Biostatistics

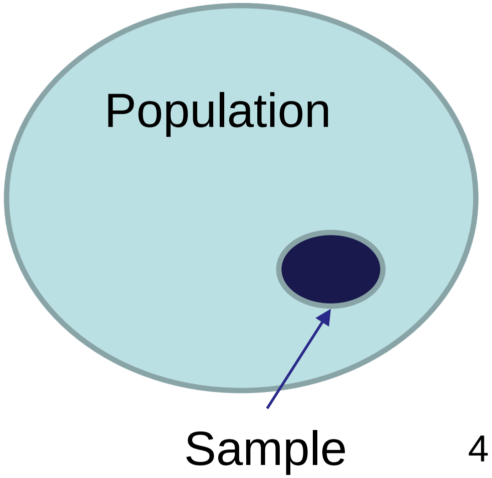
Contents

- Data ……………………………………..3
- Quantitative data ..………………………6
- Variables……………………………….9
- Probability distributions……..…………13
- Measures of central tendency…………21
- Measures of dispersion.………………..33
- Effect size………………………….…40
- Confidence interval………………………43
- Statistical hypothesis…………………..53
- Parametric tests…………………….67
- Nonparametric tests…..……………84
- Association analysis ……………….......101
- Risk and odds………………………...111
- Qualitative data……………………....117
- Supplementary materials……………….....135

Data

Data

- Characterize the observations of one or more variables
- Are obtained from a **sample** representing a **population**
- They have different forms



Data

General population - a set of elements having at least one property common to all its elements qualifying them to this set and at least one property because of which the elements of this set may differ from each other

Quantitative data

Quantitative data

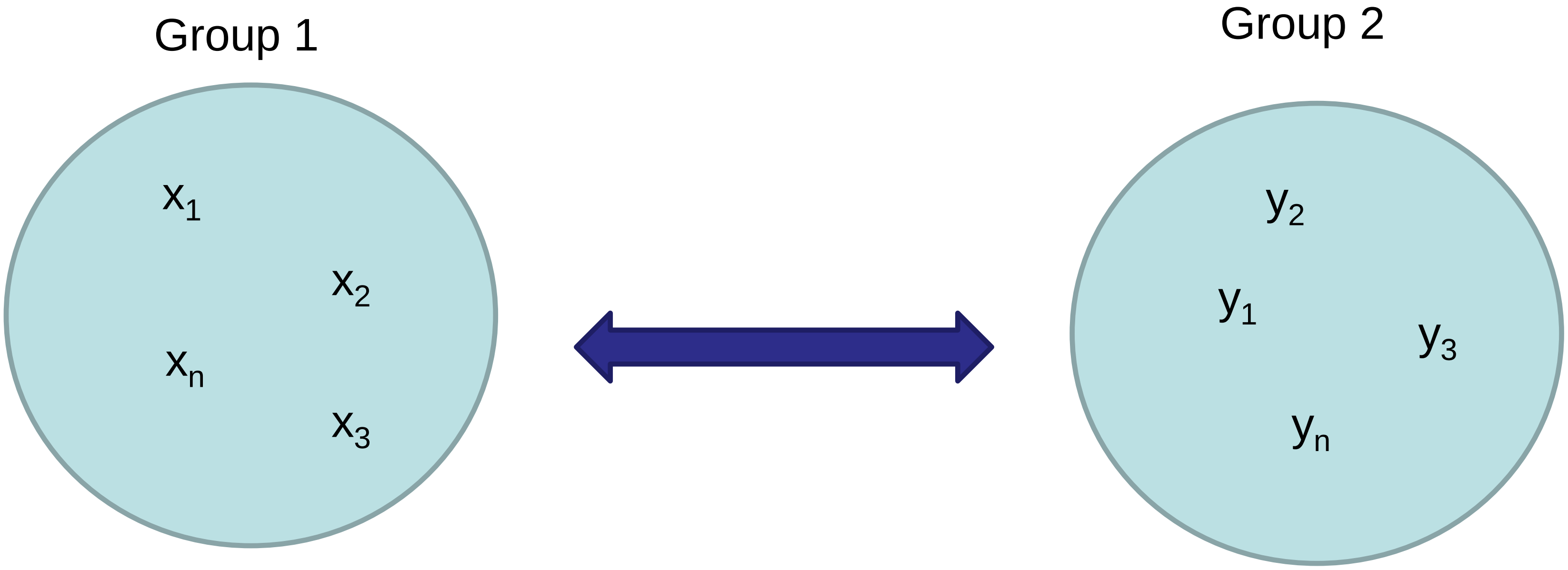
Quantitative data - a feature that is a variable takes numerical values expressed in units on a specific scale

Quantitative data

- **Continuous data** - the variable feature takes any numerical value
- **Discrete data –** the variable feature takes integer values
- **Proportions** – may be used in certain situations, e.g. BMI

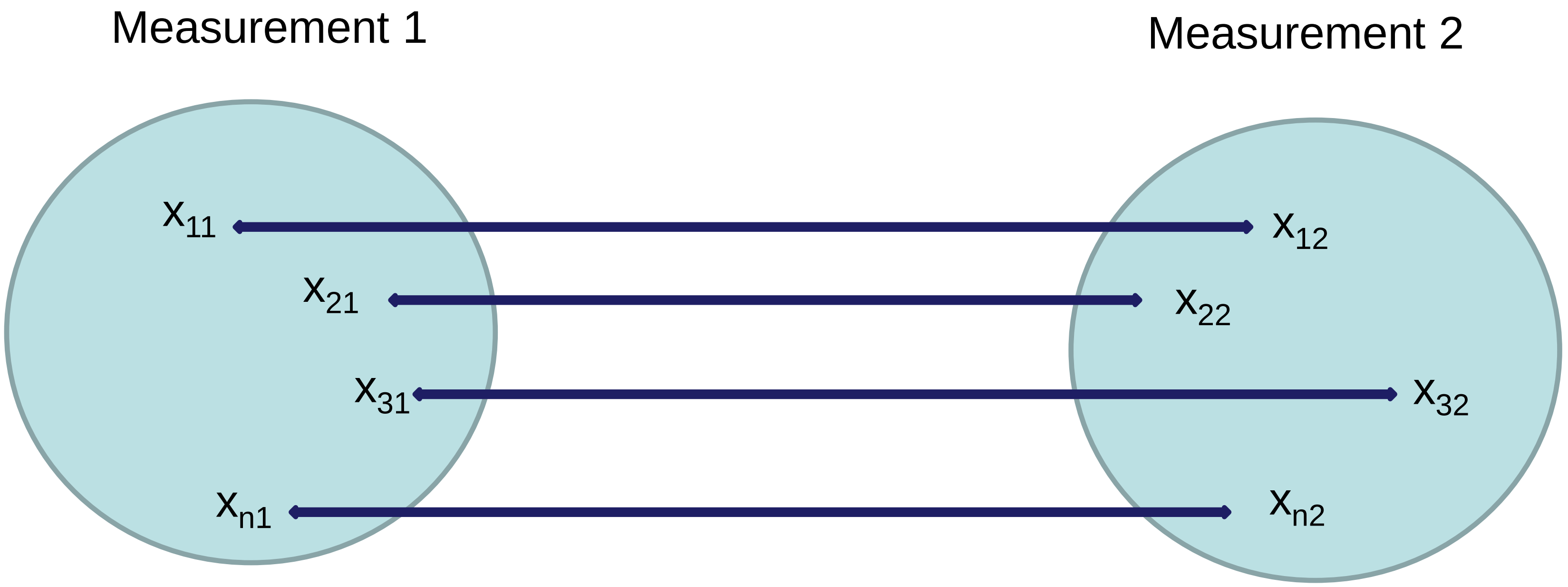
Variables

Unpaired Variables



e.g. we give two different drugs to lower the blood pressure for two groups of patients and we look at the level of lowering the blood pressure

Paired Variables



Most often, "before-after" situations, e.g. blood pressure measurement before and after the treatment, etc.

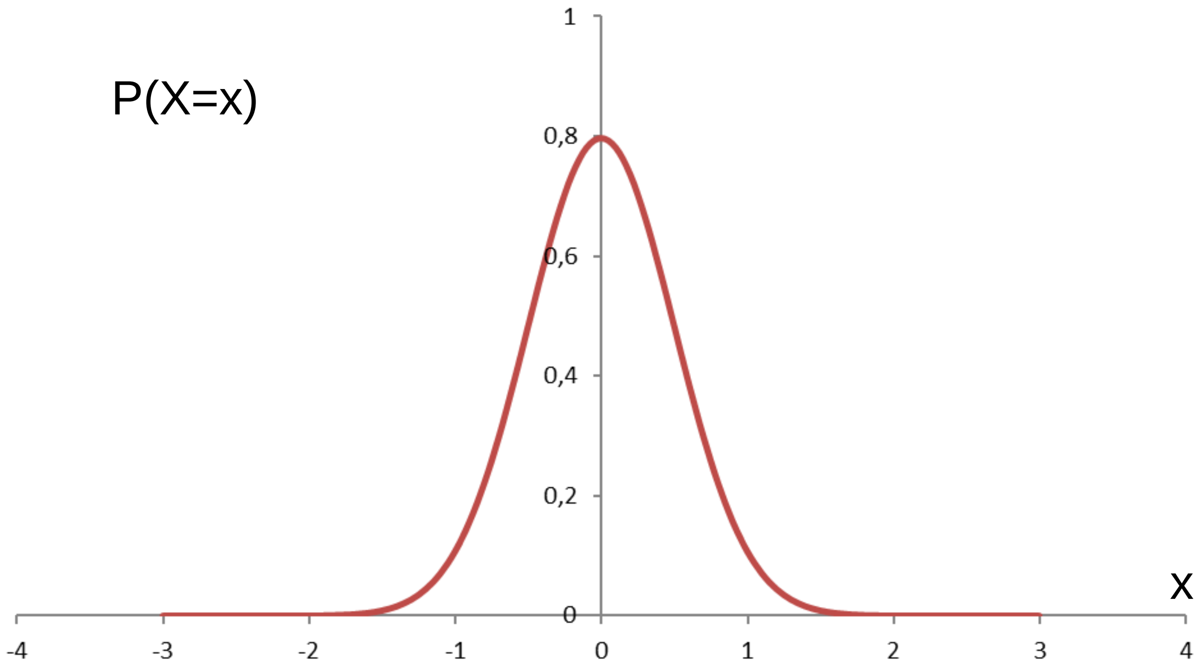
Outliers

They differ greatly from most of the data and are inconsistent with the rest of the data. They can be real, but they can also be the result of a wrong measurement, e.g. woman fr height 204 cm. Before analyzing such the reliability of the data should be confirmed checking the source data.

Probability distributions

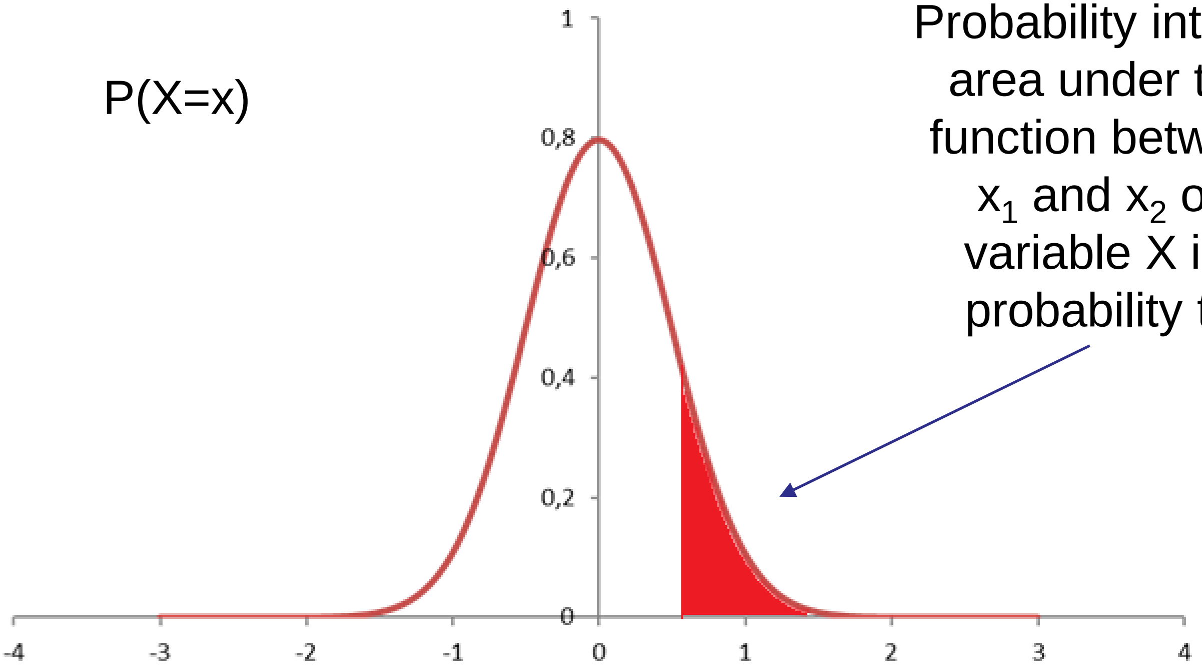
Probability distributions

The probability distribution P(X=x) of a given random variable X is a function assigning to X the probability P that X takes the value x (X=x).



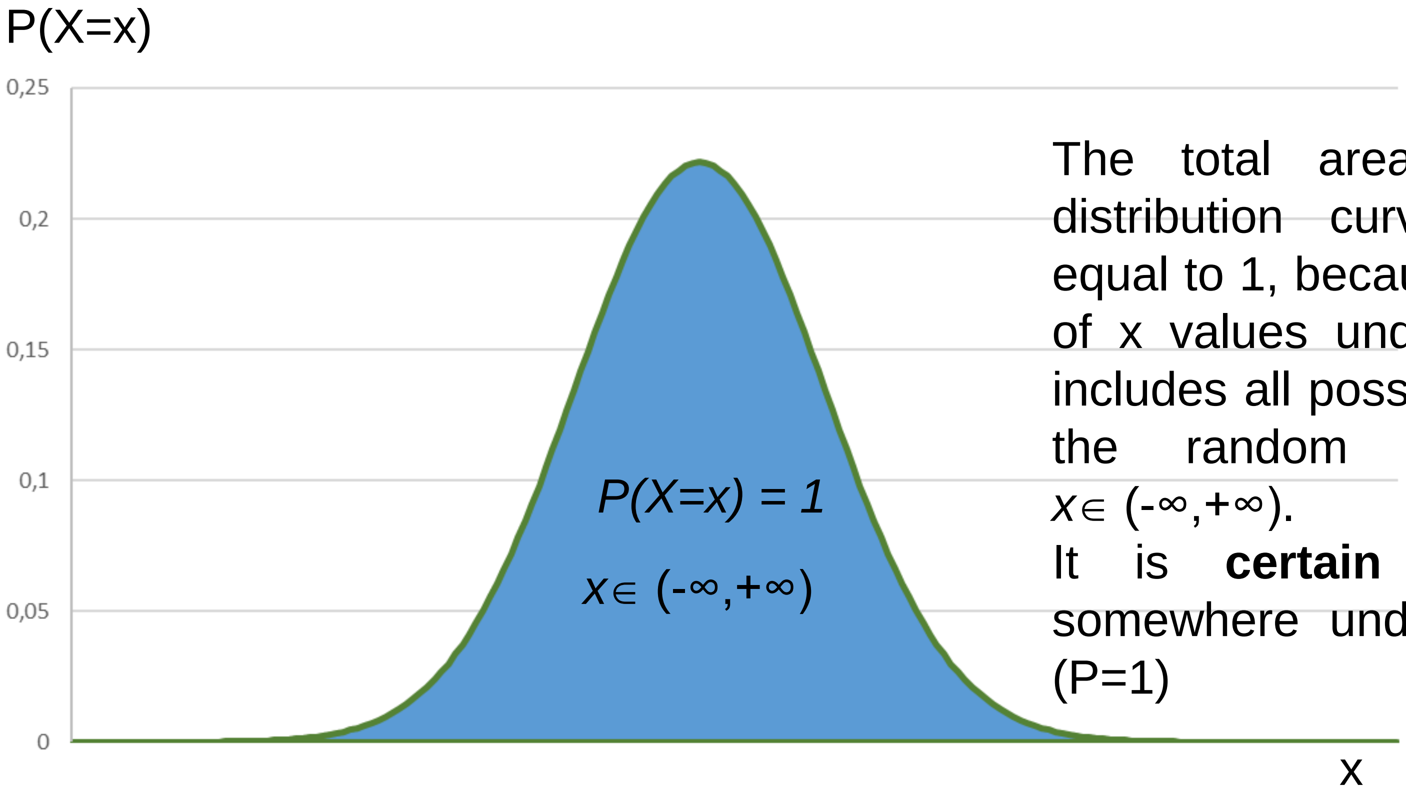
Probability distributions

The probability distribution P(X=x) of a given random variable X is a function assigning to X the probability P that X takes the value x (X=x).



Probability interpretation - the area under the distribution function between the values x_1 and x_2 of the random variable X is equal to the probability that $x_1 < X < x_2$

Probability distributions

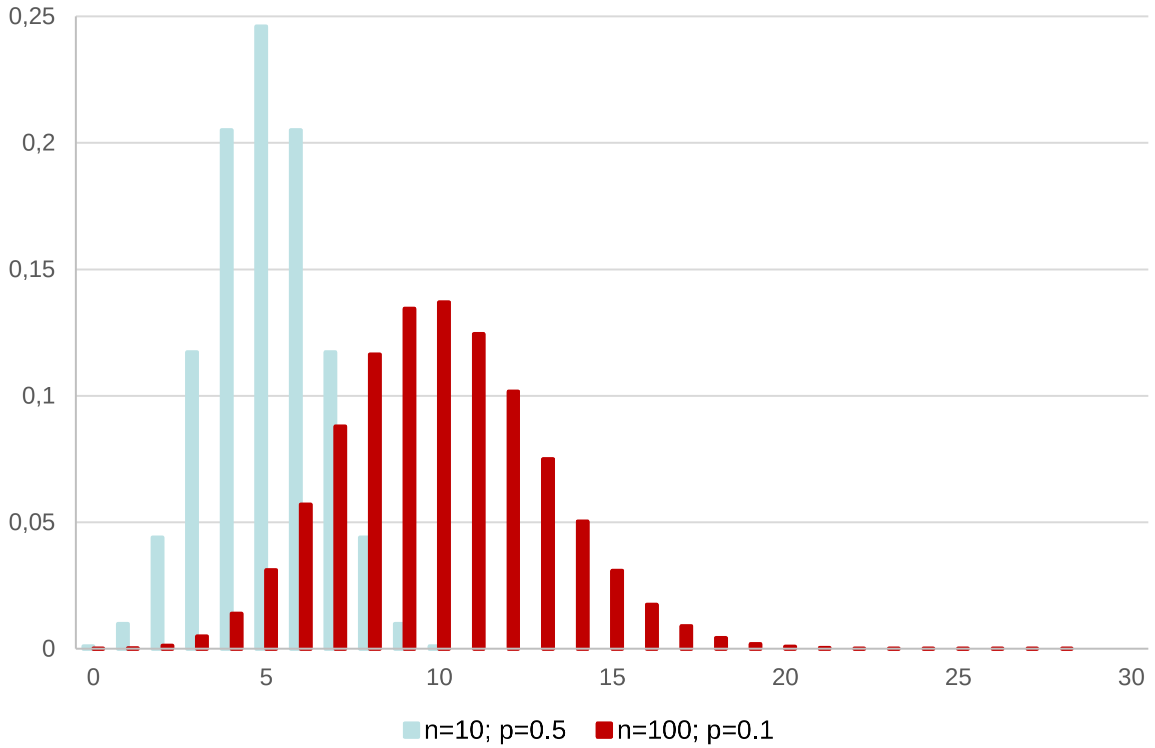


The total area under the distribution curve must be equal to 1, because the range of x values under the curve includes all possible values of the random variable X: *x*∈ (-∞,+∞).
It is **certain** that X is somewhere under the curve (P=1)

Probability distributions

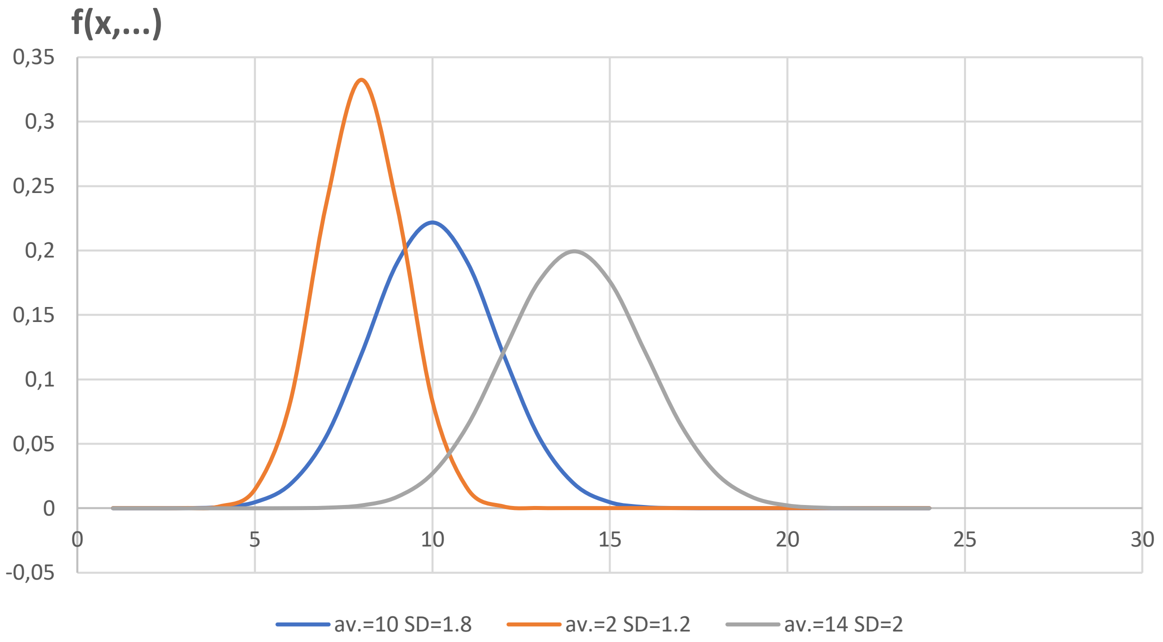
Binomial - A discrete probability distribution that describes the number of successes k over N independent trials, each of which has a constant success probability of p.

The coin toss is a very good example

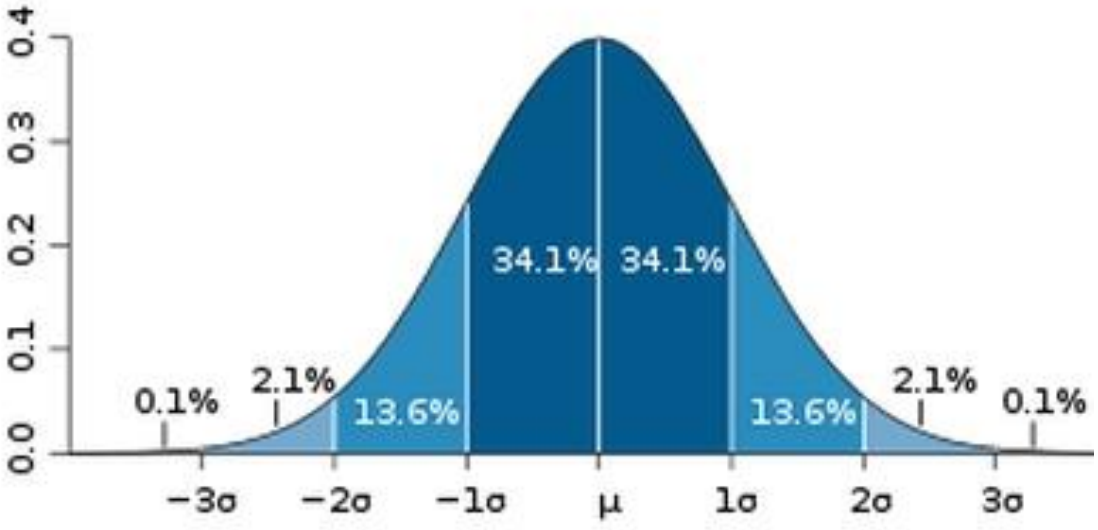


Probability distributions

Normal Distribution - the most intuitive symmetric statistical distribution. It describes a situation in the world where most of the cases are close to the average result, and the more a given result deviates from the average, the less represented it is.



Probability distributions
(the standard deviation rule)



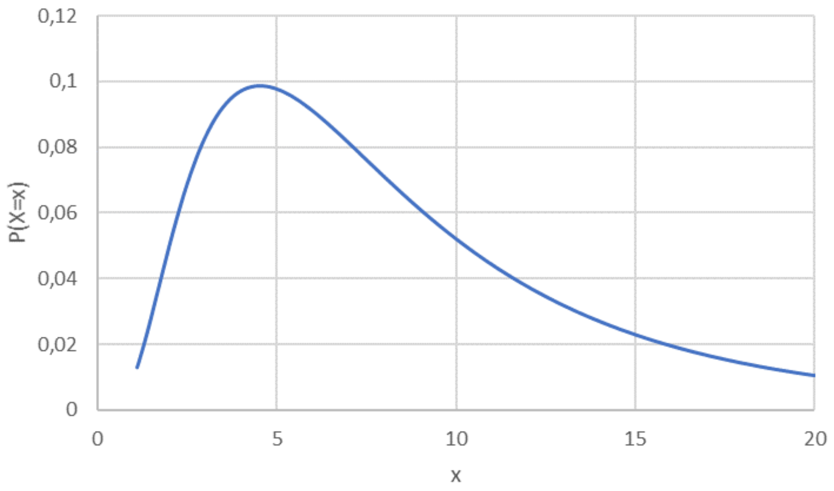
interval	% of Observations	
$[\mu-\sigma,\mu+\sigma]$	68.2%	
$[\mu-\mathbf{2}\boldsymbol{\sigma},\mu+\mathbf{2}\boldsymbol{\sigma}]$	**95.4%**	→ approx. 5% of the value outside the range
$[\mu-3\sigma,\mu+3\sigma]$	99.7%	

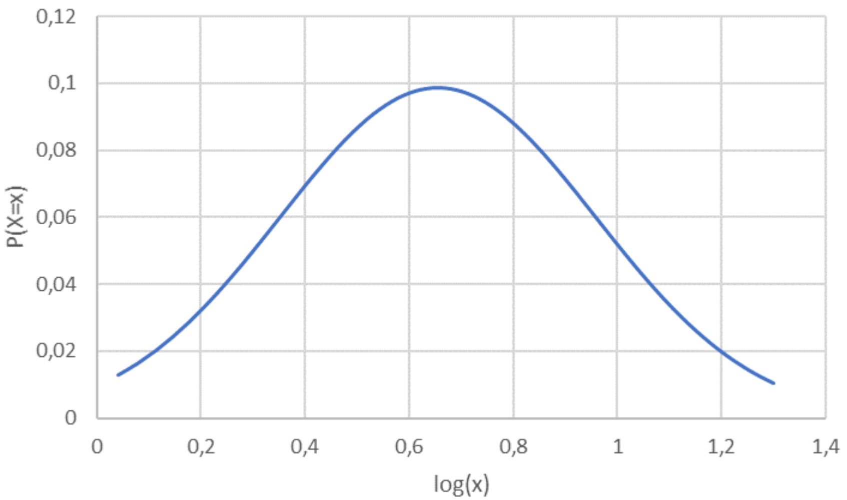
The interval $[\mu-\mathbf{2}\boldsymbol{\sigma},\mu+\mathbf{2}\boldsymbol{\sigma}]$ is most frequently used in medical sciences

Lognormal distribution

the continuous probability distribution of a positive random variable whose logarithm is normally *distributed*

(log-normal distribution is often a better approximation than the normal distribution of features in which the ratios and not differences between the values are important) The concentrations of a variety of chemical compounds in body fluids and tissues are often lognormal in the population. Their presentation on a logarithmic scale is then optimal)





Measures of central tendency

Measures of central tendency

Characterizes the „central” measurement. In other words, it ias a **measure of the location of a representative value**.

Depending on the nature of the data, several „central” value measures are used alternatively

Measures of central tendency

Arithmetic mean

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

or:

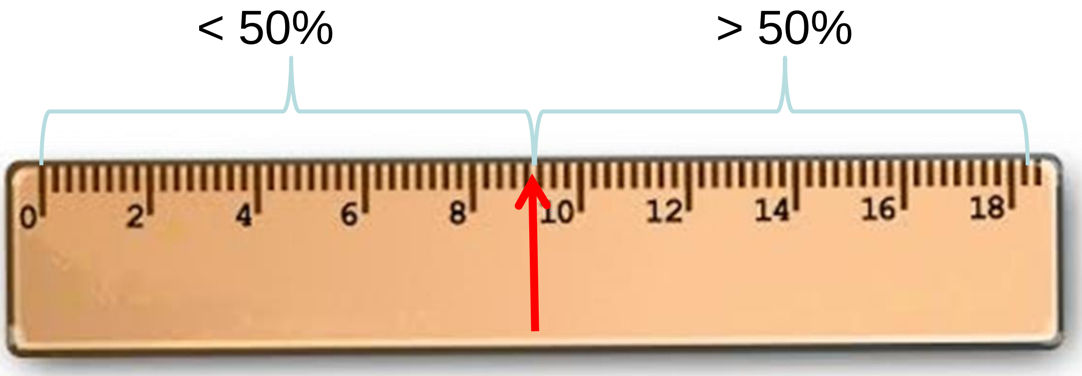
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{or} \quad \bar{x} = \frac{\sum x_i}{n} \quad \text{or} \quad \bar{x} = \frac{\sum x}{n}$$

Measures of central tendency

The **median** is the value in the middle of a series of ordered data if the number of observations is odd

or

is the arithmetic mean of the two adjacent middle observations when the number of these observations is even



< 50%

> 50%

Measures of central tendency

Mode is the most common value in the set.

There can be several modes - when two or more values occur the same number of times and other values occur less number of times.

Mode may not exist - when each value occurs only once

Measures of central tendency

Example 1:

Determine mode based on the number of children in 20 families.

Data - number of children in the 20 families:

1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5

Mode is 3

Measures of central tendency

Geometric mean

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

for right-skewed distribution is close to the median and less than the arithmetic mean.
Condition: the data distribution must be symmetrical. When the data is skewed, it must be prepared so that the geometric mean can be calculated

$$\boxed{\log \bar{x}_g = \log \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n} = (\log x_1 + \log x_2 + \log x_3 + \ldots + \log x_n)/n}$$

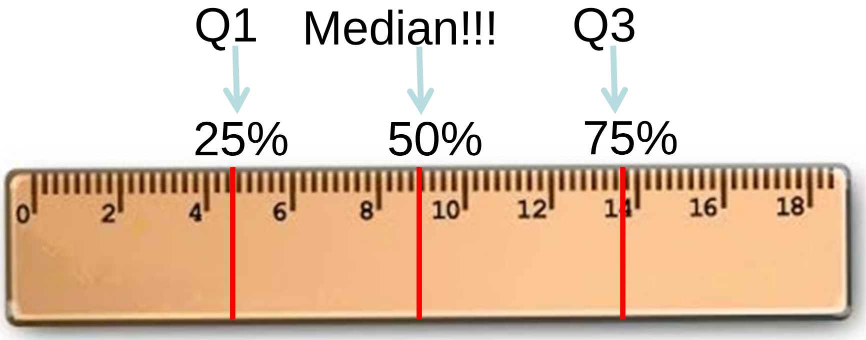
Measures of central tendency

Weighted average

the mean of the items which are assigned different weights in such a way that the items of a greater weight have a greater effect on the mean

$$\bar{x} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + \ldots + w_n \cdot x_n}{w_1 + w_2 + w_3 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Quartile. Values of Q1 (25%), Q2 (50%), Q3 (75%), the variable x that they divide ordered series into 4 equal pod in terms of the number of parts it is called quartiles



Q1 Median!!! Q3

25% 50% 75%

Measures of central tendency

Quartiles: the first quartile *Q1* divides the observations in such a way that 25% of the observations are lower or equal to the value of this quartile, and 75% of the observations are equal to or greater than the value of this quartile

Measures of central tendency

Percentile:

The values of the percentiles divide the ordered series (from smallest to largest value) of variable x into 100 equal in number parts.

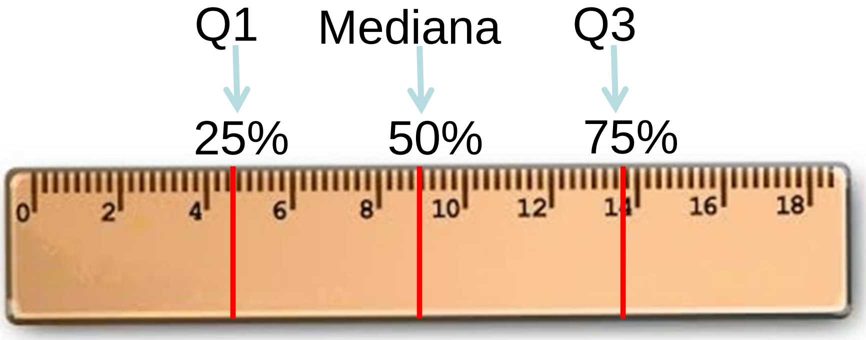
The value of the variable x below which is 1% of this series is called the first percentile. Similarly, the second percentile and third, ….

Measures of central tendency

What is the twenty-fifth percentile?

What is the 50th percentile?

What is the seventy-fifth percentile?



Measures of dispersion

Measures of dispersion

Measures of dispersion are also known as **variability**

Measures of dispersion

The **range** is the difference between the smallest and largest value in the data set

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

Measures of dispersion

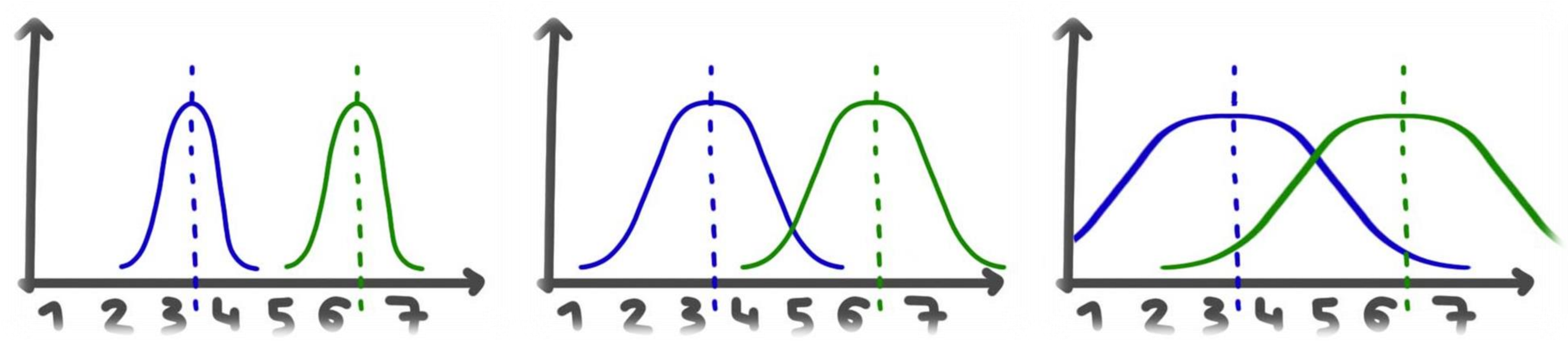
Variance of n measurements (n≥2)

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

Standard deviation

$$s = \sqrt{s^2}$$

Standard deviation



https://pogotowiestatystyczne.pl/odchylenie-standardowe-czyli-nie-sama-srednia-czlowiek-zyje/

Measures of dispersion

Coefficient of variation

$$w = \frac{s}{\bar{x}} \cdot 100\%$$

s – standard deviation; $\bar{x}$ - arithmetic mean

is a relative value expressed as a percentage

Used when the analyzed variable has only non-negative values (e.g. concentration). High values (> 100%) indicate the asymmetry of the variable distribution.

Measures of dispersion

IQR – interquartile range

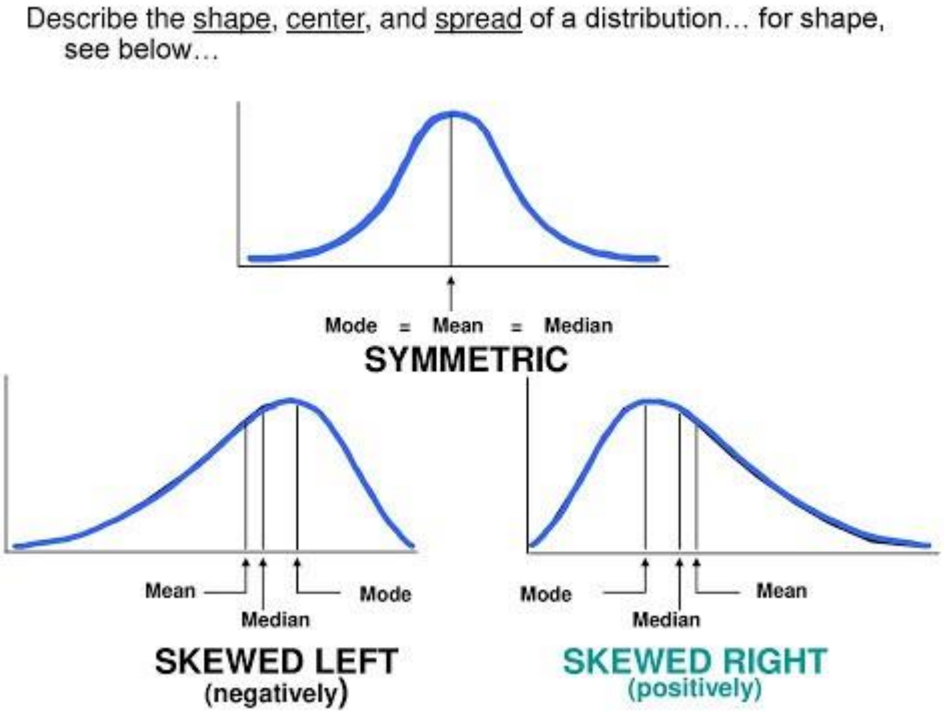
$$IQR = Q_3 - Q_1$$

Measures of symmetry of distribution

Skewness:

- left-skewed distribution
- symmetric distribution
- right skew distribution - often found in medical research when a certain part of the population has much greater values of the variable than the rest, e.g. parameters of inflammation

Example: log-normal distribution



Describe the shape, center, and spread of a distribution… for shape, see below…

https://lamduy-nguyen.medium.com/skewness-definition-problem-and-reducing-methods-a5ee25a8b6ca

Effect size

We give two drugs (A- the drug and B - placebo) to lower blood pressure to two different groups of patients. Are the effects of the two 'drugs' significantly different in reducing blood pressure?

Difference between averages:
-3,7 - 1,3 = - 5 mmHg

	A[mmHg]	B[mmHg]
	-5	2
	-6	1
	-12	2
	-9	-1
	-8	1
	5	2
	-7	-2
	8	4
	1	3
average	-3,7	1,3
difference	-5,0	

Standardized mean difference
SMD

e.g. Cohen coefficient

$$SMD = \frac{\overline{x_1} - \overline{x_2}}{s}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Thanks to standardization, the size of the effect does not depend on the unit in which the measured parameter is expressed (e.g. mmHg or kPa)

s_i^2 $- variances; s$ $- standard\ deviation\ \bar{x}_i$ $- arithmetic\ means;\ n_i$ $- number\ of\ probes$

Standardized mean difference SMD

Absolute values of SMD> 1 are interpreted as a significant effect,
> 2 as high,
> 3 as very high.
The SMD sign indicates the direction of the effect: positive when the mean in the group with the active substance is higher than in the control group

$$SMD = 5$$

$$s = 4{,}97$$

SMD = 5 / 4,97 = -1,01

Interpretation: the active substance has a significant blood pressure lowering effect

Confidence interval

Confidence interval

If the trait X in the **general population** has a normal distribution

$$X : N\left[\overline{X}, \sigma\right]$$

then the arithmetic mean of the **sample from this group** has a normal distribution with parameters:

$$\bar{x} : N\left[\overline{X}, \frac{\sigma}{\sqrt{n}}\right] \qquad \frac{\sigma}{\sqrt{n}} - standars\ error$$

Confidence interval

Standardized value of the arithmetic mean

$$u = \frac{\bar{x} - \bar{X}}{\sigma}\sqrt{n}$$

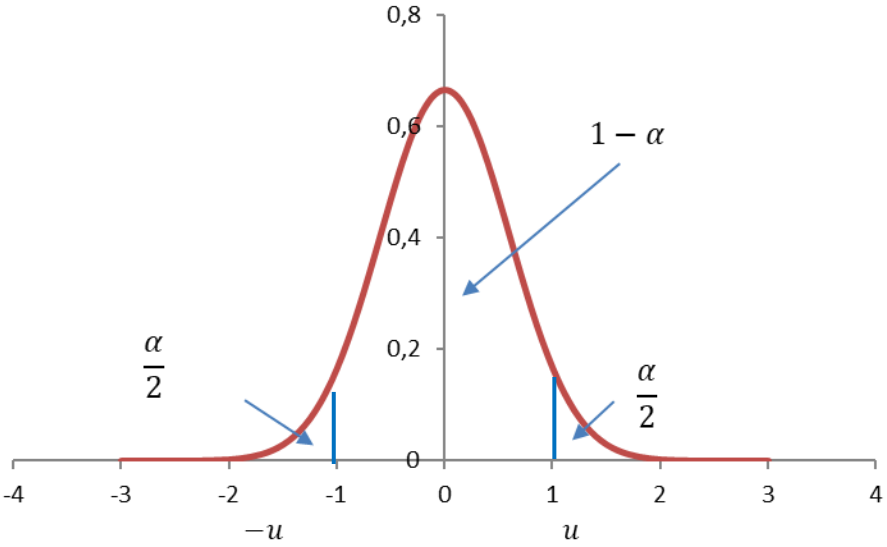
has a normal distribution with parameters:

$$u : N[0,1]$$

Confidence interval

The probability that the standardized value of the arithmetic mean is within the range (-u_α, u_α) is equal to:

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



Confidence interval

Confidence interval for the arithmetic mean:

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

confidence coefficient:

$$1 - \alpha$$

expresses the probability that the population mean estimated on the basis of the sample falls within the above range. The α / 2 value expresses the probability that the population mean is below the lower limit of this range. The same α / 2 value expresses the probability that the population mean is above the upper limit of this range)

The sum of these three probabilities (1-α) + α / 2 + α / 2 = 1 which means that the population mean must be within the confidence interval or below the lower limit of this interval or above the upper limit of this interval

Confidence interval

The most commonly used values in biostatistics:

$$1- \alpha = 0{,}95, \quad \alpha = 0{,}05$$

Confidence interval

It is known about the population of people that the response time to a certain stimulus has a normal distribution, with a standard deviation of 12 minutes. 36 healthy people were drawn. The average reaction time for this test is 23 minutes.

Calculate with a probability of 0.95 the average response time to this stimulus in all healthy people.

Confidence interval

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

$$P\left(\bar{23}-1{,}96\cdot\frac{12}{\sqrt{36}}<\bar{X}<\bar{23}+1{,}96\cdot\frac{12}{\sqrt{36}}\right)=\ 0{,}95$$

$$P(19.1<\bar{X}<26.9)=0.95$$

Confidence interval

If the trait X in the **general population** has a normal distribution

$$X : N\left[\overline{X},\sigma\right]$$

and the population standard deviation is unknown, confidence intervals are calculated on the basis of the so-called Student's t-distribution

Confidence interval

Confidence interval for the arithmetic mean when the population standard deviation is unknown:

$$P\left(\bar{x} - t_{\alpha} \cdot \frac{s}{\sqrt{n}} < \bar{X} < \bar{x} + t_{\alpha} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

s – *standard deviation of the sample*

Statistical hypotheses

Statistical hypotheses

The statistical hypothesis is any supposition of the general population regarding its statistical characteristics:

- the distribution,
- measures of central tendency,
- measures of dispersion

Statistical hypotheses

Statistical hypotheses can be divided into:

- **parametric** - the hypothesis concerns the value of the distribution parameters,
- **nonparametric** - the hypothesis concerns the form of the distribution function

Statistical hypotheses

The statistical hypothesis that is subject to verification is called **null hypothesis** H_0.

It is the opposite to the **alternative hypothesis** H_1.

Statistical hypotheses

- Traditionally, the null hypothesis H0 is formulated as no relationship hypothesis (e.g., means of two populations do not differ), and the alternative hypothesis H1 as the existence hypothesis relationship (e.g., the means of the two populations differ). The lack of relationship assumed by H0 can be interpreted as the zero value of the effect size defined e.g. as the difference of the mean distributions of the parameter compared populations: if the means of the distributions parameter of these populations do not differ from each other, it the difference of the means is zero.

Statistical hypotheses

Verification of statistical hypothesis is based on data from a sample.

Thus, conclusions can be formulated with some probability

Statistical hypotheses

It is possible to make two types of errors:

- **type I error with probability α -** reject the hypothesis H_0 although it is true
- **type II error with probability β -** accept the H_0 hypothesis even though it is false

Statistical hypotheses

The probability of type I error is called **the level of significance** α.

The level of significance is determined *a priori*. In biological and medical sciences is usually α = **0,05**.

Statistical hypotheses

In practice, the null hypothesis H_0 – no relationship (the means of the two populations do not differ) is improbable, because the probability that a given continuous parameter (e.g., mean age) determined with an accuracy of an arbitrarily large number of decimal places is identical in two populations is infinitely small.

Statistical hypotheses and confidence interval

The verification of the null hypothesis H_0 should be treated only as a theoretical tool useful for statistical inference about the values of the distribution parameters in the populations and the size of the effects on the basis of the distribution parameters in the samples from these populations.

The significance level **α** (the probability of rejection of the H_0 hypothesis if it is true) is closely related to the **1-α** confidence level for the confidence interval reflecting the probability that the parameter value in the population estimated on the basis of the sample falls within the above interval.

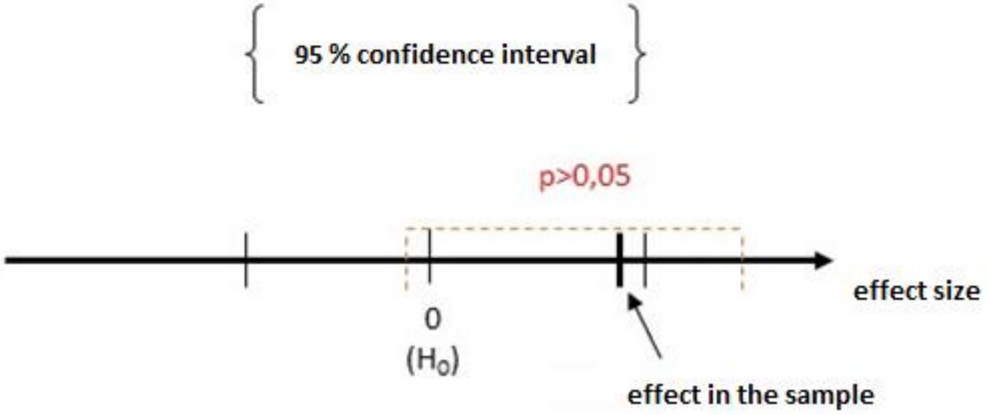
Interpretation of p-values based on a confidence interval

A p-value <0.05 for the verification of the H_0 (no effect) hypothesis means that the „no effect” is outside the 95% confidence interval for the effect size, so we can exclude with high probability „no effect” or the opposite effect to that observed in the sample in the population.

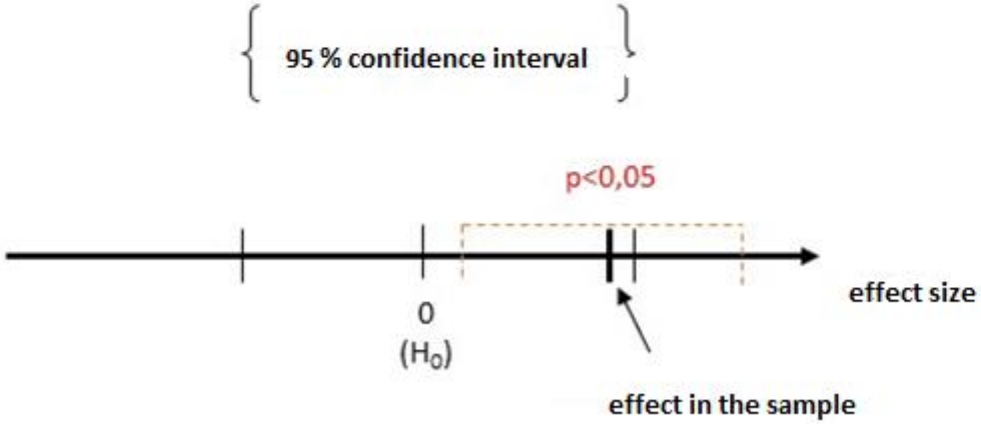
A p-value > 0.05 for the verification of the H_0 hypothesis means that the „no effect” is within the 95% confidence interval for the effect size, so we cannot exclude with high probability „no effect” or the opposite effect to that observed in the trial in the population.

Effect size = 0 - no effect

„no effect” is within the 95% confidence interval for effect size



„no effect” is outside the 95% confidence interval for effect size



Statistical tests

The selection of the test is determined by the size of the compared groups, the type of variables (paired, unpaired) and the shape of the probability distribution:

- normal distribution - parametric test
- non-normal distribution - non-parametric test

Statistical tests

Testing the shape of the distribution on the basis of a sample can be carried out using, for example,

Shapiro-Wilk test

*The lower the value of the **W** statistic, the more different from the normal distribution*

If the p value corresponding to the W statistic is <0.05, it proves a significant deviation of the analyzed distribution from the normal distribution

Parametric tests

Parametric tests

- The examined feature has a normal distribution (Gauss distribution)
- The difference in variances of the studied populations is not statistically significant (variances are homogeneous)

Parametric tests

t (Student’s) test for two unpaired random samples when variances in populations are unknown but equal:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{K \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

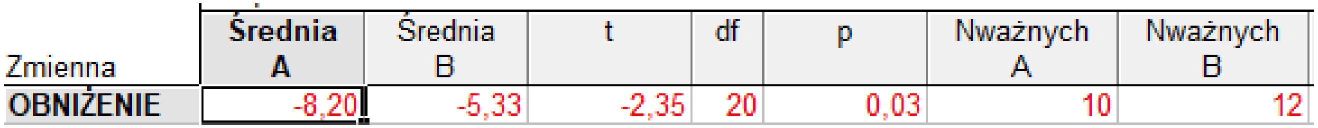
$$K = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}$$

$s_i^2 \ - sample\ variances\ (i = 1{,}2)$

Example: We give two drugs (A to 10 subjects and B to 12 subjects) to lower blood pressure in two different groups of patients. Are the effects of drugs significantly different?

pressure reduction	drug
-5	Lek A
-6	Lek A
-12	Lek A
-9	Lek A
-8	Lek A
-5	Lek A
-7	Lek A
-8	Lek A
-15	Lek A
-7	Lek A
-6	Lek B
-5	Lek B
-11	Lek B
-5	Lek B
-3	Lek B
-4	Lek B
-6	Lek B
-6	Lek B
-4	Lek B
-9	Lek B
-3	Lek B
-2	Lek B

averages



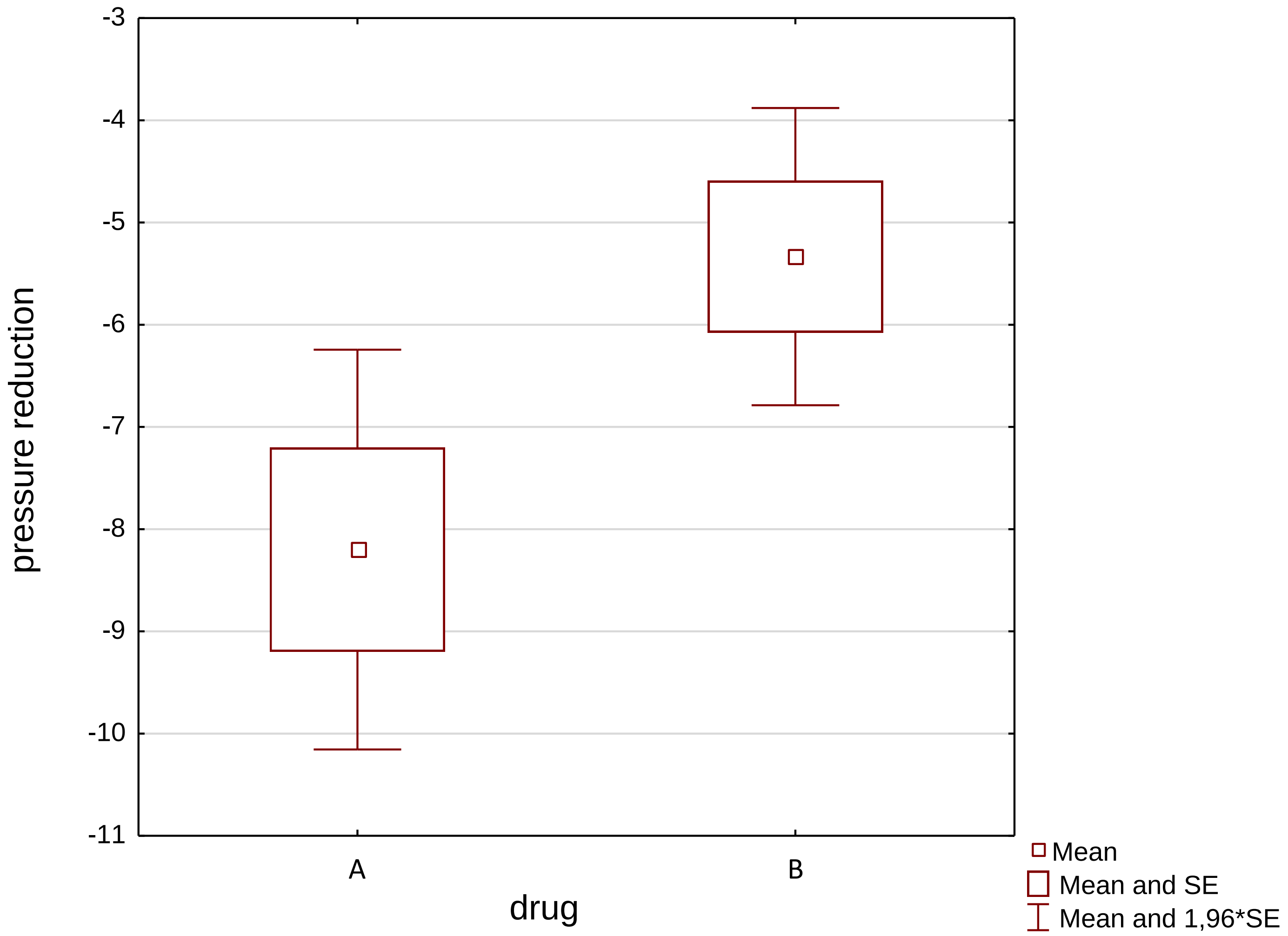
Zmienna	Średnia A	Średnia B	t	df	p	Nważnych A	Nważnych B
OBNIŻENIE	-8,20	-5,33	-2,35	20	0,03	10	12

pressure reduction

$p<0,05$

⇩

Statistically significant differences (different effect of the drugs)



We observe that drug A "showed" greater (from the mathematical point of view, but not necessarily clinically) effectiveness in lowering blood pressure

Parametric tests

If the variances differ statistically, non-parametric tests should be used

Parametric tests

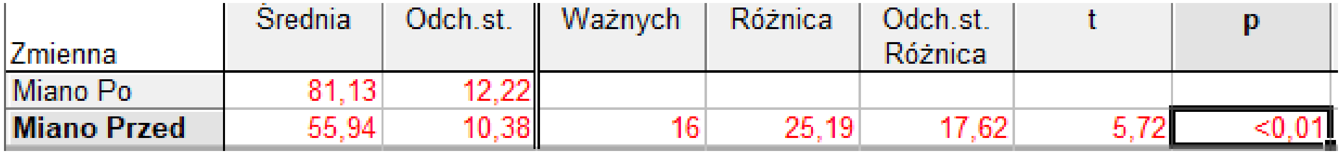
t (Student’s) test for two paired random samples:

$$t = \frac{\overline{d}}{s} \cdot \sqrt{n}$$

$$d_i = x_{1i} - x_{2i} \qquad \overline{d} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \overline{d}\right)^2}{n-1}}$$

Example: In one experiment, the antibody titer was compared before and after vaccination. Are the antibody titres before and after vaccination statistically different?

averages SD difference

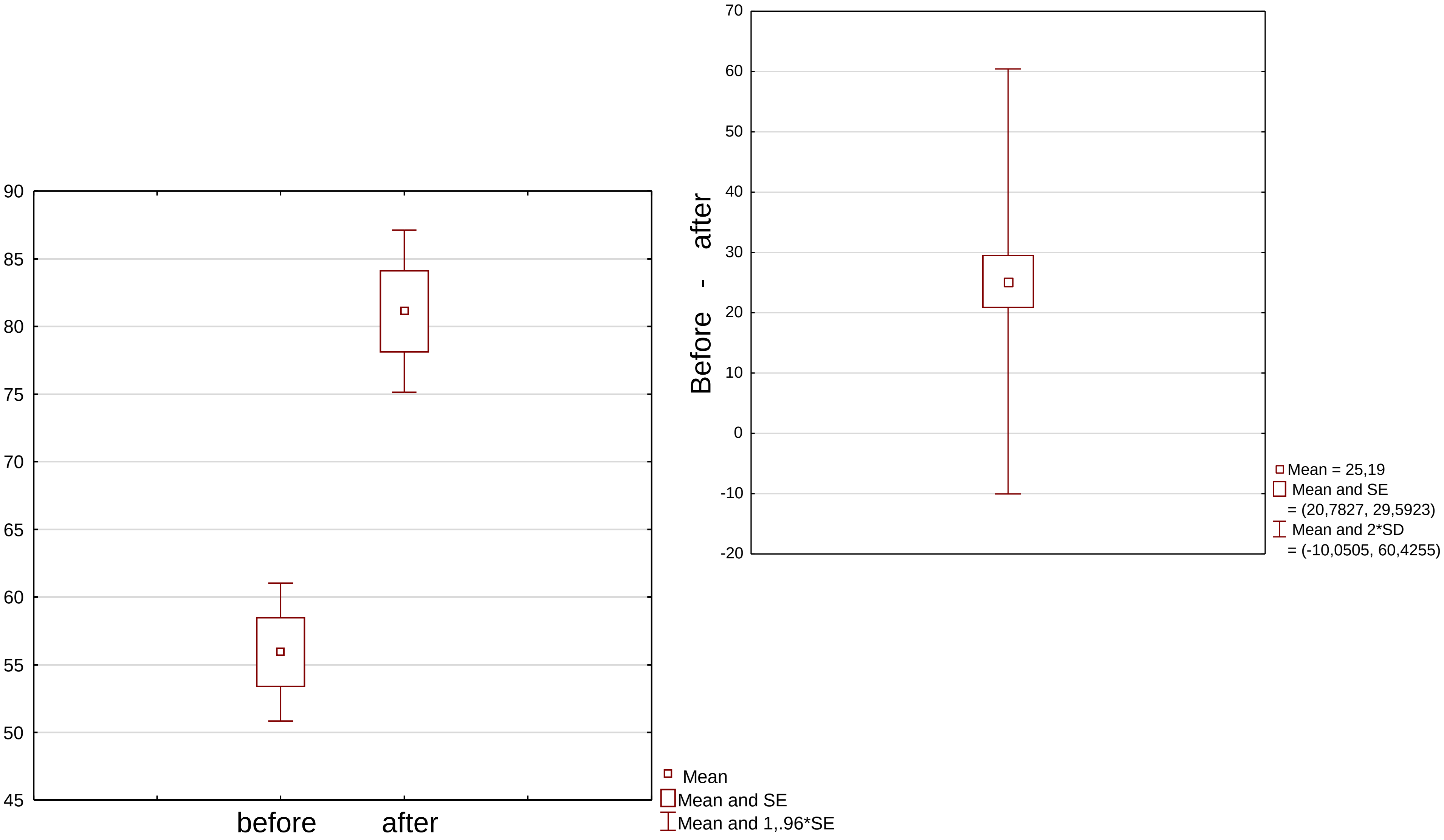


Zmienna	Średnia	Odch.st.	Ważnych	Różnica	Odch.st. Różnica	t	p
Miano Po	81,13	12,22					
Miano Przed	55,94	10,38	16	25,19	17,62	5,72	<0,01

before	after	difference
74	63	-11
49	90	41
55	64	9
75	96	21
49	66	17
41	97	56
75	72	-3
58	86	28
52	88	36
47	91	44
55	68	13
50	92	42
57	91	34
54	80	26
46	68	22
58	86	28

p<0,05

Statistically significant differences: the antibody titer before and after vaccination differs significantly



We observe that the titer of antibodies after vaccination increased significantly (mathematically, not necessarily clinically)

ANOVA
(more then two groups – unpaired samples)

ANOVA (analysis of variance) - one factor:

- each population must have a normal distribution,
- samples/groups drawn from each population have to be independent random samples/groups,
- variances in populations have to be homogeneous

GENERALIZATION OF STUDENT T-TEST FOR UNPAIRED SAMPLES

ANOVA
(more then two paired samples)

ANOVA with repeated measures:

- each variable must be normally distributed,
- the variances of the variables are similar.

GENERALIZATION OF STUDENT T-TEST FOR PAIRED SAMPLES

ANOVA

if we reject H_0 (there are statistically significant differences between the means, but we do not know between which) then the next step of the analysis is a multiple *post hoc* comparison for all pairs of means

- Bonferroni procedure - *post hoc* test: we divide the significance level by the number of comparisons made (in this case, the p-value of the Student's t-test for a given pair of means by the number of all analyzed pairs of means)
- in medical science, Tukey's post hoc test is often used, which has more statistical power than the Bonferroni procedure

Multiple comparisons are associated with an increase in the type I error (rejection of the null hypothesis when it is justified) - eg Bonferroni correction reduces this risk, but at the cost of a decrease in the test's power.

Conclusion: if multiple comparisons are made, the sample size of the population should be increased.

ANOVA

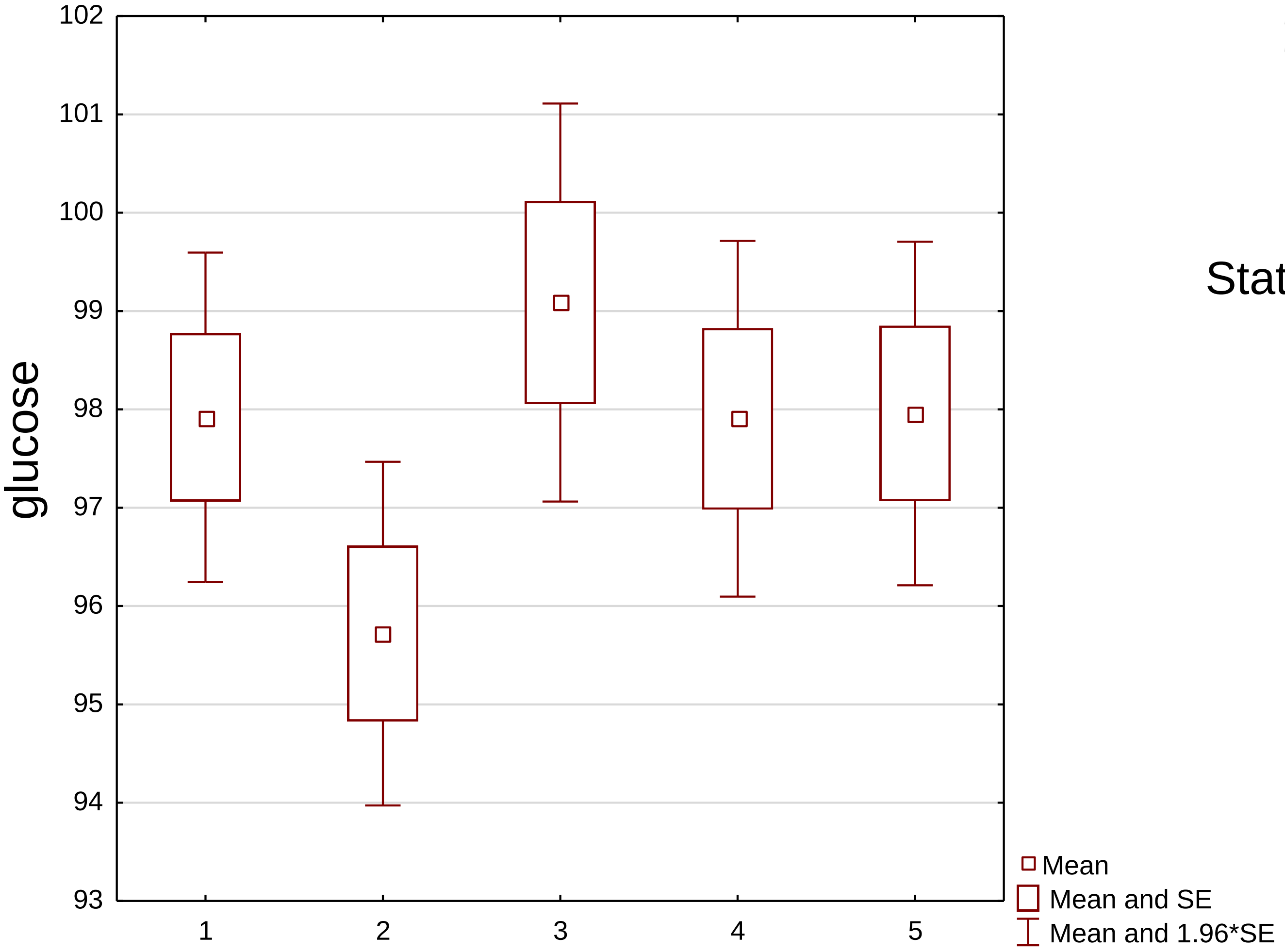
Example 1

The table shows the results of measuring blood glucose in people on different diets. Does blood glucose level depend on the diet?

1	2	3	4	5
101	93	102	92	90
105	100	104	104	97
98	93	93	102	97
96	94	103	101	97
90	98	103	100	104
100	102	96	98	93
101	90	105	95	97
102	100	99	105	99
96	103	91	100	99
91	101	101	97	97
91	92	92	97	103
100	92	104	101	103
97	96	101	92	104
94	104	93	100	92
101	93	103	91	100
102	90	96	99	100
95	93	101	104	93
99	95	103	97	98
92	95	94	94	95
99	91	95	93	103
96	92	105	94	101
104	95	91		104
94	90	104		90
102	99			95
102	102			

The verification of the normality of distributions and the homogeneity of variance showed that ANOVA can be used

ANOVA



$p_{ANOVA} < 0,05$

⇩

Statistically significant differences

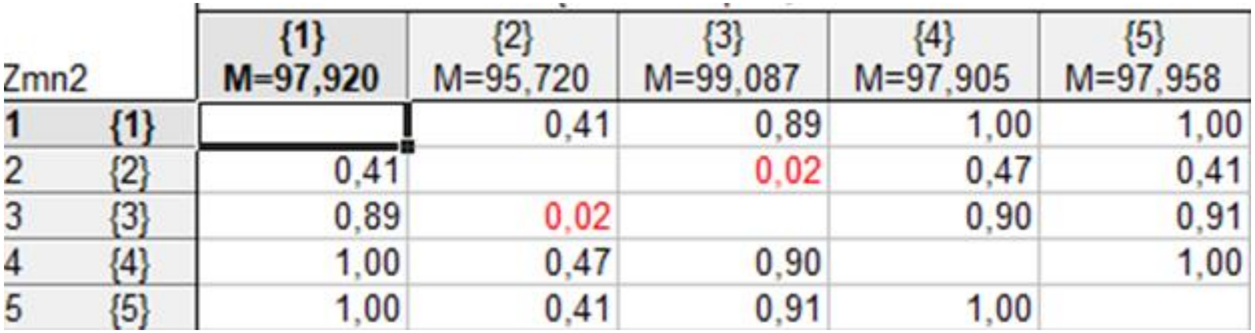
ANOVA

Where?

↓

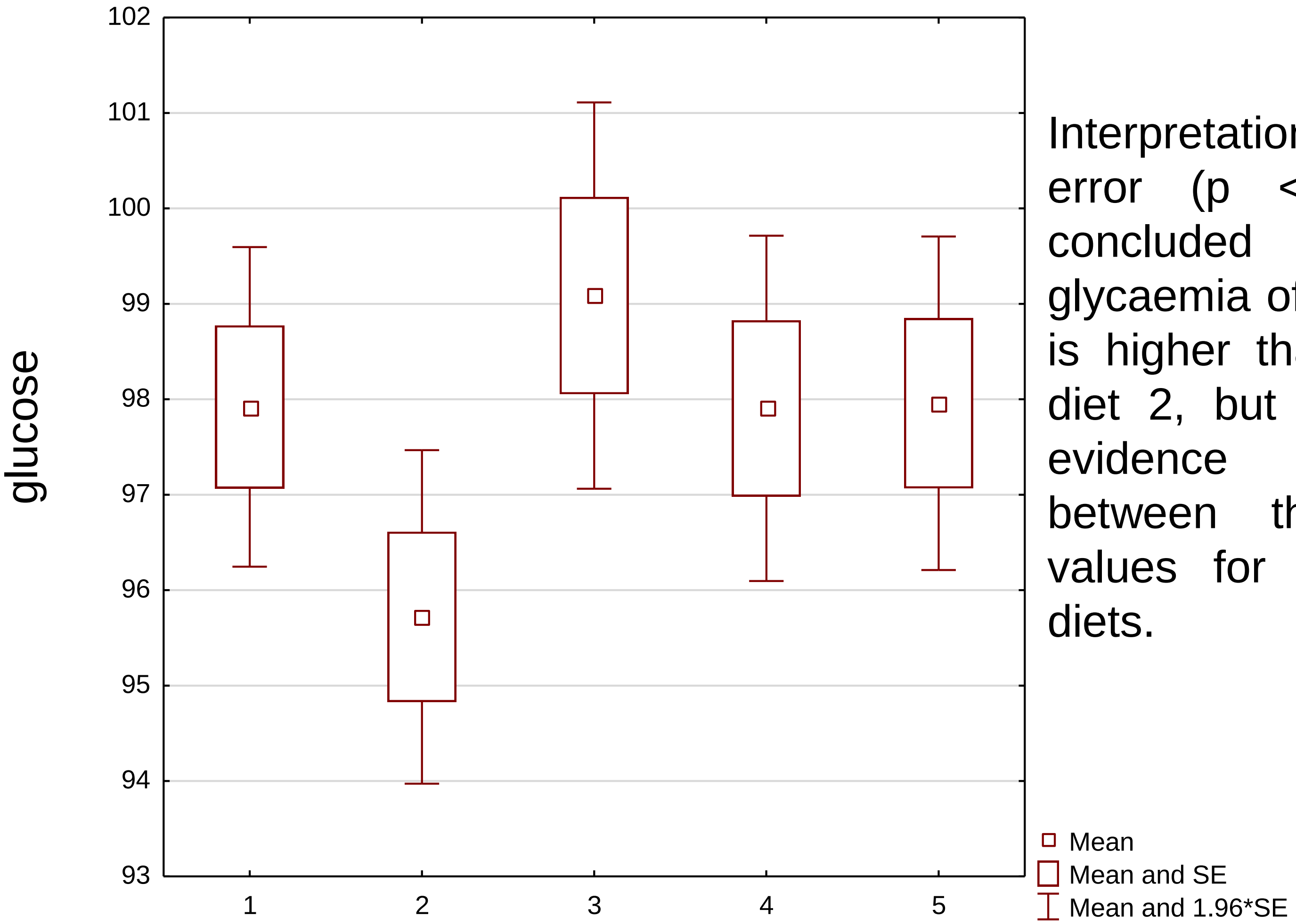
The Tukey post – hoc test

Only the difference between groups 2 and 3 is statistically significant (p = 0.02)



Zmn2		{1} M=97,920	{2} M=95,720	{3} M=99,087	{4} M=97,905	{5} M=97,958
1	{1}		0,41	0,89	1,00	1,00
2	{2}	0,41		0,02	0,47	0,41
3	{3}	0,89	0,02		0,90	0,91
4	{4}	1,00	0,47	0,90		1,00
5	{5}	1,00	0,41	0,91	1,00	

ANOVA+ Tukey’s post-hoc



Interpretation: With a low risk of error (p <0.05), it can be concluded that the mean glycaemia of the group on diet 3 is higher than in the group on diet 2, but there is insufficient evidence for a difference between the mean glucose values for the other pairs of diets.

Nonparametric tests

Nonparametric tests

- used when random variables significantly deviate from the normal distribution or are expressed on an ordinal (rank) scale
- do not require assumptions about the distribution of the population from which the sample is chosen
- they are less sensitive to outliers
- they formally require consistent distributions in the compared populations, but they are not very sensitive to the differences between these distributions

Nonparametric tests

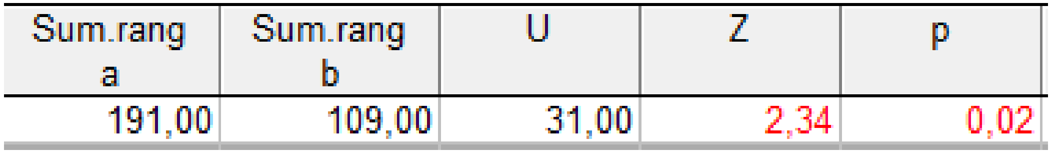
Two unpaired samples?

Mann-Whitney test

- two populations are investigated
- alternative to the Student's t-test for unpaired samples
- compares the ranks of the variables rather than values

Example: In two groups of patients with a certain neurological disease, tests of the concentration of adrenaline in the blood serum were carried out. Is the adrenaline concentration in both groups the same?

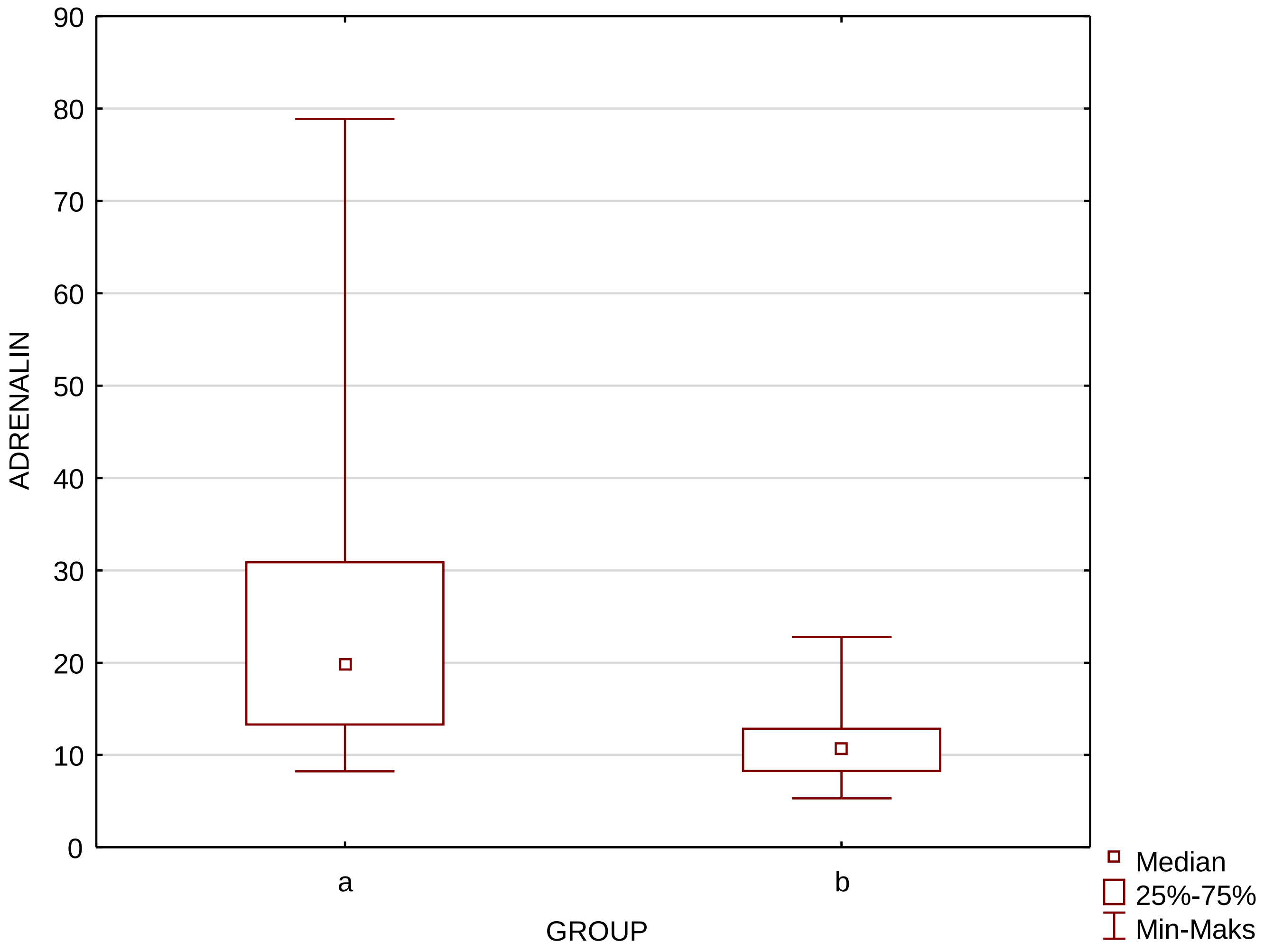
ADRENALIN	GROUP
14,34	a
20,33	a
18,79	a
8,22	a
31,5	a
12,08	a
22	a
9,22	a
19,5	a
78,89	a
30,48	a
45,86	a
5,33	b
22,5	b
11,74	b
7,39	b
12,34	b
13,22	b
8,53	b
22,8	b
12,7	b
7,78	b
9,63	b
8,9	b



Sum.rang a	Sum.rang b	U	Z	p
191,00	109,00	31,00	2,34	0,02

p<0,05

Statistically significant differences: the concentration of adrenaline differs significantly in the two studied groups



The concentration of adrenaline is significantly higher in group a, in which all measures of position shown in the figure (median, quartiles, minimum and maximum) are higher than in group b

Nonparametric tests

More than two unpaired samples?

Kruskal-Wallisa test

- An alternative to ANOVA
- compares the ranks of the variables rather than values

Nonparametric tests

Kruskal-Wallis:

- a random variable is a quantitative variable but has no normal distribution or a random variable is expressed on an ordinal (rank) scale
- formally, it requires consistent distributions in the compared populations, but it is not very sensitive to differences between these distributions

TIME	TEST
9,10	I
8,90	I
8,40	I
10,00	I
8,70	I
9,20	I
7,60	I
8,60	I
8,90	I
7,90	I
10,00	II
10,20	II
9,80	II
11,60	II
9,50	II
9,20	II
8,60	II
10,30	II
9,40	II
8,50	II
10,00	III
9,90	III
9,80	III
12,90	III
11,20	III
9,90	III
8,50	III
9,80	III
9,20	III
8,20	III
10,90	IV
11,10	IV
12,20	IV
14,40	IV
9,80	IV
12,00	IV
8,50	IV
10,90	IV
10,40	IV
10,00	IV

The clotting time of the blood plasma was tested by four different methods in a randomly selected group of patients. Does the blood clotting time measured differ significantly?

Kruskal-Wallis: H = 14,03 p =0,001

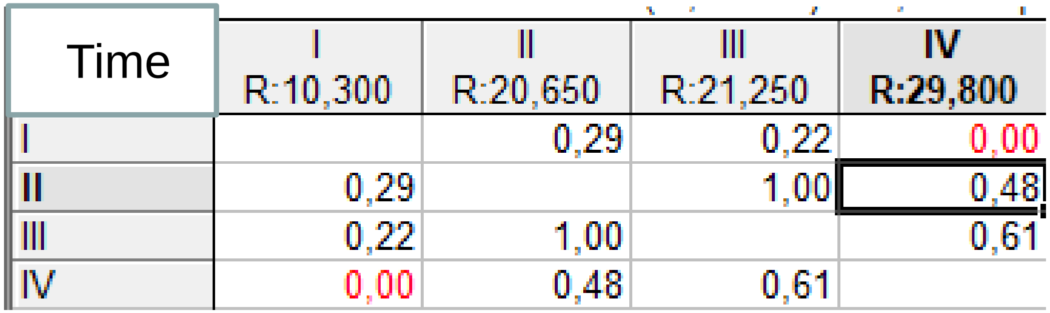
p<0,05

⇩

Statistically significant differences: the blood clotting time determined by the selected measurement methods is significantly different

But between which groups?

post – hoc test

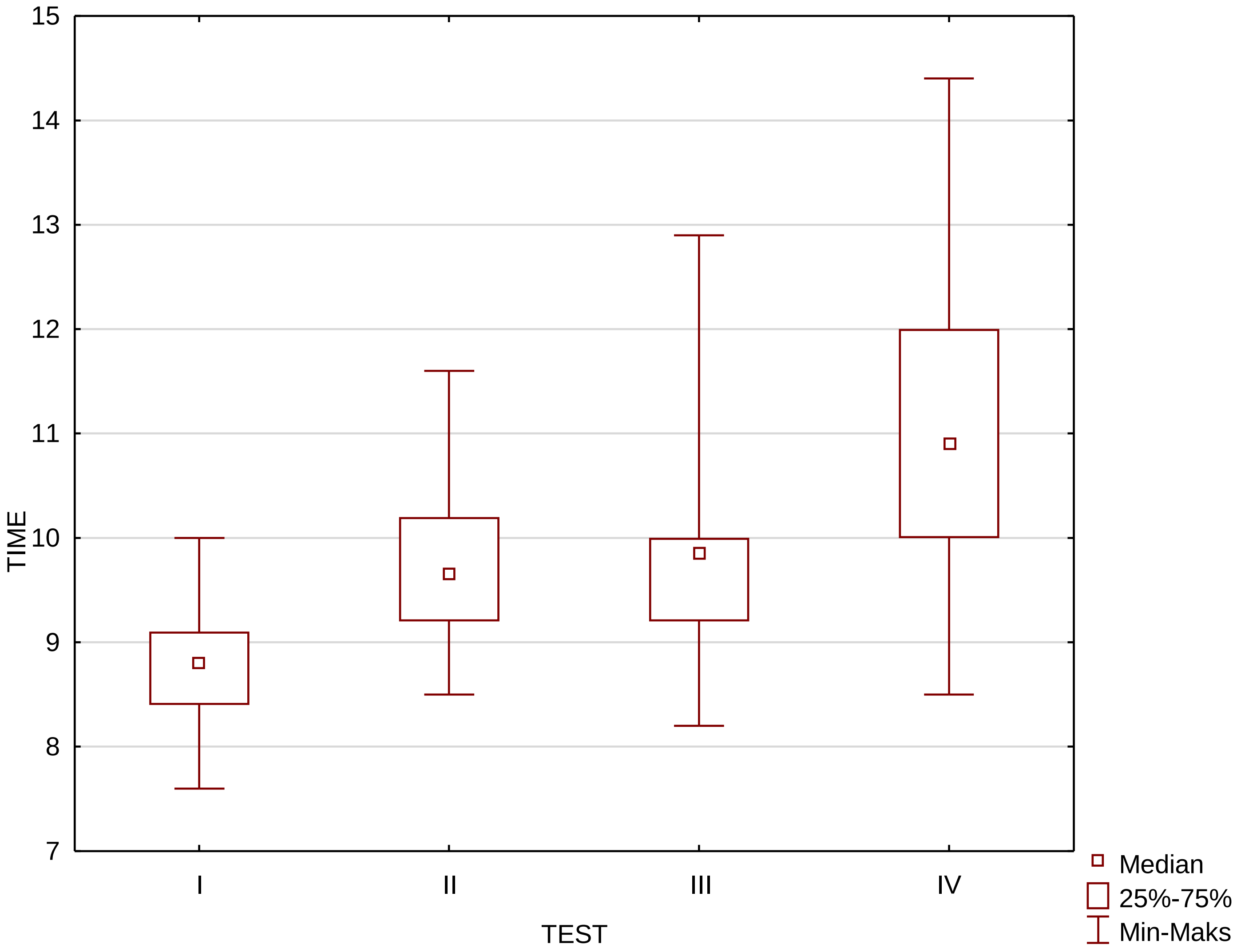


Time	I R:10,300	II R:20,650	III R:21,250	**IV** **R:29,800**
I		0,29	0,22	0,00
II	0,29		1,00	0,48
III	0,22	1,00		0,61
IV	0,00	0,48	0,61	

p<0,05

⇓

Statistically significant differences: blood clotting time differs significantly between methods I and IV



Interpretation: With a small risk of error (p <0.05), it can be concluded that the clotting time measured by method IV is longer than that measured by method I, but there is insufficient evidence for a difference between the times measured by the other pairs of methods.

Nonparametric tests

Two paired samples?

↓

Wilcoxon test

Example 1: The table shows the results of a body mass test of 15 people before and after the body slimming treatment.

Does the treatment work?

Nonparametric tests

patient	**before** [kg]	**after** [kg]	**difference** [kg]	**Median of differences**
1	96	90	-6	-5
2	87	82	-5	
3	90	83	-7	
4	101	97	-4	
5	97	95	-2	
6	97	92	-5	
7	85	88	3	
8	92	86	-6	
9	103	97	-6	
10	88	89	1	
11	97	93	-4	
12	101	95	-6	
13	95	91	-4	
14	91	90	-1	
15	88	82	-6	

Nonparametric tests

Example 1 (based on D. Miller, St. Orzeszyn Elements of medical statistics):

The significance level α = 0.05.

Probability p (p-value) related to the test result p = 0.002

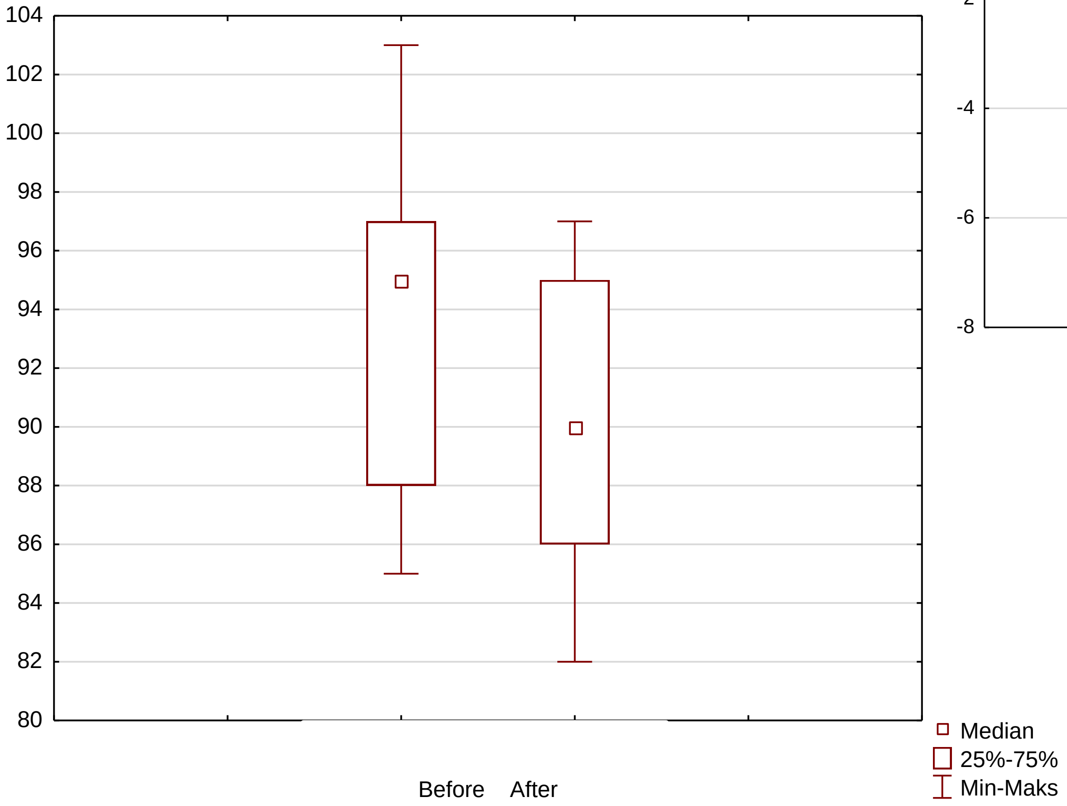
The **result of the test** is statistically significant (**p <α**) - **statistically significant** differences were noted in body masses before and after the treatment

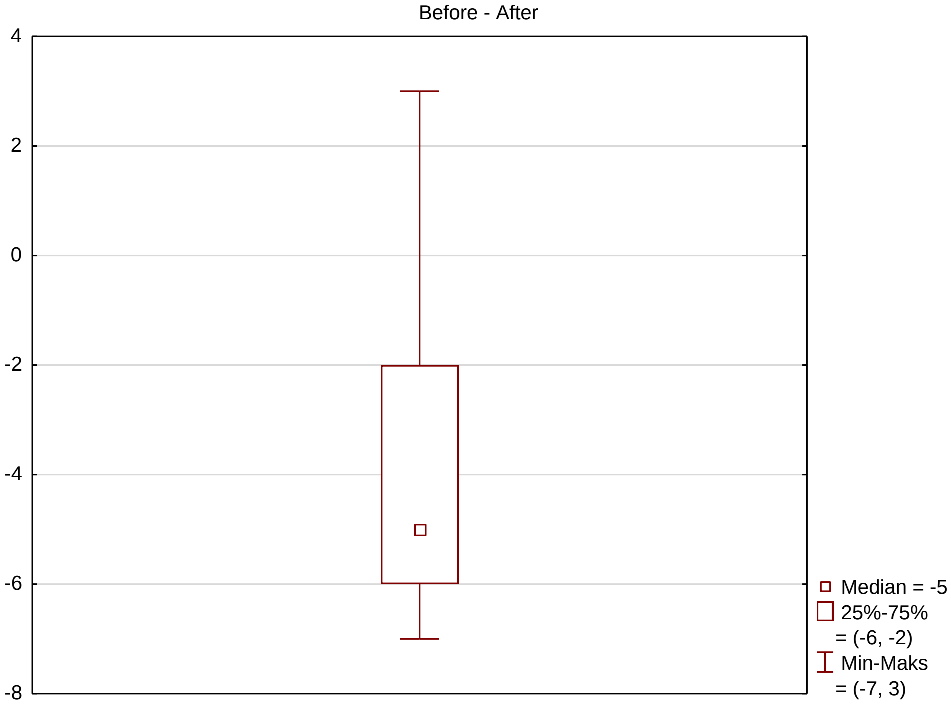
Effect direction: by interpreting the value of the median of differences, it is possible to determine the **loss** of body mass after the treatment

Effect size expressed by the median value is **-5 kg**

Summary: the treatment resulted in a statistically **significant decrease** in body mass, the **median** of which was **-5 kg**

Nonparametric tests





Nonparametric tests

More than two paired samples?

↓

Friedman test

an alternative to the Wilcoxon test to more than two samples

One clinic was evaluating a drug for pernicious anemia. During the treatment, the concentration of vitamin B12 in the patient's serum was measured five times. Did vitamin B12 levels change significantly with treatment?

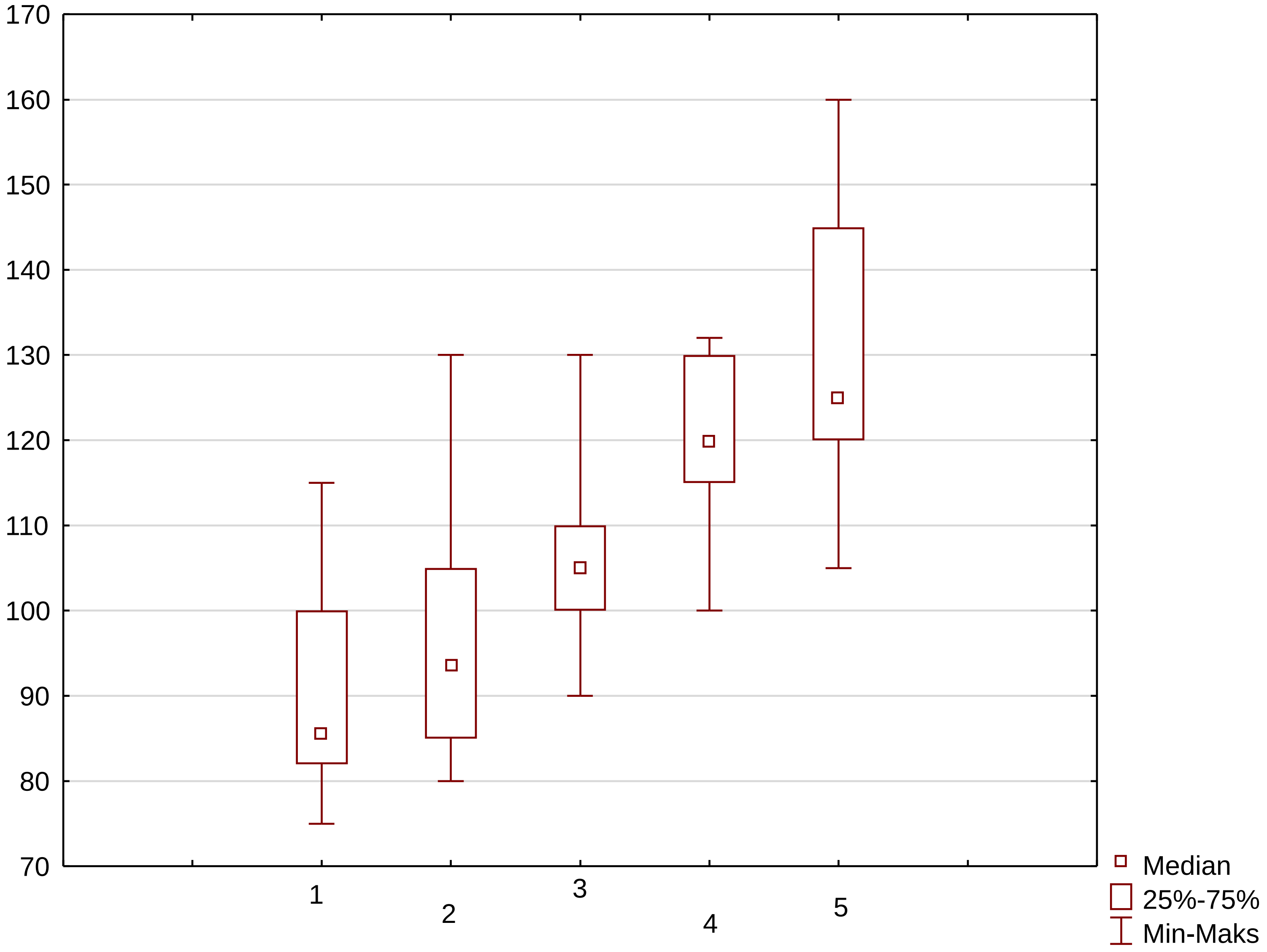
Friedman test = 36,22   p < 0,001

1	**2**	**3**	**4**	**5**
82	90	105	115	125
86	80	100	130	160
100	105	115	130	150
75	80	90	105	115
82	92	110	120	125
90	110	110	115	130
85	85	100	120	125
115	130	130	132	145
105	100	105	125	120
85	95	90	100	105

p<0,05

↓

Statistically significant differences: serum vitamin B12 concentration changes significantly with time



We observe a highly significant (mathematically: p <0.001, not necessarily clinically) increase in the concentration of vitamin B12

Association analysis

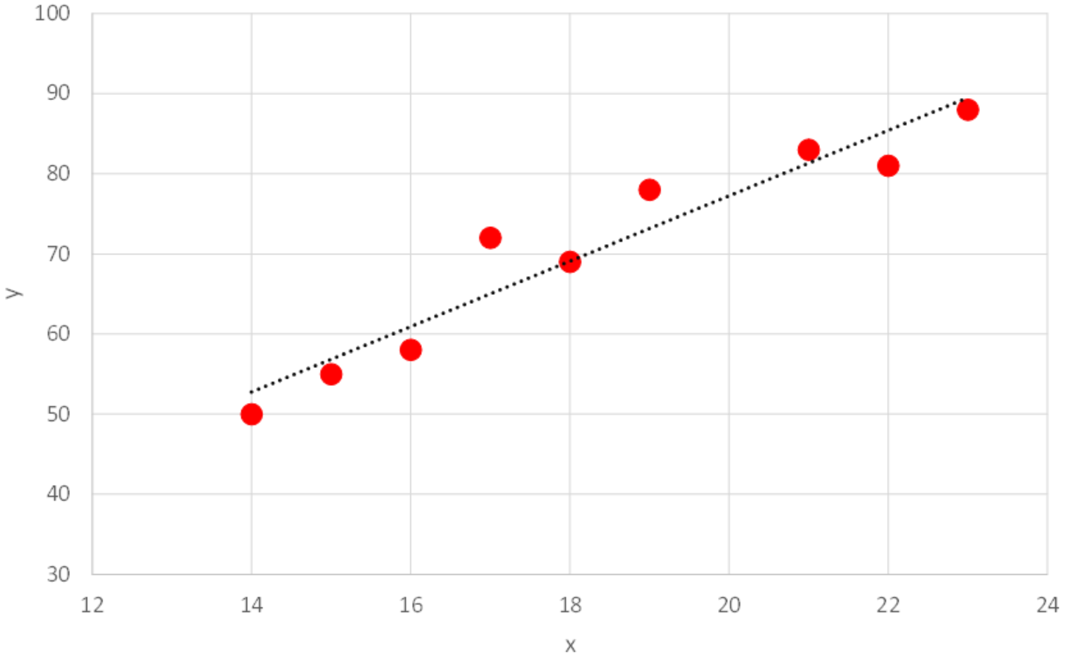
Association analysis between two measurable features

Functional relationship of the dependent variable to the independent variable:

- linear relationship
- non-linear relationship

Association analysis between two measurable features

The linear relationship is described by a linear regression equation, or first degree regression equation:



$$y = a \cdot x + b$$

Association analysis between two measurable features

The first degree regression equation allows the prediction of values within the observation interval of the independent variable!

The prediction outside this range is an extrapolation and may cause significant errors when e.g. the relationship becomes nonlinear

Association analysis between two measurable features

Measure of the linear correlation - Pearson's correlation coefficient:

$$r = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right) \cdot \left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2}}$$

Main condition: the distributions of both features are normal and the relationship between them is linear

r varies from -1 (-100%) to +1 (+ 100%)

Association analysis between two measurable features

- r > 0 positive correlation - the increase in the independent variable is related to the increase in the dependent variable
- r < 0 negative correlation - the increase in the independent variable is related to the decrease in the dependent variable

Association analysis between two measurable features

- absolute value equal to 1 - full linear relationship
- value equal to 0 - complete lack of a linear relationship
- values between 0 and 1 - weaker or stronger linear relationship

Association scale

$r_{xy} = 0$	no correlation
$0 < r_{xy} < 0,1$	very weak correlation
$0,1 \leq r_{xy} < 0,3$	weak correlation
$0,3 \leq r_{xy} < 0,5$	average correlation
$0,5 \leq r_{xy} < 0,7$	strong correlation
$0,7 \leq r_{xy} < 0,9$	very strong correlation
$0,9 \leq r_{xy} < 1$	almost „full” correlation
$r_{xy} = 1$	„full” correlation

Association analysis between two rank variables

The distributions are not close to the normal?

Dependence is not linear?

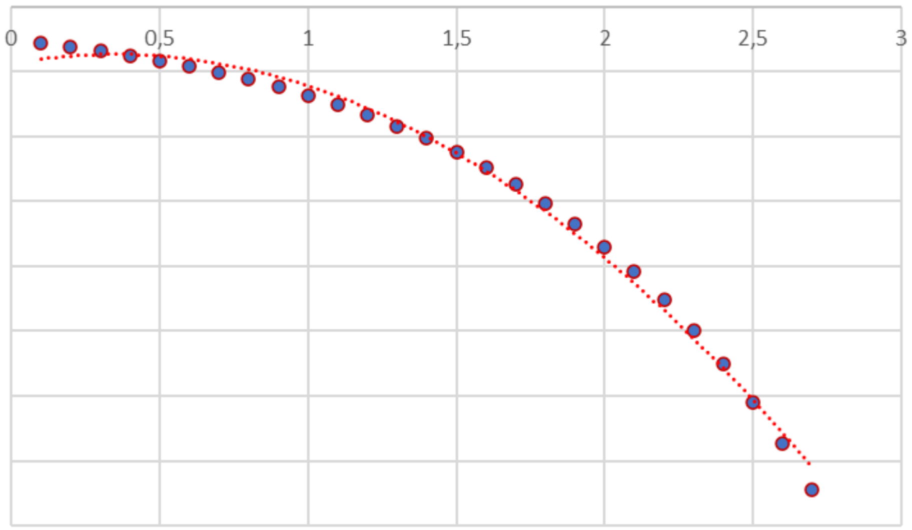
⇓

The Spearman's rank correlation coefficient (ρ)

The relationship may be non-linear, but it must be monotonic (with an increase in the independent variable, the dependent variable either always increases or always decreases)

Association analysis between two rank variables

$$e.g.\rho = -0{,}95$$



The relationship is strong but non-linear - the linear correlation coefficient would be inadequate

A negative rank correlation coefficient indicates that a higher dose of the administered drug (OX axis) is associated with a decrease in blood pressure (OY axis). An absolute value close to 1 indicates a very strong relationship between the drug dose used and the blood pressure level.

Risk and odds

Risk and odds

Risk *– in general: indicator of a condition or event that could lead to losses or gains. It is proportional to the likelihood of this event occurring and the size of the loss or profit it may cause.*

In biostatistics: the probability that a specific clinical condition occurs

Odds *- in general: the probability, the possibility of success in some matter, the achievement of success or the occurrence of some desired circumstances.*
In biostatistics: the ratio of probability that a specific clinical condition occurs to the probability that it does not occur

Relative risk (RR)

„Relative risk” *- the ratio of the probability of an event (e.g. illness, death) occurring in one study group compared to the other group. Relative risk is used to assess how much more or less is the probability of a given event occurring in group A compared to group B. "*

$$RR_{AB} = \frac{P(A)}{P(B)}$$

P(A) - probability of an event occurring in group A
P(B) - probability of an event occurring in group B

Relative risk(RR)

„Relative risk” *- the ratio of the probability of an event (e.g. illness, death) occurring in one study group compared to the other group. Relative risk is used to assess how much more or less is the probability of a given event occurring in group A compared to group B. "*

RR takes values from 0 to + ∞

•**RR≈1:** determines the situation in which the probability of an event in group A is close to the probability of an event in group B

•**RR>1:** the probability of an event occurring in group A is greater than in group B

•**RR = 2:** the occurrence of an event in group A is twice (100%) more likely than in group B (RR = 3 means that it is 3 times)

•**RR = 1.5:** 50% greater in A than in B

•**RR <1:** the opposite situation - in group A it is less than in B

•**RR = 0.75:** the occurrence of an event in group B is 25% less likely in group A than in group B, but 33% more likely in group B than in group A, because 1 / 0.75 = 1.33

Odds ratio (OR)

„Odds ratio” - the ratio of the odds of a given event occurring in a given group to the odds of the same event occurring in another comparable group. Therefore, using the OR indicator, we determine how much greater or less is the odds of an event, e.g. disease, death in one group compared to another group. "

$$OR_{AB} = \frac{S(A)}{S(B)}$$

$S(A) = \frac{P(A)}{1-P(A)}$ - odds of occurrence of the phenomenon in group A

$S(B) = \frac{P(B)}{1-P(B)}$ - odds of occurrence of the phenomenon in group B

Odds ratio(OR)

„Odds ratio” - the ratio of the odds of a given event occurring in a given group to the odds of the same event occurring in another comparable group. Therefore, using the OR indicator, we determine how much greater or less is the chance of an event, e.g. disease, death in one group compared to another group. "

•**OR≈1:** the odds of an event occurring in both groups is similar

•**OR <1:** in the study group (compared to the reference group) there is a lower odds of an event

•**OR> 1:** in the study group (compared to the reference group) there is a greater odds of a given event occurring.

Qualitative data

Qualitative data

The feature being a variable is characterized by a verbal description (can be subjective) and belongs only to one of the categories concerned. Categories are mutually exclusive

Association analysis between two qualitative features

Spearman's coefficient (ρ), Kendall tau, V Cramer

data are collected in the so-called cross table

Association analysis between two qualitative features

	B		
A	I	II	sum
I	a	b	a+b
II	c	d	c+d
sum	a+c	b+d	a+b+c+d

Association analysis between two qualitative features

Interpretation of Spearman ρ coefficient for cross table:

- ρ = 0 no correlation
- ρ> 0 the first variant of feature A coexists with the first variant of the B feature, and the second variant of the A feature coexists with the second variant of the B feature
- ρ <0 the first variant of the feature A coexists with the second variant of the B feature, and the second variant of the A feature coexists with the first variant of the B feature

Association analysis between two qualitative features

Example: Is there a relationship between vaccination and population resistance to disease?

The data is presented in the table below:

	sick	not sick	sum
vaccinated	13 (19,1%)	55 (80,9%)	68 (100%)
unvaccinated	28 (51,9%)	26 (48,1%)	54 (100%)
sum	41 (33,6%)	81 (66,4%)	122 (100%)

Association analysis between two qualitative features

Spearman:

$$\rho = 0{,}54;\ p < 0{,}001$$

ρ> 0 – average positive relationship between absence of disease and vaccination; correlation coefficient statistically significant (p <0.05)

Spearman can be used to assess the strength of the relationship between measurable, rank and dichotomous features

Association analysis between two qualitative features

	sick	not sick
vaccinated	13 (19,1%)	55 (80,9%)
unvaccinated	28 (51,9%)	26 (48,1%)

vaccinated

unvaccinated

risk of sick= 28/(28+26) = 0,5185

risk of sick = 13/(13+55) = 0,1912

chance of sick = 28/26 = 1,0769

chance of sick = 13/55 = 0,2364

RR of sick: vaccinated vs unvaccinated
= 0,1912/0,5185 = 0,369

OR of sick: vaccinated vs unvaccinated
= 0,2364/1,0769 = 0,220

Qualitative data

Chi-square test:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(E_i - T_i\right)^2}{T_i}$$

E_i – experimental (calculated) number in the category i

T_i – theoretical number in the category i

Qualitative data

Chi-square test:

- number in absolute values,
- the minimum allowable size of any category is 1,
- up to 1/5 of the category may be less than 5

Qualitative data

Example 1: (D. Schwartz, P. Lazar Elements of medical and biological statistics):

The table shows the results of a study of 298 people diagnosed with the gastric cancer.

Is there a relationship between the location of the cancer and sex?

	pylorus	stomach	sum
men	53 (44,5%)	66 (55,5%)	194 (100%)
women	48 (59,3%)	33 (40,7%)	104 (100%)
sum	101 (50,5%)	99 (49,5%)	298 (100%)

Qualitative data

An alternative to the chi test - square (more powerful / reliable)

χ^2=4,18

Fisher exact test
(2x2 table)

p=0,041

p=0,043

The result is statistically significant (there was a statistically significant relationship between the location of the tumor and sex ($p < 0.05$): in women group, the tumor is located in the pylorus more often than in men)

Qualitative data

	pylorus	stomach
men	53 (44,5%)	66 (55,5%)
women	48 (59,3%)	33 (40,7%)

Men

Women

Risk (pylorus) = 53/(53+66) = 0,4453

Risk (pylorus) = 48/(48+33) = 0,5926

RR (pylorus vs stomach) men vs women
= 0,4453/0,5926 = **0,752**

Odds (pylorus) = 53/66 = 0,8030

Odds (pylorus) = 48/33 = 1,4545

OR (pylorus vs stomach) men vs women
= 0,8030/1,4545 = **0,552**

95% confidence interval (95% CI) for the odds ratio of gastric tumor location in the pylorus versus the stomach (body) in males versus females: 0.31–0.98 does not contain an OR = 1 (no effect) value. Conclusion: in **men**, the tumor is located in the **pylorus less often** than in women. The direction and magnitude of the effect that gender (male) exerts on the location of the tumor (pylorus) is expressed by the **OR** (**95% CI**) = **0.552** (**0.31** – **0.98**) value, and the confidence interval not containing the value of **1** is an indirect evidence of **statistical significance** (p <0.05) of the observed effect.

Qualitative data

Example 1 (D. Schwartz, P. Lazar Elements of medical and biological statistics):

The group of 348 children was randomly divided into two subgroups. One was vaccinated with BCG vaccine from Company A and the other with BCG vaccine from Company B. The table below shows the results of the vaccine reactions.

Do vaccines work the same?

Qualitative data

Example 2:

Experimental data

vaccine	no reaction or weak reaction	Average reaction	severe reaction	sum
A	156 (88,1%)	12 (6,8%)	9 (5,1%)	177 (100%)
B	135 (78,9%)	29 (17,0%)	7 (4,1%)	171 (100%)

Qualitative data

χ^2=58,08

p=0,008

↓

Statistically significant result (statistically significant difference was noted in the effect of vaccines (p <0.05))

Qualitative data

vaccine	no reaction or weak reaction	severe reaction
A	156 (88,1%)	9 (5,1%)
B	135 (78,9%)	7 (4,1%)

A

Risk of strong reaction= 9/(9+156) = 0,9455

B

Risk of strong reaction = 7/(7+135) = 0,9507

RR of strong reaction : A vs B
= 0,9455/0,9507 = 0,9945

Odds of strong reaction = 9/156 = 0,0577

Odds of strong reaction = 7/135 = 0,0516

OR of strong reaction : A vs B
= 0,0577/0,0516 = 1,1127

The 95% confidence interval for the odds ratio of a severe reaction with vaccine A versus B of 0.63 - 1.97 includes an OR = 1, meaning no effect.

Qualitative data

vaccine	no reaction or weak reaction	average reaction
A	156 (88,1%)	12 (6,8%)
B	135 (78,9%)	29 (17,0%)

A

Risk of average reaction= 12/(12+156)=0,9286

B

Risk of average reaction = 29/(29+135) = 0,8232

RR average reaction: A vs B
= 0,9286/0,8232 = 1,1280

Chance of average reaction = 12/156 = 0,0769

Chance of average reaction = 29/135 = 0,2148

OR average reaction: A vs B
= 0,0769/0,2148 = 0,3581

The 95% confidence interval for the odds ratio of the average reaction after vaccine A versus B of 0.20 - 0.63 does not include an OR = 1 (no effect) value.

Supplementary materials

Inaccuracies in the statistical nomenclature

unpaired variable

≠ independent variable *

* in the Statistica software the concept of the independent variable is incorrectly applied to unrelated variables (applies to most statistical tests)

$$y = \boxed{x}$$

Correct definition: Independent variable - a variable whose values are being changed

Inaccuracies in the statistical nomenclature

paired variable

≠ dependent variable *

* in the Statistica software the concept of the dependent variable is incorrectly applied to related variables (applies to most statistical tests)

$$\boxed{y} = x$$

Correct definition: Dependent variable - the variable that we measure

Inaccuracies in the statistical nomenclature

Variance for the sample and variance for the population

When calculating the variance or standard deviation in Excel, select the function VARIABLE () or STANDARDDev (), respectively, which calculate these scatter measures "based on a sample of the population".

Do not use the VARIANCE.POPUL () or STANDARD.Dev.POPUL () functions to calculate these measures "based on the entire population", because the obtained value is lower than the real estimate of variance in the population by a coefficient (n-1) / n for the value 'for a population sample', where n is the size of the sample under study. For the standard deviation, the obtained value is underestimated by a factor equal to the root of z (n-1) / n.