

Lecture 6

Concepts from statistical testing

Types of tests

Data normality and its importance for testing

Parametric vs. non-parametric tests

- **Parametric tests**

- Have assumptions about the distribution of the input data (e.g. normal distribution)
- Given the same N and assumptions, they have higher test power than non-parametric tests
- If the assumptions of parametric tests are not met, then the power of the test drops sharply and the test result may be completely wrong and meaningless

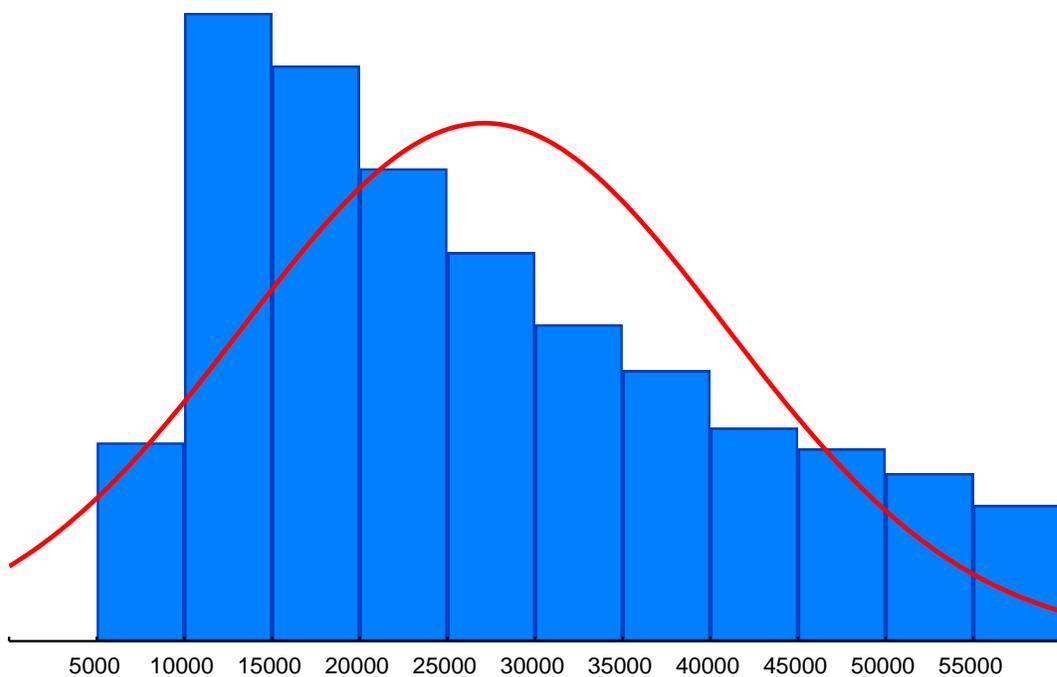
- **Non-parametric tests**

- They have no assumptions about the distribution of the input data, so they can be used even with asymmetric distributions, outliers, or non-detectable distributions
- The reduced power of these tests is due to the reduction of the information value of the original data, where non-parametric tests do not use the original values, but most often only their order

Problems of parametric and non-parametric tests

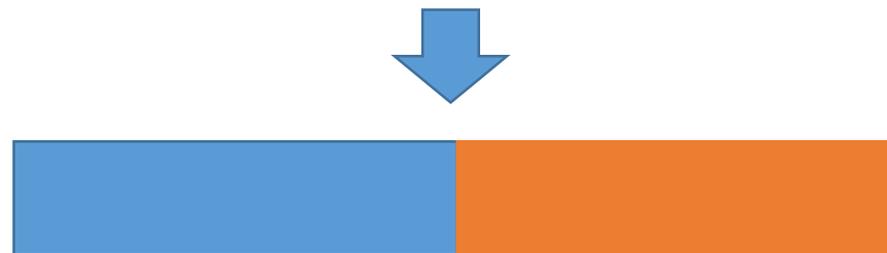
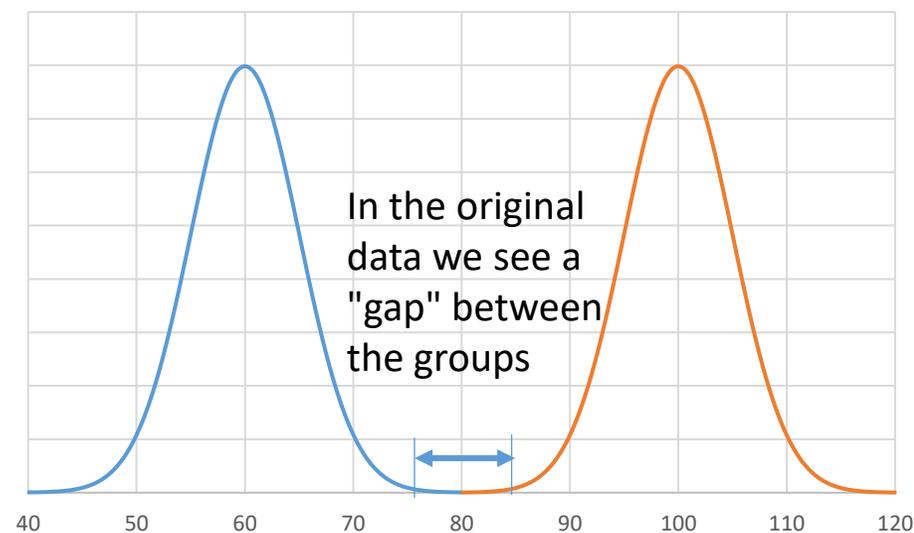
Parametric tests

- Real data do not fit the model distribution



Non-parametric tests

- By converting the data to order, we lose some of the information



We lose this information when we convert to the order.

One- and two-sample tests

- **One-sample tests**

- They compare a single sample (one sample, one-sample tests) with a reference value (or statistical parameter of the target population)
- Thus, the test compares the distribution of values (sample) with a single number (reference value, target population value)
- The question asked in the test can be related to the mean, variance, proportion of values and other statistical parameters describing the sample

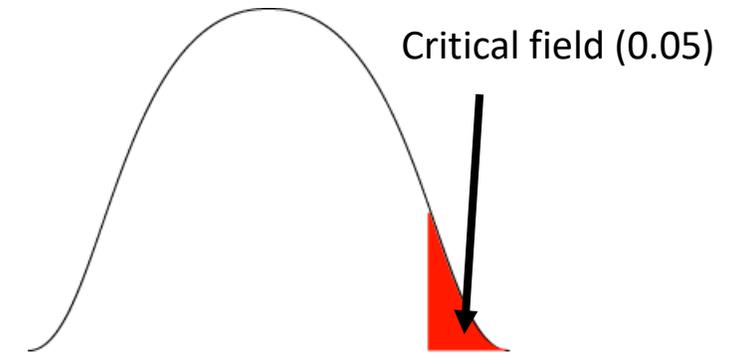
- **Two-sample tests**

- They compare two samples with each other (two sample, two-sample tests)
- The test compares two distributions of values
- The question asked in the test can again be related to the mean, variance, proportion of values and other statistical parameters describing the sample
- In addition to tests for two groups of values, there are of course tests for more groups of data

Unilateral and bilateral hypotheses

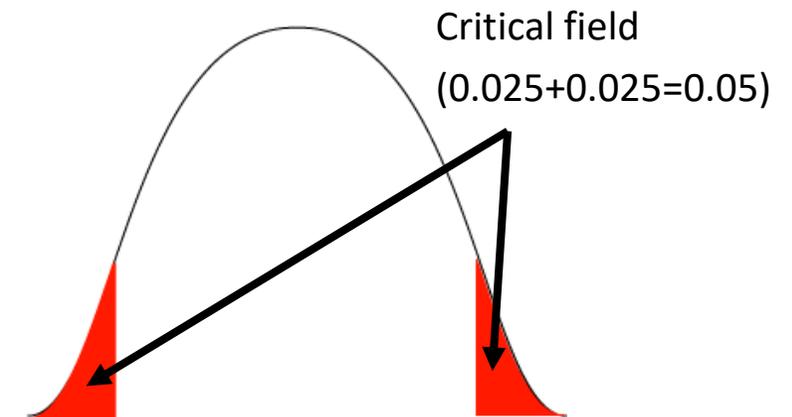
- **One-sided tests (one-tailed)**

- The hypothesis of the test is constructed asymmetrically, i.e. we ask for greater than/less than
- The test can only have a double output - one of the values is larger (smaller) than the other and all other cases
- Only if there is a clear hypothesis - otherwise it can be challenged for being purposive (easier to confirm significance with appropriate choice of testing direction)



- **Two-tailed tests**

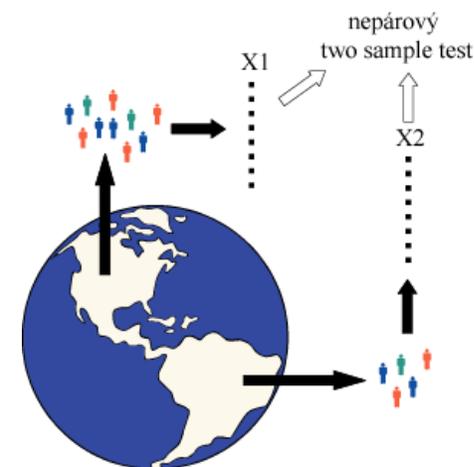
- The hypothesis of the test asks an equal/not equal question
- The test can have a triple output - less than - equal to - greater than
- The situation does not equal is therefore the sum of two possible outcomes of the test (smaller+larger)
- Meaning-neutral



Unpaired vs. paired design

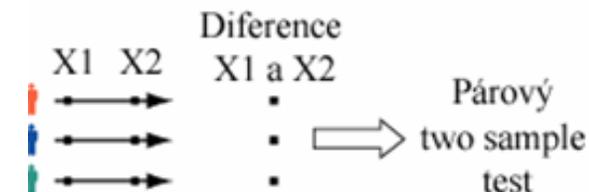
• Unpaired design

- The groups of data being compared are completely independent of each other (also independent design), e.g. people from different countries, independent groups of patients with different treatments, etc.
- When calculating, it is necessary to take into account the characteristics of both data sets



• Pair design

- There is a link between the objects in the groups being compared, e.g. human before and after surgery, reaction of the same rat strain, etc.
- The binding may be either directly given or merely assumed (in which case it must be verified)
- The test is essentially performed on the groups' differences, not on their original data



Important notes on hypothesis testing

- **Not rejecting the null hypothesis does not automatically mean accepting it!** It may be a situation where we do not have enough information to reject the null hypothesis.
- **The level of significance achieved in a test (whether 5%, 1% or 10%) must not be taken blindly as a threshold for the existence/non-existence of the effect being tested.**
- **A small p-value does not necessarily mean a large effect.** The value of the test statistic and the p-value may be affected by the large sample size and the small variability of the observed data.
- **The results of the testing must be viewed critically** - it is a conclusion based "only" on one sample.
- **Statistical significance** indicates that the observed difference is not due to chance, but may not mean that it is significant in reality. **Practical (clinical) significance is also important.**

Statistical tests and normality

- Normality of data is one of the assumptions of so-called parametric tests (tests based on the assumption of a distribution) - e.g. t-tests
- In general, any statistical method whose algorithm includes the calculation of the mean or standard deviation has the assumption of a normal distribution
- If the data are not normal, they do not fit the model distribution that is used for the calculation (t-distribution) and the test may lie
- So the solution is:
 - Data transformation to achieve normality of distribution
 - Nonparametric tests - these tests have no (or minimal) assumptions about the distribution of the data

Type of comparison	Parametric test	Non-parametric test
2 groups of data unpaired:	Unpaired t-test	Mann-Whitney test
2 groups of data in pairs:	Paired t-test	Wilcoxon test, sign test
Multiple groups unpaired:	ANOVA (analysis of variance)	Kruskal-Wallis test
Correlation:	Pearson coefficient	Spearman's coefficient

Normality tests

- Normality tests work with the null hypothesis that there is no difference between the processed distribution and the normal distribution. However, it is always a good idea to look at the histogram as some deviations from normality, such as bimodality, are not detected by some tests.

Chi-square goodness-of-fit test

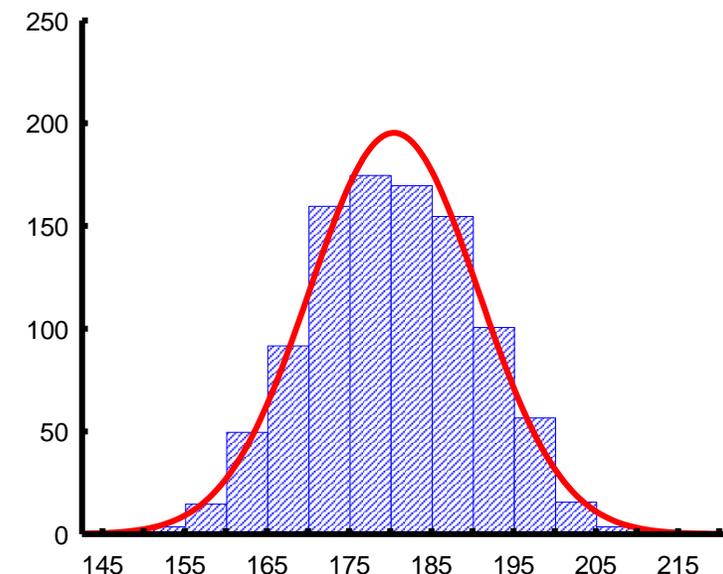
- In a goodness-of-fit test, the data are categorised (similar to histogram generation), these intervals are normalised (converted to a normal distribution) and the expected values at the intervals are calculated according to the general normal distribution formulae if the distribution were normal. The observed normalised frequencies are then compared with the expected frequencies using the χ^2 goodness-of-fit test. The test gives good results, but is demanding on n , the amount of data, to produce a sufficient number of classes of values.

Kolmogorov - Smirnov test

- This test is often used, it can find outliers well, but it assumes symmetry of values rather than normality directly. It is a non-parametric test for comparing the difference between two distributions. It is based on finding the difference between the real cumulative distribution (sample) and the theoretical cumulative distribution. It should only be calculated if we know the mean and standard deviation of the hypothetical distribution; if we do not know these values, a modification of the Liliefors test should be used.

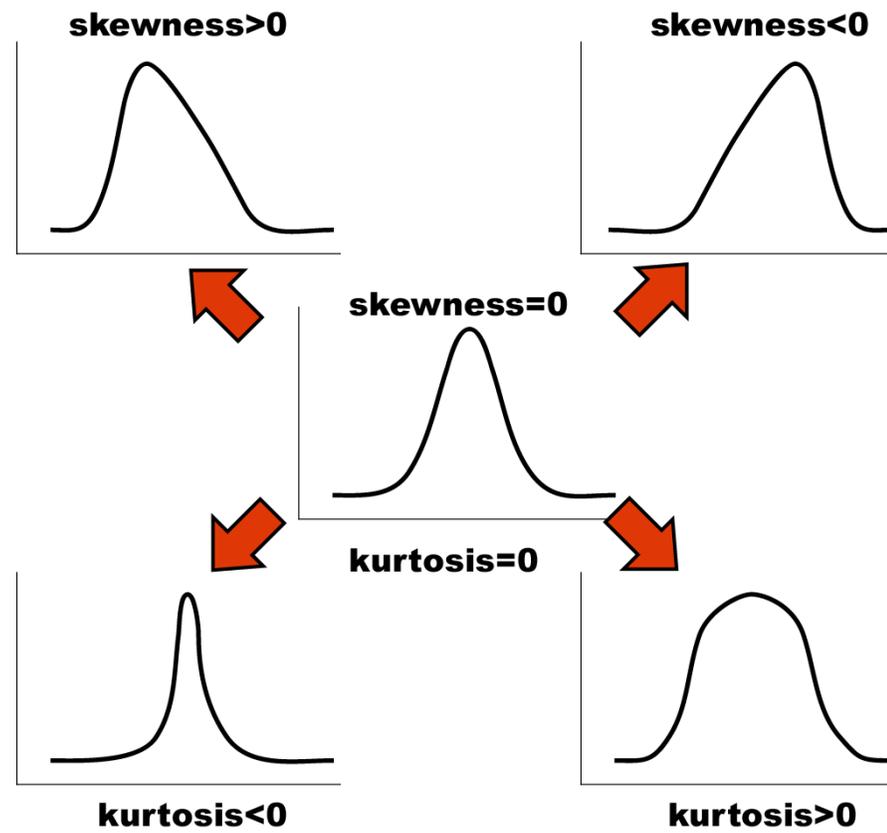
Shapiro-Wilk test

- It is a non-parametric test applicable even at very small n (10) with good test power, especially compared to alternative test types, it is aimed at testing symmetry.

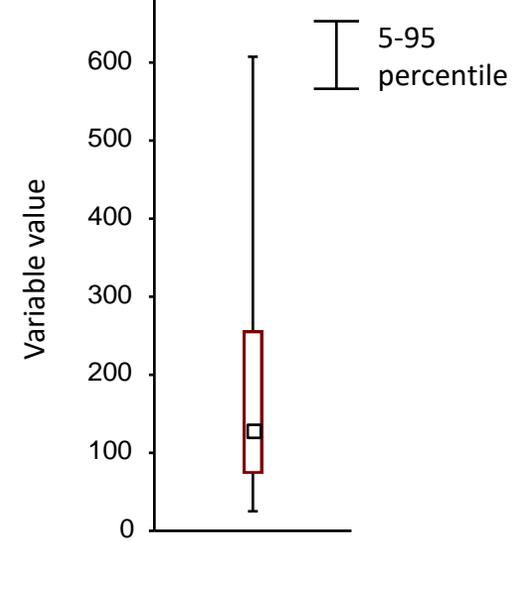
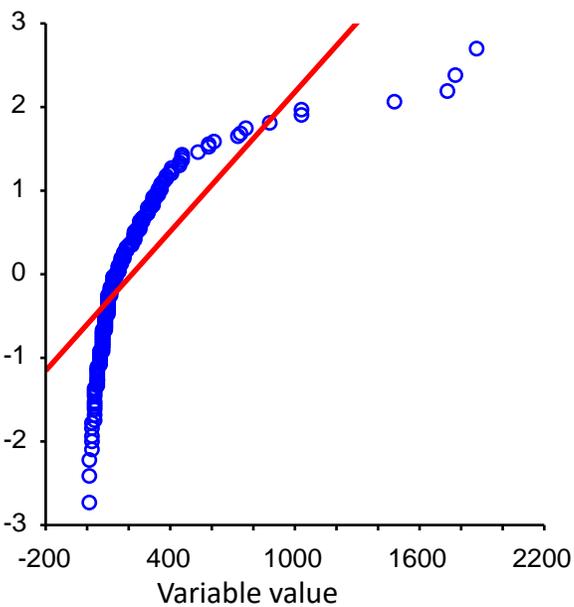
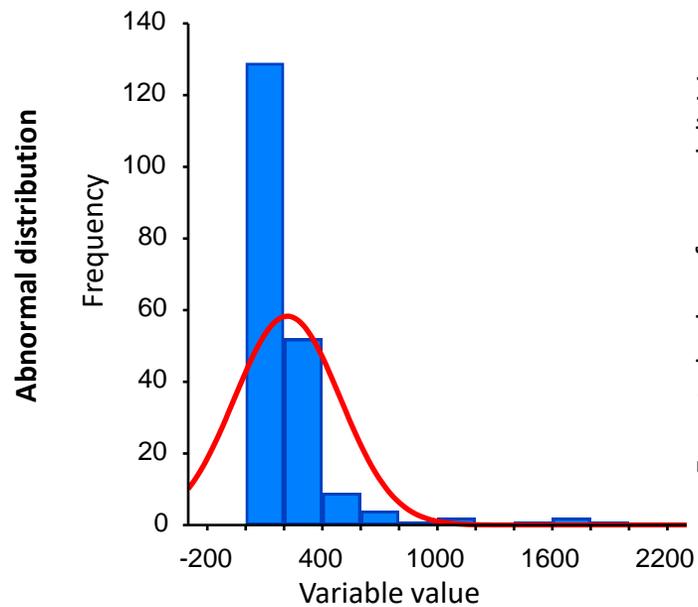
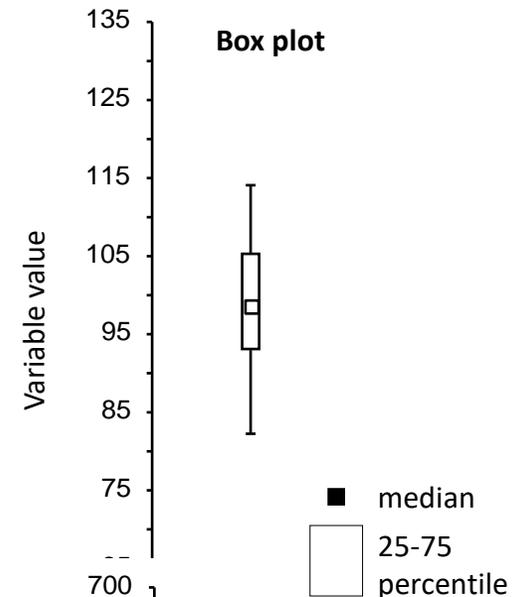
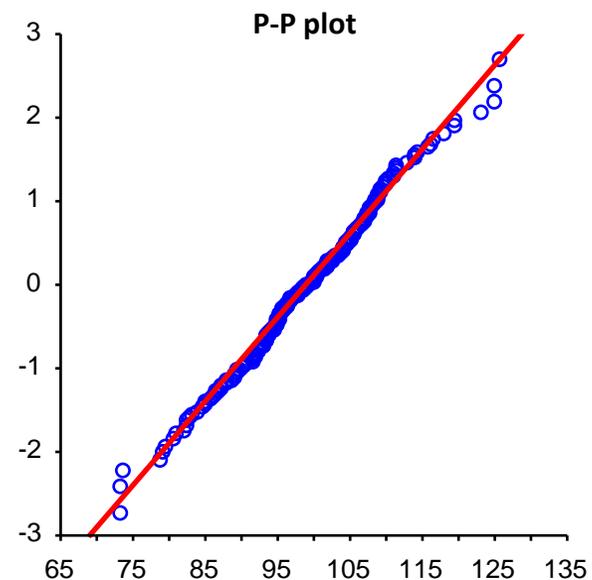
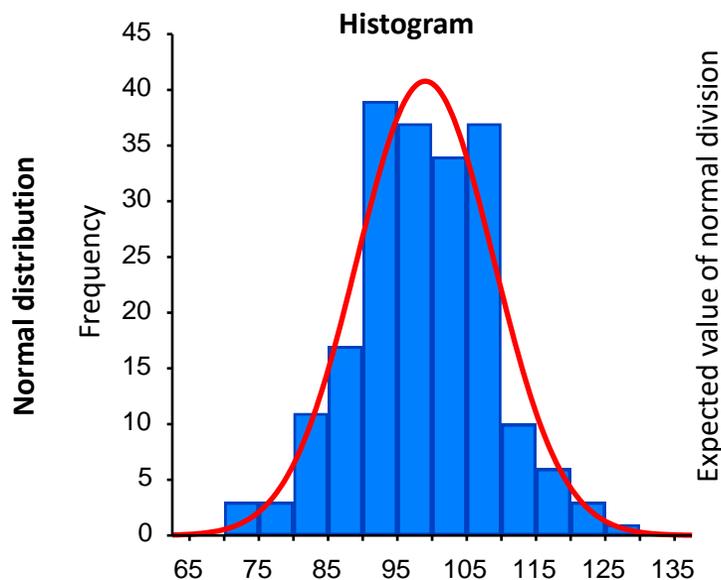


Skewness and pointedness as tests of normality

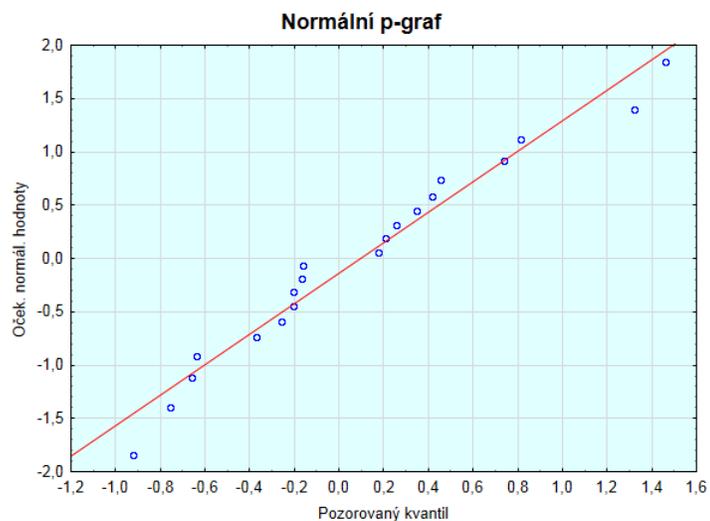
- The normal distribution, skewness and kurtosis parameters can be used for normality testing, but only for large samples (skewness - 100, skewness - 500).



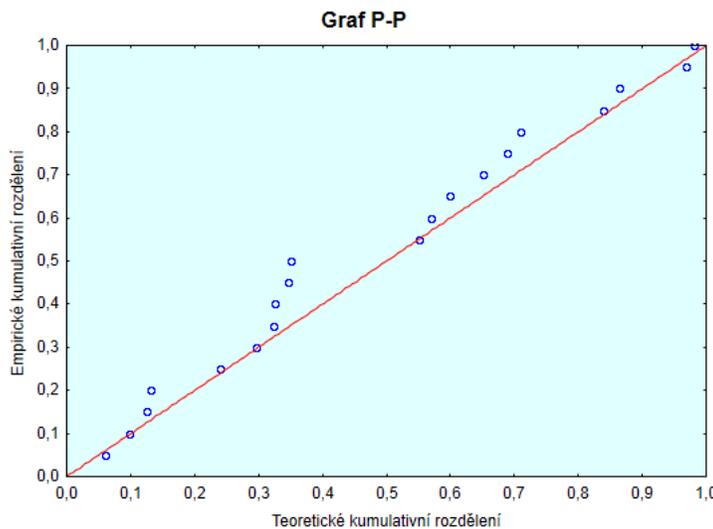
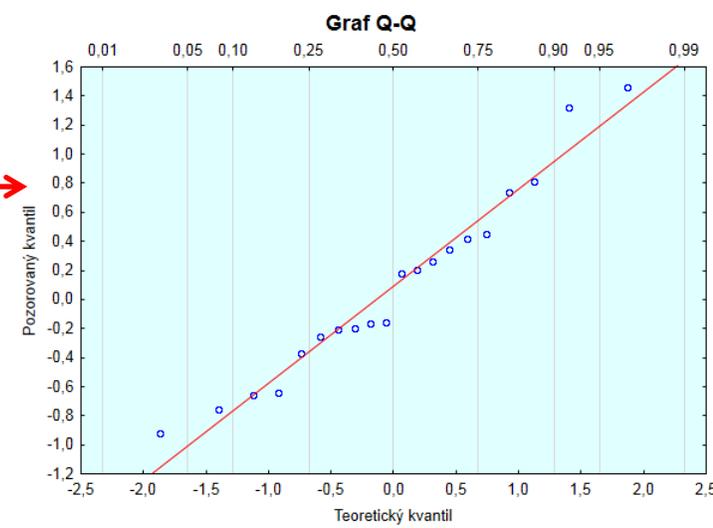
Visual assessment of normality I



Visual assessment of normality II



- Axle replacement only
- Observed and theoretical quantile shown



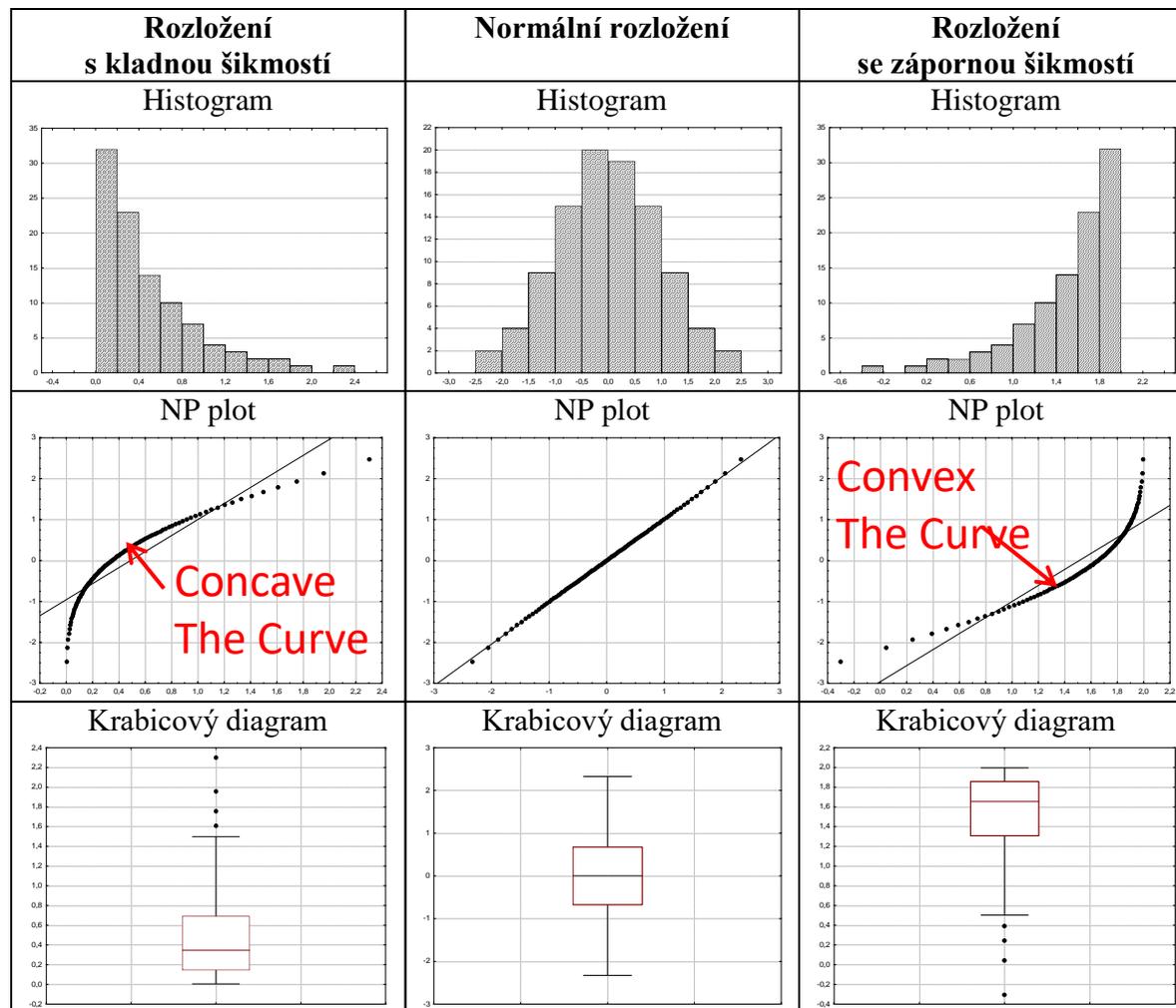
- Cumulative distribution plotted

REMEMBER:

If the data come from a normal distribution, then the points will lie around the line

A warning sign icon consisting of a red triangle with a white exclamation mark inside.

Visual assessment of normality III



Educational materials: Computational Statistics, RNDr. Marie Budíková, Dr., 2011

Parametric one-sample statistical tests

One-sample t-test

One-sample test of variance

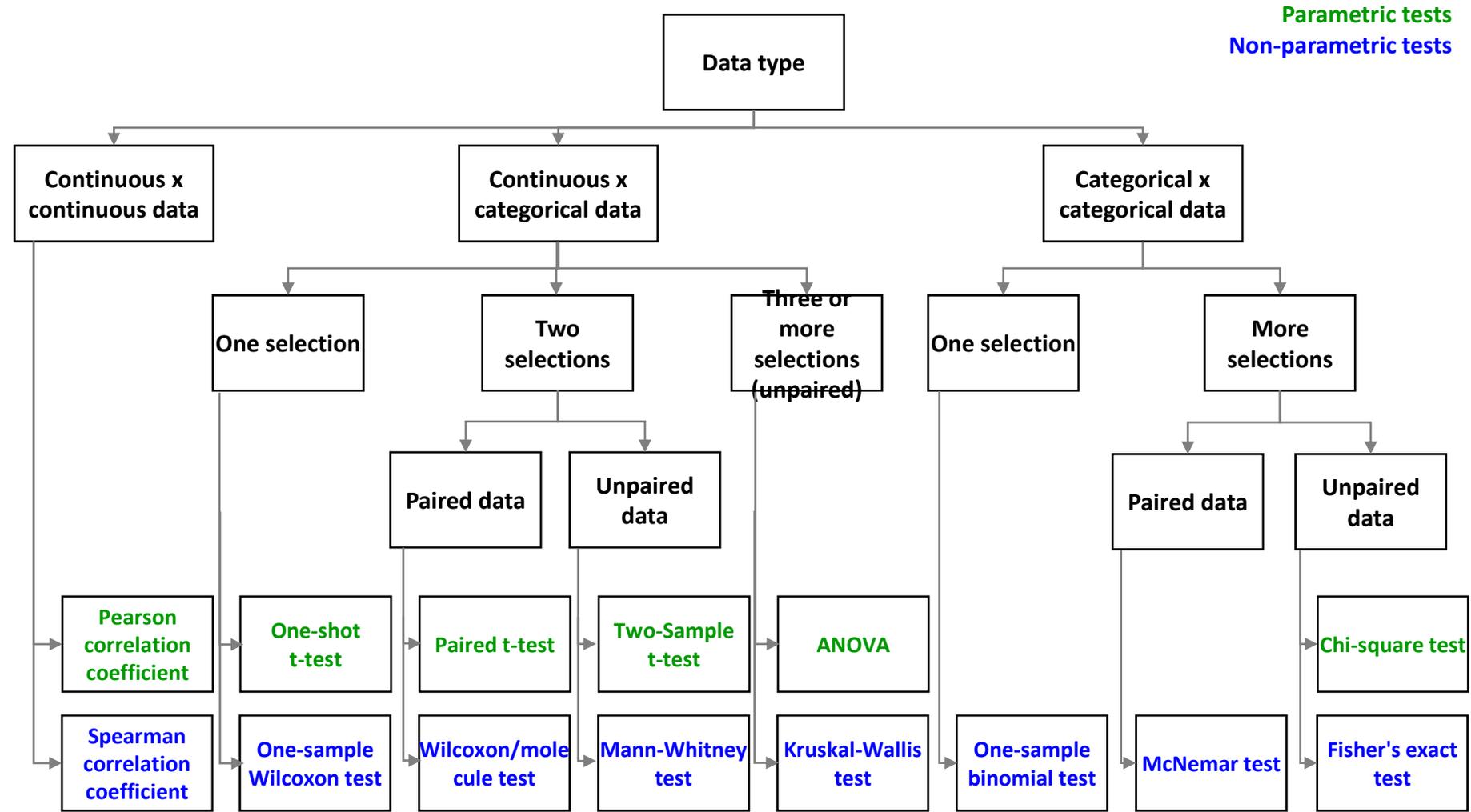
Annotation

- One-sample statistical tests compare some descriptive sample statistic (mean, standard deviation) with a single number whose significance is statistically the value of the target population
- From the point of view of statistical theory, it is about verifying whether a given sample comes from the target population being tested.

Summary of statistical tests

Type of comparison	The null hypothesis	Parametric test	Non-parametric test
1 data selection vs. reference value	The mean value is equal to the chosen reference value.	single-line t-test / z-test	One-sample Wilcoxon test
2 independent data sets (test of agreement of means)	Mean values do not differ between groups.	unpaired t-test	Mann-Whitney test
2 independent data groups (test for homoskedasticity)	The dispersion of both groups is identical.	F-test	Levene's test
2 pairwise dependent data selections	The difference (differential) of the paired values is zero.	paired t-test	Wilcoxon test; sign test
Congruence of the sampling distribution with the theoretical distribution	The data distribution corresponds to the theoretical (selected) distribution.	goodness of fit test (χ^2 test)	Shapiro-Wilk test; Kolmogorov-Smirnov test; Lilliefors test
3 or more groups unpaired (test of agreement of means)	Mean values do not differ between groups.	ANOVA	Kruskal-Wallis test
Correlation	There is no relationship between the values of the two selections.	Pearson correlation coefficient	Spearman's correlation Coefficient

Basic decision making on the selection of statistical tests



Parametric one-sample tests

- Prerequisite: **data normality**
- **One-sample z-test** (comparison of the baseline and sample, we know the mean and variance of the baseline)
- **Student's one-sample t-test** (testing differences between two means) - (comparing the baseline and sample, we know the mean but do not know the variance of the baseline; we replace it with the sample variance of our data)
- **Chi-square test** (testing the difference between target vs. sample population)

One-sample z and t test

- In the case of one-sample tests, it is a comparison of the data selection (i.e. one sample) with the target population. For parametric tests, the data set must have a normal distribution.
- The difference between a one-sample z-test and a t-test is knowing the variance of the underlying population (z-test) or replacing it with the sample variance of our data (t-test):

$$z = \frac{\bar{x} - \mu}{\rho} \sqrt{N}$$

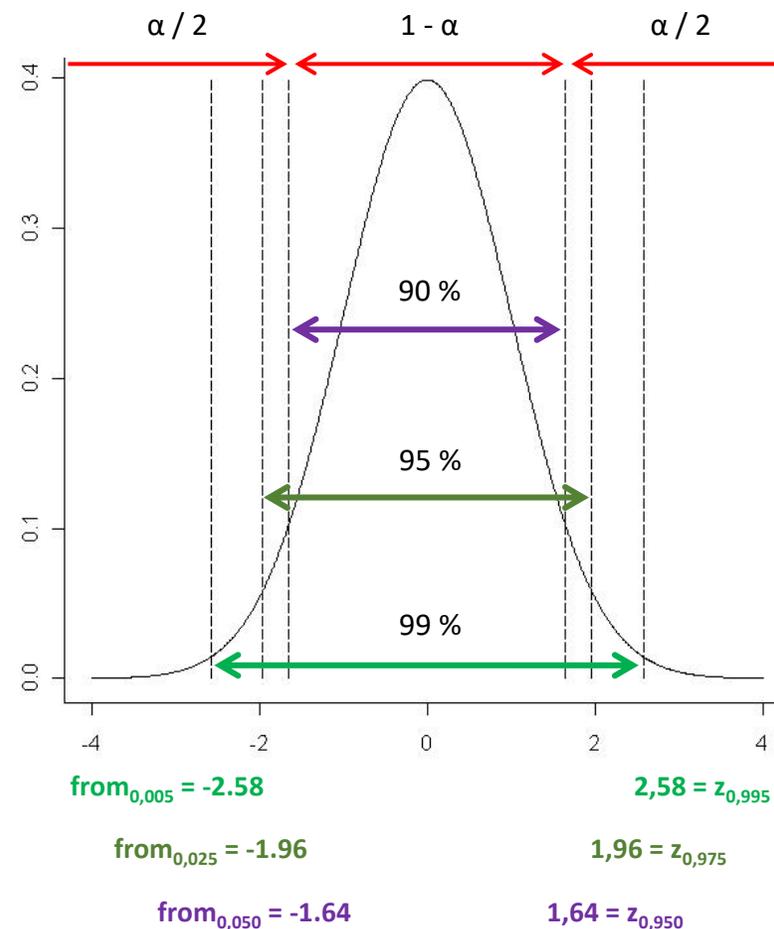
t-test:

$$t = \frac{\bar{x} - \mu}{s} \sqrt{N}$$

H_0	H_A	Test statistics	Critical value
$\bar{x} \leq \mu$	$\bar{x} > \mu$	z / t	$z > z_{1-\alpha} / t > t_{1-\alpha}^{N-1}$
$\bar{x} \geq \mu$	$\bar{x} < \mu$	z / t	$z < z_{\alpha} / t < t_{\alpha}^{N-1}$
$\bar{x} = \mu$	$\bar{x} \neq \mu$	z / t	$ z > z_{1-\alpha/2} / t > t_{1-\alpha/2}^{N-1}$

Example: z-test for one selection 1

- In a population-based epidemiological survey, the mean prostate volume in men was found to be 32.73 ml (SD = 18.12 ml).
- At a significance level of $\alpha = 0.05$, we want to test whether men over 70 differ from the general population.
- We have a random sample of size $n = 100$ and a sample mean of 36.60 ml.
- We want to check the validity:
 - $H_0 : \mu = 32.73$
 - $H_A : \mu \neq 32.73$



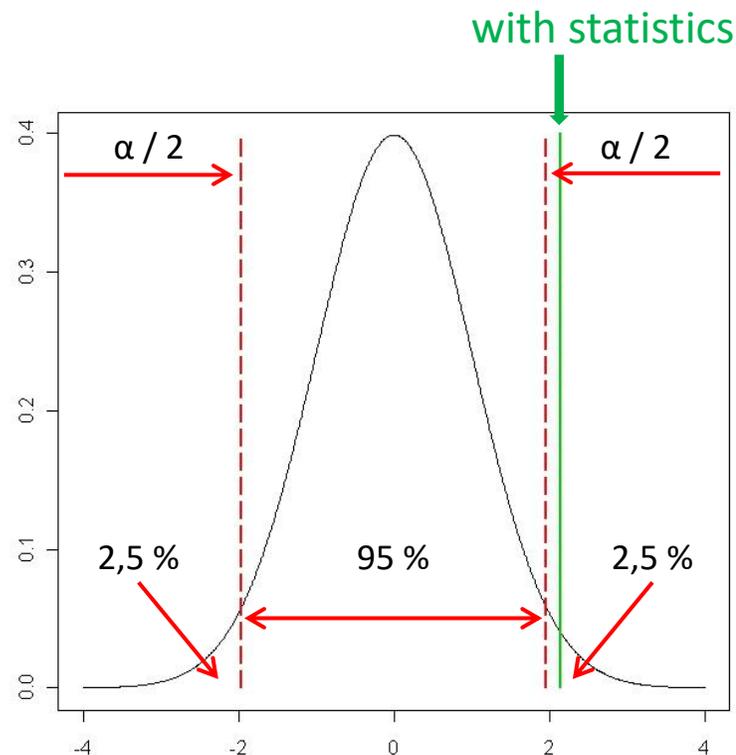
Example: z-test for one selection 2

- The value of the test statistic:
$$z = \frac{\bar{x} - \mu}{\rho} \sqrt{N} = \frac{36,60 - 32,73}{18,12} \sqrt{100} = 2,14$$

- Can we reject the null hypothesis at the significance level of the $\alpha = 0.05$ test or not?

$$z = 2,14 > 1,96 \left(z_{1-\alpha/2} = z_{0,975} \right)$$

- We reject the null hypothesis of equality of prostate volume in men over 70 years of age to the population value of 32.73 ml at the significance level $\alpha = 0.05$ because the resulting value from the statistic is greater than the critical value (the corresponding quantile) of the $N(0,1)$ distribution.



Example: t-test for one selection



- A certain urban bus line has an average speed of 8 km/h during peak hours. It was considered whether a change of route would lead to a change in the average speed. The new route was therefore driven on ten randomly selected days and the following average speeds were found: 8.4; 7.9; 9.0; 7.8; 8.0; 7.8; 8.5; 8.2; 8.2; 9.3. Decide whether the change of route leads to a change in the average speed. Assume a normal distribution and $\alpha=0.05$.
- Procedure:
 1. At the 0.05 significance level, we test the hypothesis $H_0 : \mu = 8$, against $H_A : \mu \neq 8$
 2. Calculate the arithmetic mean and variance of the sample.
 3. Calculate the test statistic t:
$$t = \frac{\bar{x} - \mu}{s} \sqrt{N} = \frac{8,310 - 8}{0,507} \sqrt{10} = 1,934$$
 4. We compare the calculated t with the critical value:
$$t_{1-\alpha/2}^{N-1} = t_{0,975}^9 = 2,262$$
 5. If $- >$ statistically insignificant difference of the tested parameters at the chosen α ; **we do not reject the null hypothesis**, at the significance level $\alpha=0.05$ we could not show that the change of the route would result in a change of the average speed.

Typical SW outputs (Statistica, similar in others)

Samplng average
(average of observed data)

Scope of selection

Standard error

Test criterion value

Degree of freedom

Test of means against reference constant (value) (04_doprava.sta)								
Variable	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
rychlost	8,310000	0,506513	10	0,160174	8,000000	1,935401	9	0,084934

Samplng standard deviation
(observed data)

Reference constant-assumed magnitude of the mean value

ATTENTION: Valid for both sides of the test!!!



Example to solve: t-test 1

Data - antibiotic concentration in the target organ

- In 1000 measurements of the antibiotic, the mean concentration in the target organ was found to be 202.5 units and the standard deviation was 44 units.
- The required antibiotic concentration is 200 units.

Research questions

1. Is the difference of 2.5 significant given the trait variability at the 5% significance level?
2. What is the true level of significance?

Example to solve: t-test 1

Data - antibiotic concentration in the target organ

- In 1000 measurements of the antibiotic, the mean concentration in the target organ was found to be 202.5 units and the standard deviation was 44 units.
- The required antibiotic concentration is 200 units.

Research questions

1. Is the difference of 2.5 significant given the trait variability at the 5% significance level?
2. What is the true level of significance?

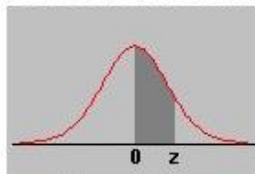
$$t = \frac{\bar{x} - \mu}{s} \sqrt{N} = \frac{202,5 - 200}{44} \sqrt{1000} = 1,797$$

Example to solve: t-test 1

$$t = \frac{\bar{x} - \mu}{s} \sqrt{N} = \frac{202,5 - 200}{44} \sqrt{1000} = 1,797 \sim 1,8$$

$$t_{1-\alpha/2}^{N-1} = t_{0,975}^{999} = 1,960$$

Area between 0 and z



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706

Research questions

1. Is the difference of 2.5 significant given the trait variability at the 5% significance level?
 - We do not reject the null hypothesis
2. What is the true level of significance?
 - $p = 2*(1-0,4641)=0,072$

Example to solve: t-test 2

Data - enzyme activity in cells

- The mean of 3.5 units and standard deviation of 1 were found when the enzyme activity was detected in a sample of 25 cells.

Research questions

1. The question is, do the measured values of our sample differ from the results of an earlier large study on the whole target population, where an average activity of 2.5 units was found?
2. question - what is the minimum deviation of X from another value we would capture at the given values?
3. assuming that in practical terms a significant deviation is already 0.2 units, what is the minimum number of measurements we need to make to be able to demonstrate it ?

Example to solve: t-test 2

Data - enzyme activity in cells

- The mean of 3.5 units and standard deviation of 1 were found when the enzyme activity was detected in a sample of 25 cells.

Research questions

1. The question is, do the measured values of our sample differ from the results of an earlier large study on the whole target population, where an average activity of 2.5 units was found?

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{3,5 - 2,5}{1} \sqrt{25} = 5$$

$$t_{0,975}^{24} = 2,064 \Rightarrow t > t_{1-\alpha/2}^{24} \Rightarrow H_0 \text{ rejected at } \alpha = 0.05$$

Example to solve: t-test 2

Data - enzyme activity in cells

- The mean of 3.5 units and standard deviation of 1 were found when the enzyme activity was detected in a sample of 25 cells.

Research questions

2. question - what is the minimum deviation of X from another value we would capture at the given values?

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{d}{s} \sqrt{n} \rightarrow d = \frac{t_{1-\alpha/2}^v}{\sqrt{n}} s \rightarrow d = \frac{2,064}{5} \cdot 1$$

Example to solve: t-test 2

Data - enzyme activity in cells

- The mean of 3.5 units and standard deviation of 1 were found when the enzyme activity was detected in a sample of 25 cells.

Research questions

3. assuming that in practical terms a significant deviation is already 0.2 units, what is the minimum number of measurements we need to make to be able to demonstrate it ?

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{d}{s} \sqrt{n} \rightarrow n = \left(\frac{t_{1-\alpha/2}^v}{d} s \right)^2$$

One-sample test for variance

- In the case of one sample tests, it is a comparison of the data selection (i.e. one sample) with the target population. For parametric tests, the data set must have a normal distribution.

Chi-squared test:

$$\chi^2 = \frac{(N - 1)s^2}{\sigma^2}$$

H_0	H_A	Test statistics	Critical value
$s^2 \leq \sigma^2$	$s^2 > \sigma^2$	χ^2	$\chi^2 > \chi_{1-\alpha}^{2(N-1)}$
$s^2 \geq \sigma^2$	$s^2 < \sigma^2$	χ^2	$\chi^2 < \chi_{\alpha}^{2(N-1)}$
$s^2 = \sigma^2$	$s^2 \neq \sigma^2$	χ^2	$\chi^2 > \chi_{1-\alpha/2}^{2(N-1)}$ nebo $\chi^2 < \chi_{\alpha/2}^{2(N-1)}$

Nonparametric one-sample statistical tests

One-sample t-test

One-sample test of variance

Parametric vs. non-parametric tests

Parametric tests

- Have assumptions about the distribution of the input data (e.g. normal distribution)
- Given the same N and assumptions, they have higher test power than non-parametric tests
- If the assumptions of parametric tests are not met, then the power of the test drops sharply and the test result may be completely wrong and meaningless



Non-parametric tests

- They require fewer assumptions about the distribution of the input data, so they can be used even with asymmetric distributions, outliers, or non-detectable distributions
- The reduced power of these tests is due to the reduction of the information value of the original data, where non-parametric tests do not use the original values, but most often only their order
- Related to small file size (we are not able to verify the normality of the data)

Why don't parametric and non-parametric tests come out the same?

One-sample Wilcoxon test

- The assumption is that the data are symmetrically distributed around the median.
- Tests whether the **median of** one selection is equal to c (in the case of a pairwise design, $x_{0.5}$ is represented by the median difference of the values)

$$H_0 : x_{0.5} = c \text{ versus } H_1 : x_{0.5} \neq c.$$

Procedure:

1. We calculate the differences of the selection values with the median value tested.
2. We arrange the absolute values of the differences in ascending order and assign them a ranking.
3. We compute the statistics S_w^+ and S_w^- , which correspond to the **sum of the order of positive (S_w^+) and negative differences (S_w^-)**. We take the minimum of S_w^+ and S_w^- as the final value of the test statistic. We reject the null hypothesis if the value of the test statistic is less than or equal to the tabulated critical value (at a given significance level and number of non-zero differences).

or

3. For $N > 30$, the asymptotic normality of the S statistic can be used.

$$E(S_{w+}) = \frac{n(n+1)}{4} \quad D(S_{w+}) = \frac{n(n+1)(2n+1)}{24} \quad Z = \frac{S_{w+} - E(S_{w+})}{\sqrt{D(S_{w+})}} \approx N(0,1)$$

- If $|Z| \geq z_{1-\alpha/2}$ we reject the null hypothesis that the median selection is equal to c .

One-sample sign test

- It can be used in situations where the assumption of symmetry of the distribution around the median is not satisfied.
- Tests whether the median of one selection is equal to c (in the case of a pairwise design, $x_{0.5}$ is represented by the median of the difference of the values)

$$H_0 : x_{0.5} = c \text{ versus } H_1 : x_{0.5} \neq c.$$

Procedure:

1. We calculate the differences of the selection values with the median value tested.
 2. Calculate the statistic S_z^+ , which corresponds to the number of positive differences → **the test does not use the rank values of the original data but only the information whether the value is realized above or below the median** → the power of the test is reduced
 3. We reject the null hypothesis if the statistic S_z^+ realizes in the critical range of values $W=(0, k_1) \cup (k_2, n)$, where n corresponds to the number of non-zero differences and the values of k_1 and k_2 can be traced in mathematical tables.
- **or**
 - 3. For $N > 20$, the asymptotic normality of the S statistic can be used.

$$E(S_{z^+}) = \frac{n}{2} \quad D(S_{z^+}) = \frac{n}{4} \quad Z = \frac{S_{z^+} - E(S_{z^+})}{\sqrt{D(S_{z^+})}} \approx N(0,1)$$

If $|Z| \geq z_{1-\alpha/2}$ we reject the null hypothesis that the median selection is equal to c .

Example: one-sample test

- For 15 randomly selected patients, the amount of time they had to spend in the waiting room before being invited by a nurse to the office was assessed. At the 5% significance level, test the null hypothesis that the median waiting time is equal to half an hour.



Example: one-sample test - Wilcoxon test

- For 15 randomly selected patients, the amount of time they had to spend in the waiting room before being invited by a nurse to the office was assessed. At the 5% significance level, test the null hypothesis that the median waiting time is equal to half an hour.

Patient no.	waiting time (min)	median	difference	difference	order
1	1	30	-29	29	15
2	45	30	15	15	10
3	25	30	-5	5	3.5
4	15	30	-15	15	10
5	34	30	4	4	2
6	19	30	-11	11	8
7	31	30	1	1	1
8	25	30	-5	5	3.5
9	8	30	-22	22	14
10	12	30	-18	18	12
11	20	30	-10	10	6
12	15	30	-15	15	10
13	40	30	10	10	6
14	20	30	-10	10	6
15	10	30	-20	20	13



$$S = 19_{w^+}$$

$$S = 101_{w^-}$$

$$\min(S_{w^+}, S_{w^-}) = 19$$

Critical value $w_{15}(0.05) = 25$

The value of the test static is smaller than the critical value \rightarrow **reject H_0**

Example: one-sample test - Sign test

- For 15 randomly selected patients, the amount of time they had to spend in the waiting room before being invited by a nurse to the office was assessed. At the 5% significance level, test the null hypothesis that the median waiting time is equal to half an hour.

Patient no.	waiting time (min)	median	difference	Greater than the median?
1	1	30	-29	No
2	45	30	15	Yes
3	25	30	-5	No
4	15	30	-15	No
5	34	30	4	Yes
6	19	30	-11	No
7	31	30	1	Yes
8	25	30	-5	No
9	8	30	-22	No
10	12	30	-18	No
11	20	30	-10	No
12	15	30	-15	No
13	40	30	10	Yes
14	20	30	-10	No
15	10	30	-20	No



$$S = 4_z^+$$

Critical region: $W=(0,3) \cup (12,15)$

The value of statistics is realized outside of critical range of values → **do not reject H_0**

Example: software solution

1) Wilcoxon test output

Test statistics: $\min(S_w^+, S_w^-)$

Wilcoxon Matched Pairs Test (v cekarne.sta)				
Marked tests are significant at p < .05000				
Pair of Variables	Valid N	T	Z	p-value
doba & median	15	19.00000	2.328644	0.019879

Statistics and p-value for the asymptotic variant of the test (use only for $N > 30$)

Number of non-zero differences

2) Sign test output

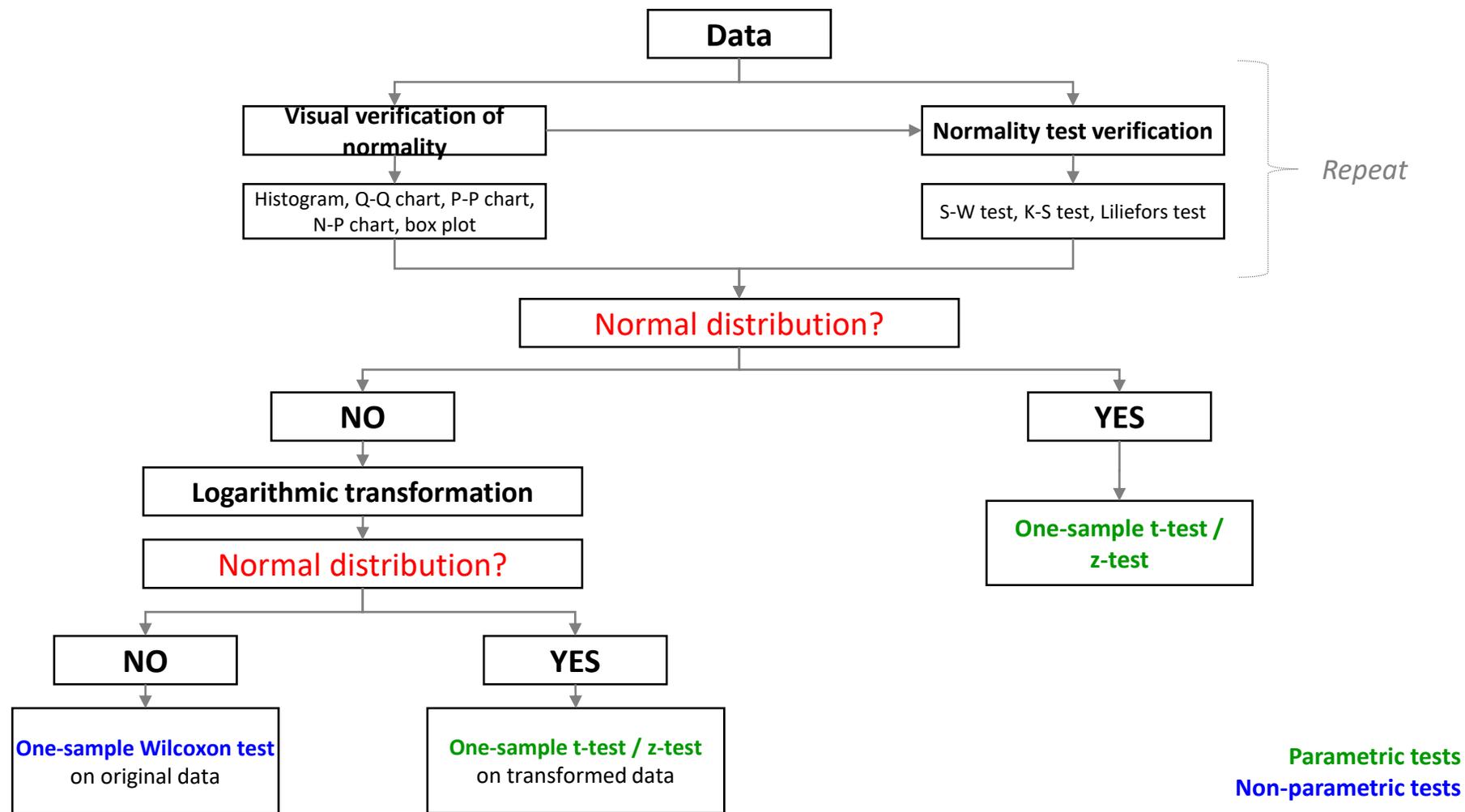
Proportion of values less than the tested median

Sign Test (v cekarne.sta)				
Marked tests are significant at p < .05000				
Pair of Variables	No. of Non-ties	Percent v < V	Z	p-value
doba & median	15	73.33333	1.549193	0.121335

Statistics and p-value for the asymptotic variant of the test (use only for $N > 20$)

Number of non-zero differences

Flowchart for testing with one-sample tests



Parametric two-sample statistical tests

Two-sample unpaired t-test

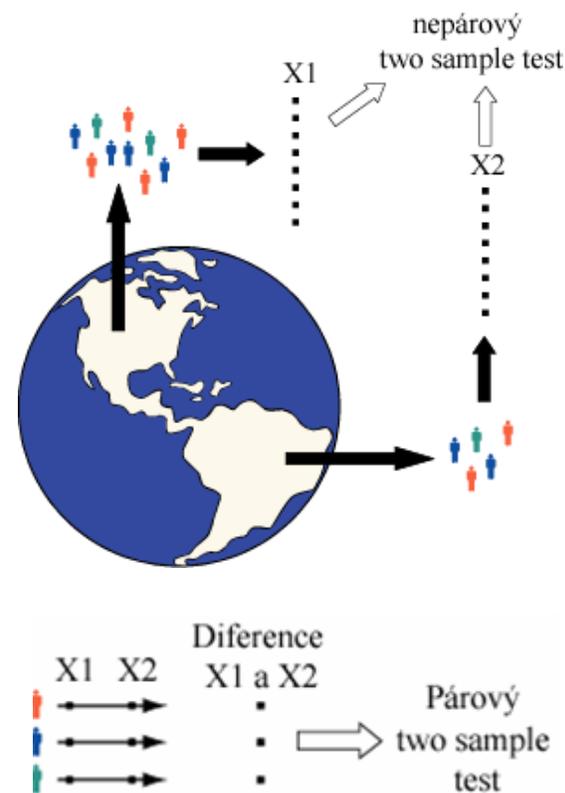
Two-sample paired t-test

Annotation

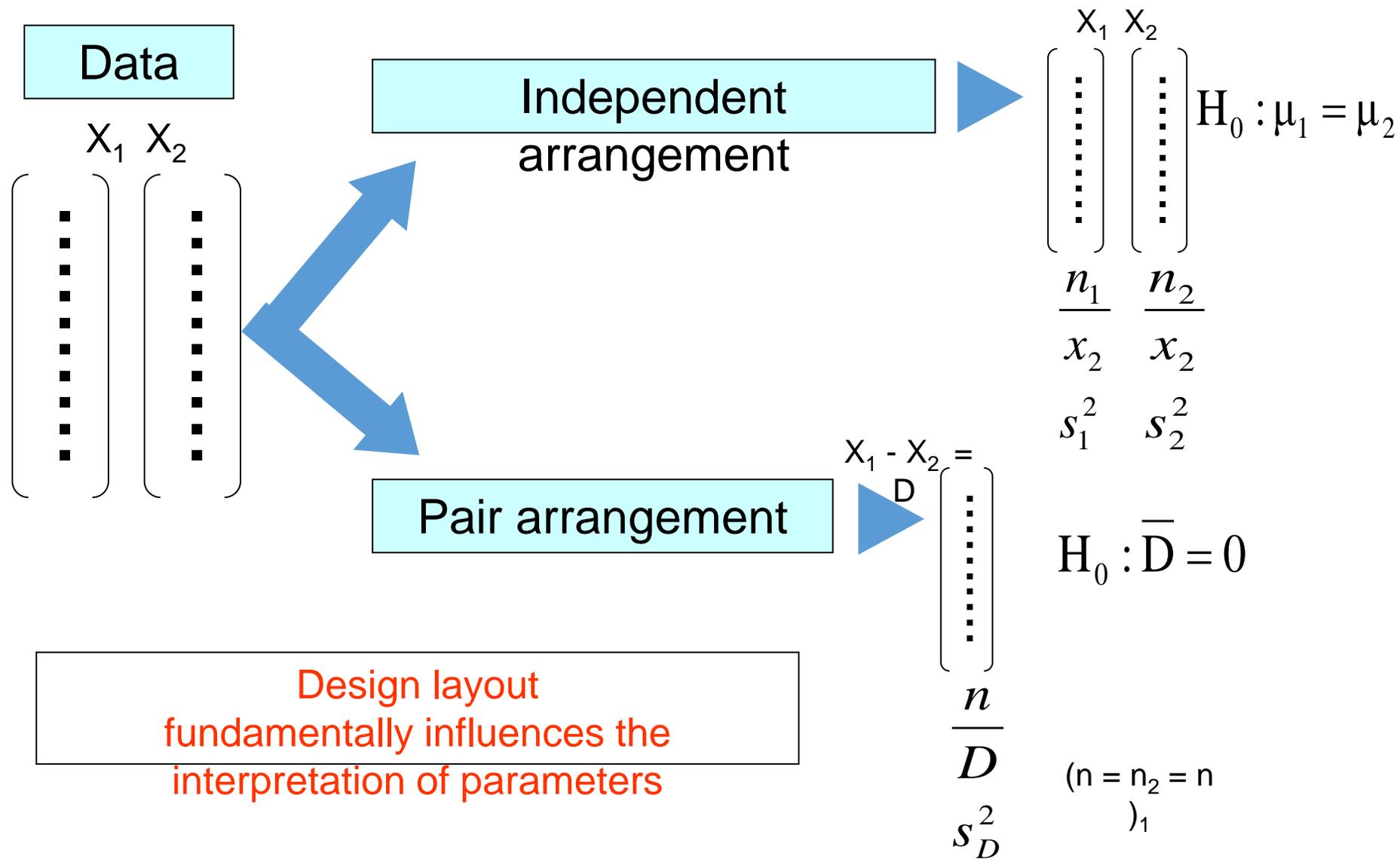
- One of the most common tasks in statistical data analysis is the comparison of continuous data in two groups of patients.
- There is a range of tests to choose from, the choice of a particular test then depends on whether the comparison is pairwise or unpaired and whether it is appropriate to use a parametric test (has assumptions about the distribution of the data) or a non-parametric test (does not have assumptions about the distribution of the data, but has lower predictive power).
- The best known tests in this group are the so-called t-tests used to compare the averages of two groups of values

Two-sample tests: paired and unpaired I

- Using two sample tests, we compare two distributions together. Their basic division is into paired and unpaired tests according to the design of the experiment.
- The basic test for comparing two independent distributions of continuous numbers is the unpaired two-sample t-test
- The basic test for comparing two dependent distributions of continuous numbers is the paired two-sample t-test

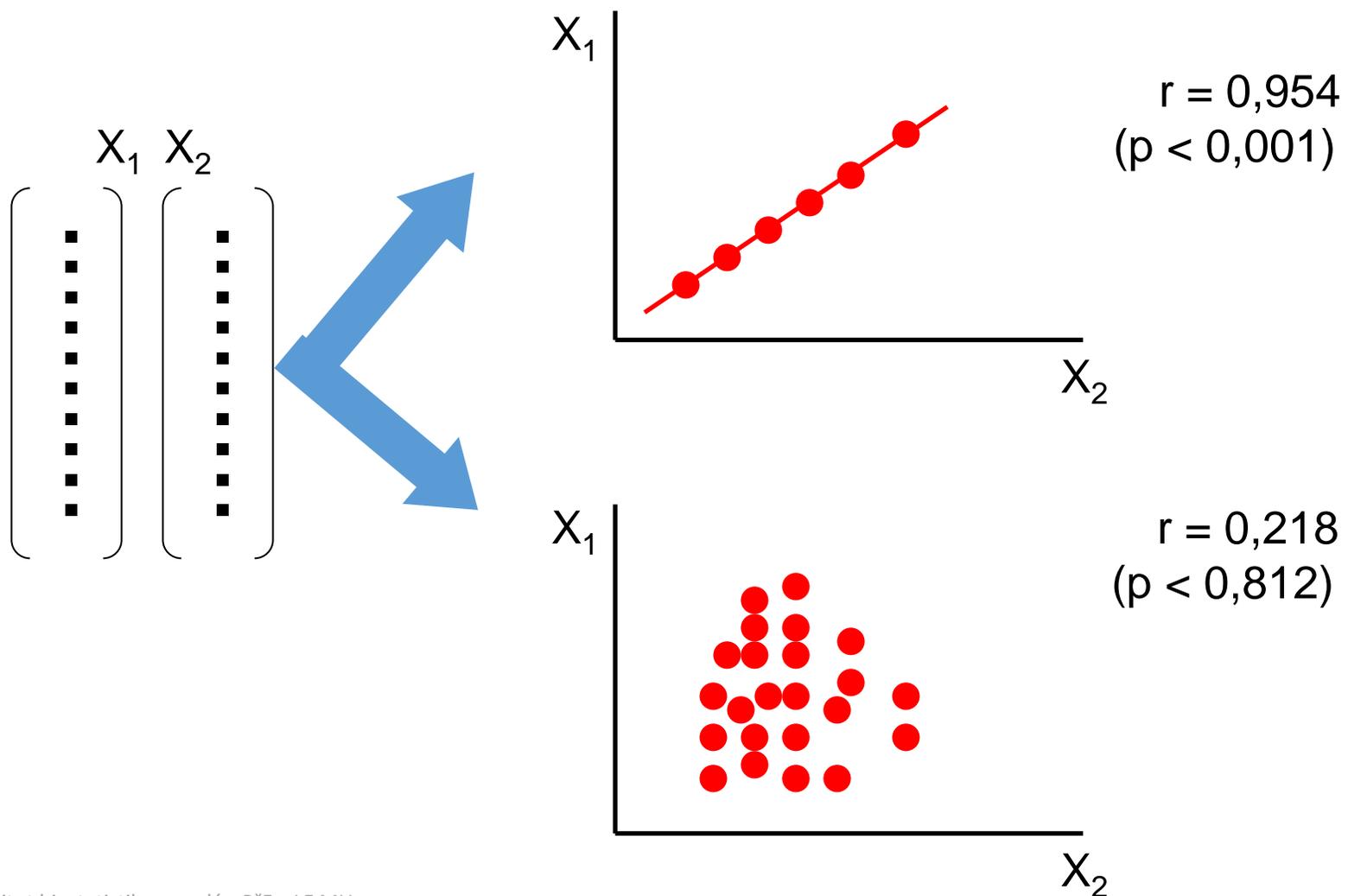


Two-sample tests: paired and unpaired II



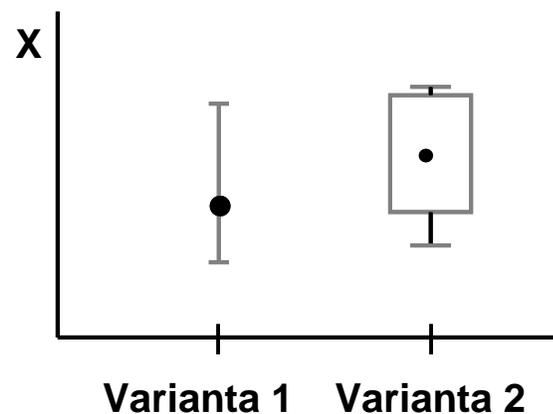
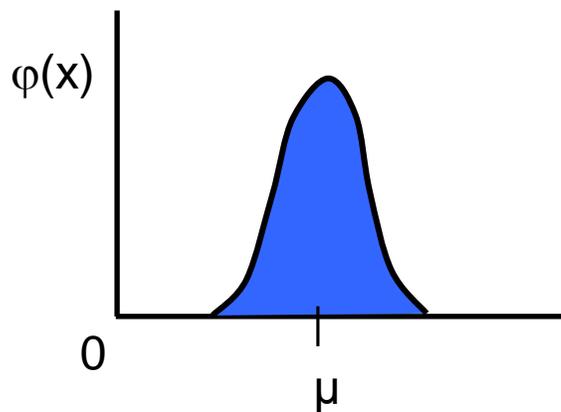
Two-sample tests: paired and unpaired III

- Pairwise identification (Correlation, Covariance)



Unpaired two-sample t-test assumptions

- Random selection of subjects of each group from their target populations
- Independence of the two compared samples
- Approximately normal distribution of the variable in the samples, however, small deviations from normality are not critical, the test is robust to small deviations from this assumption, normality can be tested with normality tests
- The variance in both samples should be approximately the same (homoscedastic). This assumption is tested by several possible tests - Levene's test or F-test.
- It is always advisable to look at histograms of the variable in each sample for octo-metric comparison and to check assumptions of normality and homogeneity of variance - it will not replace statistical tests, but will give an initial idea.



Unpaired two-sample t-test - calculation I

- null hypothesis: means of both groups are the same, alternative hypothesis is that they are not the same, two tailed test
- inspect the data, mean, median, etc. for deviations from normality and inhomogeneity of variance, perform F -test
- F-test for comparison of two sample variances
 - It is used to compare the variance of two groups of values, often to check the homogeneity of the variance of these groups of data.
- In the case of homogeneity testing, the hypothesis of matching variances (two tailed) is tested; if the variances are the same, everything is fine and the t-test can be continued, otherwise it is not appropriate to calculate the test.

H_0	H_A	Test statistics
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F = \frac{s_1^2}{s_2^2}$
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F = \frac{s_2^2}{s_1^2}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{\max(s_1^2; s_2^2)}{\min(s_1^2; s_2^2)}$

Unpaired two-sample t-test - calculation II

- Calculation of the test statistic (degrees of freedom are $n_1 + n_2 - 2$):

$$t = \frac{\text{Rozdíl} - \text{průmě}}{\text{SE}(\text{rozdíl} \text{prů} \text{ěrů})} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
 weighted estimate of variance

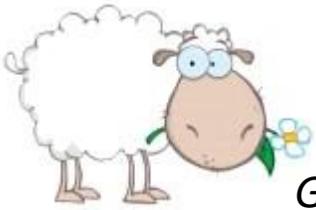
- Compare the resulting t with the tabulated t value for the given degrees of freedom and α (usually $\alpha=0.05$)
- The confidence interval for the difference of means (e.g. 95%), the number of degrees of freedom and s^2 can be calculated to match the previous formulae

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975} SE(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{0,975} \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Unpaired two-sample t-test - calculation summary

- Null hypothesis: the means of the two groups are the same
 - The alternative hypothesis is that they are not identical.
- View the data waveform, average, median, etc.
 - Verify normality of data (e.g. Shapiro-Wilk test)
 - Verify homogeneity of variances (F-test)
 - In the case of homogeneity verification, the hypothesis of matching variances is tested; in the case of matching variances, everything is fine and it is possible to continue calculating the t-test, otherwise it is not appropriate to calculate the test.
- Calculate the value of the test statistic and the p-value. When the calculated p-value is less than 0.05, reject the null hypothesis.

Lecture 7



Example : Unpaired two-sample t-test

Group 1, N=30

- The average weight of sheep at the time of mating was compared for the control group and the group fed the increased diet. The control group comprised 30 sheep, and the enhanced feed group comprised 24 sheep.
- The actual experiment was conducted by starting with 54 sheep (ideally of the same breed, same age, etc.), which we randomly divided into two groups (randomizing objects into experimental groups is the subject of an entire specialized branch of statistics called randomization). After the experiment has been run, we must first test the theoretical assumption for the use of the unpaired t-test. Graphs are plotted for the two groups (we can also calculate the basic descriptive statistics) on which we can assess normality and homogeneity of variance; in addition to the octo-metric view, we can use normality tests to test normality and the F-test to test homogeneity of variance.
- If all the assumptions of the two-sample unpaired t-test are valid, we can calculate the test statistic, the resulting t is 2.43 with 52 degrees of freedom, according to the tables, a $t_{0,975(52)} = 2.01$, so $|t| > t_{0,975(52)}$ and we can reject the null hypothesis, the true probability is 0.018. The difference between the groups is 1.59 kg in favour of the group with increased income.

$$t = \frac{\text{Rozdil.prumeru}}{SE(\text{rozdil.prumeru})} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad v = n_1 + n_2 - 2$$

- The 95% confidence intervals for the difference between the two sets are calculated as $1.59 \pm 2.01 * (0.655)$ kg, corresponding to a range of 0.28 to 2.91 kg. The fact that the confidence interval does not include 0 is further confirmation that there is a significant difference between the groups - this is another way of testing the significance of the differences between the data sets - the null hypothesis that the difference in the means of the t to some value is rejected if the 95% confidence interval for the difference does not include that value

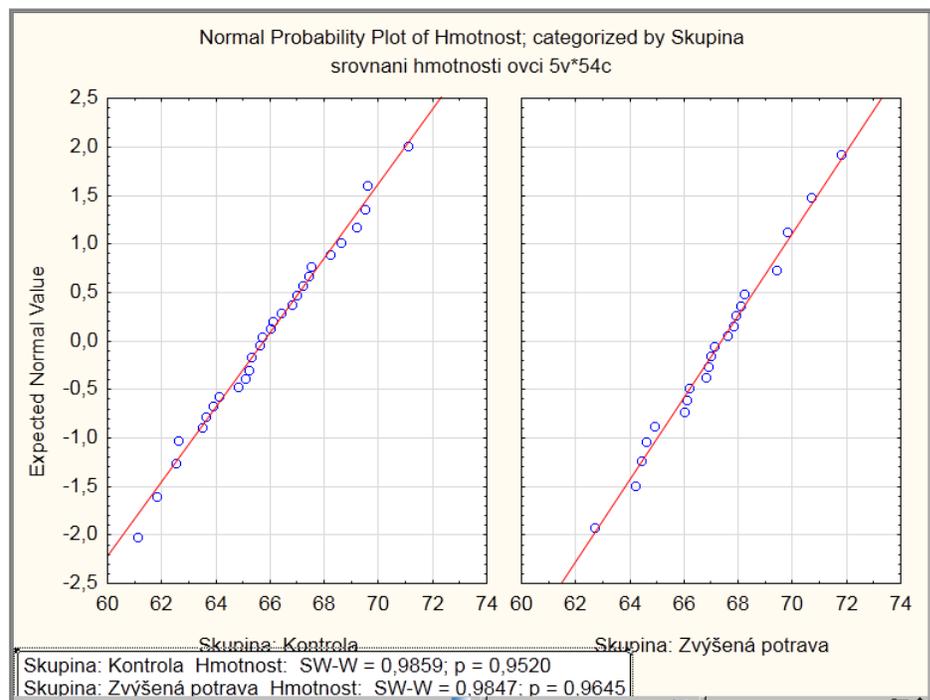
$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975} SE(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{0,975} \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$



Group 2, N=24

Example : Unpaired two-sample t-test

- First, verify the normality of the weight in both the control group and the group with increased food



- In both cases, the dots deviate only slightly from a straight line and the p-values of the S-W test exceed 0.05. The assumption of normal distribution of data in both groups is justified.

Example : Unpaired two-sample t-test

•ATTENTION: The output table is evaluated from the back!!!

Sampling average for group 1

Sampling average for group 2

Sampling standard deviation for group 2

Scope of selection of Group 1

Scope of selection of Group 2

T-tests; Grouping: Skupina (srovnani hmotnosti ovcí)											
Group 1: Kontrola											
Group 2: Zvýšená potrava											
Variable	Mean Kontrola	Mean Zvýšená potrava	t-value	df	p	Valid N Kontrola	Valid N Zvýšená potrava	Std.Dev. Kontrola	Std.Dev. Zvýšená potrava	F-ratio Variances	p Variances
Hmotnost	65,77333	67,36667	-2,43226	52	0,018483	30	24	2,497162	2,252470	1,229066	0,617383

Value of the test statistic
(for the test of agreement of means)

Number of degrees of freedom

Test statistic for the test of agreement of variances
(F-test)

**These columns can only be interpreted
if the difference between the variances was inconclusive !!!**

Paired two-sample t-test

- Data sets are linked via the object of measurement, an example of this would be the measurement of patient parameters before and after treatment (not necessarily the same object, another example could be rats from the same line).
- Both files must have the same number of values, because all measurements in one file must be paired with the measurements in the other file. The actual calculation then takes into account the change in the values (differences) of the subjects in both files.
- In the case that the measurements are not on the same subject, it is advisable to check whether there is a relationship between the two groups - plotting on a graph, correlation - before the paired test.

There are several possible experimental designs, briefly summarized:

- the experiment is paired, and the pairing is
- paired experimental design - paired experiments do not occur
 - maybe the pairing isn't
 - poorly performed experiment - small n, high variability, poor selection of individuals
- we expected independent and they are
- we were expecting independent and they are not
 - Binding
 - Coincidence

Paired two-sample t-test

- This test makes no assumptions about the distribution of the input data because it is calculated only on the basis of their differences.
- These differences should be normally distributed and the question in a paired t-test is whether the mean of the differences equals some number, typically a comparison to zero as evidence of no change between the two paired groups.
- Basically, it is a one sample t-test, where instead of the difference between the sample mean and the target population mean, the mean of the differences and the number being compared is given (0 in the case of the question whether there is no difference between the samples).

- For comparison with 0 (the test statistic is the t distribution)
$$t = \frac{\bar{D}}{s} \sqrt{n} \quad \nu = n - 1$$

- Sometimes it is difficult to decide whether or not it is a pairwise arrangement, a paired t-test should only be used if we can confirm the association (correlation, plotting), one of the reasons for checking this is that in the case of a paired t-test it is not necessary to take into account the variability of the original two sets, but this assumption is only valid in the case of an association between variables. In fact, the calculation of the two types of tests differs in the s used, in one case it is with the difference, in the other case it is a composite estimate of the variance of the two sets.
- Whether a pairwise arrangement is more efficient can be determined by:
 - Binding forces
 - If s_D is significantly smaller than $s_{x_1-x_2}$

- The dependence can be broken down using the formula
$$s_D^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 - 2Cov(x_1; x_2)$$

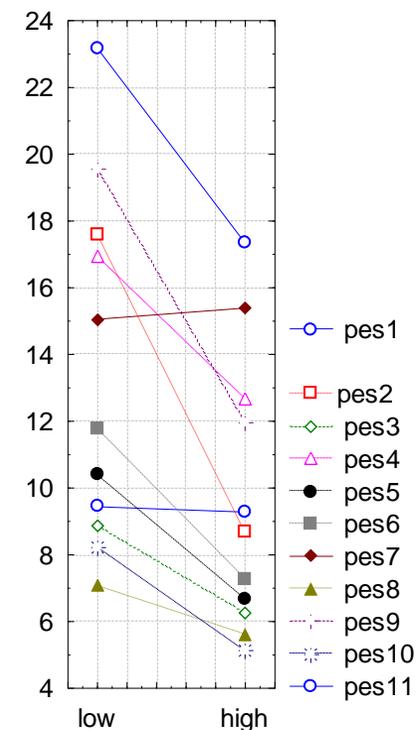
- in the case $Cov=0$, i.e. in the absence of coupling, then s_D^2 corresponds to the sum of the original variances, i.e. approximately $S_{x_1-x_2}$.

Example 1: Paired two-sample test

- An experiment was conducted with the diet of 11 diabetic dogs, each dog was exposed to two diets with different type of carbohydrates (easily absorbed X slowly decomposing into glucose), blood glucose values during each diet were to be compared to determine the effect of diet on blood glucose levels. Since each dog completed both diets, this is a paired design where the results of the values in both experiments are pooled across the experimental animal.
- The null hypothesis is that the true average difference between the two diets is 0; the alternative hypothesis is that it is not 0.
 - For each dog, the difference between its glucose levels on the two diets is calculated and the assumptions for a one sample t-test - i.e. at least an approximately normal distribution - should be verified.
 - The test characteristic is calculated, the calculation is actually done as a one-sample t-test, where the significance of the mean of the differences of the two sets is determined as the difference between this value and zero (zero is the value that the mean difference should take if the null hypothesis holds). $T=4.37$ with 10 degrees of freedom, true value $p=0.0014$ and therefore at the $p=0.05$ level we can reject the null hypothesis

$$t = \frac{\text{rozdíel}_\text{průměru}_\text{vzorku}_\text{a}_\text{populace}}{SE(\text{průměru})} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

- In conclusion, the null hypothesis of no difference between the two diets was rejected, which means that the high-fibre diet has a significant effect on lowering blood glucose levels.



Example 2: Paired two-sample test

- A diet experiment was conducted in 18 diabetic rats, each rat was exposed to two diets (one novel special diet and one control diet). Because each rat was exposed to both diets, this is a paired design, where the values in both experiments are pooled across the experimental animal. Determine whether the test diet causes a change in weight in the rats (whether the weight of the rats differs after the new special diet and after the control diet).
1. The null hypothesis is that the actual mean difference in weight between rats on the special and control diets is zero (the special diet did not lead to a change in weight compared to the control diet), the alternative hypothesis is that the difference in weight is different from zero (the special diet led to a change in weight compared to the control diet).
 2. For each rat, the difference in weights measured after the two diets is calculated and the assumptions for a one-sample t-test - at least an approximately normal distribution of differences - should be verified.
 3. A test statistic is calculated, the calculation is actually done as a one-sample t-test, where the significance of the mean of the differences of the two sets is tested as the difference between this value and zero (0 is the value that the mean difference should take if the null hypothesis holds). $T=-1.72$ with 17 degrees of freedom, true $p\text{-value}=0.102$ and therefore at the $\alpha=0.05$ significance level we cannot reject the null hypothesis.

$$t = \frac{\text{rozdíl}_\text{průměru}_\text{vzorku}_\text{a}_\text{populace}}{SE(\text{průměru})} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

4. In conclusion, the null hypothesis of no difference in the effect of on weight loss between the two diets has not been rejected.



Example: paired two-sample test

Sampling average

Sampling standard deviation

Number of observations

T test for Dependent Samples (efektivita diety pro krysy)								
Marked differences are significant at $p < ,05000$								
Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
testovaná dieta	186,0556	59,52011						
kontrolní dieta	191,7222	69,65022	18	-5,66667	13,91994	-1,72714	17	0,102266

Average value of differences

Test criterion value

Sampling standard deviation of differences

Nonparametric two-sample statistical tests

Unpaired Mann-Whitney test

Paired Wilcoxon and sign test

Mann-Whitney U test

- A non-parametric alternative to the two-sample t-test.
- Counts the order of the data in the files instead of the original data.
- Assumption: the probability distribution of a variable in groups can only differ by a shift.

Procedure:

1. Determine the null and alternative hypotheses ($F(x)$ =distribution function):

$$H_0 : F(x_1) = F(x_2)$$

$$H_1 : F(x_1) \neq F(x_2).$$

2. The numbers of both files are merged and their order in this merged file is determined.
3. For both selections separately, the sum of the rankings (T_1 and T_2) is computed.
4. The final value of the test statistic U is determined from the sums of the rankings in the groups.

$$U_1 = n_1 n_2 + \frac{n_1 - (n_1 + 1)}{2} - T_1$$

$$U_2 = n_1 n_2 + \frac{n_2 - (n_2 + 1)}{2} - T_2$$

$$U = \min(U_1, U_2)$$

5. We compare the value of the test statistic U with the critical value of the test, if this value is less than the critical value of the test, we reject the null hypothesis of agreement between the distribution functions of the two groups.

Mann-Whitney U test - asymptotic variant

5. The asymptotic normality of the U statistic can be used for large n_1 and n_2 (>30).

$$E(U) = \frac{n_1 n_2}{2} \quad D(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

6. Z-statistics can be used for testing:

$$Z = \frac{U - E(U)}{\sqrt{D(U)}} \approx N(0,1)$$

7. If $|Z| \geq z_{1-\alpha/2}$ we reject the null hypothesis of identity of distribution functions

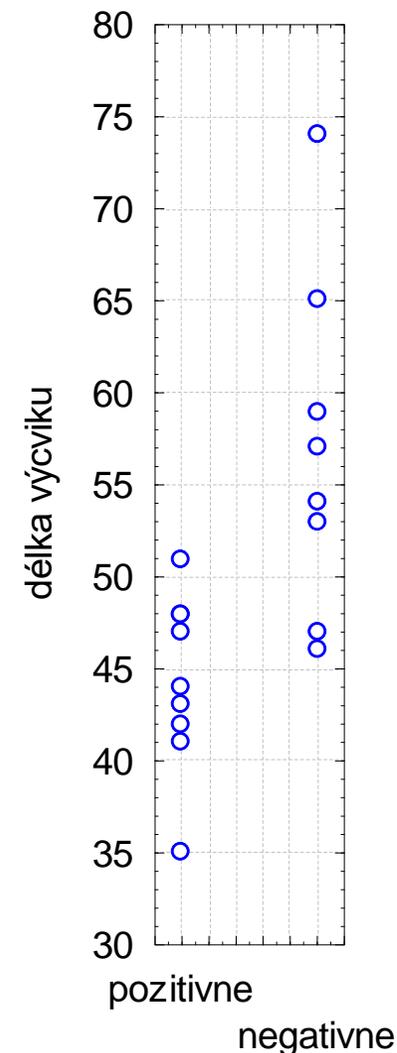
Mann-Whitney U test

- Like many other non-parametric tests, this test also considers the order of the data in the files instead of the original data. It is a non-parametric analogue of the unpaired t-test and has the highest test power of these non-parametric tests (95% paired t-test).
- In the case of the Mann-Whitney test, first the numbers of both files are merged and their order is created in this merged file, then the values are returned to the original files and only their order is worked with.
- Thus, a rank sum is produced for both sets and the smaller of the two sums is compared to the critical test value; if this value is less than the critical test value, we reject the null hypothesis of a match between the distribution functions of the two sets.

X1	X2	ALL	Rank ALL	X1 rank	X2 rank
27	25	25	5	6	5
35	29	29	7,5	11	7,5
38	31	31	9	13	9
37	23	23	4	12	4
39	18	18	2	14	2
29	17	17	1	7,5	1
41	32	32	10	15	10
	19	19	3		3
		27	6		
		35	11		
		38	13		
		37	12		
		39	14		
		29	7,5		
		41	15		

Example: the Mann-Whitney U test

- 17 puppies were trained to go to the toilet using positive motivation (praise when going to the toilet outside) or negative motivation (punishment when going to the toilet at home). As a parameter, the number of days a puppy is trained was measured.
- The null hypothesis is that there is no difference in the training methods, i.e. that the puppy is trained in the same amount of time by both methods.
- After comparing the + distribution due to the low number of values, it is appropriate to use a non-parametric test.
- The order of values in the complete file is created.
- The value of the test statistic is determined from the rank order of the values in each group.



- **How will the testing go?**

Example: software solution

The sum of the order T_1

The sum of the order T_2

Value From statistics

Mann-Whitney U Test (Spreadsheet15)										
By variable skupina										
Marked tests are significant at $p < .05000$										
variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-value	Z adjusted	p-value	Valid N Group 1	Valid N Group 2	2*1sided exact p
delka	49,50000	103,5000	135,0000	-2,11695	0,034265	-2,11955	0,034045	8	9	0,027396

Value of the test statistic

Asymptotic p-value

Exact p-value

(use if the selection range is less than 30)

Paired Wilcoxon and sign test

- We start from pairwise differences and move to a one-sample test design
- Tests whether the **median of the pairwise differences (D)** is equal to c $H_0 : D_{0.5} = c$ versus $H_1 : D_{0.5} \neq c$.

Wilcoxon matched pairs test

1. Calculate the differences of the selection **differences** with the tested median = c .
2. We arrange the absolute values of the differences in ascending order and assign them a ranking.
3. We compute the statistics S_w^+ and S_w^- , which correspond to the sum of the order of positive (S_w^+) and negative differences (S_w^-). We take the minimum of S_w^+ and S_w^- as the final value of the test statistic. We reject the null hypothesis if the value of the test statistic is less than or equal to the tabulated critical value (at a given significance level and number of non-zero differences).

Sign Pair Test

1. Calculate the differences of the selection **differences** with the tested median = c .
2. Calculate the statistic S_z^+ , which corresponds to the number of positive differences \rightarrow the test does not use the rank values of the original data but only the information whether the value is realized above or below the median \rightarrow the power of the test is reduced
3. We reject the null hypothesis if the statistic S_z^+ realizes in the critical range of values $W=(0, k_1) \cup (k_2, n)$, where n corresponds to the number of non-zero differences and the values of k_1 and k_2 can be traced in mathematical tables.

Example 2: Paired two-sample test

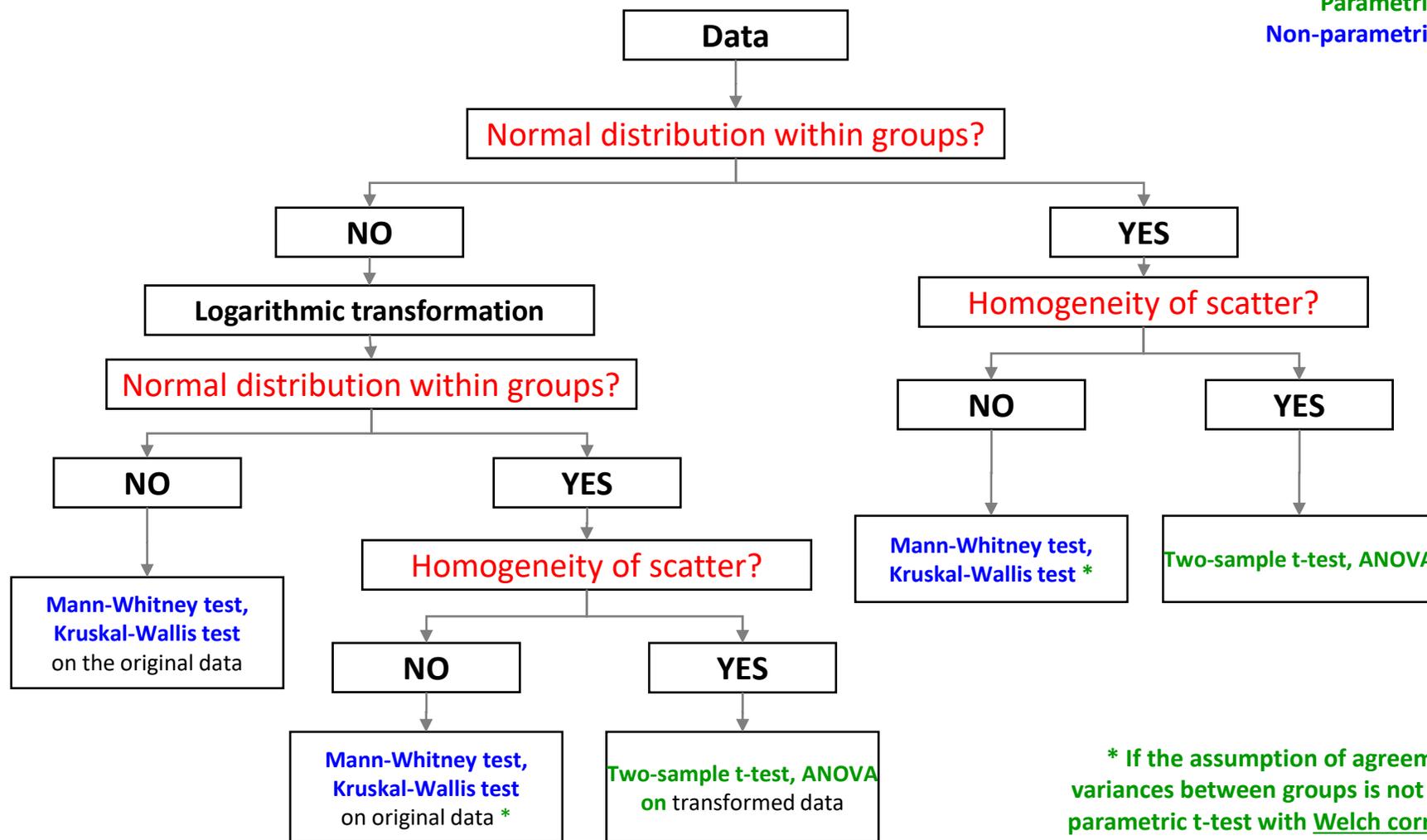
A new diet for laboratory rats was tested, and its effect on weight in different rat lines was investigated, so a pair arrangement was chosen where rats in both diets are connected through their line, i.e. at the beginning there were pairs of rats of the same line, one of them was randomly assigned to the diet, the other of the pair to the other diet.

1. the null hypothesis is that the weight of the rats is not affected by the diet used, the alternative hypothesis is that there is an effect of diet
2. calculate the differences - these differences are non-normal and therefore it is appropriate to use a non-parametric test
3. Calculate the sum of the order of positive and negative differences, here is the smaller sum of negative differences - 31
4. the result of the calculation is $p > 0.05$ and therefore we do not have sufficient evidence to reject the null hypothesis, we cannot say that the new diet is more effective than the old one
5. to complement the results, it is also useful to find out the actual magnitude of the difference in weights between the groups, e.g. in the form of a median



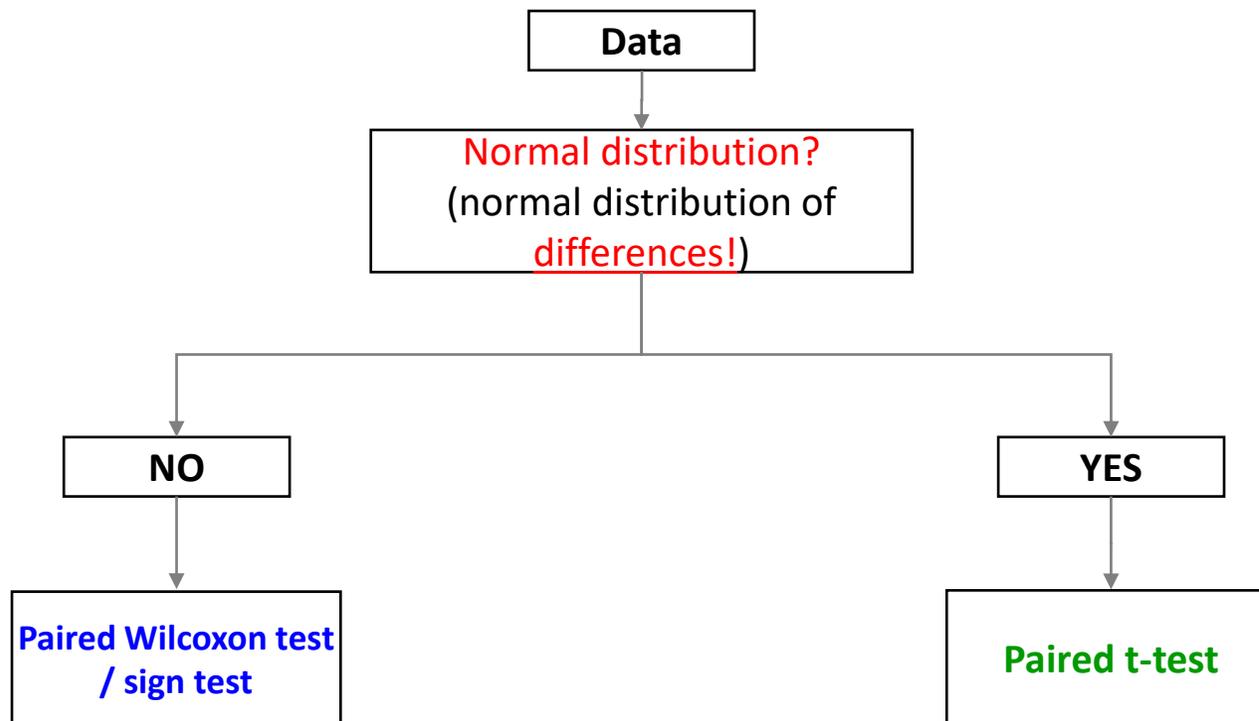
Scheme when testing 2 or more groups

Parametric tests
Non-parametric tests



* If the assumption of agreement of variances between groups is not met, a parametric t-test with Welch correction can be used

Paired testing scheme



Parametric tests
Non-parametric tests

Binomial distribution

Description of the binomial distribution

Hypothesis testing of binomial distributed data

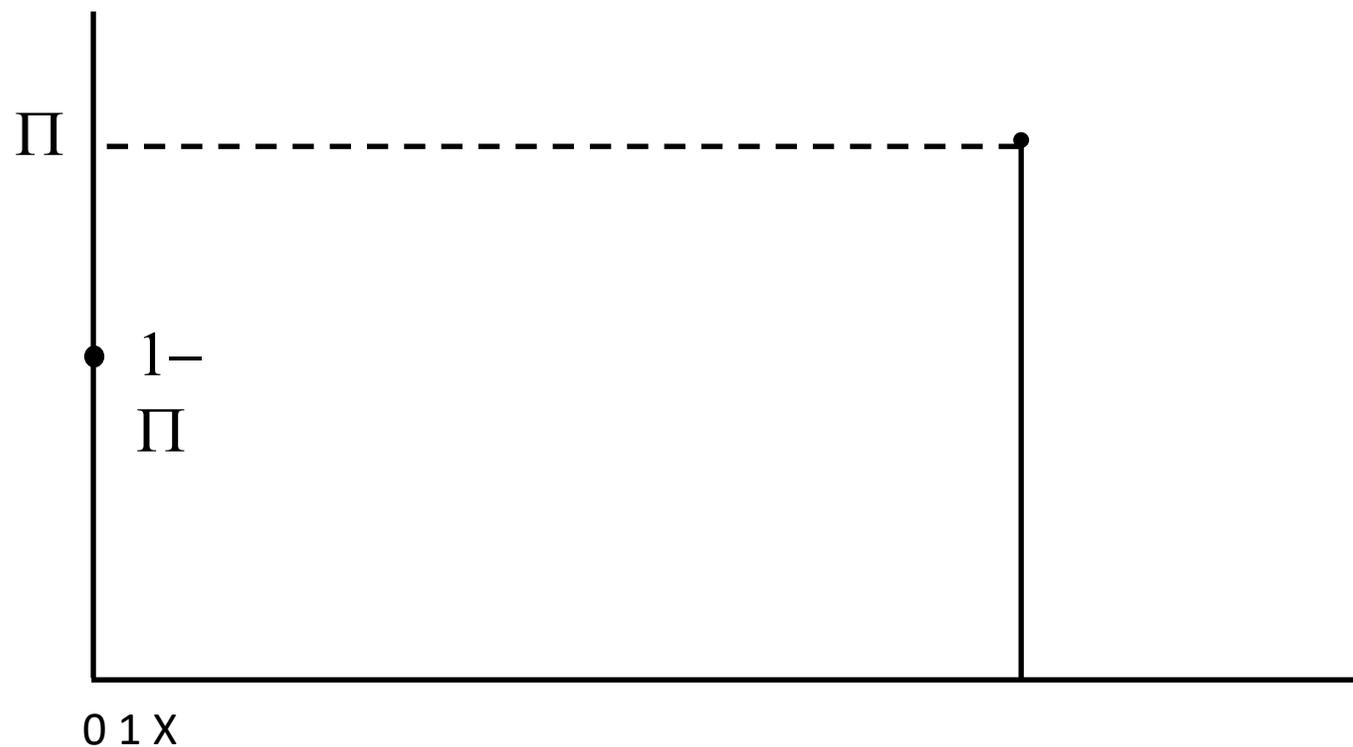
Annotation

- In addition to continuous data, we also encounter categorical data, the simplest case of which is binary data.
- The binary data are described by a binomial distribution, the descriptive statistics of the binary data (percentage of occurrence of a phenomenon), its confidence interval and binomial tests for comparing the percentage of occurrences of phenomena in different groups are derived from the behaviour of the binomial distribution.

Alternative distribution

- One of two possible scenarios will occur

$$\begin{aligned} \pi(x) &= \Pi \text{ for } X = 1 \\ \pi(x) &= 1 - \Pi \text{ for } X = 0 \\ \pi(x) &= 0 \text{ otherwise} \end{aligned} \quad \left. \vphantom{\begin{aligned} \pi(x) &= \Pi \text{ for } X = 1 \\ \pi(x) &= 1 - \Pi \text{ for } X = 0 \\ \pi(x) &= 0 \text{ otherwise} \end{aligned}} \right\} X = 1 \dots j \text{ev}$$



Binomial distribution

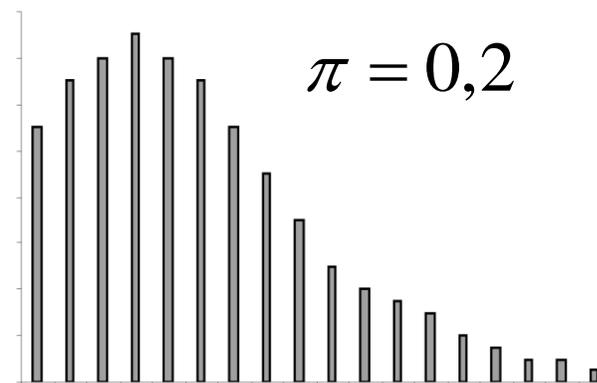
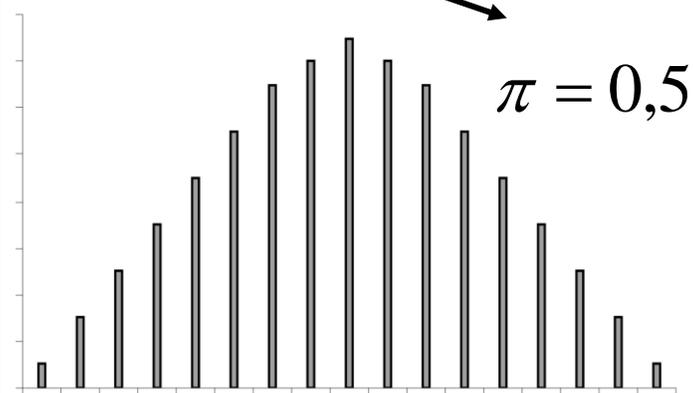
X total number of occurrences in n independent experiments

$$E(x) = n \cdot \pi$$

$$D(x) = n \cdot \pi (1 - \pi)$$

$$\pi \sim p$$

single distribution parameter
determines the shape of the distribution



Binomial distribution as a model for investigating the occurrence of the observed phenomenon

n number of independent repetitions (queries)

X number of people with a certain symptom

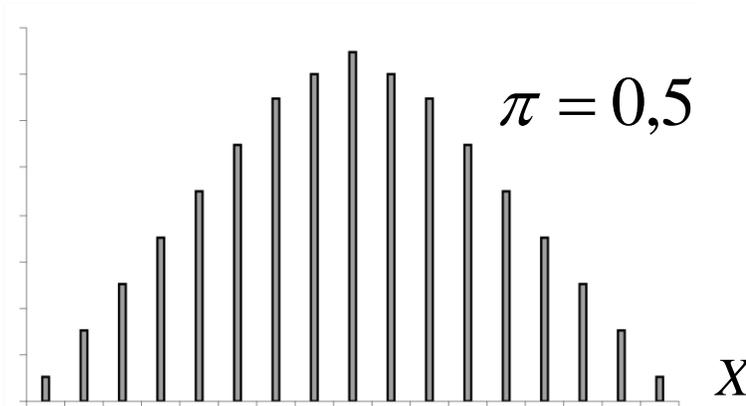
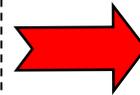
r means the total number of occurrences of the phenomenon in n independent experiments

$r : 0 \dots n$

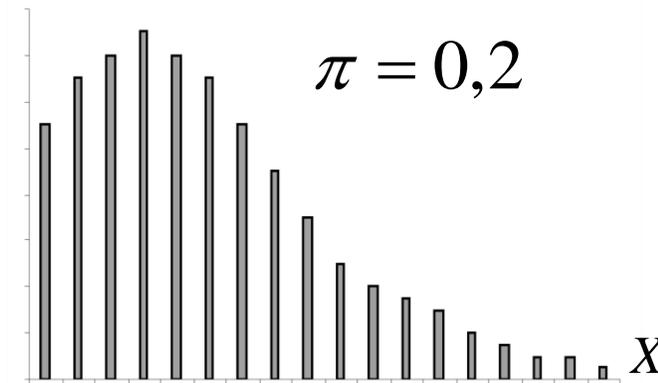
$p \sim \pi$.. the only parameter of the binomial distribution

p relative frequency of occurrence

p determines the shape of the distribution



$$p = \frac{r}{n}$$



Binomial variable X

Binomial distribution as a model

Phenomenon: birth of a boy $P = 0.5$

n : family with 5 children

r : 0,1,2,3,4,5 boys

$$P(\mathbf{r}) = \binom{\mathbf{n}}{\mathbf{r}} \cdot \mathbf{p}^{\mathbf{r}} \cdot (\mathbf{1} - \mathbf{p})^{(\mathbf{n} - \mathbf{r})} = \frac{\mathbf{n}!}{\mathbf{r}!(\mathbf{n} - \mathbf{r})!} \cdot \mathbf{p}^{\mathbf{r}} \cdot \mathbf{q}^{(\mathbf{n} - \mathbf{r})}$$

$$r = 0: \frac{5!}{(0! 5!)} \cdot (0,5)^0 \cdot (0,5)^5 = 0,031$$

$$r = 1: \frac{5!}{(1! 4!)} \cdot (0,5)^1 \cdot (0,5)^4 = 0,15625$$

$$r = 2: P(r) = 0.3125$$

$$r = 3: P(r) = 0.3125$$

$$r = 4: P(r) = 0.15625$$

$$r = 5: P(r) = 0.031$$

X : Binomial variable

Center of the layout:

Dispersion: $E(x) = n \cdot p$

$$D(x) = n \cdot p \cdot (1 - p)$$

Example: $n = 100$ respondents
 $r = 20$ has a symptom



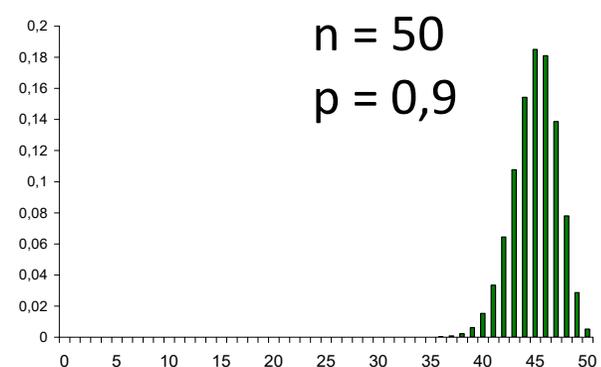
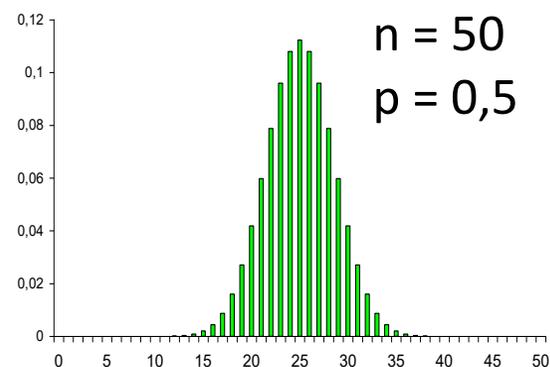
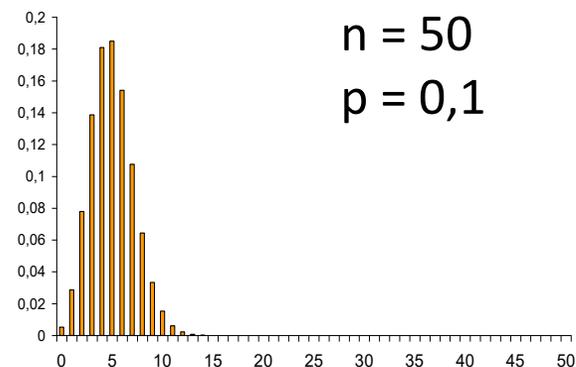
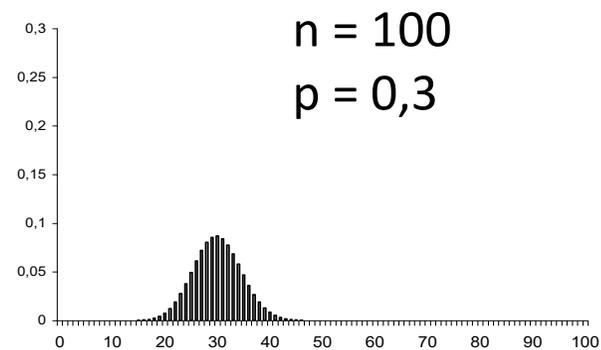
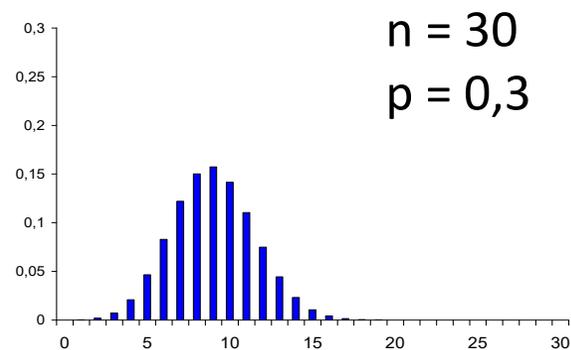
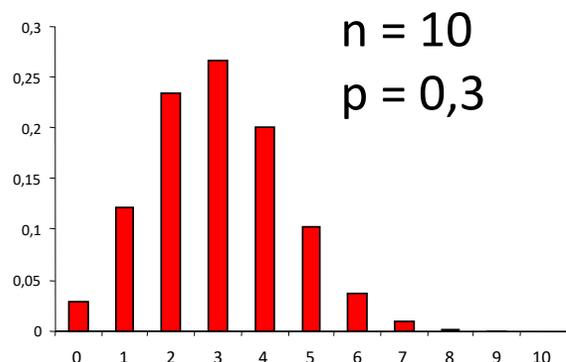
$$E(x) = n \cdot p = 20$$

is the centre of the
distribution
and the most likely
value

Binomial distribution as a model

$$P(x = r) = \frac{n!}{r!(n-r)!} \cdot p^r \cdot q^{(n-r)}$$

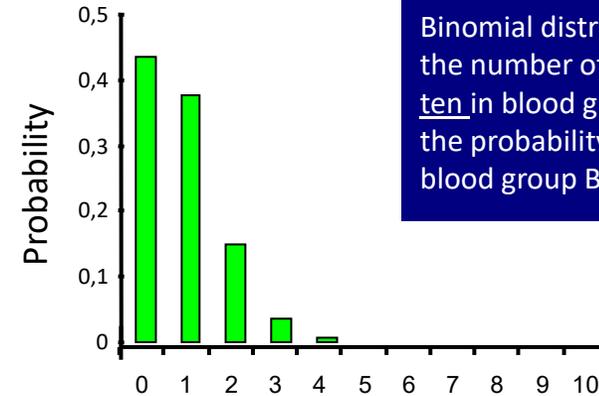
$$q = 1 - p$$



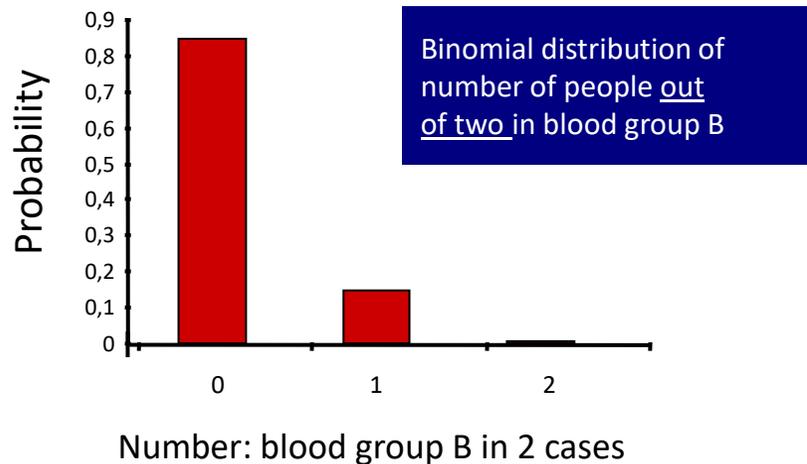
Application of the binomial distribution

- Prevalence of blood group B in a population: $p = 0.08$

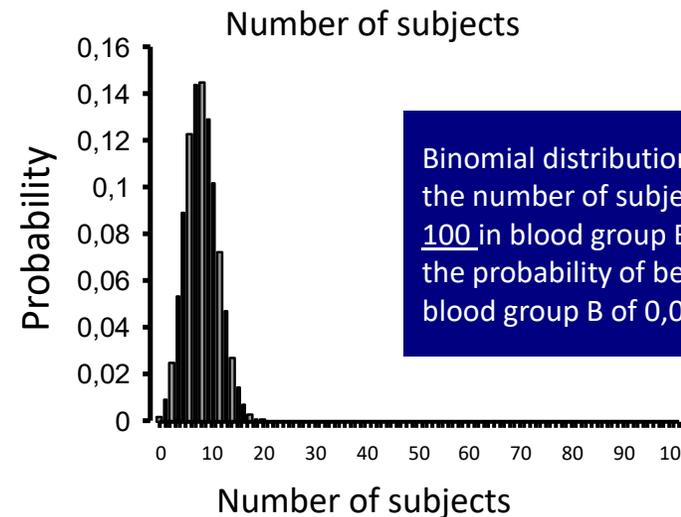
		Number in blood group B	Probability
B	B	2	0,0064
not B	B	1	0,0736
B	not B	1	0,0736
not B	not B	0	0,8464



Binomial distribution showing the number of subjects out of ten in blood group B based on the probability of being in blood group B of 0,08.



Binomial distribution of number of people out of two in blood group B



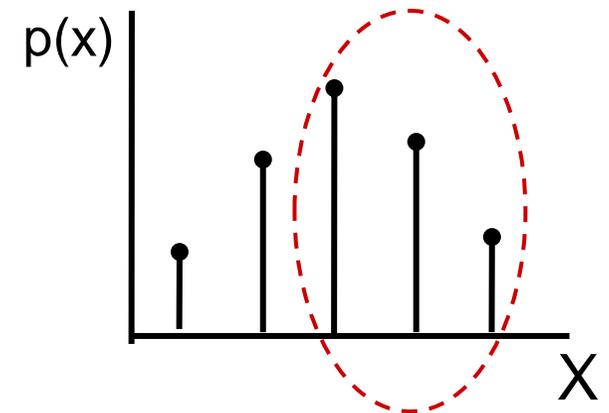
Binomial distribution showing the number of subjects out of 100 in blood group B based on the probability of being in blood group B of 0,08.

Application of the binomial distribution

- Population: 60% of individuals have elevated cholesterol; sample size: 5 people
- How many people have higher cholesterol levels in the selection ?
 - $n \cdot p = 5 \cdot 0.6 = 3$ people $\sim E(x)$
- What is the P that just 3 people will have higher cholesterol levels ? \sim That is, the selection exactly matches the population ?

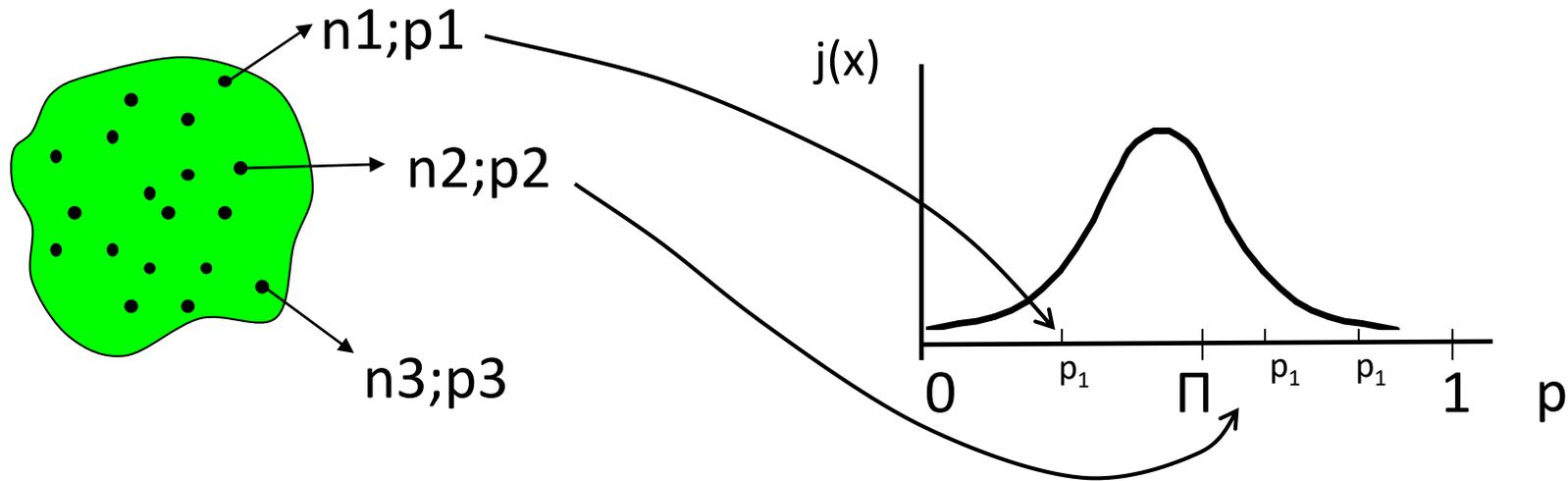
$$P_{(3)} = \frac{5!}{3!(5-3)!} \cdot (0,6)^3 \cdot (0,4)^2 = 0,346 = 35\%$$

- What is the P that most individuals (i.e. at least 3) have higher cholesterol levels ? \sim I.e., a selection of at least generally corresponds to the population under study ?
 - $P(X > 3) = P(3) + P(4) + P(5) = 0.346 + 0.259 + 0.078 = 68 \%$

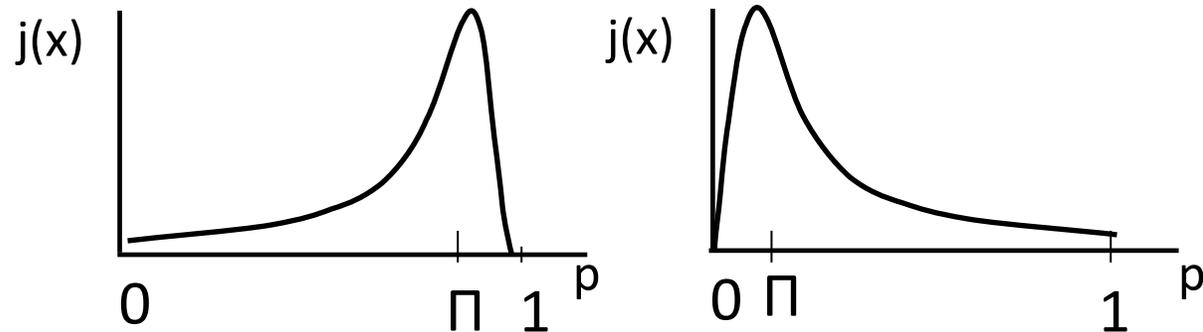


Estimation of the parameter Π of the binomial distribution

- In multiple estimation, the parameter Π behaves as a normally distributed



For small or large values of p (Π), however, the normality assumption is limited



Estimation of the parameter Π of the binomial distribution

$$\pi \approx \hat{p}; \quad \hat{p} = r/n$$

1) Spot

$$\hat{p}; \quad s_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1}$$

2) Interval - approximation

$$\hat{p} - Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} \leq \pi \leq \hat{p} + Z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

$$\pi: \hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}}$$

Estimation of the parameter π of the binomial distribution: example I

X: % of individuals with a given trait

n = 100 individuals

r = 60; $\hat{p} = 0,6$

$s_{\hat{p}} = 0,049$

Confidence Interval : 95%

FROM_{0,975} = 1.96

$$0,6 - 1,96 \cdot 0,049 \leq \pi \leq 0,6 + 1,96 \cdot 0,049$$

$$0,504 \leq \pi \leq 0,697$$



$$P(0,504 \leq \pi \leq 0,697) \geq 0,95$$

Estimation of the parameter π of the binomial distribution

- Interval estimation without approximation to normal distribution

$$L_1 = \frac{r}{r + (n - r + 1) \cdot F_{\alpha/2}^{(v_1; v_2)}}$$



lower limit of the interval

$$v_1 = 2(n - r + 1); \quad v_2 = 2r$$

$$L_2 = \frac{(r + 1) \cdot F_{\alpha/2}^{(v'_1; v'_2)}}{n - r + (r + 1) \cdot F_{\alpha/2}^{(v'_1; v'_2)}}$$



upper limit of the interval

$$v'_1 = 2(r + 1) = v_2 + 2$$

$$v'_2 = 2(n - r) = v_1 - 2$$

$$P(L_1 \leq \pi \leq L_2) \geq 1 - \alpha$$

Estimation of the parameter Π of the binomial distribution: example II

Random sample $n = 200$ individuals.

Only $r = 4$ individuals without a particular trait were found.

$$\hat{p} = \frac{4}{200} = \underline{\underline{0,02}}$$

95% confidence interval = ?

Lower boundaries

$$v_1 = 2(n - r + 1) = 2(200 - 4 + 1) = 394$$

$$v_2 = 2r = 2 \cdot 4 = 8$$

$$F_{1-\alpha/2}^{(394;8)} = \underline{\underline{3,67}}$$

$$L_1 = \frac{4}{4 + (200 - 4 + 1) \cdot 3,67} = \underline{\underline{0,0055}}$$

Upper limit

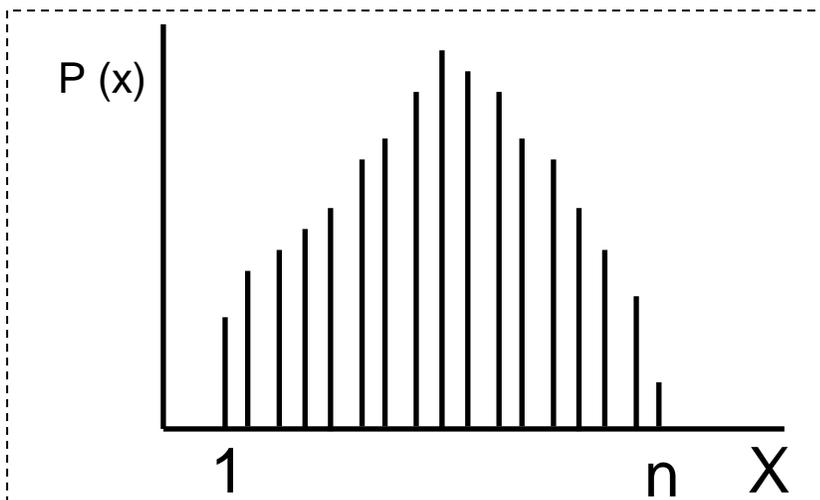
$$v'_1 = 2(r + 1) = 10$$

$$v'_2 = 2(n - r) = 2(200 - 4) = 392$$

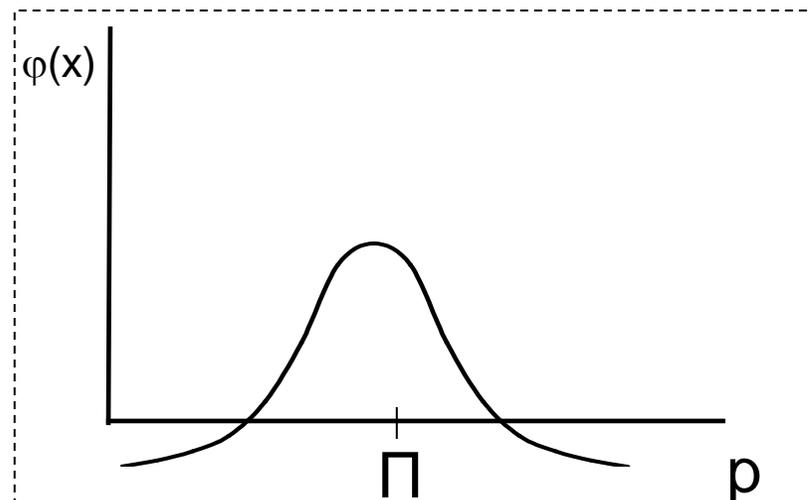
$$F_{1-\alpha/2}^{(10;392)} = \underline{\underline{2,08}}$$

$$L_2 = \frac{(4 + 1) \cdot 2,08}{200 - 4 + (4 + 1) \cdot 2,08} = \underline{\underline{0,051}}$$

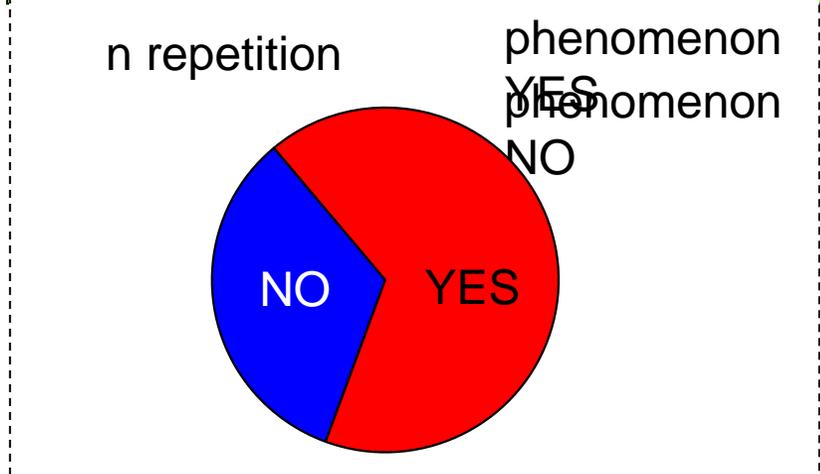
Binomial distribution in data: visualization



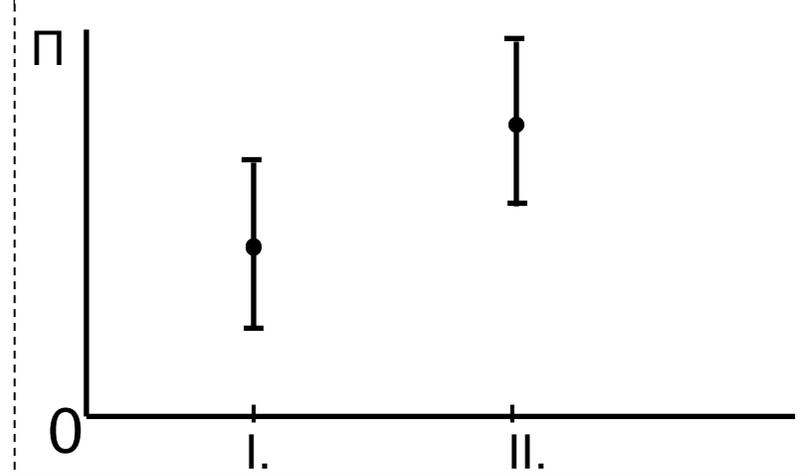
Probability of occurrence of X values



Model distribution of the estimated parameter



Binary nature of the original values



Confidence interval for P

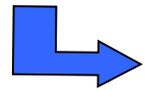
Statistical testing of binomial data

I.

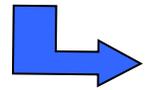
Does the estimate of \underline{p} differ from the expected value of P ?

II.

Do two or more estimates differ \underline{p} ?



- dependent estimates -



- independent estimates -

III.

Is the occurrence of categories of two phenomena independent ?

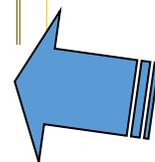
IV.

An assessment of the relative risk of a phenomenon occurring within a group of people

One-sample binomial test

H_0	H_A	Test statistics	Confidence interval
$p \leq \Pi$	$p > \Pi$	z	$z > z_{1-\alpha}$
$p \geq \Pi$	$p < \Pi$	z	$z < z_{\alpha}$
$p = \Pi$	$p \neq \Pi$	z	$\frac{1}{2}z > z_{1-\alpha/2}$

$$Z = \frac{n \cdot \hat{p} - n \cdot \pi}{\sqrt{n \cdot \hat{p}(1 - \hat{p})}} \cong \frac{|n \cdot \hat{p} - n \cdot \pi| - 0,5}{\sqrt{n \cdot \hat{p}(1 - \hat{p})}}$$



Correction on
Continuity

H_0	H_A	Test statistics	Confidence interval
$p \leq \Pi$	$p > \Pi$	$L_1 = \frac{(r + 1)F_{\alpha, v_1', v_2'}}{n - r + (r + 1)F_{\alpha, v_1', v_2'}}$	$p = r / n > L_1$
$p \geq \Pi$	$p < \Pi$	$L_2 = \frac{r}{r + (n - r + 1)F_{\alpha, v_1', v_2'}}$	$p < L_2$
$p = \Pi$	$p \neq \Pi$	$L_1 ; L_2 (F_{\alpha/2} ; F)_{1-\alpha/2}$	$p < L_2 \vee p > L_1$

Test $\pi ? p$: Example 1

- Trees with altered crown shape
 - $n = 9\ 000$ individuals
 - $r = 2\ 250$ altered individuals
- How likely is change in up to 1/3 of individuals?

$$Z = \frac{n \cdot p - n \cdot \pi}{\sqrt{p(1-p) \cdot n}} = \frac{2250 - 3000}{\sqrt{0,25 \cdot 0,75 \cdot 9000}} = \underline{\underline{-18,26}}$$

$$\alpha = 5\%; Z_{1-\alpha/2} = 1.96; Z_{1-\alpha} = 1.645$$

$$Z > Z_{1-\alpha/2} \dots\dots\dots \text{reject } H_0 : p < 0.01$$

- 95% confidence interval ... p : (0.241; 0.258)

Test π ? p : Example 2

- The probability of a boy being born is about 1/2.
- You are to evaluate the results of a survey of a population living in a severely degraded environment.
- The survey covers 1000 randomly selected families and the observed proportion of boys born is 0.41.

What are your conclusions about this population (are the same proportion of boys born as in the general population?)

How does your estimate become more precise if you use a sample of $n = 10,000$ families while keeping the estimate of $p = 0.41$?

Test π ? p : Example 2

- We use a one-sample binomial test with null hypothesis $H_0: p=\pi$, significance level $\alpha=0.05$
- Test statistics: and the corresponding quantile:
- Because the null hypothesis ????
- Confidence interval:
- If we use $n=10\ 000$, the int. reliability will be ????:

Test π ? p : Example 2

- We use a one-sample binomial test with null hypothesis $H_0: p=\pi$, significance level $\alpha=0.05$

- Test statistics: $Z = \frac{n \cdot \hat{p} - n \cdot \pi}{\sqrt{n \cdot p(1-p)}}$ and the corresponding quantile: $Z = \frac{1000 \cdot 0,41 - 1000 \cdot 0,5}{\sqrt{1000 \cdot 0,41 \cdot 0,59}} = -1,96$

$$Z_{1-\frac{\alpha}{2}} = Z_{0,975} = 1,96$$

- Because $|Z| > Z_{0,975}$ we reject the null hypothesis. Boys are not born in the study population with probability 0.5.

- Confidence interval: $\pi: \hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0,4 \pm Z_{0,975} \cdot 0,046 = 0,41 \pm 1,96 \cdot 0,016 = 0,41 \pm 0,03$

- If we use $n=10,000$, the int. reliability will be narrower:

$$\pi: \hat{p} \pm Z_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n-1}} = 0,41 \pm 1,96 \cdot 0,005 = 0,41 \pm 0,01$$

Test π ? p : Example 3

- Example of a test without approximation to the normal distribution

12 individuals were examined for the presence of a particular trait,
10 individuals had no sign

?

How much does this result differ from the 6 - 6 result: i.e. from a situation where half of the individuals have the trait?

a) Use of the distribution function

r	0	1	2	3	4	5	6	7	8	9	10	11	12
P(r)	0,0002 4	0,0029 3	0,0161 1	0,0537 1	0,1208 5	0,1933 5	0,2255 9	0,1933 6	0,1208 5	0,0537 1	0,0161 1	0,0029 3	0,0002 4

$$P(r \geq 10) = 0.01611 + 0.00393 + 0.00024 = 0.01928$$

$H_0 : p = 0.5$ is therefore highly improbable

b) The observed exceeded the upper limit of the 95% interval
reliability for p : $\hat{p} = 10/12 = 0,833$

$$p = 0,5 : L_2 = \frac{(6+1) \cdot 2,64}{12 - 6 + (6+1) \cdot 2,64} = \underline{\underline{0,755}}$$

Two-sample binomial test ($p_1 ? p_2$)

$$Z = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

$$\bar{p} = \frac{n_1 \cdot \bar{p}_1 + n_2 \cdot \bar{p}_2}{n_1 + n_2}$$

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{(1-\alpha/2)} \cdot \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}$$

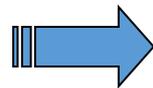
Two-sample binomial test ($p_1 ? p_2$)

- This example is an example of testing differences between two binomial populations (i.e., comparing two estimates of p).
- A total of 49 experimental mice were used to test the toxic preparation during a two-month culture period. The following table contains the original data along with the test of the null hypothesis: the proportion of surviving individuals is the same in the affected population.

	Alive	Dead	Total	Proportion alive	Proportion dead
Treated	15	9	24	$\hat{p}_1 = 0,625$	$\hat{q}_1 = 0,375$
Not Treated	10	15	25	$\hat{p}_2 = 0,400$	$\hat{q}_2 = 0,600$
Total	25	24	49	$\hat{p} = 0,510$	$\hat{q} = 0,490$

$$Z = \frac{0,625 - 0,400}{\sqrt{\frac{(0,510)(0,490)}{24} + \frac{(0,510)(0,490)}{25}}} = \frac{0,225}{\sqrt{0,010413 + 0,009996}} = 1,573$$

$$Z_{0.05(2)} = t_{0.05(2)} = 1.96$$

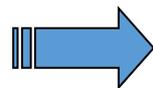


We do not consider $H_0 : 0.10 < P < 0.20$

With correction on continuity:

$$Z = \frac{\frac{15 - 0,5}{24} - \frac{10 + 0,5}{25}}{0,143} = \frac{0,604 - 0,420}{0,143} = 1,287$$

$$Z_{0.05(2)} = t_{0.05(2)} = 1.96$$



We do not consider $H_0 : 0.10 < P < 0.20$

Lecture 8

Contingency tables

Goodness of fit test

Fisher's exact test

McNemar test

Odds ratio and relative risk

Annotation

- The contingency table analysis allows to analyze the relationship between two categorical variables.
- The basic method of testing is the so-called chi-square test, which compares the observed frequencies of combinations of categories against the expected frequencies based on the theoretical situation where the relationship between the variables is random.
- The goodness-of-fit test is also used to compare observed frequencies against expected frequencies given by a rule (a typical example is Hardy-Weinberg equilibrium in genetics)
- A specific type of outputs derived from contingency tables are odds ratios and relative risks, often used in medicine to identify and describe at-risk patient groups.

What is a contingency table ?

- Frequency summation of two categorical variables (binary, nominal or ordinal variables).
- In general: R x C contingency table (R - number of categories of one variable, C - number of categories of the other variable).
- Special case: 2 x 2 table = four-field table.
- Contingency tables: absolute frequencies, total percentages, row/column frequencies
- Example: summary of examined persons by sex and diagnostic test result.

Gender	Result of the examination		Total
	Sick	Healthy	
Man	45	11	56
Woman	25	6	31
Total	70	17	87

Sample contingency table

- Relationship between sex and disease incidence (beware of assessing a nonsensical relationship)

	Sick	Healthy	Total
Man	a	b	a + b
Woman	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

	Sick	Healthy	Total
Man	45	11	56
Woman	25	6	31
Total	70	17	87



Are men or women sicker?

Goodness-of-fit test - basic theory

Test statistics:

$$\chi^2 = \sum \frac{\left[\begin{array}{c} \text{viewed at} \\ \text{frequency} \end{array} - \begin{array}{c} \text{awaited} \\ \text{frequency} \end{array} \right]^2}{\text{expected frequency}}$$

$$\chi^2 = \underbrace{\frac{\left[\begin{array}{c} \text{viewed at} \\ \text{frequency} \end{array} - \begin{array}{c} \text{awaited} \\ \text{frequency} \end{array} \right]^2}{\text{expected frequency}}}_{\substack{1. \\ \text{phenomenon}}} + \underbrace{\frac{\left[\begin{array}{c} \text{viewed at} \\ \text{frequency} \end{array} - \begin{array}{c} \text{awaited} \\ \text{frequency} \end{array} \right]^2}{\text{expected frequency}}}_{\substack{2. \\ \text{phenomenon}}} + \dots$$

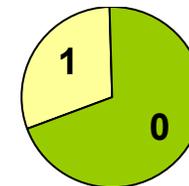
$$\chi^2 > \chi^2_{(1-\alpha)} \text{ (s.v.)} \quad \dots \text{ we reject } H_0$$

1 - significance level

degrees of freedom

Goodness-of-fit test: example

Binomial phenomena (1/0)



$$\chi^2_{(1)} = \frac{\left[\frac{\text{viewed at frequency}}{\text{expected frequency}} - \frac{\text{awaited frequency}}{\text{expected frequency}} \right]^2}{\frac{\text{viewed at frequency}}{\text{expected frequency}}} + \frac{\left[\frac{\text{viewed at frequency}}{\text{expected frequency}} - \frac{\text{awaited frequency}}{\text{expected frequency}} \right]^2}{\frac{\text{viewed at frequency}}{\text{expected frequency}}}$$

Phenomen

phenomen

on 1
✓

on 2

Example

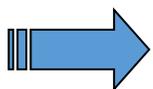
10,000 people flip a coin
 flip: 4,000 cases (R)
 face: 6 000 cases (L)

?

Can the result be considered statistically significantly different (or not different) from the expected ratio R : L = 1 : 1 (i.e., the outcome of the coin toss is random)?

$$\chi^2 = \frac{(4000 - 5000)^2}{5000} + \frac{(6000 - 5000)^2}{5000} = 400$$

Table value: $\chi^2_{(0,95)} (v = k - 1 = 1) = \underline{3,84} \quad (0,95 = 1 - \alpha)$

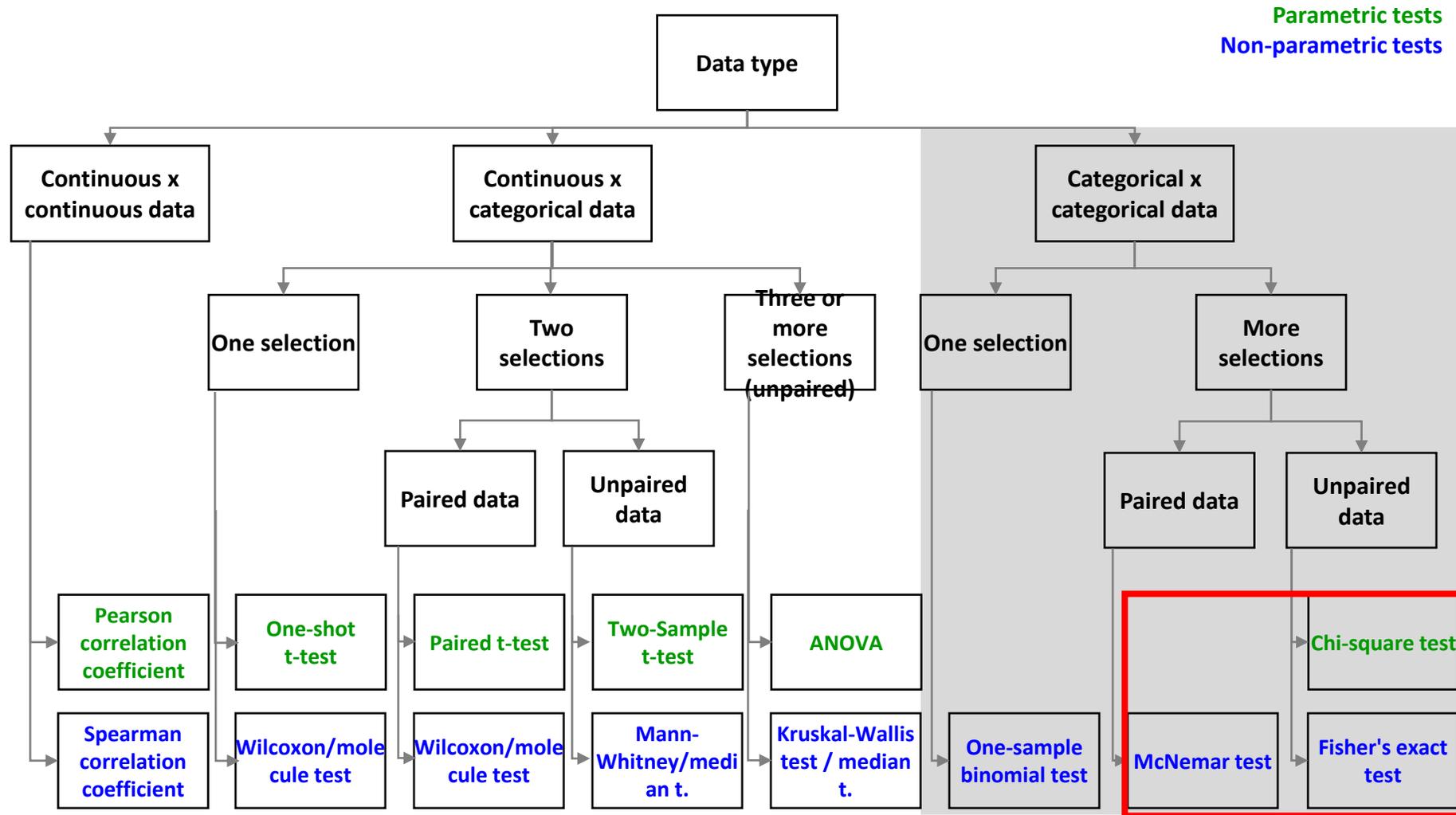


The difference is highly statistically significant (p < 0.001)

Contingency table - hypotheses

- **INDEPENDENCE** (Pearson chi-square test, Fisher's exact test)
 - One selection, 2 characteristics - similar to unpaired arrangement
 - E.g.: the existence of a relationship between eye colour and students' biostatistics grades
- **STRUCTURE agreement** (Pearson chi-square test, Fisher's exact test)
 - The so-called homogeneity test
 - Multiple selections, one characteristic - similar to unpaired arrangement
 - E.g.: age structure of diabetic patients in K hospitals (i.e. K selections)
- **SYMMETRY** (McNemar test)
 - One selection, repeatedly one characteristic - similar to a pairwise arrangement
 - E.g.: assessment of tree condition in two seasons

Basic decision making on the selection of statistical tests - contingency table analysis



Contingency table - general

- We have two nominal quantities, X (has r variations) and Y (has s variations)
- Contingency table of type r x s

$x_{[j]} \backslash y_{[k]}$	$y_{[1]}$	$y_{[s]}$	$n_{j.}$
$x_{[1]}$	n_{11}	n_{1s}	$n_{1.}$
.
.
$x_{[r]}$	n_{r1}	n_{rs}	$n_{r.}$
$n_{.k}$	$n_{.1}$.	.	$n_{.s}$	n

 Marginal absolute frequency
 Marginal absolute Frequency
 Simultaneous absolute Frequency

- Designation:
 - n_{jk} - simultaneous absolute frequency,
 - $n_{j.}$ - marginal absolute frequency

Contingency tables H0 :Independence of two phenomena A and B

Contingency table
2 x 2

↘ B / ↗ A	+	-	Share (+)
+	a	b	$\frac{a}{(a+b)}$ p_1
-	c	d	$\frac{c}{(c+d)}$ p_2
Share (+)	$\frac{a}{(a+c)}$	$\frac{b}{(b+d)}$	

$$N = a + b + c + d$$

$$P(B^+) = \frac{(a+b)}{N}$$

$$P(B^-) = \frac{(c+d)}{N}$$

Expected frequencies:

$$F_{(A)} = \frac{(a+b)(a+c)}{N}$$

$$F_{(C)} = \frac{(a+c)(d+c)}{N}$$

$$\chi^2_{\nu=1} = \sum_{i=1}^4 \frac{(f_i - F_i)^2}{F_i}$$

$$F_{(B)} = \frac{(a+b)(b+d)}{N}$$

$$F_{(D)} = \frac{(b+d)(c+d)}{N}$$

$$\nu = 1 = (r-1) * (c-1)$$

$$P_{(A)}; P_{(B)}$$

$$\chi^2_c = \sum \sum \frac{(|f_{ij} - F_{ij}| - 0,5)^2}{F_{ij}}$$

Expected frequencies in the contingency table

- The expected frequencies for calculating a goodness-of-fit test in a contingency table correspond to a table that has no relationship between rows and columns (random row-column relationship)

$$\chi^2_{(1)} = \frac{\left[\begin{array}{c} \text{viewed at} \\ \text{frequency} \end{array} - \begin{array}{c} \text{awaited} \\ \text{frequency} \end{array} \right]^2}{\text{expected frequency}}$$

Calculated for each cell of the table

	☠	😊
A	10	0
B	0	10

Observed table

	☠	😊
A	5	5
B	5	5

Expected table

Independence testing - Pearson chi-square test

- Is the occurrence of two nominal traits measured on a single sample related?
- Example: the eye colour (blue, green, brown) and hair colour (brown, black, blonde) of a sample of 30 students are independent.
- **The null hypothesis:** the characters X and Y are independent random variables.
- **Alternative hypothesis:** the characters X and Y are dependent random variables.
- Test: **Pearson chi-square**

$$K = \sum_{j=1}^r \sum_{k=1}^s \frac{(n_{jk} - e_{jk})^2}{e_{jk}} \approx \chi^2((r-1)(s-1)) \quad \text{H}_0 \text{ applies}$$

- Expected (theoretical) frequencies $e_{jk} = \frac{n_{j.} \cdot n_{.k}}{n}$
- We reject H0 at the α significance level if $K \geq \chi_{1-\alpha}^2((r-1)(s-1))$

- **Assumptions of the test ?**

Independence testing - Pearson chi-square test

Pearson chi-square test assumptions:

- **Individual observations** summarized in the contingency table are **independent**, i.e. each element belongs to only one cell of the contingency table, it cannot belong to two at the same time.
- **Good approximation conditions:** expected (theoretical) frequencies are greater than or equal to 5 at least 80% of the time and must not be less than 2 100% of the time (if this assumption is not met, it is appropriate to merge categories with low frequencies).

$$V = \sqrt{\frac{K}{n(m-1)}}, \text{ kde } m = \min\{r, s\}, V \text{ je z intervalu } (0,1)$$

- **Measuring the strength of dependence:**

Cramer's coefficient:

Significance of values: 0-0.1....negligible dependence

0.1-0.3...weak dependence

0.3-0.7...medium dependence

0.7-1 strong dependence

Contingency tables: example

gen \ †	Yes	No	Σ
e Yes	20	82	102
No	10	54	64
Σ	30	136	166

$$F_A = 102 * 30 / 166 = 18.43$$

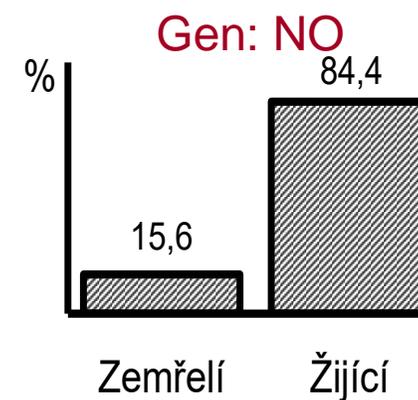
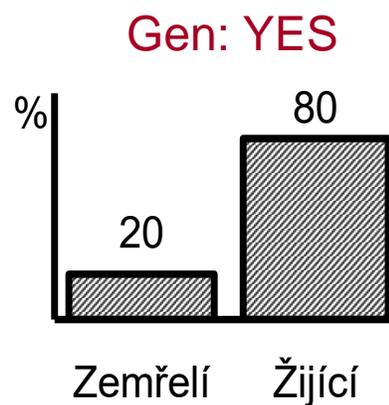
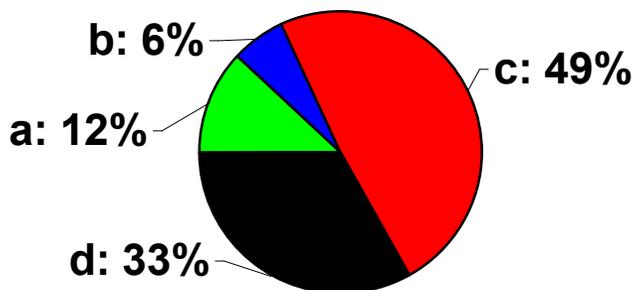
$$F_B = 102 * 136 / 166 = 83.57$$

$$F_C = 11.57$$

$$F_D = 52.43$$

$$\chi^2_{(1)} = \frac{(20 - 18,43)^2}{18,43} + \frac{(82 - 83,57)^2}{83,57} + \frac{(10 - 11,57)^2}{11,57} + \frac{(54 - 52,43)^2}{52,43} = 0,423 \quad 0,423 < \chi^2_{0,95}^{(1)} = 3,84$$

Contingency table in picture



Output of the solution in SW

Table 1: Observed frequencies

Summary Frequency Table (07 priklad_z_prednasky_K
Marked cells have counts > 10
(Marginal summaries are not marked)

Gen	Stav_pacienta úmrtí	Stav_pacienta žijící	Row Totals
přítomen	20	82	102
nepřítomen	10	54	64
All Grps	30	136	166

Table 2: Expected frequencies

Summary Table: Expected Frequencies (07 priklad_z_pre
Marked cells have counts > 10
Pearson Chi-square: ,421322, df=1, p=,516278

Gen	Stav_pacienta úmrtí	Stav_pacienta žijící	Row Totals
přítomen	18,43373	83,5663	102,0000
nepřítomen	11,56627	52,4337	64,0000
All Grps	30,00000	136,0000	166,0000



Are the conditions for a good approximation met?

Table 3: Paerson chi-square

The value of the test statistic

Number of degrees of freedom

p-value

Statistic	Chi-square	df	p
Pearson Chi-square	,4213223	df=1	p=,51628
M-L Chi-square	,4277117	df=1	p=,51311
Phi for 2 x 2 tables	,0503794		
Tetrachoric correlation	,0949754		
Contingency coefficient	,0503156		

R x C contingency table

Sample: N people from a sociological survey (offenders)

Phenomenon A: Origin from broken homes

Phenomenon B: Crime rate I < II < III < IV

A \ B	I.	II.	III.	IV.	Σ
YES	a	b	c	d	number 1
NO	e	f	g	h	
	Σ number2				

Degrees of freedom:
 $(R-1) * (C-1) = 1 * 3 = 3$

$$F_a = \frac{\text{číslo 1} \cdot \text{číslo 2}}{N}$$

Tables: $\chi^2_{(1-\alpha)}^{(v)}$

Expected frequencies:

$$p_a = \frac{a}{a+e}$$

$$p_b = \frac{b}{b+f}$$

$$p_c = \frac{c}{c+g}$$

$$p_d = \frac{d}{d+h}$$

Recoding categorical variables to binary

Original	Dummies				Given the reference			
	NYHA	NYHA I	NYHA II	NYHA III	NYHA IV	NYHA II ref	NYHA III ref	NYHA IV ref
I	1	0	0	0	0	0	0	0
I	1	0	0	0	0	0	0	0
I	1	0	0	0	0	0	0	0
II	0	1	0	0	0	1		
II	0	1	0	0	0	1		
III	0	0	0	0	0		1	
III	0	0	0	0	0		1	
IV	0	0	1	1				1
IV	0	0	1	1				1

Recoding categorical variables to binary

- Categorical and ordinal data can enter the analysis as binary variables
- Categorical data (cannot be sorted) -> dummies
- Ordinal data (sortable)
 - Dummies
 - Definition of the reference category (usually the category with the lowest risk for the endpoint being evaluated)
- Example: the New York Heart Association (NYHA) Functional Classification

Original NYHA	Dummies				Given the reference		
	NYHA I	NYHA II	NYHA III	NYHA IV	NYHA II ref	NYHA III ref	NYHA IV ref
I	1	0	0	0	0	0	0
I	1	0	0	0	0	0	0
I	1	0	0	0	0	0	0
II	0	1	0	0	1		
II	0	1	0	0	1		
III	0	0	0	0		1	
III	0	0	0	0		1	
IV	0	0	1	1			1
IV	0	0	1	1			1

Goodness-of-fit test: example I

? Using data from an experiment with 100 flowers of a particular species, verify that flower colour is genetically split in

ratio yellow : red = 3 : 1

H_0 : The observed frequencies for each flower color are a sample of the population having a ratio of yellow and red flowers 3:1.

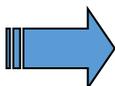
✓ The sum of the frequencies for both flower colors (f_i) equals 100 and the observed frequencies for the color categories

will be compared with the expected frequencies (given in brackets):

	Categories of colour		
	Yellow	Red	n
$f_{poz.}$	84	16	100
$f_{oček.}$	75	25	

$$\chi^2 = \sum \frac{(f_{poz.} - f_{oč.})^2}{f_{oč.}} = \frac{(84 - 75)^2}{75} + \frac{(16 - 25)^2}{25} = 4,320$$

St. of freedom = $n = k - 1 = 1$



We reject the hypothesis of the matching of the compared frequencies

In testing H_0 , we used the mathematical notation ($0.025 < P < 0.05$). From the χ^2 distribution tables, we see that the probability of crossing the 2.706 threshold is 0.1 (10%), which can be written succinctly as $P(\chi^2 \geq 2.706) = 0.10$.

Furthermore, we can find for $P(\chi^2 \geq 3.841) = 0.05$. In the problem we solved, we arrived at the value of the test statistic $\chi^2 = 4.320$. Thus, for this case, we can write $0.025 < P(\chi^2 \geq 4.320) < 0.05$; and more simply, $0.025 < P < 0.05$. This is essentially an approximation of the bounds of a type 1 error.

Goodness-of-fit test: example II

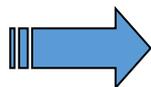
This example is an extension of the problem in Example 1 to compare observed and expected frequencies for multiple categories of the observed trait:

- ✓ A total of 250 seeds of a particular plant species were examined and classified into the following categories: yellow/smooth; yellow/wrinkled; green/smooth; green/wrinkled. The following table contains the original observation data and the procedure for testing H_0 .

	yellow/smooth	yellow/wrinkled	green/smooth	green/wrinkled	n
$f_{\text{poz.}}$	152	39	53	6	250
$f_{\text{oček.}}$	140,6250	46,8750	46,8750	15,6250	

$$v = k - 1 = 3$$

$$\chi^2 = \frac{11,3750^2}{140,6250} + \frac{7,8750^2}{46,8750} + \frac{6,1250^2}{46,8750} + \frac{9,6250^2}{15,6250} = 8,972$$



We reject the hypothesis that the observed frequencies are consistent with the expected

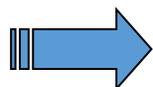
Goodness-of-fit test: example III

More complex examples solved by frequency comparisons can be divided into partial hypothesis testing:

✓ For the data from the previous problem, suppose we want to test the hypothesis of the existence of a 9 : 3 : 3 splitting ratio for the first three seed categories:

	yellow/smooth h	yellow/wrinkle d	green/smooth	n
$f_{\text{poz.}}$	152	39	53	244
$f_{\text{oček.}}$	146,400	48,800	48,800	

$n = k - 1 = 2$

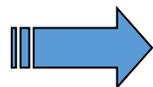
$$\chi^2 = \frac{5,600^2}{146,40} + \frac{9,800^2}{48,80} + \frac{4,200^2}{48,80} = 2,544$$


We do not reject the hypothesis that the observed frequencies match the expected frequencies

✓ We now test the hypothesis of a splitting ratio of the categories green/wrinkled:other types = 1:15

	green/wrinkled	Other	n
$f_{\text{poz.}}$	6	244	25
$f_{\text{oček.}}$	15,625	234,375	

$n = k - 1 = 1$

$$\chi^2 = \frac{9,625^2}{15,625} + \frac{9,625^2}{234,375} = 6,324$$


We reject the hypothesis that the observed frequencies match the expected frequencies

Homogeneity testing (conformity of structure)

- Motivation: we are interested in the occurrence of a nominal trait in r independent samples from r different populations.
- Example: is the interest in sport the same for girls as for boys?
- Null hypothesis: the probability distribution of the categorical variable is the same in different populations
- Test: **Pearson chi-square**

		Girls	Guys	
Interest about sport	Yes	a	b	$a+b$
	No	c	d	$c+d$
		$a+c$	$b+d$	n

Some marginal frequencies (either column or row) are fixed in advance

Homogeneity test for binomial distributions

- ✓ Phenomenon: Leukaemia mortality
- Assumption: $\Pi = 0.6$
- Absolute frequency of the phenomenon denoted by r_i

$$\bar{p} = \frac{\sum p_i}{S}$$

Followed with authors from s

Author	n_i	r_i	p_i
1			
2			
⋮			
⋮			
⋮			
s	$\sum n_i = N$		

➡ Homogeneity test for binomial distributions

➡ After possible merger with selections

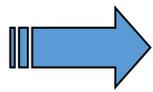
Real r consistency test $(\sum r_i) \ a \ n \cdot \Pi$

$$\chi^2_{s-1} = \frac{(\sum r_i p_i - \bar{p} \sum r_i)}{\bar{p} (1 - \bar{p})}$$

$$\chi^2_{(1)} = \frac{\left(\left| \sum r_i - N \cdot \Pi \right| - \frac{1}{2} \right)^2}{N \cdot \Pi \cdot (1 - \Pi)}$$

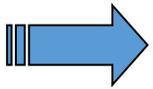
Test for homogeneity of binomial frequencies: an example

The χ^2 distribution can also be used to assess the homogeneity of multiple independent experiments testing the same hypothesis.



Six independent samples were taken from a population of young men who developed severe meningitis in childhood.

H_0 : In this population, right-handed and left-handed individuals occur in a 1 : 1 ratio.



Find the appropriate relationships in the literature to test the homogeneity of all six sample populations and decide on a course of action based on the results of this test.

The following table shows the original data and the result of the testing (expected frequencies are in brackets):

Sample	The Right	Lefties	n	c2	St. Liberty
1	3 (7)	11 (7)	14	4,5714	1
2	4 (8)	12 (8)	16	4,000	1
3	15 (10)	5 (10)	20	5,000	1
4	14 (9)	14 (9)	18	5,5556	1
5	13 (8,5)	4 (8,5)	17	4,7647	1
6	17 (11)	5 (11)	22	6,5455	1

$$\chi^2_{heterogenia} = 30,2$$

$$\nu = s - 1 = 5$$

$$P < 0,001$$

Simple testing reveals that all tests for each sample are significant, meaning that in no case was the agreement between the expected and observed frequencies confirmed. The test for homogeneity of the splitting ratio in the populations examined also led to the rejection of the possibility to merge the individual samples and consider them as a whole (thus, apart from the tested 1 : 1 ratio, there is no other uniform splitting ratio between the two traits in the data).

In the event that this test did not show variation between the sample populations, the individual samples could be pooled and treated as a homogeneous sample.

χ^2 test - example of fractionation of more complex contingency table I

The aim of the larger population survey was to investigate the relationship between two types of diseases and blood types in humans. The specific data are shown in the table:

Blood group	Stomach ulcers	Cancer of the stomach	Check	Total
0	983	383	2892	4258
A	679	416	2625	3720
B	134	84	570	788
Total	1796	883	6087	8766

Calculate the test characteristic for this contingency table and test the null hypothesis of independence of the phenomena ($\chi^2 = 40.54$; 4 degrees of freedom)

χ^2 test - example of fractionation more complex contingency table

II

A more detailed exploration of the more complex tables is greatly aided by transcribing the original table into a percentile representation of the categories:

Blood group	Stomach ulcers	Cancer of the stomach	Check
0	983	383	2892
A	679	416	2625
B	134	84	570
Total	1796	883	6087

This table shows:

1. There are only small differences in the distribution of blood groups in the control and stomach cancer groups.
2. Patients with ulcers are much more likely to have blood type 0.

Based on these findings, it is possible to construct a smaller contingency table that tests the hypothesis of an identical distribution of blood groups for cancer patients and healthy people.

Construct this table and test the null hypothesis.

($\chi^2 = 5.64$ (2 deg. v.), P is approximately equal to 0.06)

χ^2 test - example of fractionation more complex contingency table III

- This subtest suggests the possibility of merging the group of cancer patients and healthy people because they behave as a homogeneous population due to the distribution of blood groups.
- The next logical step in the detailed analysis is to test the agreement of the relative frequencies of blood groups A and B between the combined sample (merged cancer and control group) and the sample of people with gastric ulcers - i.e. we do not consider blood group 0 now. The result of this test is $\chi^2 = 0.68$ (1 st. vol.); $P > 0.7$. Thus, the samples for blood groups A and B can be combined into a mixed A + B sample.
- We now test the concordance of the relative frequencies of group 0 versus A + B among the combined population (control + cancer patients) and among the sample of ulcer patients ($c^2 = 34.29$; 1 st. vol.).
- Thus, it can be summarized that the high value of the original c^2 with 4 degrees of freedom was due to the increased frequency of people with blood group 0 among gastric ulcer patients.

χ^2 test - example of fractionation more complex contingency table IV

The evaluation process can be summarised in a table:

Compare	St. Liberty	c2
0, A, B group in cancer patients (r) x control (k)	2	5,64
A, B group in patients with ulcers x combined sample (r + k)	1	0,68
0, A, B group in patients with ulcers x combined sample (r + k)	1	34,29
Total	4	40,61

The overall sum of the test statistics χ^2 (40.61) corresponds approximately to the original value of χ^2 (40.54). This is also true for the degrees of freedom (4). This fact confirms that we have exhausted the information content of the original contingency table by detailed analysis and that, apart from the described dependence (increased prevalence of blood group 0 in people with gastric ulcers), the individual categories of the phenomena under study are completely independent.

2 x 2 contingency table: solutions for insufficient sample size

Yates' correction

Fisher's exact test



H_0 : Independence
of phenomena

The test analyzes all possible 2 x 2 tables that give the same sum of rows and columns as the source table.

The algorithm assigns to each table the probability that such a situation occurs if H_0 is true.

Spectacle wearing among juvenile delinquents and non-delinquents who failed a vision test (Weindling et al., 1986)

		Juvenile delinquents	Non-delinquents	Total
Spectacle wearers	Yes	1	5	6
	From	8	2	10
	Total	9	7	16

2 x 2 contingency table: solutions for insufficient sample size

All possible variants of the table with the given sum of rows and columns

(I)	0	6	(V)	4	2
	9	1		5	5
(II)	1	5	(VI)	5	1
	8	2		4	6
(III)	2	4	(VII)	6	0
	7	3		3	7
(IV)	3	3			
	6	4			

Probability of random occurrence of table variants

	a	b	c	d	P
(I)	0	6	9	1	0,00087
(II)	1	5	8	2	0,02360
(III)	2	4	7	3	0,15734
(IV)	3	3	6	4	0,36713
(V)	4	2	5	5	0,33042
(VI)	5	1	4	6	0,11014
(VII)	6	0	3	7	0,01049
Total					0,99999

Fisher's exact test

- Use in a four-field table (currently larger due to increased computer power) with low frequencies that preclude the use of Pearson chi-square tests.
- It is a non-parametric test working with data on a nominal scale, in its simplest form in two classes: positive/negative, success/failure, etc.
- The null hypothesis assumes equal representation of the observed trait in two independent sets.
- The word exact (direct) means that the probability of rejection or validity of the null hypothesis is calculated directly.

Fisher's exact test

- Calculating the "exact" p-value, which here plays the role of a test statistic:
 - the partial probability of the four-field table p_1 is calculated:

Sledovaný jev	Skupina		Celkem
	Experimentální	Kontrolní	
Ano	a	b	$a + b$
Ne	c	d	$c + d$
Celkem	$a + c$	$b + d$	n

$$p_1 = \frac{(a+b)! \cdot (c+d)! \cdot (a+c)! \cdot (b+d)!}{N! \cdot a! \cdot b! \cdot c! \cdot d!}$$

- The probability of all possible tables is computed while keeping marginal frequencies (row and column totals) and the resulting p-value is the sum of probabilities less than or equal to p_1 that belongs to the observed table.

Test of the symmetry hypothesis (McNemar's test for a four-field table)

- Motivation: we observe a binary variable on subjects before and after the experiment, the aim is to see if there is a change in the distribution of this variable.
- **Analysis of paired dichotomous variables**

Frequency table

		after		n _{j.}
		+	-	
Before	+	a	b	a+b
	-	c	d	c+d
		a+c	b+d	n
		n _{.k}		

Table of theoretical probabilities

		after		
		+	-	
Before	+	p ₁₁	p ₁₂	p _{1.}
	-	p ₂₁	p ₂₂	p _{2.}
		p _{.1}	p _{.2}	

- Null hypothesis: the experiment has no effect on the occurrence of the trait

$$\chi^2 = \frac{(b - c - 1)^2}{b + c}$$

- Test statistic: if it is greater than the critical value of the distribution by one degree of freedom (suitable for numbers of data b+c > 8), then reject the null hypothesis

McNemar test: example I

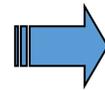
- Find out whether the success of our athletes at the Olympics or World Cup is leading to a change in pupils' attitudes towards sport.
- Null hypothesis: the number of students who change their attitude in a positive direction is only randomly different from the number of students who change their attitude in a negative direction.

		Attitude after teaching		
		+	-	
Attitude before teaching	+	5	3	8
	-	16	2	18
		21	5	26

$$\chi^2 = \frac{(|3 - 16| - 1)^2}{3 + 16} = 7,58$$

Tables: $\chi^2_{1-\alpha} (v = k(k-1)/2 = 1) = 3,84$

Degrees of freedom



H₀ rejected

- Conclusion: the success of our athletes has a positive effect on the attitude of students towards playing sports.

McNemar's test: example II

Example: comparison of 2 methods of antigen determination in blood (antigen always present)



H_0 : method 1 = method 2

Method 1	Method 2	Frequency
success	success	202
success	Failure	60
Failure	success	42
Failure	Failure	10

} $\Sigma = 102$

$$\chi_{(c)}^2 = \frac{(|60 - 42| - 1)^2}{102} = 2,83$$

Tabulky : $\chi_{1-\alpha}^2 (v=1) = 3,84$



H_0 not rejected

Application of 2 x 2 table analysis for risk assessment

I. Prospective study - estimation of relative risk

Individuals are followed prospectively to see if a trait occurs.

THE SELECTION IS GIVEN BY THE COLUMN

GENERAL

		Group 1	Group 2
Sign	YES	a	b
	NO	c	d

$$\text{Risk: } \frac{a}{(a+c)} \quad \frac{b}{(b+d)}$$

$$RR = \frac{\frac{a}{(a+c)}}{\frac{b}{(b+d)}}$$

✓ $H_0 : RR = 1$

EXAMPLE

		Fetal retardation	
		Symmetric	Asymmetric
Agpar skore	YES	2	33
	NO	14 2/16=0,13	58 33/91=0,36
$RR = \frac{2/16}{33/91} = 0,345$			

The risk in the "symmetric group" is about 35% of the risk in the asymmetric group

$$SE(\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}$$

$$\text{IS: } \ln RR - Z_{1-\alpha/2} \cdot SE(\ln RR)$$

$$\ln RR + Z_{1-\alpha/2} \cdot SE(\ln RR)$$

Application of 2 x 2 table analysis for risk assessment

II. Retrospective study - "ODDS RATIO"

A fundamentally different approach from the retrospective study
SELECTION IS GIVEN BY PROPERTY - ROW
 It is therefore not possible to analyse the relative risk because we can change the size of the controls by preparing the lines.

GENERAL

		Group	Group
		1	2
Sign	YES	a	b
	NO	c	d

odds a/c b/d

$$\text{Odds ratio: } \frac{a/c}{b/d}$$

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

EXAMPLE

		Dental defects	
		YES	NO
Swimming weekly	< 6h	32	118
	≥ 6h	17	127

$$OR = (32/17) / (118/127) = 2,026$$

$$\ln(OR) = 0,706$$

$$SE(\ln(OR)) = 0,326$$

Relative risk vs. Odds ratio ?

Relative risk
(relative risk)



Odds ratio
(odds ratio)

- The meaning of RR and OR
- Calculation
- Comparability
- Interpretation
- Advantages and disadvantages

- Applications in clinical trials

The meaning of RR and OR

- Description of the influence of the factor (treatment, clinical parameter) on the occurrence of the event (death, progression, etc.)

Relative risk
(relative risk)



Odds ratio
(odds ratio)

- ✓ Easy natural interpretation of risks expressed as a percentage of events

BUT

- ✓ Mathematical constraints for some applications

- ✓ Only a few people have the natural ability to interpret OR

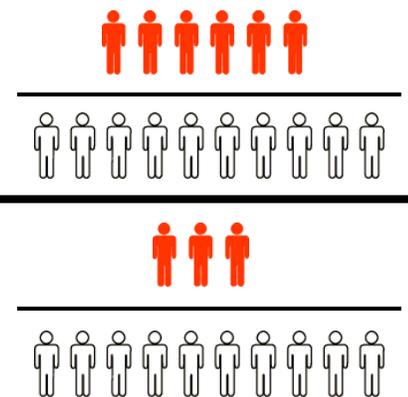
BUT

- ✓ OR more advantageous mathematical properties in many applications

Calculation

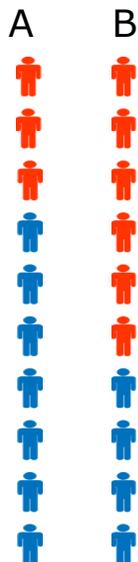
 event  without event

Relative risk
(relative risk)

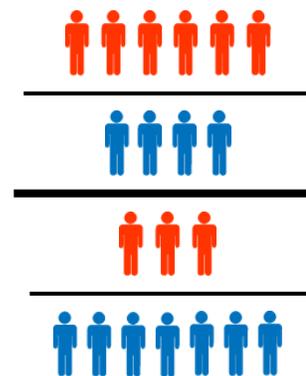
$$RR = \frac{\frac{6}{10}}{\frac{3}{10}} = \frac{6}{3} = 2$$




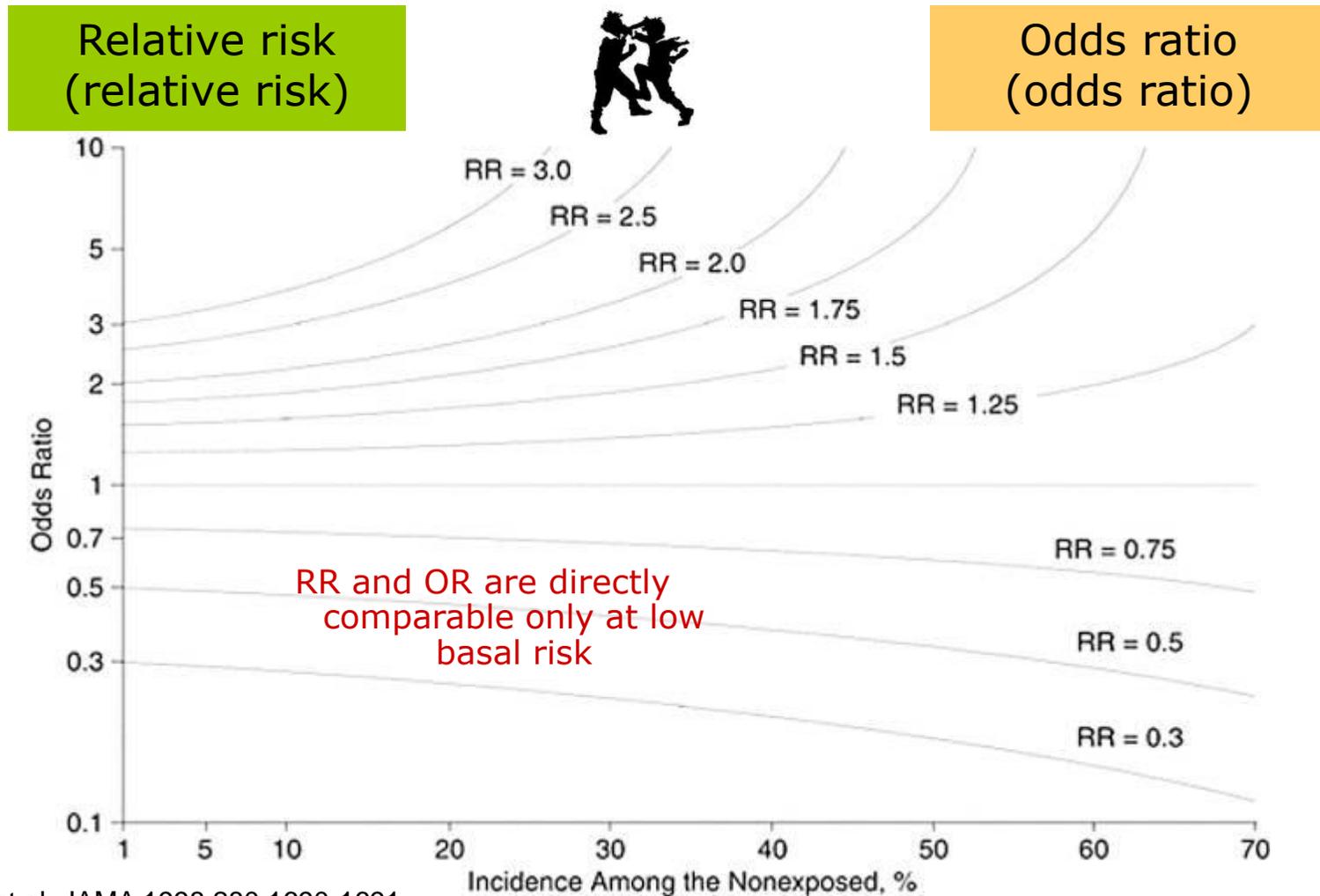
A B



Odds ratio
(odds ratio)

$$OR = \frac{\frac{6}{4}}{\frac{3}{7}} = \frac{6}{3} \cdot \frac{7}{4} = 3.5$$


Relationship between RR and OR



Zhang, J. et al. JAMA 1998;280:1690-1691.

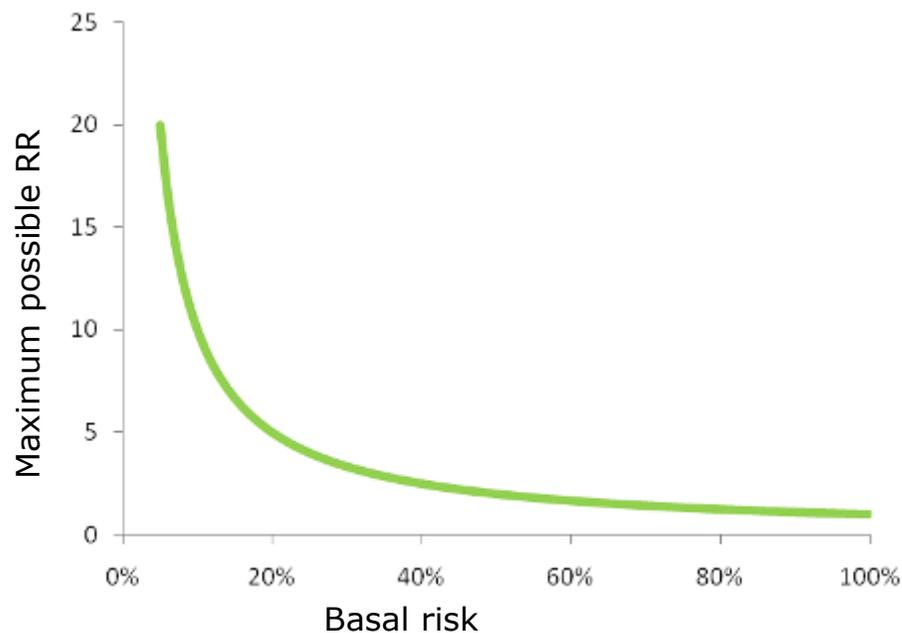
Comparability of RR and OR I: maximum

Relative risk
(relative risk)



Odds ratio
(odds ratio)

- RR varies its maximum according to the basal risk



- ✓ RRs in studies with different basal risks are incomparable !!!!

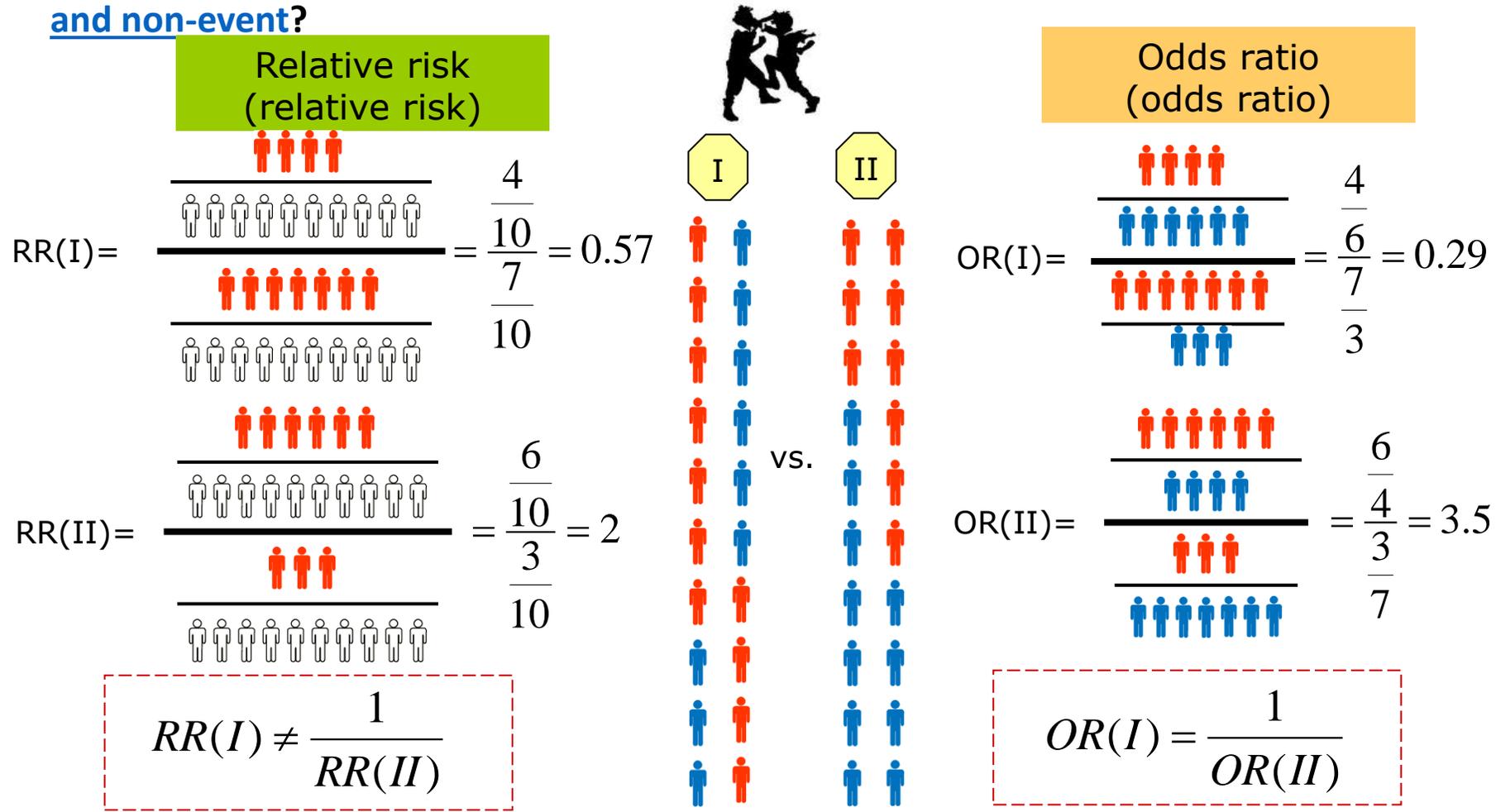
- ✓ Odds ratio always has a range from 0 to infinity
- ✓ The size of the OR is independent of the size of the basal risk

- ✓ ORs can be used to compare studies with different basal risks !!!!

- ✓ Advantageous for meta-analysis

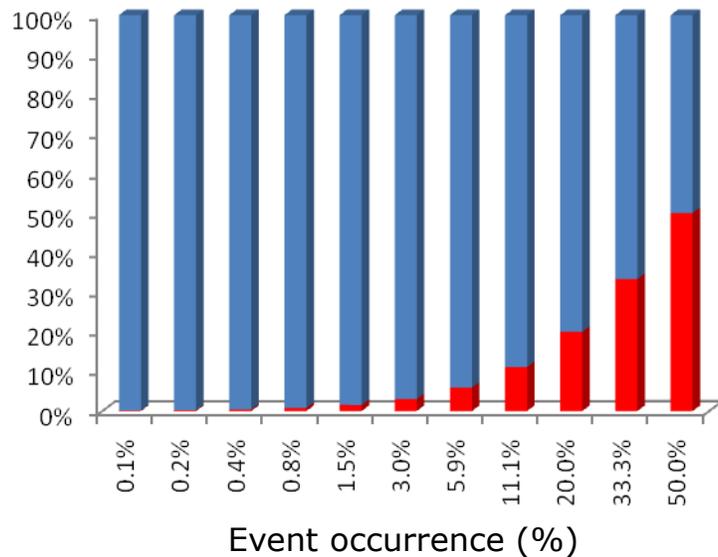
Comparability of RR and OR I: symmetry

- Is there a difference between RR and O in the case of exchanging the definition of event and non-event?

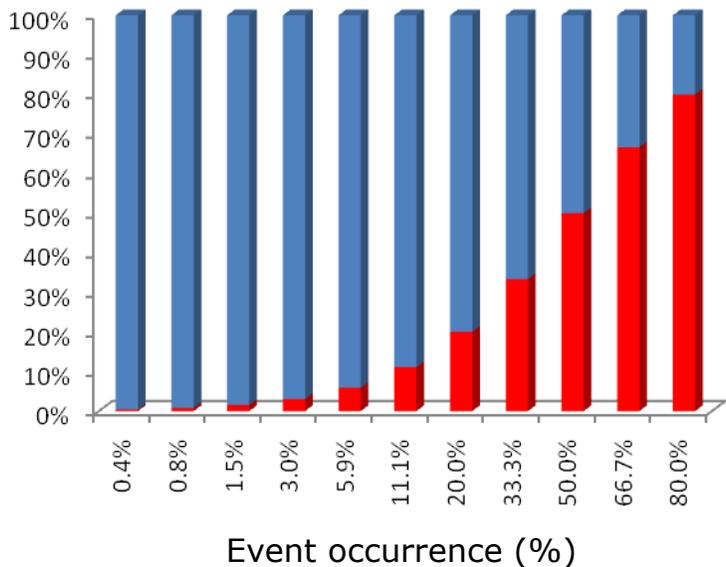


RR and OR in studies with different baseline risk levels

Control

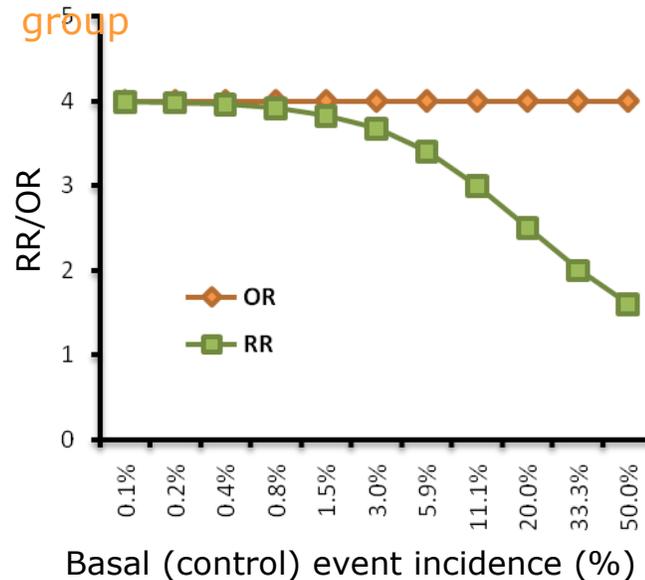


Case



Odds ratio

In the "Case" group, there are 4 times as many patients with an event per patient without an event than in the "Control" group



Relative risk

A patient in the "Case" group has an x times increased probability of having an event than a patient in the "Control" group. X times depends on the basal incidence of the event.

RR and OR in prospective and retrospective studies

Prospective studies

- ☑ Tracking the occurrence of an event and subsequent analysis of its causes
- ☑ Predominantly cohort studies



- ☑ The basal incidence of the event is determined by the characteristics of the patient cohort
- ☑ Seamless use of RR



Relative risk
(relative risk)

Retrospective study

- ☑ **Tracking back the causes of an event**
- ☑ **Mainly case-control studies**
- ☑ **Patient selection influences the basal incidence of the event**



- ☑ RR cannot be used -limited by the basal occurrence of the event
- ☑ Use of OR - not constrained by study design



Odds ratio
(odds ratio)

Relative risk vs. Odds ratio: summary

Relative risk (relative risk)



Odds ratio (odds ratio)

- ✓ **Intuitively easy to interpret**
- ✓ **For prospective studies**
- ✓ **The maximum varies according to the basal value of the event occurrence**

- ✓ Retrospective study
- ✓ Applications in meta-analysis
- ✓ Standard logistic regression output
- ✓ Range always 0 to infinity, not affected by the basal occurrence of the event
- ✓ More difficult to interpret

Lecture 9

Poisson distribution

Description of the layout and its use

Annotation

- The Poisson distribution is used to describe the frequency of occurrence of a phenomenon per experimental unit, an example being the number of bacterial mutations per petri dish or the number of heart defects per unit time

Poisson distribution

Total number of phenomena in n independent experiments

$$\left. \begin{array}{l} E(x) = n p \\ D(x) = n p \end{array} \right\} E(x) = D(x)$$

$$P(r) = \frac{e^{-\mu} \cdot \mu^r}{r!} = e^{-\lambda} \cdot \frac{\lambda^r}{r!}$$

$\mu = \lambda =$ average number of phenomena from n trials

$$\hookrightarrow P(X = 0) = e^{-\mu}$$

$$\hookrightarrow P(X = 1) = e^{-\mu} \cdot \mu^1$$

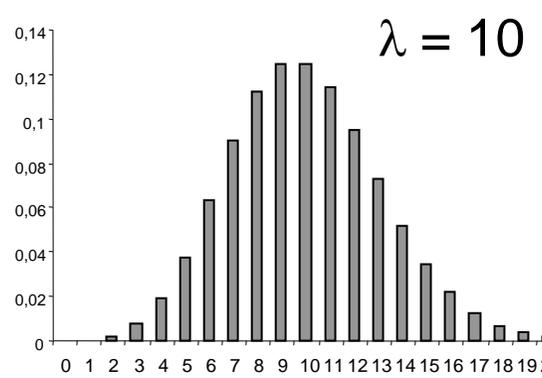
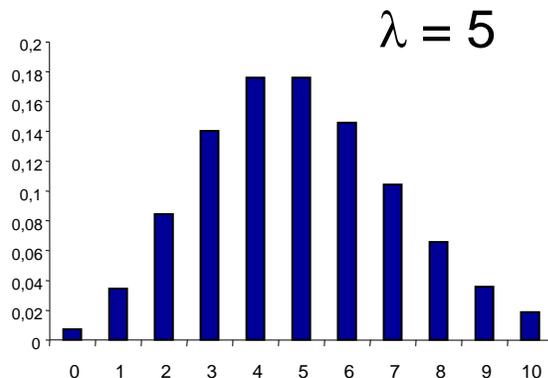
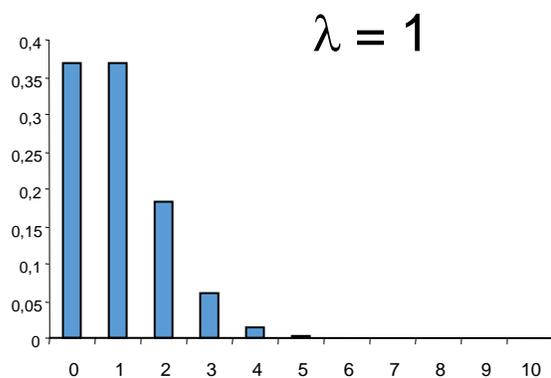
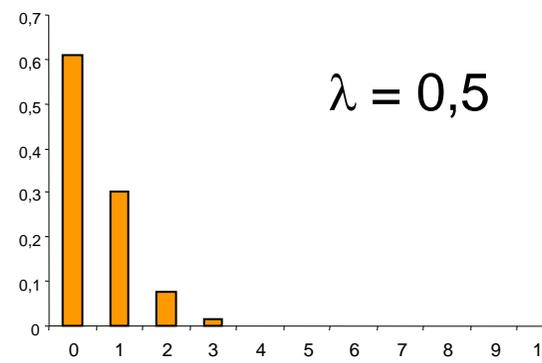
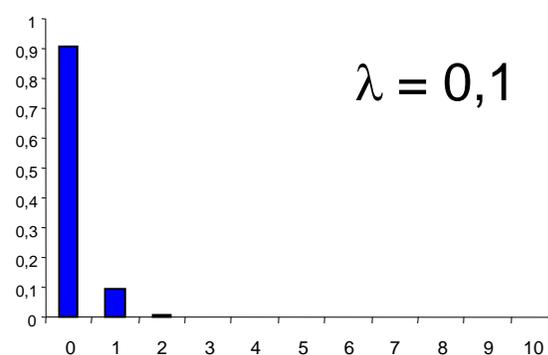
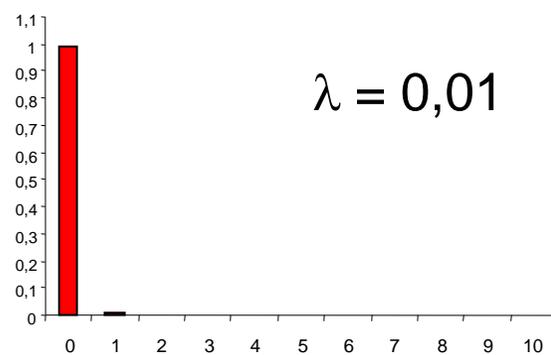
$$\hookrightarrow P(X = 3) = \frac{e^{-\mu} \cdot \mu^3}{(3)(2)}$$

$$\hookrightarrow P(X = 2) = \frac{e^{-\mu} \cdot \mu^2}{2}$$

$$\hookrightarrow P(X = 4) = \frac{e^{-\mu} \cdot \mu^4}{(4)(3)(2)}$$

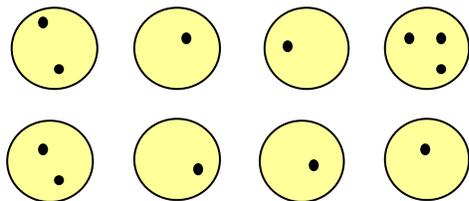
Poisson distribution as a model

$$P(x = r) = e^{-\lambda} \cdot \frac{\lambda^r}{r!}$$



Poisson distribution exists in nature

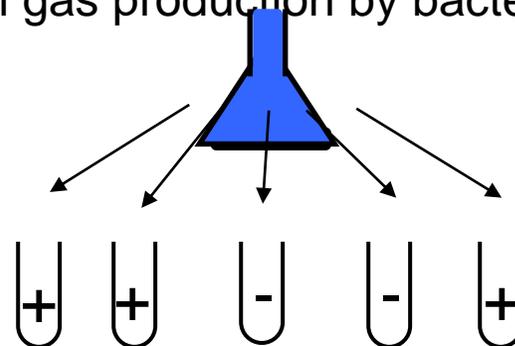
Mutation of bacteria on incubation plates



Occurrence of the phenomenon in space
(number of earthworms per certain area of the field)

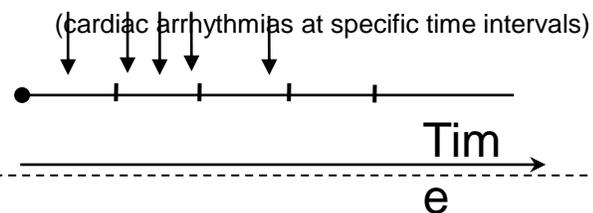


Approximate determination of the phenomenon
(in gas production by bacteria)



The most probable number technique

Occurrence of the phenomenon in time
(cardiac arrhythmias at specific time intervals)

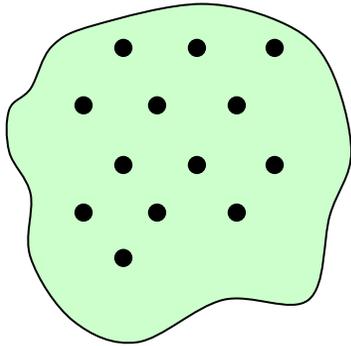


Poisson distribution as a model for random occurrence of phenomena

Assumption: random distribution of the phenomenon among the studied objects

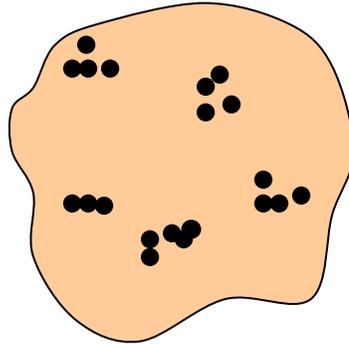
(possibly in time, in space).

$$\sigma^2 < \mu$$

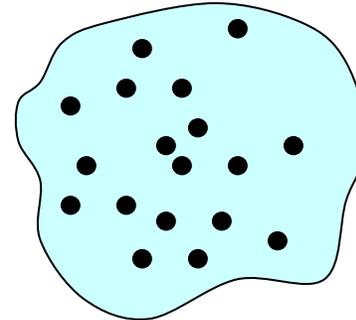


Uniform

$$\sigma^2 > \mu$$



$$\sigma^2 = \mu$$



Poisson

If λ is rather larger ($\sim 5-10$), then the Poisson corresponds to a binomial to normal distribution.

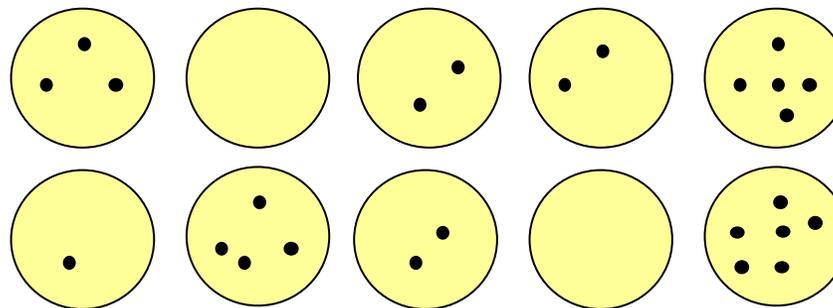
Formal presentation of the Poisson distribution

Re: experiment.....10 000 bacteria on a dish

n = 10 dishes

Phenomenon: mutation (r=25)

□average number of mutants per one bowl



$$r = 25$$

$$\bar{x} \approx \lambda = 25/10 = 2,5$$

95% IS:

$$\bar{x} - Z_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}} \leq \lambda \leq \bar{x} + Z_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}}$$

$$2,5 - 1,96 \cdot \sqrt{0,25} \leq \lambda \leq 2,5 + 1,96 \cdot \sqrt{0,25}$$

$$1,52 \leq \lambda \leq 3,48$$

Poisson random variable

- When measuring the number of blood cells altered by a particular disease (relatively rare), a diluted blood sample is observed under a microscope in a chamber divided into equal-sized fields. The observed quantity, indicating the number of blood cells in the i -th field, can be considered to follow a Poisson distribution:
- $n = 169$ = number of independent variable observations
- $r = 10$ = number of observed blood cells
- What is the value of the \square Poisson distribution parameter and what is its interpretation ?
- What is the 95% confidence interval for the parameter λ
- If we were to observe the total number of red blood cells (again in $n = 169$ independent fields), could this variable also be considered to follow a Poisson distribution ? Consider the total number of observed red blood cells as 2013.

Calculation of confidence interval for \square (without approximation to normal distribution)

IS lower limit

$$L_1 = \frac{\chi_{1-\alpha/2}^2 (f_1=2r)}{2}$$

Upper limit of IS

$$L_2 = \frac{\chi_{\alpha/2}^2 (f_2=f_1+2)}{2}$$

Poisson random variable

Constant radiator: $n = 2608$ time intervals (7.5 s each)

i : number of particles in the interval (x)

s_i : observed frequency of intervals with i particles

$$P(x = i) = \frac{\lambda^i \cdot e^{-\lambda}}{i!} \sim p_i$$

Poisson variable:

* Excellent model for experiments in which the time the number of occurrences of a certain phenomenon

i	Number of intervals with just recorded particles s_i	theoretical frequency e.g. np_i	$\frac{(s_i - np_i)^2}{np_i}$
0	57	54,399	0,1244
1	203	210,523	0,2688
2	383	407,361	1,4568
3	525	525,496	0,0005
4	532	508,418	1,0938
5	408	393,515	0,5332
6	273	253,817	1,4498
7	139	140,325	0,0125
8	45	67,882	7,7132
9	27	29,189	0,1642
10	10	17,075 (= $P\{\xi \geq 10\}$)	0,0677
11	4		
12	2		
13	0		
	$n = 2608$	2608,00	12,8849

Poisson distribution: one-sample test

$$P_{(r)} = \frac{(e^{-\lambda} \cdot \lambda^r)}{r!}$$

Př: Number of quail nests in a given area

$$\left. \begin{array}{l} n = 8\,000 \text{ "sub-sites"} \\ r = 28 \end{array} \right\} \hat{p} = 0,0035$$

Let the comparison file be (previous survey)

$$p_o = 0,0020$$

$$p_o \cdot 8\,000 = 16 = \mu = \lambda$$

$$H_o: p \leq p_o \sim \mu \leq 16 \quad ?$$

1) Take the data as coming from the population:

$$P(r = 28) = \frac{e^{-16} \cdot 16^{28}}{28!} = 0,00192$$

$$2) \left. \begin{array}{l} P(r \geq 28) = ? \\ [0,00411] \end{array} \right\} < 0,05 \Rightarrow H_o \text{ zamítnuta}$$



$r = 28$ is too large for a population with p_o



$p > p_o$ so that $r = 28$ is
More likely

Analysis of variance

Parametric analysis of variance

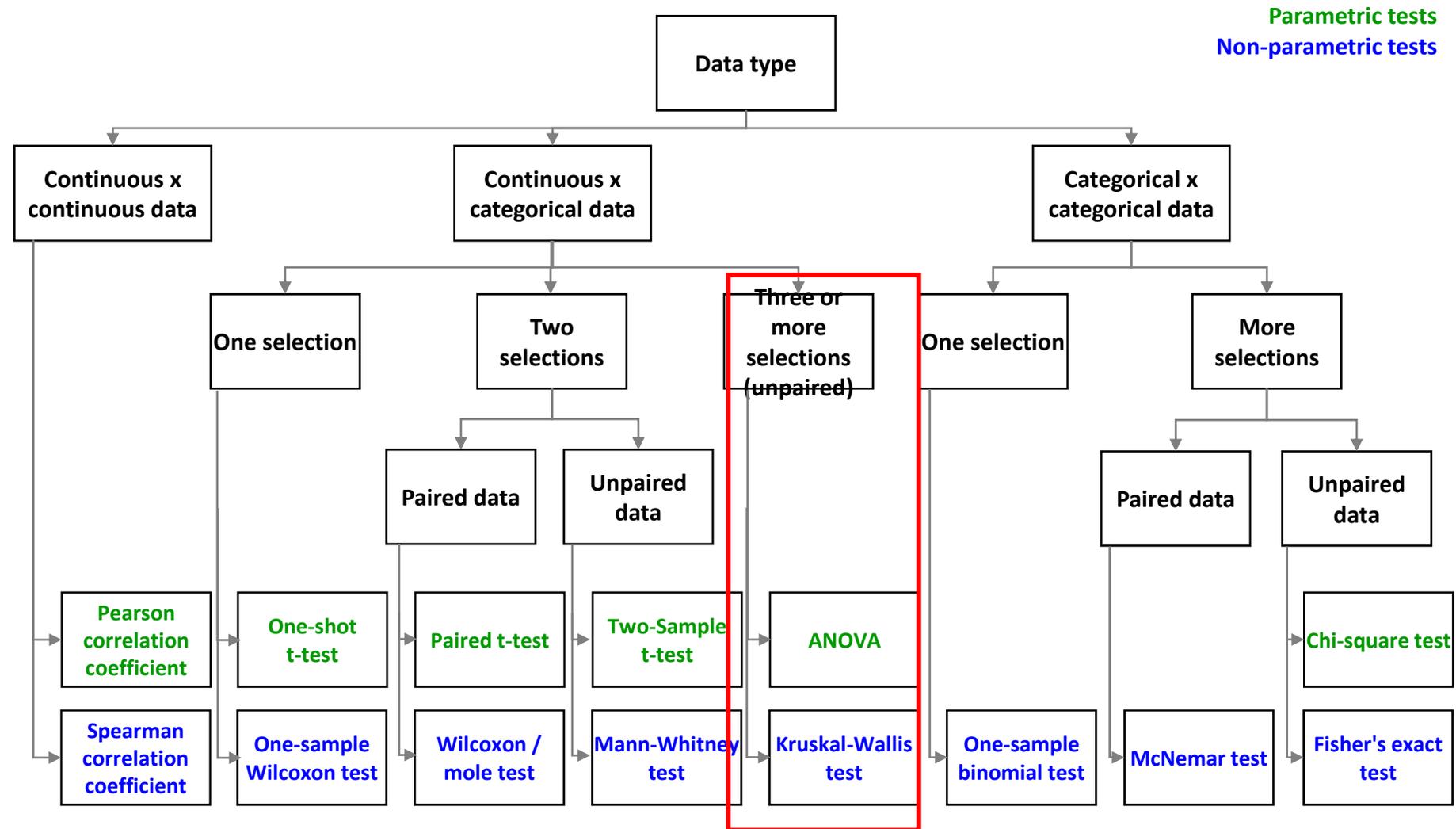
Post hoc tests

Kruskal-Wallis test

Annotation

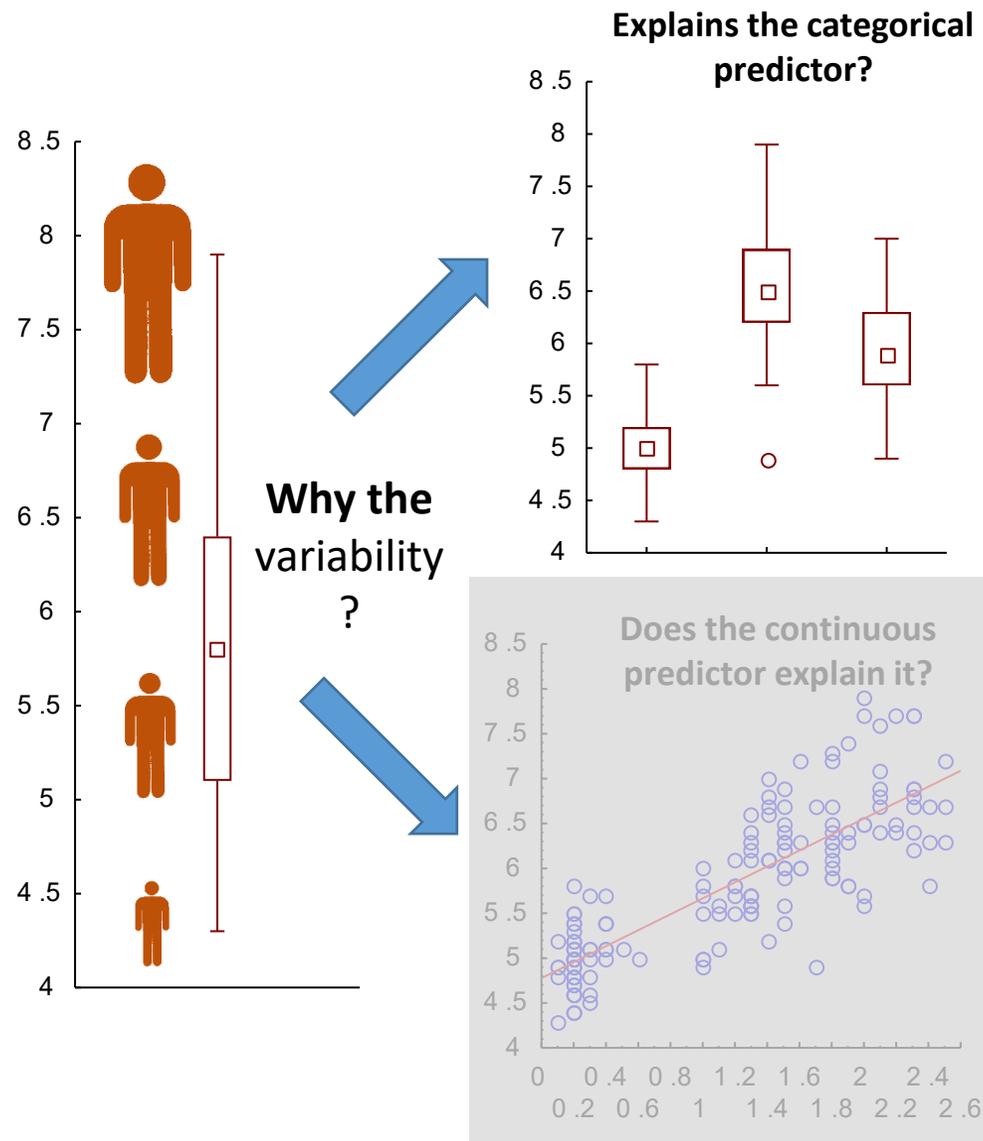
- Analysis of variance is a basic tool for analysing differences between means in several groups of objects.
- The basic idea on which ANOVA is based is to divide the total variability in the data (unknown, given only by random distribution) into a systematic part (associated with patient categories, explained variability) and a random part. If the systematic, i.e., non-random and explained part of the variability dominates, we consider the categorical factor to be important in explaining the variability in the data.
- Analysis of variance evaluates only the overall effect of a factor on variability, in the case of category-by-category analysis, post-hoc tests should be used

Basic decision making on the selection of statistical tests



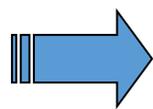
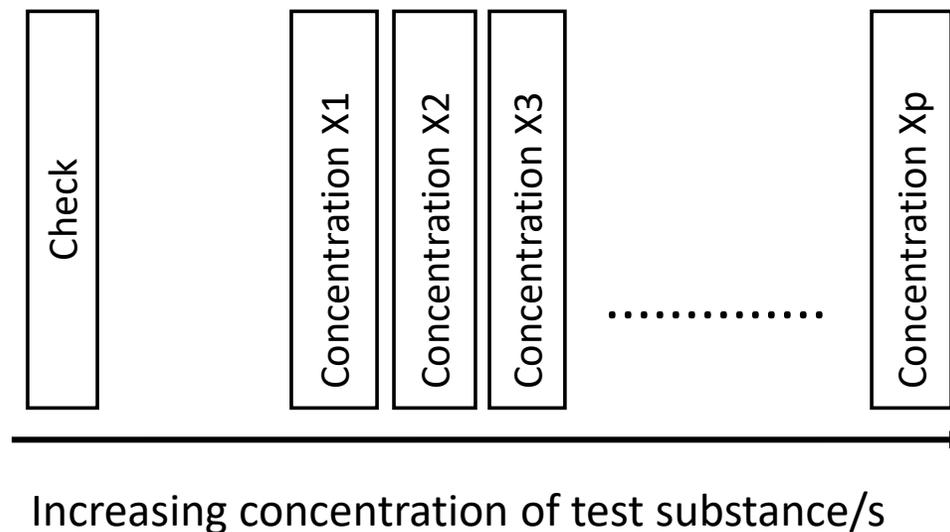
Goal of stochastic modelling

- The general goal is to try to explain the variability of the predicted variable (endpoint, Y) using predictors (explanatory variable, factor, X)
- Both the predicted variable and the predictor can be of different types
 - Binary
 - Categorical
 - Ordinal
 - Continuous
- Censored (-> survival analysis)
- The combination of the data type of the predicted variable and the predictor determines the analysis method used

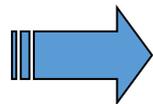


Analysis of variance - ANOVA

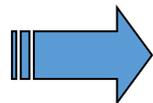
Basic technique used to assess differences between multiple levels of experimental intervention



Overall significant changes in the response of the biological system



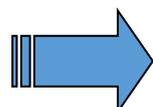
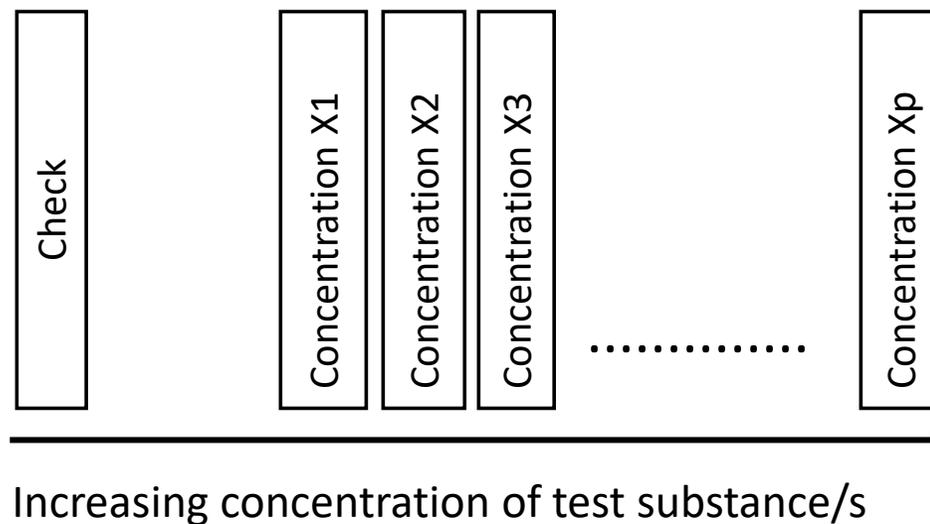
Mutual differences in the effect of individual doses



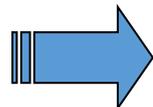
Differences in dose effect from control

Analysis of variance - ANOVA

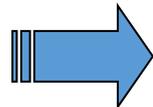
Important steps of the analysis leading to an effective comparison of options



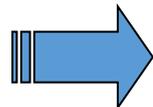
Meeting the assumptions of the analysis
Data transformation



Relevance of the control
(effect of self-application of substances)



Suitability of the ANOVA model for test purposes



Custom comparison of variants
Minimising errors in hypothesis testing

Analysis of variance - ANOVA

MEETING THE ANOVA PREREQUISITES IS A NECESSARY CONDITION
THE USE OF THIS TECHNIQUE

ANOVA
= parametric data
analysis

1. Presumption of independence
repetition of the experiment

2. Homogeneity of
variance within
experimental variants

3. Normality of distribution
within the experimental
variants

NON-PARAMETRIC METHODS ARE AN ALTERNATIVE

ANOVA - assumptions

- Symmetric distribution of values and normality of deviations from the estimated ANOVA model. Much of the data can be adequately normalized using a logarithmic transformation. Of course, the assumption of a lognormal transformation can be theoretically ruled out for many data sets containing discrete parameters where the appropriateness of another type of transformation is indicated. For asymmetrically distributed and discrete data, it is necessary to use nonparametric alternatives to analysis of variance.
- Homogeneity of variance is a prerequisite for meaningful comparisons between experimental variants. In toxicity tests, this assumption should be verified (Bartlett's test), as serious differences (up to orders of magnitude) in the units of the parameter tested may occur due to inhibition by doses of the substance. Inhomogeneity of variance is often related to non-normality (asymmetry) of the data and can be removed by an appropriate normalising transformation.
- Statistical independence of residuals evaluated by ANOVA model. If the estimation and assessment of correlation relationships between experimental variants is not directly under investigation, their influence on the evaluation can be eliminated by re-ranking the data within experimental variants - i.e. changing the order to random. However, the extent of the influence of these autocorrelations must primarily be limited by the accuracy of the experimental design.
- Additivity as an assumption concerning more complex experimental setups. The exact testing of the additivity of multiple experimental factors is a procedure quite demanding for an experimental design balanced in terms of the number of repetitions. It is also difficult to test the interaction on non-standard data, as any transformation may change the nature of the deviations of the original data from the ANOVA model being evaluated.

Limitations of ANOVA can be addressed

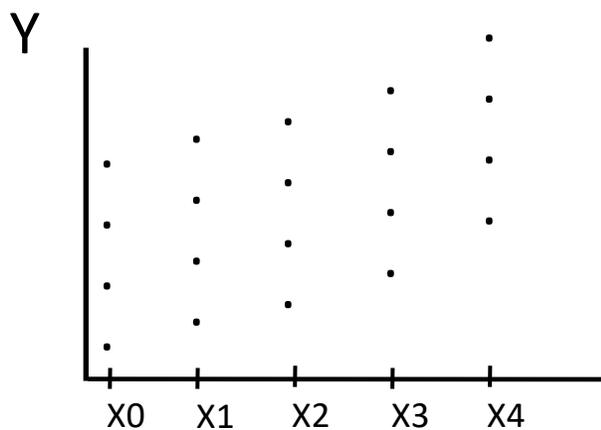
- **Missing data.** A serious problem is missing data on the whole group of combinations of test substances, for example in factorial experiments, where it is impossible to evaluate the experiment as a whole.
- **Different numbers of repetitions.** This is a typical phenomenon for experimental datasets. With different numbers of repetitions in the experimental variants, ANOVA tests are more sensitive to data non-normality. If the numbers of repetitions are completely different (except for order of magnitude differences), non-parametric techniques or analysis of variance of unbalanced trials should be used.
- **Outliers.** Isolated outliers must be excluded before parametric analysis of variance.
- **Lack of independence between model residences.** This is a serious deficiency, biasing the result of the F-test. Very often this is the result of poor execution or planning of the experiment.
- **Inhomogeneity of dispersion.** A very common deficiency in experimental data, often associated with non-normality of distribution or outliers.
- **Data abnormality.** Also in this case, the situation can be corrected by excluding outliers or by a normalizing transformation.
- **Non-additivity of the combined effect of multiple experimental interventions.** This situation can be tested either by special additivity tests or directly by the F test controlling for the significance of the effect of the interaction of the experimental interventions. When the interaction is significant, it is necessary to examine first of all its nature in an appropriate experimental setup.

Analysis of variance models

Model I. Fixed model

X_0	X_1	X_2	X_3	X_4
.
.
.
.
.
.
.
.

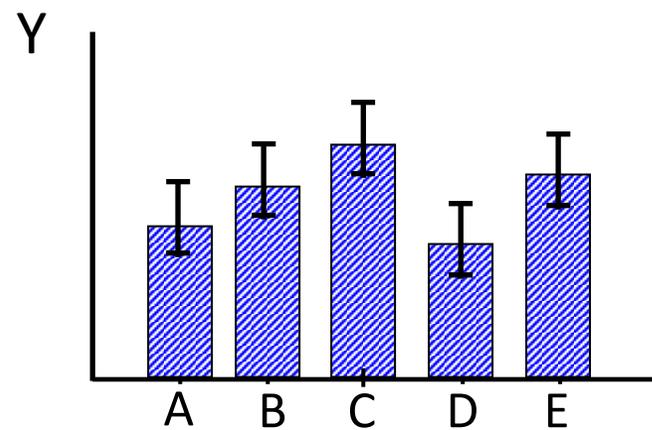
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$



Model II. Random model

A	B	C	D	E
.
.
.
.
.
.
.
.

$$y_{ij} = \mu + A_i + \varepsilon_{ij}$$



The principle of ANOVA

- The basic principle of ANOVA is to compare the variance attributable to:
 - The division of data into groups (so-called effect, variance between groups)
 - Variation of objects within groups (so-called error, variance within groups), it is assumed that it is a random variation (=error)

1. Variability between groups

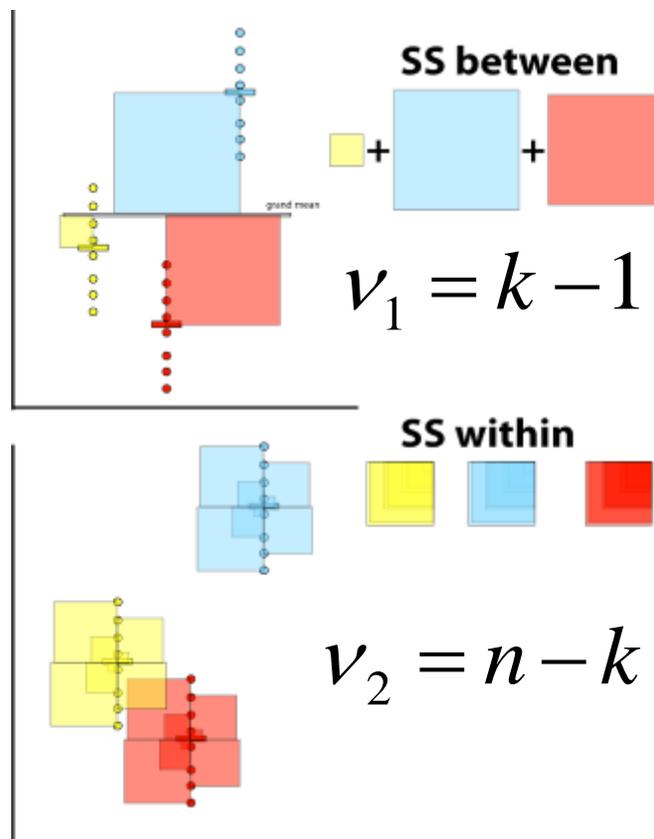
The variance is calculated for the overall mean (grand mean) and the averages in each data group

The degrees of freedom are derived from the number of groups (= number of groups -1)

2. Variability within groups

The variance is calculated for the averages of the groups and the objects within the groups, and the total variability is summed for all groups

The degrees of freedom are derived from the number of values (= number of values - number of groups)



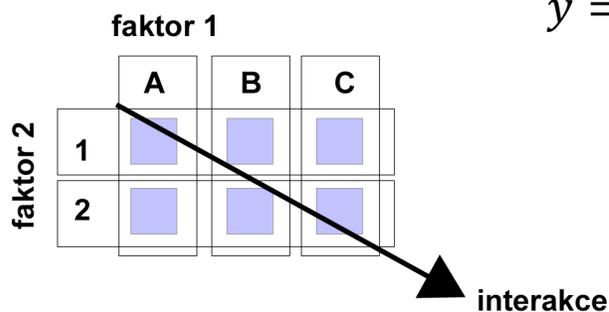
$$F = \frac{\text{between_groups}}{\text{within_groups}}$$

We compare the resulting ratio (F) with the F distribution tables for the v_1 and v_2 degrees of freedom

SS=sum of squares

Model design

- Model design means what variables and in what combinations will explain the variable being assessed
- In general, it is advisable, either expertly or as a result of preliminary analysis, to develop and test hypotheses about the interrelationships of variables and to build the final model according to these preliminary results
- Model design is closely related to concepts:
 - Analysis of main effects of variables only
 - Analysis of interactions between variables and complexity of interactions
- The model design can be expressed graphically or in an equation or matrix notation

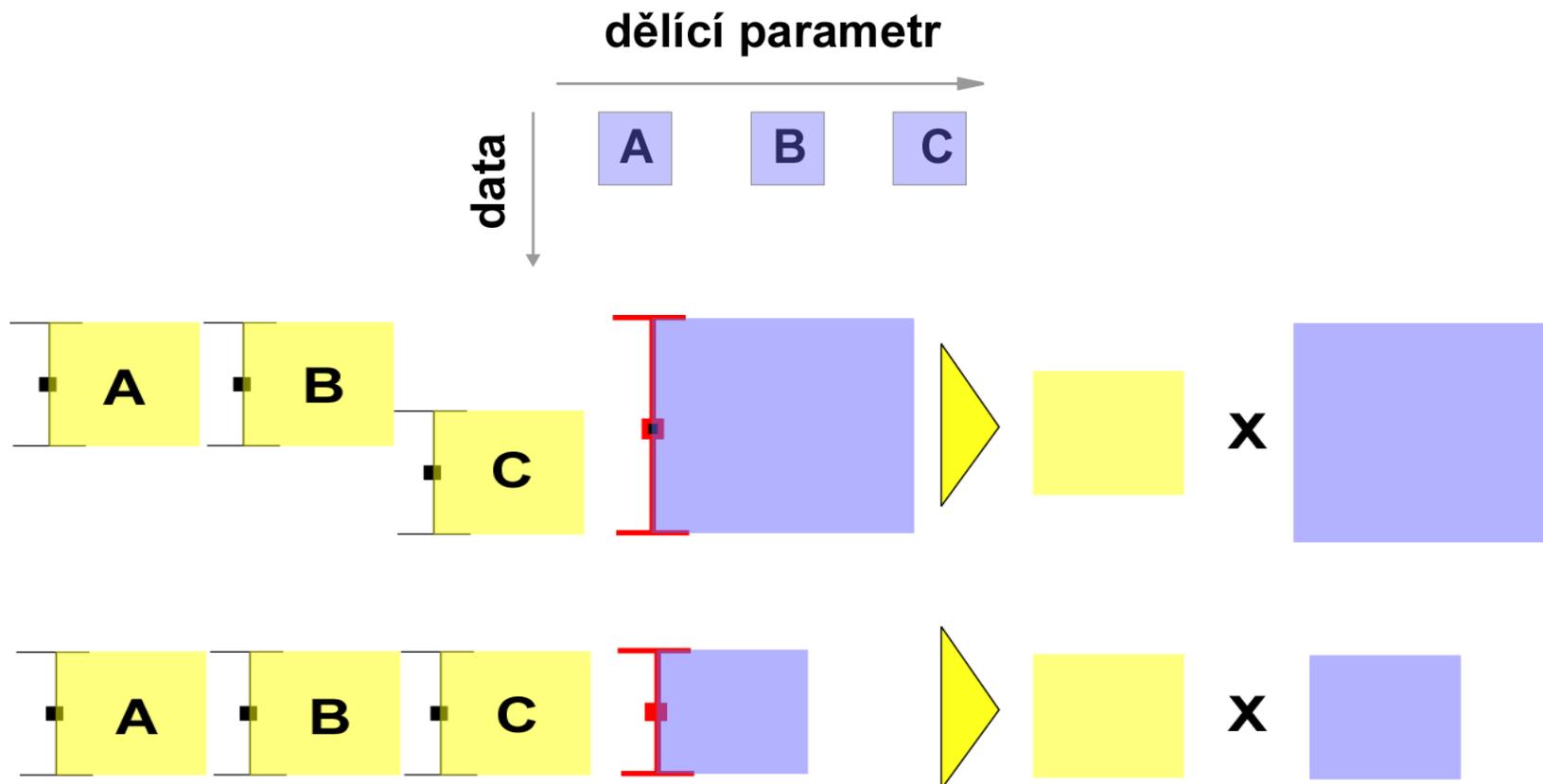


$$y = hmotnost * 1.5 + věk * 3.6 + hmotnost * věk * 1.8 + 9$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Simple ANOVA design

- The simplest case of ANOVA design is to divide into groups according to one parameter



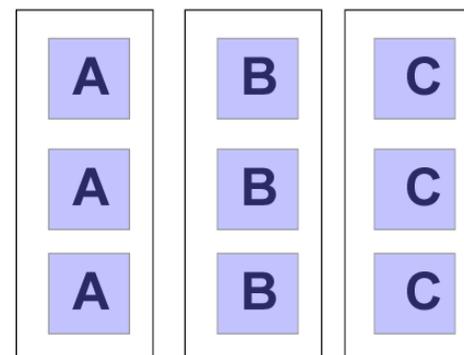
Nested ANOVA

- Splitting groups into random subgroups (e.g. repeating an experiment)
- The aim is to see if the data in one group is not a mere coincidence
- First, the agreement of subgroups in the main groups is tested,
 - if they are identical, everything is fine
 - if they are not, it is still possible to see if the variability within the main groups differs from the overall variability

jednoduchá ANOVA

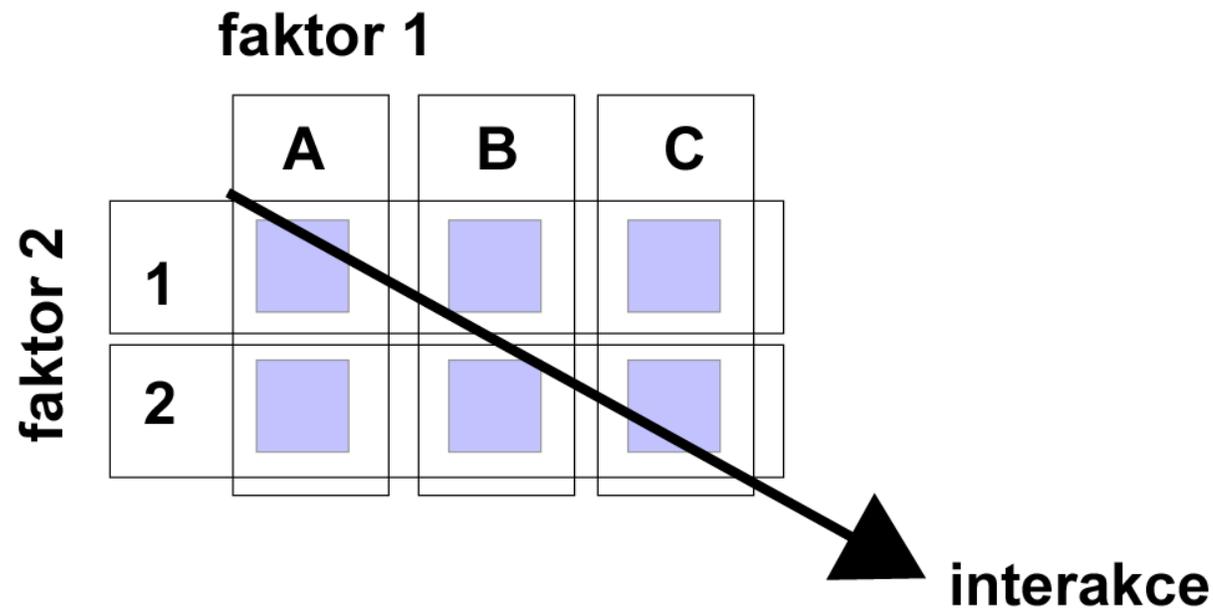


nested ANOVA



Two way ANOVA

- There are more parameters for categorization
- Unlike nested ANOVA, these are not random repetitions of the experiment, but controlled interventions (e.g. the effect of pH and O2 concentration)
- In addition to the influence of the main factors, there is also their interaction

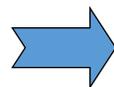


ANOVA - basic output

- The basic output of the analysis of variance is the ANOVA table - fractionation of variance components

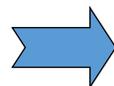
Source of dispersion	St. v.	SS	MS	F
Pok. intervention (between groups)	$a - 1$	SS_B	$SS_B / (a - 1)$	MS / MS_{BE}
Inside the groups	$N - a$	SS_E	$SS_E / (N - a)$	
Total	$N - 1$	SS_T		

Total



Quantified proportion of the difference between experimental interventions in the total variance

SS / SS_{BT}

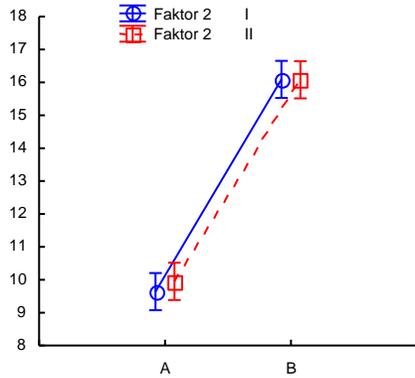


Statistical significance of the difference

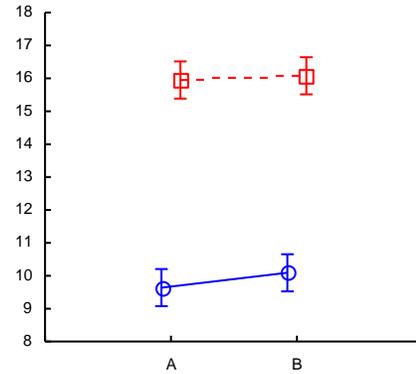
MS / MS_{BT}

Institut biostatistiky a analýz, PŘF a LF MU

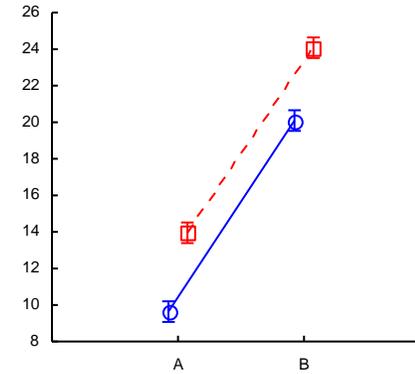
Main effects and interactions



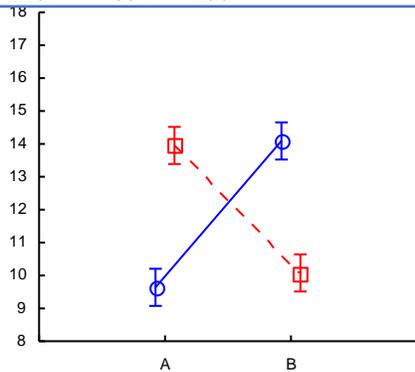
	SS	D.f.	MS	F	p
The Intercept	33487	1	33487	8165.3	0.000
Factor 1	1978	1	1978	482.2	0.000
Factor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



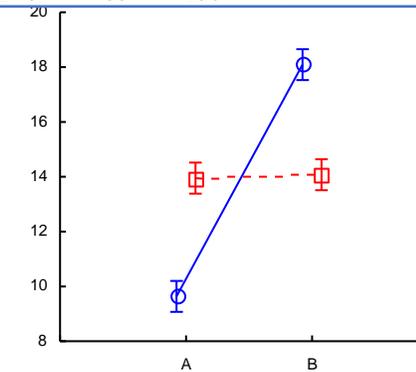
	SS	D.f.	MS	F	p
The Intercept	33487	1	33487	8165.3	0.000
Factor 1	4	1	4	1.0	0.314
Factor 2	1891	1	1891	461.1	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



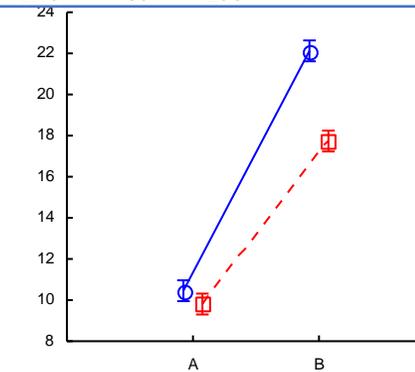
	SS	D.f.	MS	F	p
The Intercept	57391	1	57391	13993	0.000
Factor 1	5293	1	5293	1290.7	0.000
Factor 2	861	1	861	209.9	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	SS	D.f.	MS	F	p
The Intercept	28511	1	28511	6952.0	0.000
Factor 1	4	1	4	1.0	0.314
Factor 2	1	1	1	0.3	0.602
F1*F2	867	1	867	211.3	0.000
Error	804	196	4		

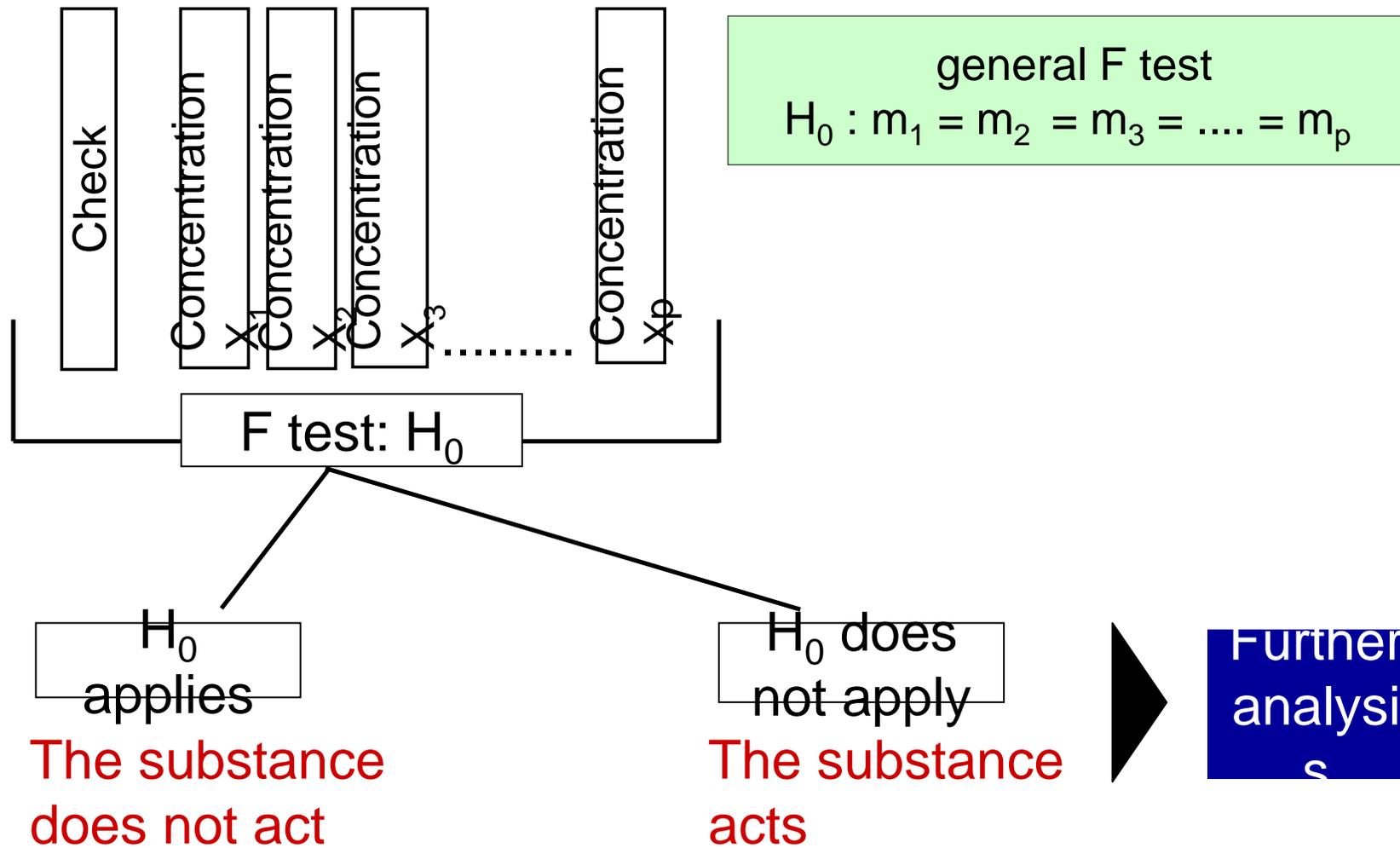


	SS	D.f.	MS	F	p
The Intercept	38863	1	38863	9476.2	0.000
Factor 1	920	1	920	224.3	0.000
Factor 2	1	1	1	0.3	0.602
F1*F2	867	1	867	211.3	0.000
Error	804	196	4		



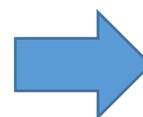
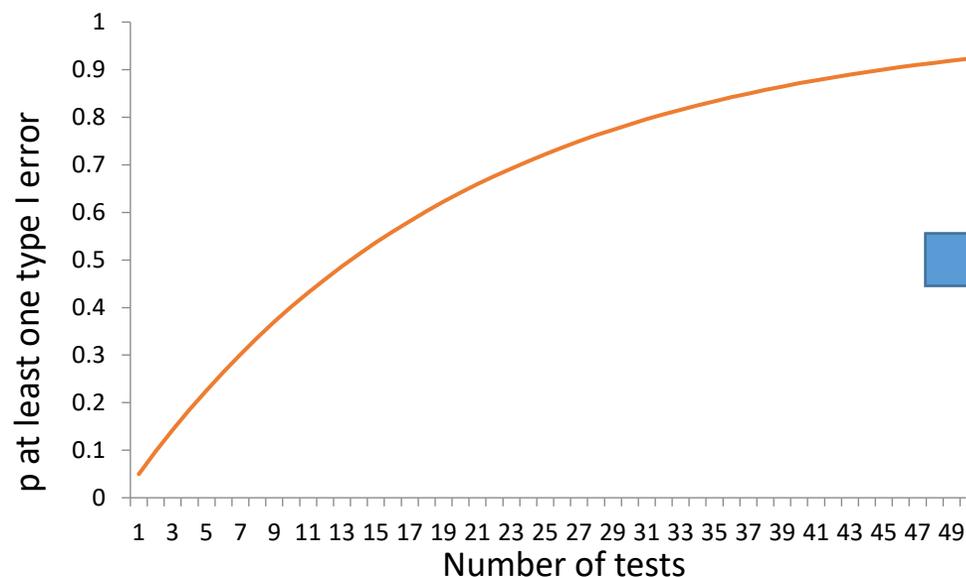
	SS	D.f.	MS	F	p
The Intercept	45203	1	45203	13596	0.000
Factor 1	4799	1	4799	1443.4	0.000
Factor 2	316	1	316	95.0	0.000
F1*F2	175	1	175	52.5	0.000
Error	652	196	3		

Analysis of variance - general F test



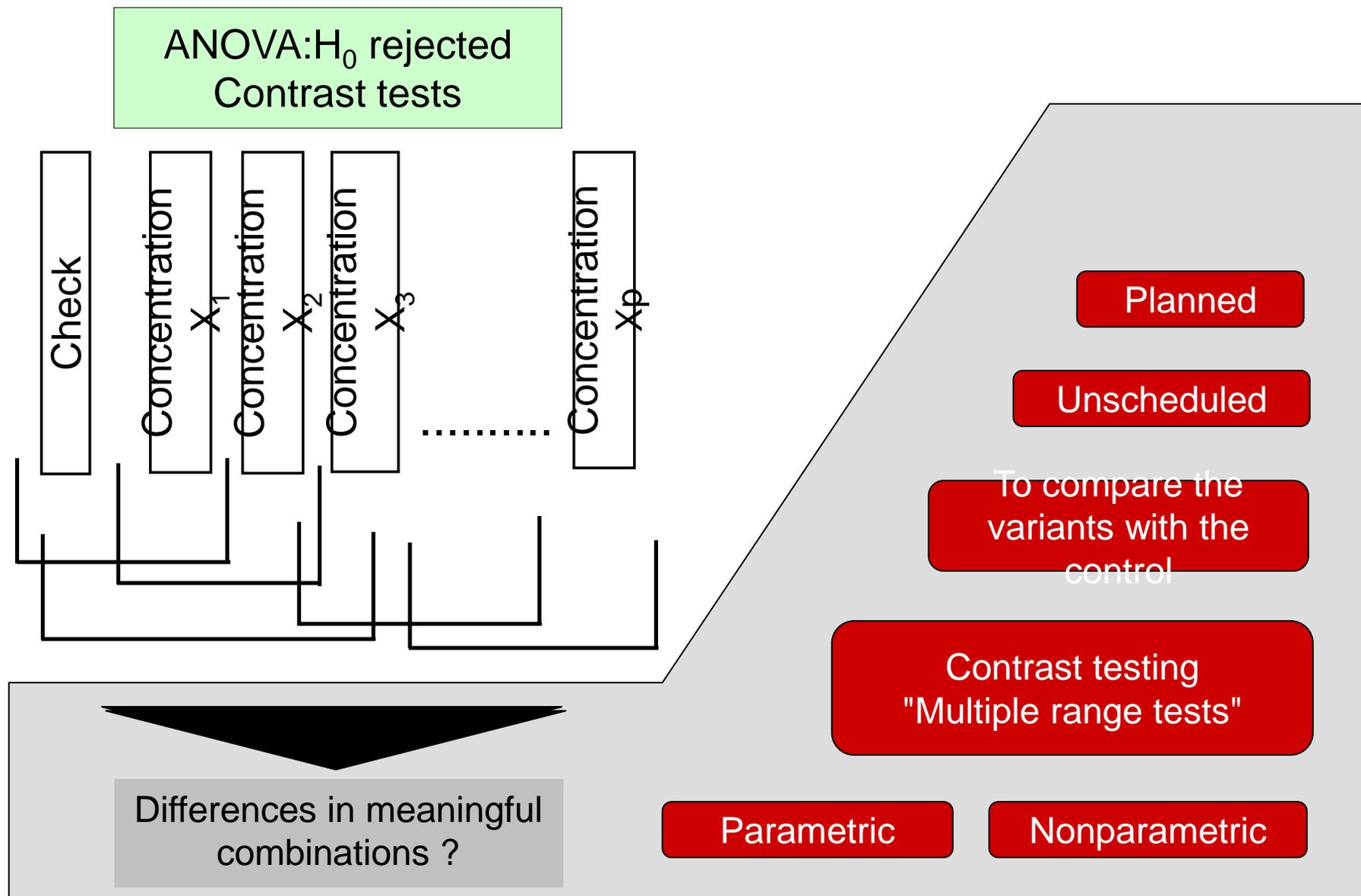
Testing partial hypotheses

- In many analyses, it is necessary to work with peer testing of multiple groups of objects in a peer-to-peer style
- The general analysis procedure is
 - Testing overall significance - all groups among groups
 - If overall significance is found, testing continues by analysing specific combinations of pairs of groups (ENG: between)
- The problem is the effect of multiple testing on the statistical significance of the tests:
 - Every single test has $\alpha=0.05$ (type I error)
 - With multiple testing, the probability that at least one test will erroneously reject the null hypothesis (i.e., a Type I error) increases

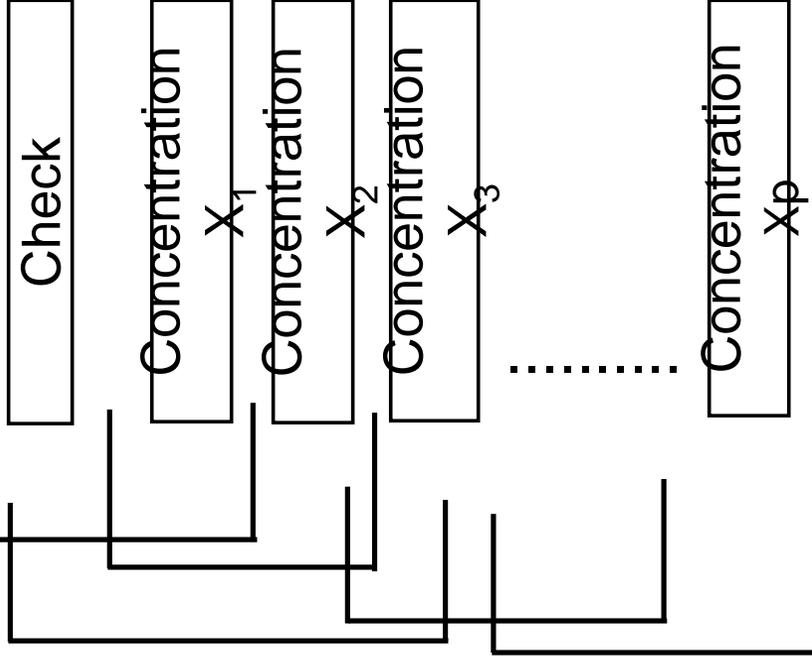


The solution is various p-value correcting procedures (e.g. Bonferroni correction, FWR, FDR procedures, etc.)

Analysis of variance - contrast tests



ANOVA: H_0 rejected
Contrast tests



Planned

Unscheduled

To compare the variants with the control

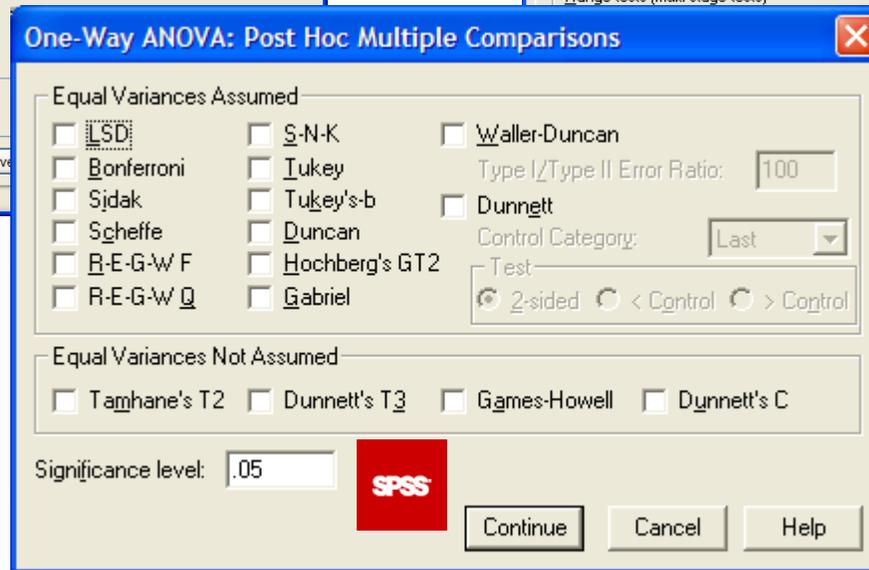
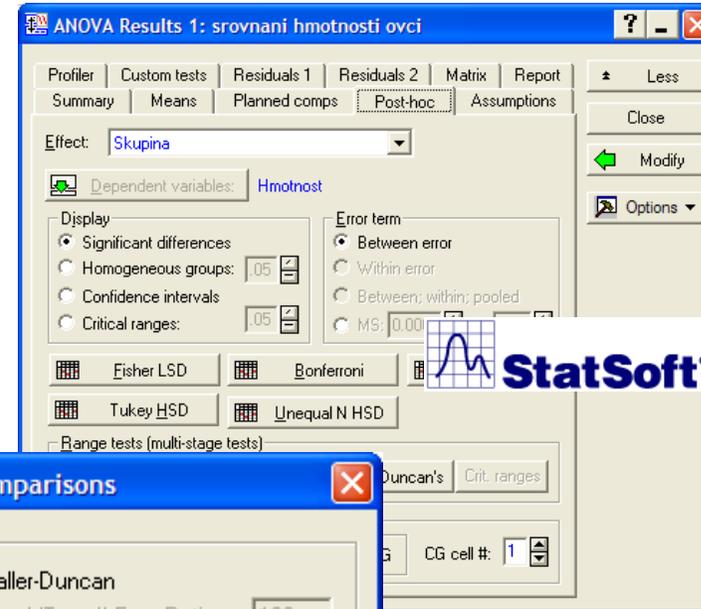
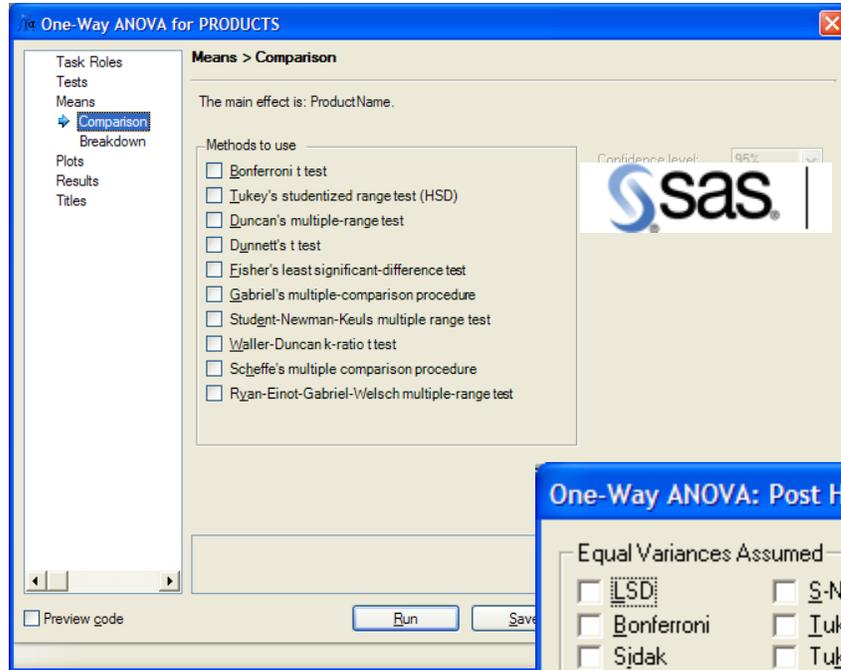
Contrast testing "Multiple range tests"

Parametric

Nonparametric

Differences in meaningful combinations ?

A number of different post-hoc tests



Anova - One way

Plant stimulator dose (0, 4, 8, 12 mg/l)

A = 4 ; n = 8

I. ANOVA

Bartlett's test: P = 0.9847

K-S test: P = 0.482 - 0.6525 for each category

Source	D.f.	SS	MS	F	p
Between	3	305.8	101.9	8.56	<0.001
Within	28	322.2	11.9		
Total	31	638			

II. Multiple Range Test (NKS -test)

Level	Average	Homogeneous groups		
0	34.8	x		
4	41.4		x	
12	41.8		x	
8	52.6			x