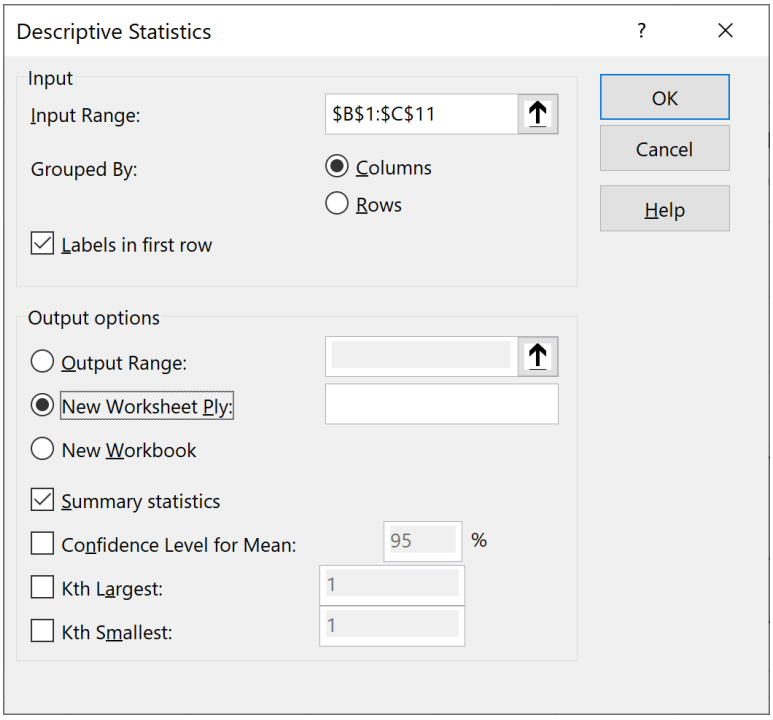


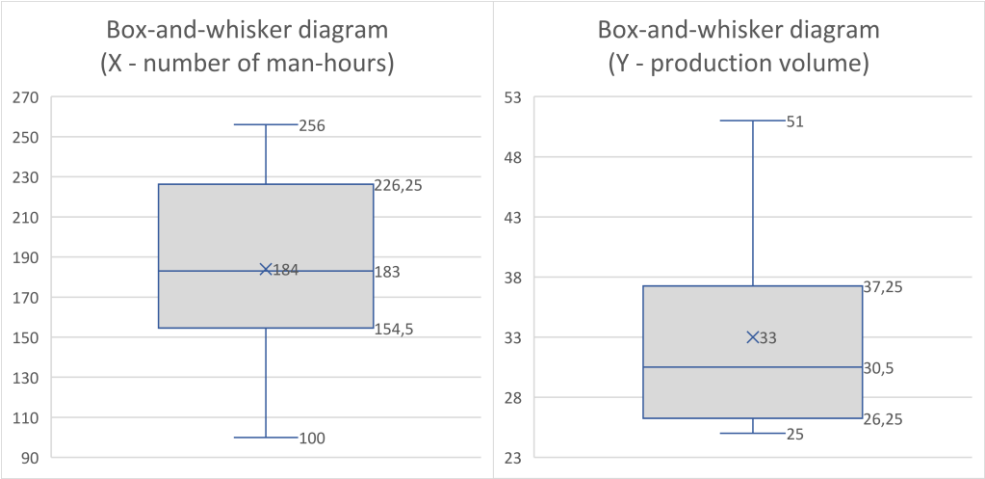
Figure 8.

Finally, we get the set of statistics for both variables.

	A	B	C	D
1	*Number of man-hours [in thousands] (X)*		*Production volume [in thousands] (Y)*	
2	Mean	184	Mean	33
3	Standard Error	16,38	Standard Error	2,64
4	Median	183	Median	30,5
5	Mode	#N/A	Mode	26
6	Standard Deviation	51,79	Standard Deviation	8,35
7	Sample Variance	2682	Sample Variance	69,78
8	Kurtosis	-0,84	Kurtosis	1,04
9	Skewness	-0,01	Skewness	1,20
10	Range	156	Range	26
11	Minimum	100	Minimum	25
12	Maximum	256	Maximum	51
13	Sum	1840	Sum	330
14	Count	10	Count	10

Figure 9.

MS Excel also allows us to generate box-and-whisker diagrams.



After analysing the results obtained, it can be concluded that both variables are characterised by average variability, however the presented factories are characterised by slightly greater variation in terms of the number of man-hours (X, coefficient of variation of X variable is equal to 28%) than the production volume (CV(Y) = 25%).
The distribution of variable X is almost symmetric (slightly negative skewness), while the distribution of variable Y is right skewed (positive skewness). It is also visible on the box-and-whisker diagrams.
The distribution of variable X is a platykurtic distribution (negative kurtosis), while the distribution of variable X is a leptokurtic distribution (concentrated toward the mean, positive kurtosis).
There is no mode in the case of variable X, whereas the most frequent value for variable Y is 26 thousand pairs of shoes.
The median for variable X is equal to 183 thousand, which means that in half of the factories the monthly number of man-hours does not exceed 183 thousand, while in half of them it is not less than 183 thousand of hours.
The lower and upper quartile values can be interpreted similarly. The lower quartile is 154.5 thousand, which means that in 25% of the factories the number of man-hours is no more than 154.5 thousand, while in 75% the number of hours is not less than 154.5 thousand. The upper quartile is equal to 226.25 thousand, which means that in 75% of the factories the number of man-hours is equal to no more than 226.25 thousand, while in 25% the number of hours is not less than 226.25 thousand.
The median and quartiles for variable Y are interpreted in the same way.
No outliers are observed for both variables. None of the values of variable X fall outside the range of [46.875; 333.875]. Similarly for variable Y – there is no value out of the range [9.75; 53.75]. The extreme values of the variable X are far from the ends of the specified interval. Whereas, for variable Y, the maximum value is close to the upper value of the given range. Therefore the kurtosis for variable Y is much higher than for variable X.

Example of a task – discrete random variable

Task content

There are 200 tickets prepared for a cash lottery, including two winning tickets for €1000, eight for €500, ten for €200, twenty for €100 and sixty for €10. The rest of the tickets are empty. A random variable X means the amount won in the lottery. Present the distribution of a random variable X. Determine: expected value, variance, standard deviation.

Solution – probability distribution of a random variable

The random variable is a discrete variable taking the values: 0, 10, 100, 200, 500, 1000. From the content of the task, it follows that exactly half of the lottery tickets are empty. Therefore, the probabilities of individual realisations of the random variable can be determined as follows:

$$P(X = 0) = \frac{100}{200} = \frac{50}{100} = 0.50$$

$$P(X = 10) = \frac{60}{200} = \frac{30}{100} = 0.30$$

$$P(X = 100) = \frac{20}{200} = \frac{10}{100} = 0.10$$

$$P(X = 200) = \frac{10}{200} = \frac{5}{100} = 0.05$$

$$P(X = 500) = \frac{8}{200} = \frac{4}{100} = 0.04$$

$$P(X = 1000) = \frac{2}{200} = \frac{1}{100} = 0.01$$

Probability distribution of a random variable:

x_i	0	10	100	200	500	1000
p_i	$\frac{50}{100}$	$\frac{30}{100}$	$\frac{10}{100}$	$\frac{5}{100}$	$\frac{4}{100}$	$\frac{1}{100}$

Solution – expected value, variance, standard deviation

The expected value (average value, mean) for a discrete variable is determined from the formula:

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

$$E(X) = \sum_{i=1}^{6} x_i p_i = 0 \times \frac{50}{100} + 10 \times \frac{30}{100} + 100 \times \frac{10}{100} + 200 \times \frac{5}{100} + 500 \times \frac{4}{100} + 1000 \times \frac{1}{100}$$
$$= 53$$

The variance is determined from the formula:

$$D^2(X) = Var(X) = \sum_{i=1}^{n} \big(x_i - E(X)\big)^2 p_i$$

$$D^2(X) = \sum_{i=1}^{6} \big(x_i - E(X)\big)^2 p_i$$

$$= (0-53)^2 \times \frac{50}{100} + (10-53)^2 \times \frac{30}{100} + (100-53)^2 \times \frac{10}{100}$$

$$+ (200-53)^2 \times \frac{5}{100} + (500-53)^2 \times \frac{4}{100} + (1000-53)^2 \times \frac{1}{100} = 20221$$

or

$$D^2(X) = E^2(X) - \big(E(X)\big)^2 = \sum_{i=1}^{n} x_i^2 p_i - \big(E(X)\big)^2$$

$$D^2(X) = E^2(X) - \big(E(X)\big)^2 = \sum_{i=1}^{6} x_i^2 p_i - \big(E(X)\big)^2$$

$$0^2 \times \frac{50}{100} + 10^2 \times \frac{30}{100} + 100^2 \times \frac{10}{100} + 200^2 \times \frac{5}{100} + 500^2 \times \frac{4}{100}$$

$$+1000^2 \times \frac{1}{100} - 53^2 = 23030 - 2809 = 20221$$

Standard deviation:

$$D(X) = \sqrt{D^2(X)}$$

$$D(X) = \sqrt{20221} \approx 142.2$$

Example of a task – continuous random variable (uniform distribution)

Task content

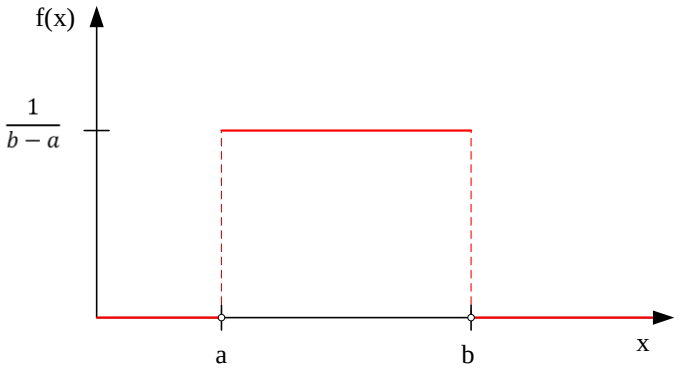
A bus departs from a bus stop every 10 minutes. The waiting time for a passenger at the bus stop is a random variable with a uniform distribution.

1) Determine the density function of the random variable.
2) Determine the distribution of the random variable.
3) Determine the expected value, variance, and standard deviation of the random variable.
4) Calculate the probability that a passenger will wait:
   a) less than 3 minutes,
   b) no less than 5 but also no more than 7 minutes,
   c) no less than 8 minutes.

Solution

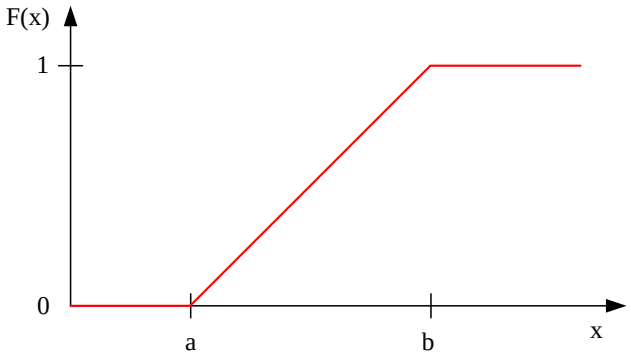
We are dealing with a continuous uniform distribution in the interval [a; b], where a and b are the smallest and largest values of the random variable, respectively.
The general form of the density function of a variable described by a uniform distribution:

$$f(x) = \begin{cases} 0 & for\ x < a \\ \dfrac{1}{b-a} & for\ a \le x \le b \\ 0 & for\ x > b \end{cases}$$



The general form of the distribution of the variable:

$$F(x) = \begin{cases} 0 & for\ x < a \\ \dfrac{x-a}{b-a} & for\ a \le x \le b \\ 1 & for\ x > b \end{cases}$$



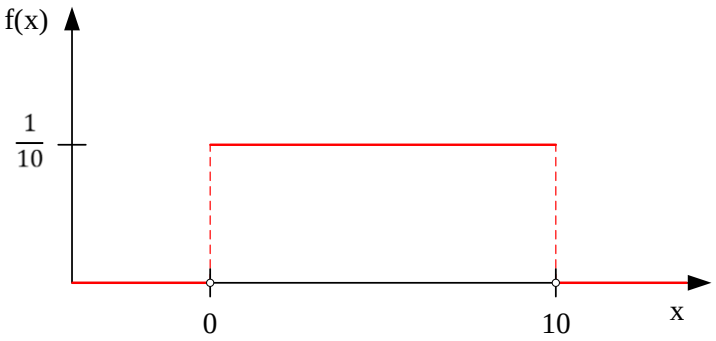
The expected value EX and the variance D^2X (VarX) of the random variable for a uniform distribution:

$$EX = \frac{a+b}{2}$$
$$D^2X = \frac{(b-a)^2}{12}$$

1) Density function of the random variable

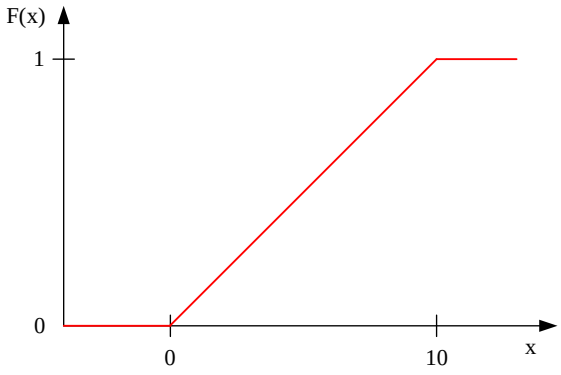
We are dealing with a uniform distribution in the interval [0; 10].

$$f(x) = \begin{cases} 0 & for\ x < 0 \\ \frac{1}{10} & for\ 0 \le x \le 10 \\ 0 & for\ x > 10 \end{cases}$$



2) Distribution of the random variable

$$f(x) = \begin{cases} 0 & for\ x < 0 \\ \frac{1}{10} & for\ 0 \le x \le 10 \\ 0 & for\ x > 10 \end{cases}$$



3) The expected value EX, variance D^2X and standard deviation DX of the random variable

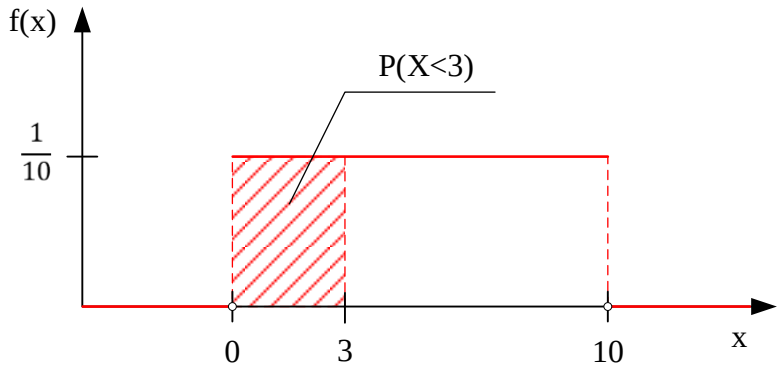
$$EX = \frac{0 + 10}{2} = 5$$

$$D^2X = \frac{(10 - 0)^2}{12} \approx 8.33$$

$$DX = \sqrt{D^2X} \approx \sqrt{8{,}33} \approx 2.89$$

4a) Probability that a passenger will wait less than 3 minutes

The measure of the probability sought is the highlighted area under the density function:

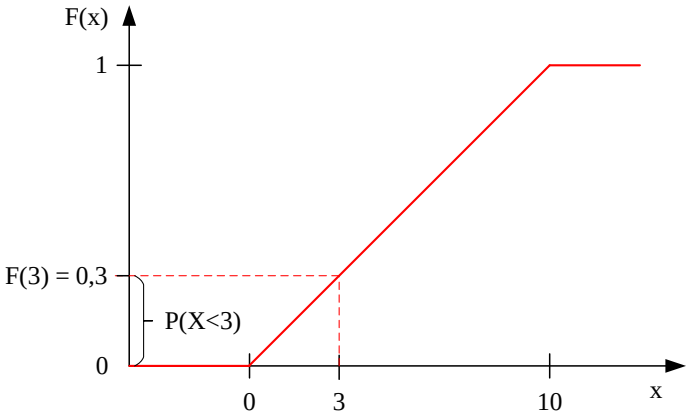


$$P(X < 3) = \int_0^3 \frac{1}{10} dx = \left[\frac{1}{10}x\right]_0^3 = \frac{3}{10}$$

The probability sought can also be determined as the area of the marked rectangle:

$$P(X < 3) = 3 \times \frac{1}{10} = \frac{3}{10}$$

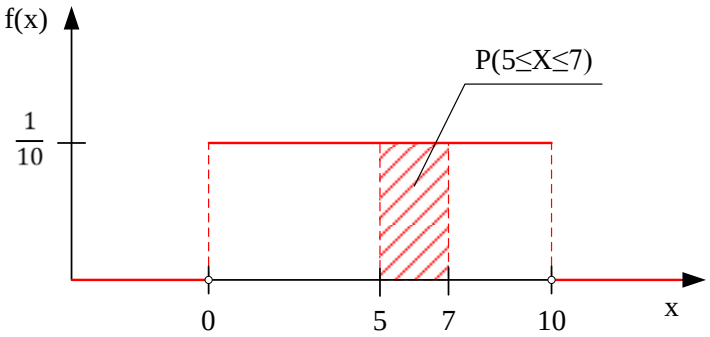
Also the distribution of the random variable can be used:



$$P(X < 3) = F(3) = \frac{1}{10} \times 3 = \frac{3}{10}$$

4b) Probability that a passenger will wait no less than 5 but also no more than 7 minutes

The measure of the probability sought is the highlighted area under the density function:

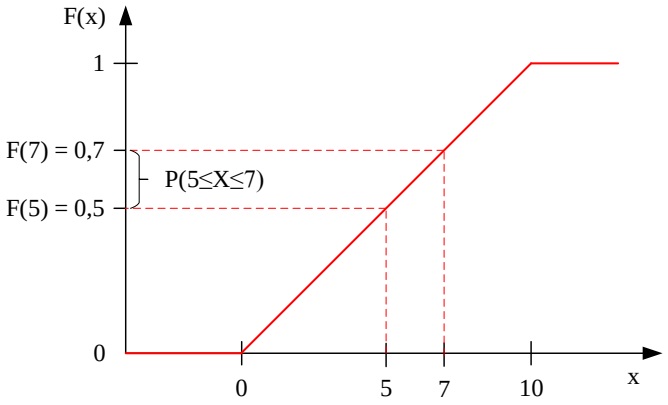


$$P(5 \le X \le 7) = \int_5^7 \frac{1}{10} dx = \left[\frac{1}{10}x\right]_5^7 = \frac{1}{10} \times 7 - \frac{1}{10} \times 5 = \frac{7}{10} - \frac{5}{10} = \frac{2}{10}$$

The probability sought as the area of the marked rectangle:

$$P(5 \le X \le 7) = 2 \times \frac{1}{10} = \frac{2}{10}$$

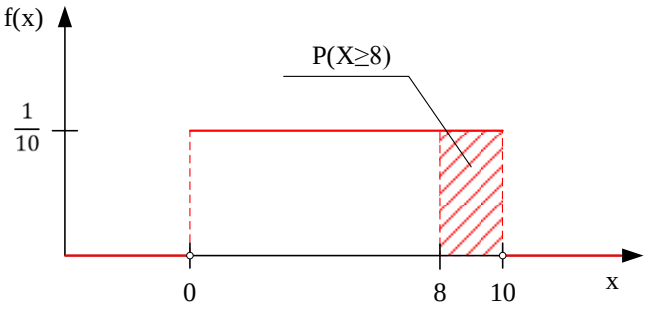
Using the distribution of the variable:



$$P(5 \le X \le 7) = F(7) - F(5) = \frac{7}{10} - \frac{5}{10} = \frac{2}{10}$$

4c) Probability that a passenger will wait no less than 8 minutes

The measure of the probability sought is the highlighted area under the density function:

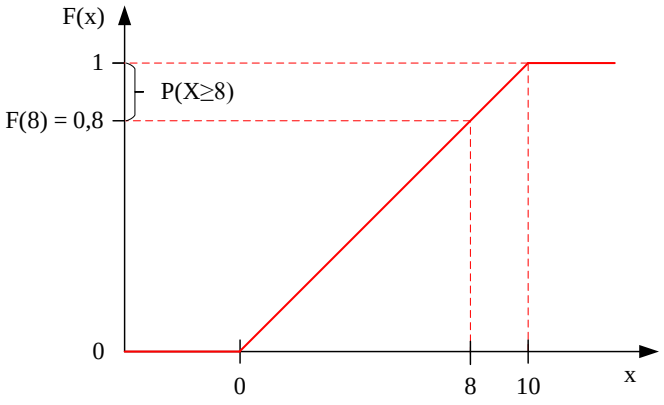


$$P(X \ge 8) = \int_8^{10} \frac{1}{10} dx = \left[\frac{1}{10}x\right]_8^{10} = \frac{1}{10} \times 10 - \frac{1}{10} \times 8 = 1 - \frac{8}{10} = \frac{2}{10}$$

The probability sought as the area of the marked rectangle:

$$P(X \geq 8) = 2 \times \frac{1}{10} = \frac{2}{10}$$

Also the distribution of the random variable can be used:



$$P(X \geq 8) = 1 - F(8) = 1 - \frac{1}{10} \times 8 = 1 - \frac{8}{10} = \frac{2}{10}$$

Example of a task – continuous random variable (normal distribution)

Task content

The diameter of a machine shaft is equal to $\emptyset 110^{+0,1}_{-0,2} mm$. The distribution of the random variable denoting the diameter follows a normal distribution with an expected value of 109.98 mm and a standard deviation of 0.05 mm. Calculate the probability of:

1) receiving a product that conforms to the quality requirements,
2) receiving a product which does not comply with the quality requirements,
3) exceeding the upper tolerance limit,
4) exceeding the lower tolerance limit.

Solution

X – random variable associated to the machine shaft diameter
μ = 109.98 mm (mean, expected value)
σ = 0.05 mm (standard deviation)
X ~ N(109.98; 0.05)

The tolerance limits are:

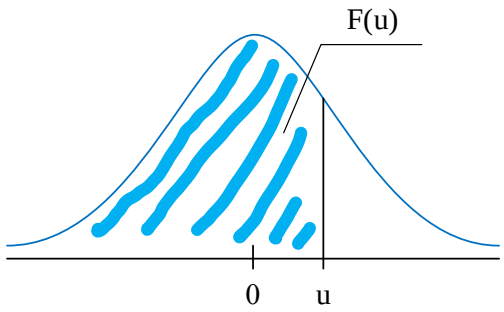
- upper tolerance limit UTL = 110 + 0.1 = 110.1 mm,
- lower tolerance limit LTL = 110 - 0.2 = 109.8 mm.

Variable X ~ N(109.98; 0.05) will be standardised to the variable U ~ N(0; 1):

$$u = \frac{x - \mu}{\sigma} = \frac{x - 109.98}{0.05}$$

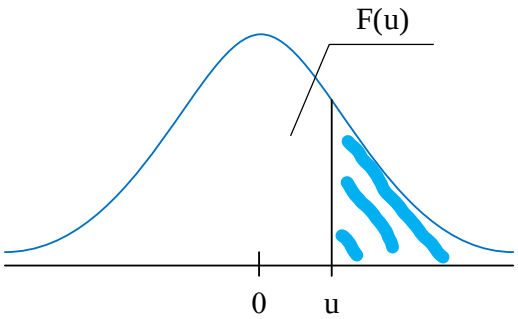
Standardising the variable allows us to use the table of the standard normal distribution.

F(u) – the distribution of a variable *U* at a point *u* is the probability that the variable *U* takes a value less than or equal to *u*:



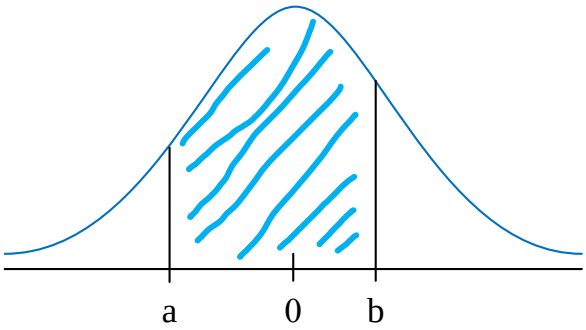
$$F(u) = P(U \leq u)$$

To determine the probability P(U ≥ u) we will use the relationship (the entire field under the curve of the normal distribution density function, or Gauss curve, is equal to 1):



$$P(U \geq u) = 1 - F(u)$$

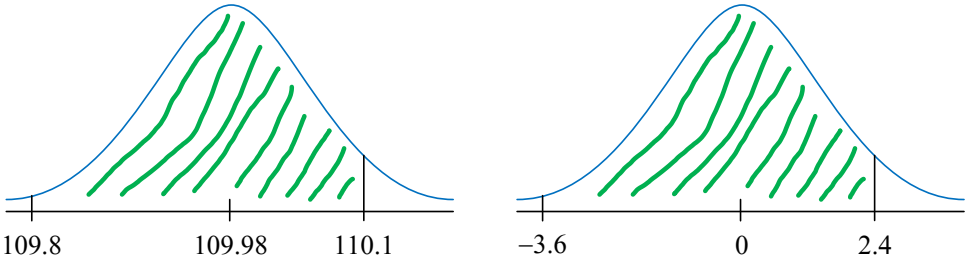
To determine the probability P(a < U < b) (equations may also be ≤), we will use the relation:



$$P(a < U < b) = F(b) - F(a)$$

1) Probability of receiving a product that conforms to the quality requirements

A product that complies with the quality requirements is one which parameters are within the accepted tolerance limits.



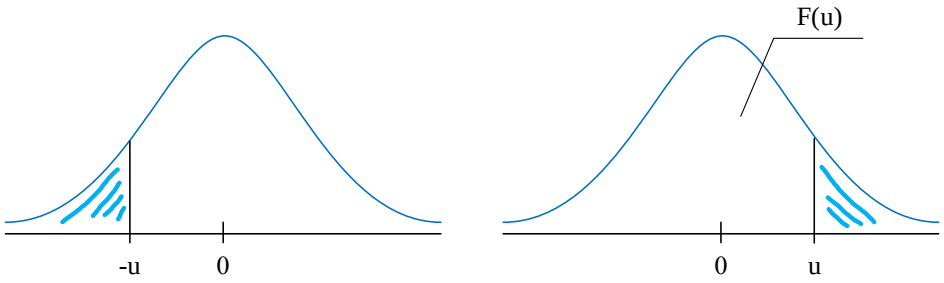
$$P(109.8 \le X \le 110.1) = P\left(\frac{109.8 - 109.98}{0.05} \le U \le \frac{110.1 - 109.98}{0.05}\right) = P(-3.6 \le U \le 2.4)$$
$$= F(2.4) - F(-3.6) = F(2.4) - [1 - F(3.6)] = F(2.4) - 1 + F(3.6)$$
$$= 0.9918 - 1 + 0.9998 = 0.9916$$

The diameter can be expected to be within tolerance limits for approximately 99.16% of all products. The value of F(2.4) is read from the standard normal distribution table. In tasks rounding to 4 decimal places is usually adopted.

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
					...					
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

u = 2.4 **F(2.4)**

Probabilities shown in the table provide the area under the Gauss curve to the left from *u*. The standard normal distribution table contains only probability values for non-negative variables $u \ge 0$. Therefore, for negative variables, we use the properties of the normal distribution, namely its symmetry with respect to the mean. The fields (probabilities) in the following graphs are equal.



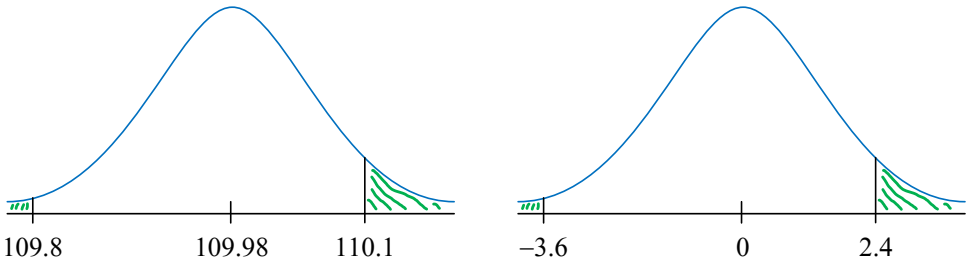
$$P(X \leq -u) = F(-u) \qquad P(X \geq u) = 1 - F(u)$$

Thus, in the case of a negative *u* variable, we use the relationship:

$$F(-u) = 1 - F(u)$$

2) Probability of receiving a product which does not comply with the quality requirements

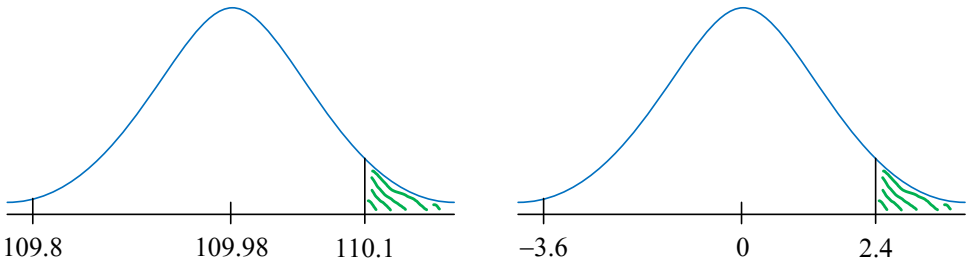
The parameters of a nonconforming product are not within the accepted tolerance limits. Thus, we use the probability of the opposite event to the event specified in (1).



$$1 - P(109.8 \leq X \leq 110.1) = 1 - 0.9916 = 0.0084$$

For approximately 0.84% of all products, the diameter will be outside the tolerance limits.

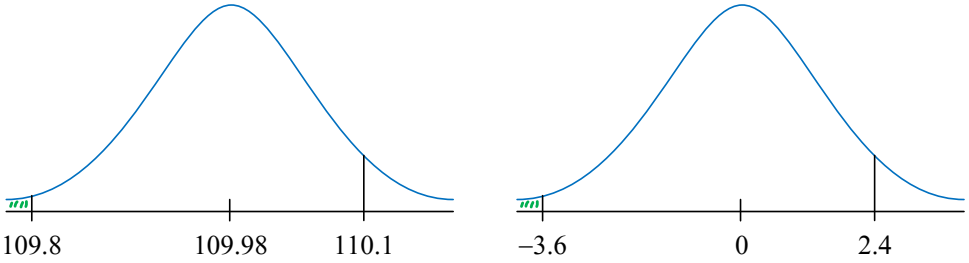
3) Probability of exceeding the upper tolerance limit



$$P(X > 110.1) = P\left(U > \frac{110.1 - 109.98}{0.05}\right) = P(U > 2.4) = 1 - F(2.4) = 1 - 0.9918 = 0.0082$$

For about 0.82% of all products, the diameter will be too large.

4) Probability of exceeding the lower tolerance limit.



$$P(X < 109.8) = P\left(U < \frac{109.8 - 109.98}{0.05}\right) = P(U < 3.6) = F(-3.6) = 1 - F(3.6)$$
$$= 1 - 0.9998 = 0.0002$$

For about 0.02% of all products, the diameter will be too small.

The calculations in (2) and (3) may also be used to determine the probability of exceeding the lower tolerance limit. If the probability of exceeding the tolerance limits on both sides (2) is subtracted from the probability of exceeding the upper tolerance limit (3), the probability of exceeding the lower tolerance limit is obtained:

$$P(X < 109.8) = 0.0084 - 0.0082 = 0.0002$$

Control tasks

1. A coin is flipped three times. Let us assume, that a random variable X is the number of times the coin comes up heads. Determine the probability distribution of the random variable X and its expected value, variance, and standard deviation.
2. Male height was measured among a student population of a selected university. A random variable expressing student height was found to have a normal distribution with an expected value of 175 cm and a standard deviation of 5 cm.
   Calculate the probability that the height of a randomly encountered student:
   a) is smaller than 180 cm,
   b) is smaller than 165 cm,
   c) is larger than 182 cm,
   d) is larger than 171 cm,
   e) belongs to the interval (180; 188),
   f) belongs to the interval (165; 168),
   g) belongs to the bracket (174; 180),
   h) belongs to the intervals: mean ±1σ; ±2σ; ±3σ.

Lecturer:
Mariusz Kołosowski, PhD Eng., Professor at University of Applied Sciences in Nysa

e-mail address: mariusz.kolosowski@pwsz.nysa.pl