# Statistical methods in biology and medicine

# Statistics

- Group of mathematical methods concerning the collection, analysis and interpretation of the data
- A complete description of the world is both impossible and impractical (statistics represent a tool for reducing the variability of the data)
- Statistics creates mathematical models of the reality that can be helpful in making decisions
- It works correctly only when the assumptions of its methods are met

# Descriptive statistics

- Population-wide – works with the data related to whole surveyed population (e.g. census, medical registry)
- Inductive – conclusions based on sample data (obtained from a part of the target population) are extrapolated to whole population (assumption: random selection of the sample)

# Statistics as a data processing tool

- „raw data“ – often difficult to grasp
- Descriptive statistics can make the data (of given sample) understandable

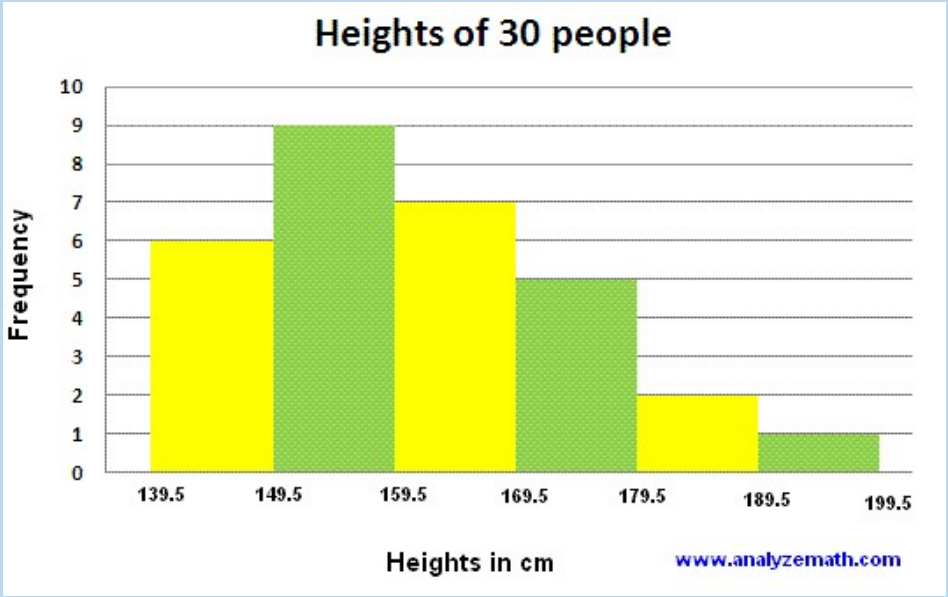
| kod | cislo | adrenalin | noradrenalin | hypokineza | ERα 397/PvuII | ERα 351/XbaI |
|---|---|---|---|---|---|---|
| TTCBI13-2013 | 1 | 354 | 3643 | baze | CT | AG |
| TTCKE14-2013 | 2 | 307 | 2955 | apex | TT | AA |
| TTCKH15-2013 | 3 | 473 | 6076 | apex | CT | AG |
| TTCAJ16-2013 | 4 | 341 | 2108 | apex | CT | AG |
| TTCCHM17-2013 | 5 | 321 | 2031 | apex | CC | GG |
| TTCCHS18-2013 | 6 | 426 | 1931 | apex | TT | AA |
| TTCRK19-2013 | 7 | 508 | 1753 | difuzni | TT | AA |
| TTCPD20-2013 | 8 | 374 | 1088 | difuzni | CT | AA |
| TTCMJ21-2013 | 9 | 597 | 1798 | apex | CC | GG |
| TTCPO22-2013 | 10 | 420 | 2856 | apex | CT | AG |
| TTVVA23-2013 | 11 | 367 | 2657 | apex | CT | AA |
| TTCNL24-2013 | 12 | 327 | 2467 | apex | CT | AG |
| TTCJF25-2013 | 13 | 395 | 3929 | apex | CC | GG |
| TTCZM26-2013 | 14 | 344 | 3706 | apex | CT | AG |
| TTCHJ27-2013 | 15 | 426 | 4225 | apex | TT | AA |
| TTCGT28-2013 | 16 | 265 | 2406 | apex | CT | AG |
| TTCSB29-2013 | 17 | 295 | 3186 | apex | CT | AG |

# Kinds of data

- Continuous (always quantitative) – the parameter can theoretically be of any value in a given interval (e.g. mean arterial pressure: 0-∞; ejection fraction: 0-100%)
  - Ratio data – both difference and ratio of two values can be computed (e.g.body weight)
  - Interval data – only differences, but not ratios of two values can be determined (e.g. IQ score)
- Categorical (usually qualitative) – the parameter can only be of some specified values (e.g. blood group: 0, A, B, AB; sex: male, female; a disease is present/absent)
  - Ordinal data – are categorical, but quantitative (they can be ordered – e.g. heart failure classification NYHA I-IV)
  - Count data – can be ordered and form a linearly increasing row (e.g. number of children in a family: 0,1,2…) - they are often treated as continuous data
  - Binary data – only two possibilities (patients / healthy controls)

# The distribution of continuous data - histograms

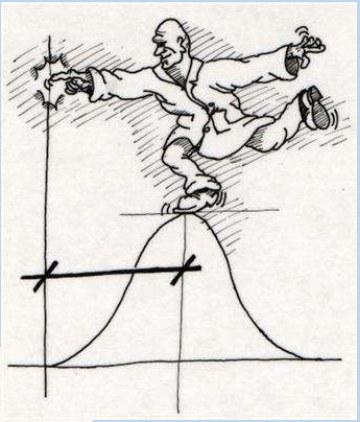
- The distribution of a continuous parameter can be visualized graphically (e.g. using histograms)
- The values usually cluster around some numbers

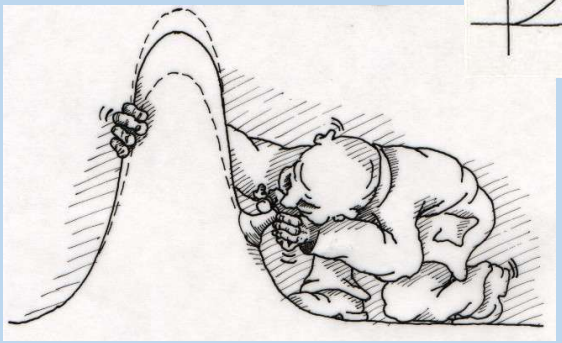


# Description of continuous data

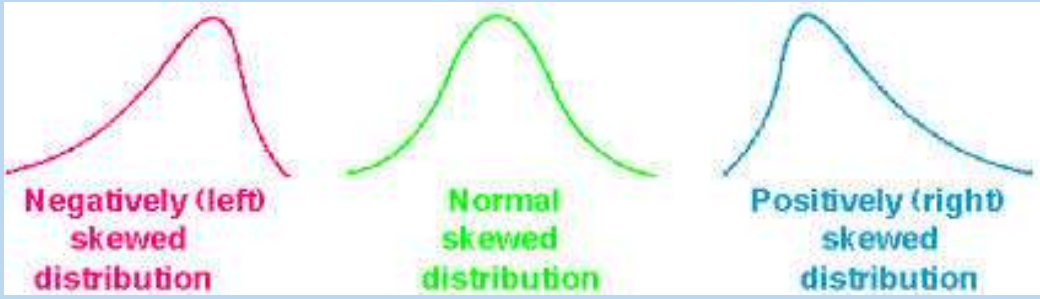
- **Measures of central tendency**
  - The arithmetic mean ($\mu$)
    - sum of values divided by their number (n)
  - The median (= 50% quantile)
    - cuts the order of values in half
  - The mode
    - most frequent value
- **Measures of variability**
  - variance ($\sigma^2$)
  - standard deviation (SD, $\sigma$)
  - coefficient of variance (CV)
    - CV = $\sigma/\mu$
  - standard error of mean (SE, SEM = $\sigma/\sqrt{n}$)
  - min-max (= range)
  - quartiles
    - upper 25%
    - median
    - lower 75%
  - skewness
  - kurtosis

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

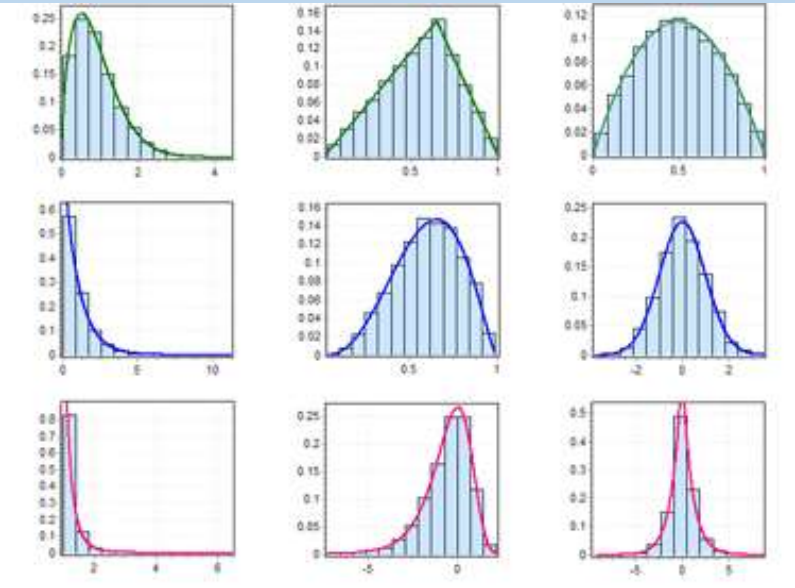


# The probability distribution of continuous random variable




- Probability density function
- In graphs each (continuously) quantifiable variable (x axis) is linked to its probability (y axis)

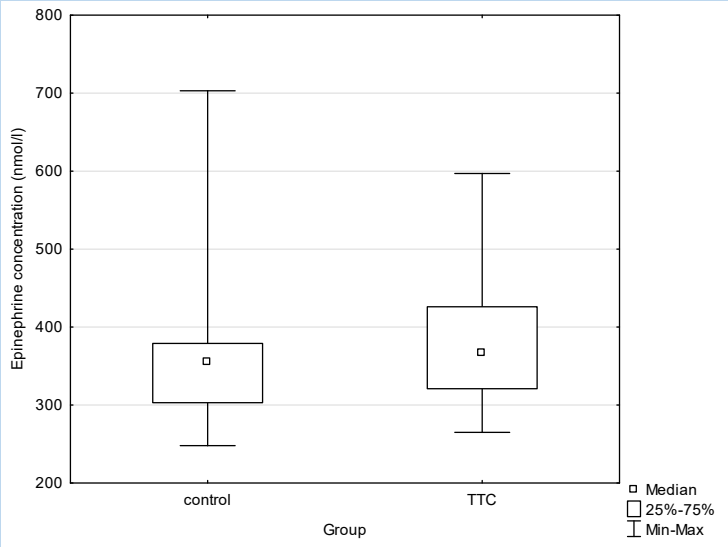
# Examples of continuous data distribution



Histograms + corresponding probability density functions

# Other ways of graphical visualisation

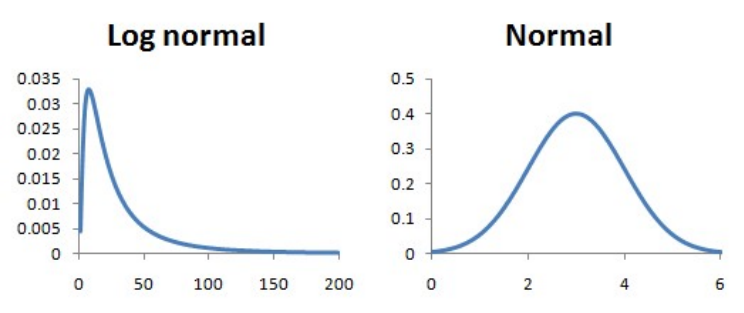
- Box and whisker plots




- Instead of median e.g. mean can be used, instead of quartiles („box“) ±σ, instead of range („whiskers“) e.g. non-outlying values… etc.

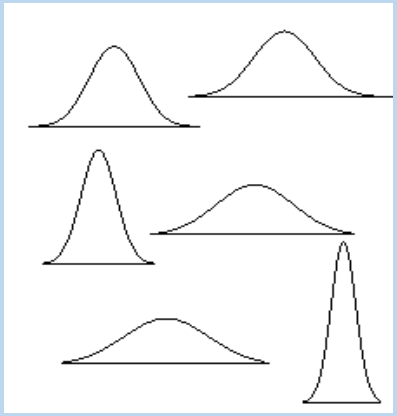
# Normal distribution of the data

- Defined by the Gaussian function $y = a^{-(x-b)^2/2c^2}+d$, where a, b, c, d are real numbers
- Graphical representation is the **Gaussian „bell“ curve**
- mean = median = mode
- A random variable **x** is normally distributed when its value can be interpreted as the sum of an infinite number of independent effects with equal absolute value
- E.g.: throwing a coin, we assign the value of +1 to a head and -1 to a tail. When throwing many times ($n \rightarrow \infty$), the probability distribution of the resulting value will be normal

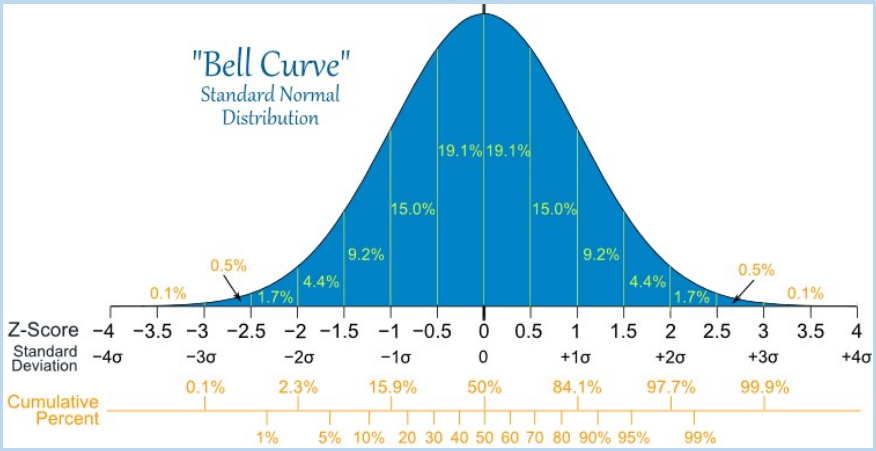



- Log-normal distribution: after logarithmic transformation of the data, we obtain the Gaussian curve (with **the geometric mean** at its peak) – an example of data transformation

# Normal (Gaussian) vs. symmetric distribution



- Not each symmetric distribution is normal
  - Meeting of several assumptions is necessary
    - interval frequency distribution
    - distribution function
    - skewness = 0, kurtosis = 0
  - Data transformation („normalization“)
    - Creating normal distribution by applying a formula
- Student distribution is an approximation of normal distribution in small datasets

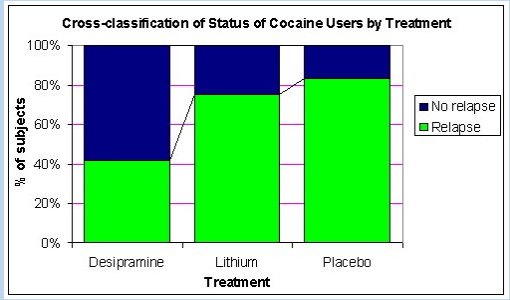


# Description of categorical data

- Sumarization of given categories of the dataset (frequency table)
- When more than one categorical parameter is available, we can create **contingency tables** (and, based on them, we can eventually create graphs)

| | | Body Image | | | |
|---|---|---|---|---|---|
| | | About Right | Overweight | Underweight | **Total** |
| Gender | Female | 560 | 163 | 37 | 760 |
| | Male | 295 | 72 | 73 | 440 |
| | **Total** | 855 | 235 | 110 | 1200 |

| | Right-handed | Left-handed | Total |
|---|---|---|---|
| **Males** | 43 | 9 | 52 |
| **Females** | 44 | 4 | 48 |
| **Totals** | 87 | 13 | 100 |

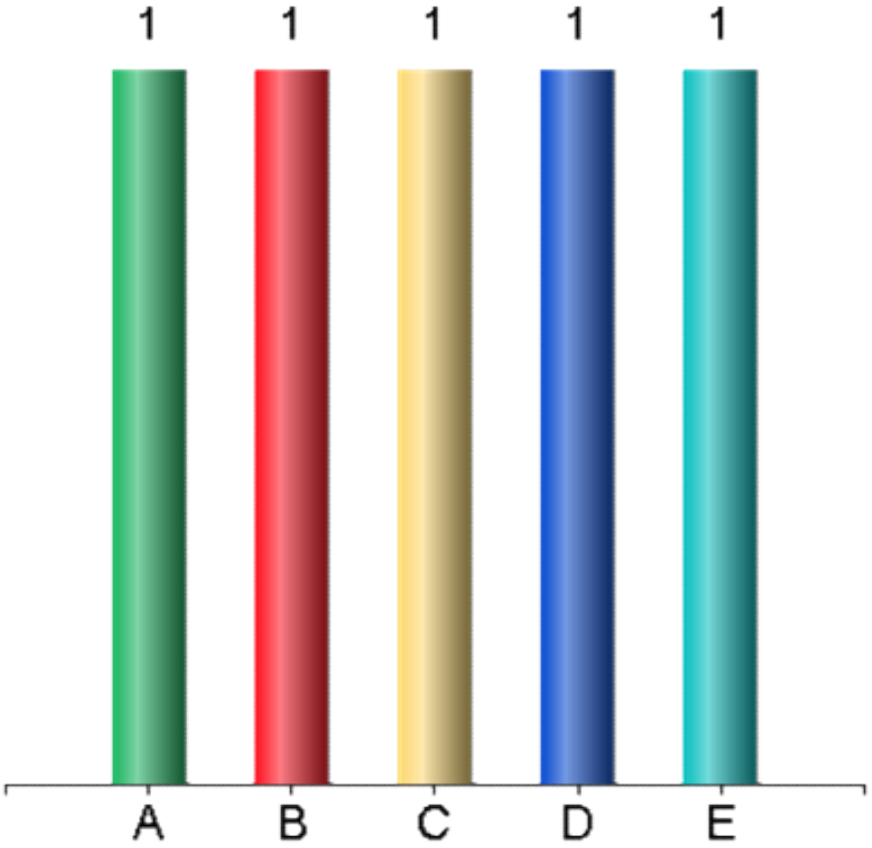


# Expression of categorical data variability - examples

- Variation ratio
  - $v = 1 – (f_m/N)$, where $f_m$ = number of cases in mode (most frequent) category, N = total number of cases
  - Ratio of all cases **except** the mode category to the total number of cases
- Shannon-Wiener diversity index
  - Expresses the uncertainty in prediction, to which category will a case belong
  - $H' = -\sum p_i * \ln(p_i)$, where $p_i$ is a proportion of category i from total sample
  - If $p_i = 100\%$ then $H' = 0$; Higher value corresponds with higher diversity
  - Widely used in ecology, common range between: 1,5 – 3,5

# The level of education (basic, high school, university) is an example of…

- ✔ A. Ordinal variable
- B. Interval variable
- C. Binary variable
- D. Continuous variable
- E. Qualitative variable



# Repetition - formulation of statistical hypotheses

- Research hypothesis (e.g. drug A has better effect than drug B, blood pressure decreases during the treatment, there is a correlation between sex and body height etc…) – can be formulated both for an experiment or for an observation
- Testing of research hypothesis uses a proof by contradiction
- For statistical hypothesis testing, a **null hypothesis $H_0$** must be defined (e.g. between two groups, there is no difference in means, there is no difference in variances, there is no correlation between two parameters, a parameter does not change in time…resp. any observed difference is only due to a chance)
- During the testing of null hypothesis, we try to refute it (or, more exactly, to show that it is highly improbable)
- If the null hypothesis, is not true, then its negation must be true – **alternative hypothesis $H_A$** (there is a difference, there is a correlation…)
- The result of hypothesis testing can thus be:
    - A) non-refutal of the null hypothesis (at certain level of statistical significance α)
    - B) refutal of the null hypothesis favouring the alternative hypothesis

# Repetitions - errors in hypothesis testing

| | Real nature of the null hypothesis | |
|---|---|---|
| Statistical decision | $H_0$ true | $H_0$ false |
| $H_0$ refuted | **type I error (α) = false pos.** | **Correctl y pos. (1-β)** |
| $H_0$ confirmed | **Correctly neg. (1-α)** | **type II error (β) = false neg.** |

- Type I error rate (α) – also **significance level**
- α must be defined before the statistical testing – 0.05 is usual in biomedicine (i.e. when $H_0$ is refuted, there is 95% certainty, that it is really false and the observed difference/correlation is real)
- 1-α = specificity of a statistical test
- 1-β – also **power of a test (sensitivity of a test)**
- P-value – probability that the observed result was obtained under the assumption that $H_0$ is true
- **When p < α, we refute the null hypothesis at a given significance level and the alternative hypothesis is valid**
- We say that the difference (effect) is **statistically significant** (that, of course, does not mean that it has to be significant practically; this depends on the power, too)

# P-value

- “A p-value doesn‘t “prove“  anything. It's simply a way to use surprise as a basis for making a reasonable decision.“

Cassie Kozyrkov

# Statistical tests

- For different statistical hypotheses, different tests are used
- The selection of the right test depends on:
    - the number of compared groups
    - the character of the data (categorical vs. continuous)
    - the distribution of the data
    - mutual dependence of the data

# Power of a test...

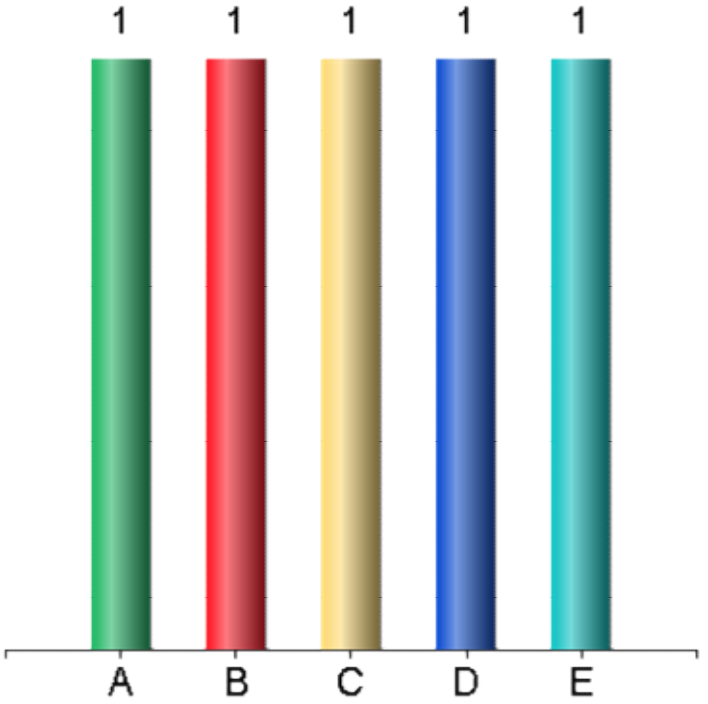
A. Expresses its practical (not statistical) significance

B. Increases with increasing variability of the data

✔ C. Expresses the ability of a test to correctly refute the null hypothesis

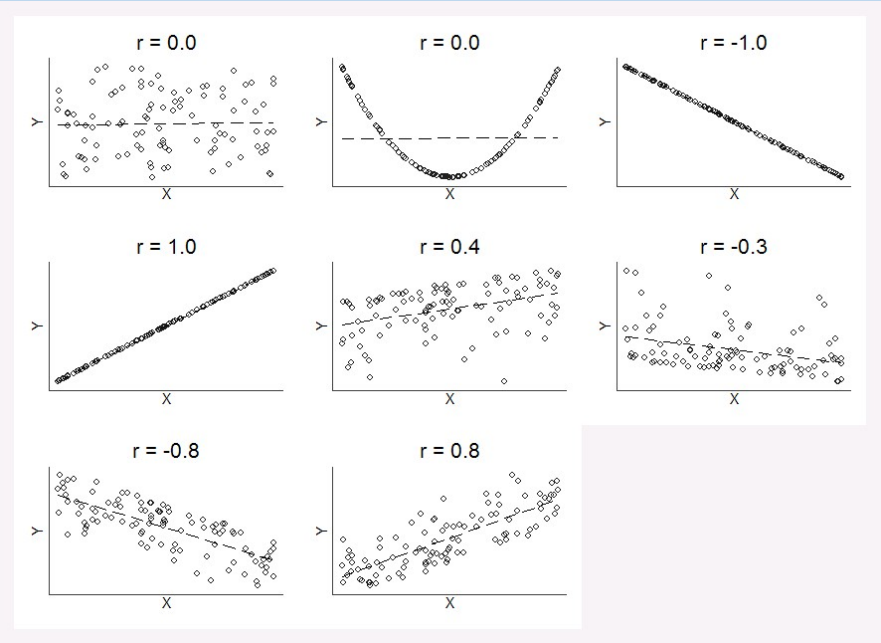
D. Is expressed by a letter p

E. Is a probability that the alternative hypothesis is true in a case when the null hypothesis is refuted



# Correlation of two continuous parameters

- Mutual dependence of two parameters - correlation
- Expressed by correlation coefficient (r)
- r generally express the size of the effect
- r can achieve values in the interval from -1 to 1, where 0 corresponds to no correlation, 1 corresponds to 100% positive correlation (when one factor increases, the other does the same) and -1 corresponds to total negative correlation



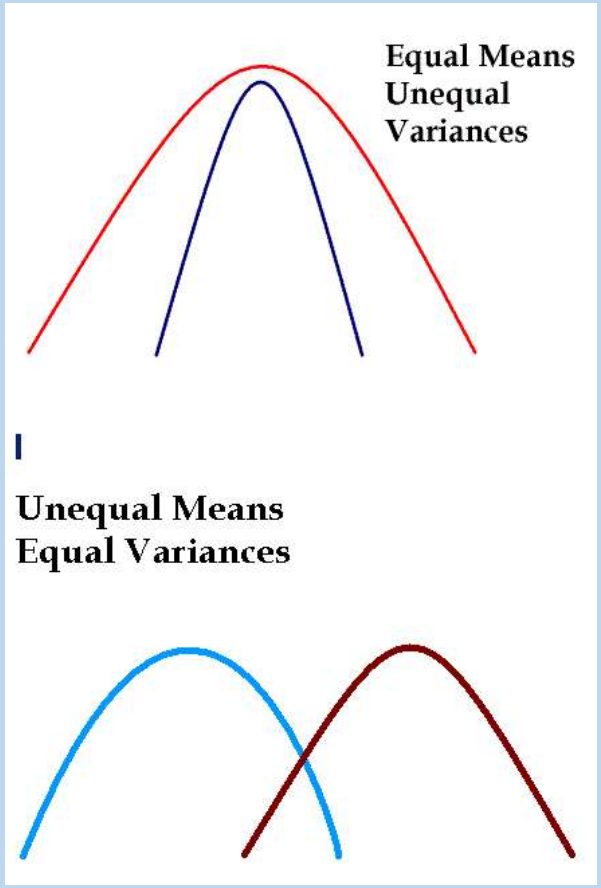

- Besides r, the corresponding p-value can be determined ($H_0$ – the variables are independent)
- Correlation of a categorical vs. continuous variable – see the „tests for continuous data“ (categorical variable defines the groups that are compared by the tests)

# Examples of correlation coefficients

- Pearson coefficient (parametric) – measures linear correlation between variables
  - The main assumption is approximately normal distribution of the data
- Spearman coefficient (non-parametric) – measures the rank correlation of the variables
- None of the coefficients can reveal e.g. U-shaped dependence

# Comparing the continuous variable in two and more samples

- $H_0$ – there is no difference in the value of the variable between the samples (or is due to chance – e.g. the concentration of glycated hemoglobin in treated and untreated diabetics is equal)
- Generally, central tendency (more often; see further) or variability (e.g. F-test, Levene test) can be tested



# Parametric vs. non-parametric tests for continuous data

**Parametric**

- Use the values
- Have higher power, but only when their assumptions are met (esp. normal distribution of the data in each sample)
- If the distribution is not normal, we can try to transform (normalize) them

**Non-parametric**

- Use ranks of values
- Power is generally lower (but the difference is small in big samples)
- They are more „robust“ – their use is not that dependent on data distribution
- They also work well with count data

The normality can be tested by normality tests (e.g. Kolmogorov-Smirnov, Shapiro-Wilks – they compare the real distribution with the normal distribution) and „by eye“ evaluation of whether the histograms correspond to Gaussian curve (in small samples, the normal probability plot is a better choice)

# Tests for continuous data - paired vs. unpaired tests

**Paired (matched samples)**

- Used when to each value from sample A, we can match one value from sample B that differs only by its membership in the sample (e.g. comparing salaries in two hospitals: director A – director B; head physician A – head physician B… up to charwoman A – charwoman B)
- **Most often, this design is used to assess the change in time** (e.g. patients‘ weight now vs. after 5 years: patient XY – and other patients – is the same person now as well as after 5 years and differs only by the time difference)
- They assess differences between the samples (or their ranks)

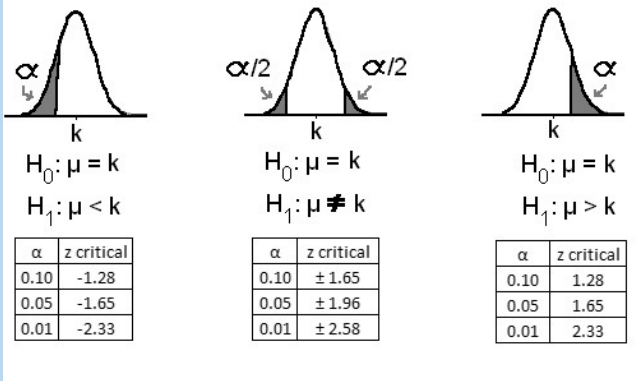
**Unpaired (unmatched samples)**

- Used in independent samples (they can differ in size)
- They compare the actual values of the variable between the samples (or their ranks)
- It is necessary to decide between the paired or unpaired design before the start of the study (pairing is technically challenging, but paired tests have higher power)

# One-tailed vs. two-tailed tests

**One-tailed**

- $H_0$ is asymmetric: e.g. drug A is not better than drug B – but we are not interested whether it is or is not worse
- They have higher power




| α | z critical |
|---|---|
| 0.10 | -1.28 |
| 0.05 | -1.65 |
| 0.01 | -2.33 |

| α | z critical |
|---|---|
| 0.10 | ± 1.65 |
| 0.05 | ± 1.96 |
| 0.01 | ± 2.58 |

| α | z critical |
|---|---|
| 0.10 | 1.28 |
| 0.05 | 1.65 |
| 0.01 | 2.33 |

**Two-tailed**

- $H_0$ is symmetric: there is no difference between drug A and drug B (i.e. A is neither better nor worse than B)
- They can reveal the differences in both ways
- They are usually more suitable – we don‘t know the result a priori, and we are interested in both possible effects

# Tests for continuous data, 2 samples – examples

| Test | Parametric | Non-parametric |
|---|---|---|
| Paired | Paired (dependent)<br>Student‘s t-test | Wilcoxon paired test<br>Sign test |
| Unpaired | Unpaired (independent)<br>Student‘s t-test | Mann-Whitney U-test *<br>Kolmogorov-Smirnov test |

- * has almost the same power as unpaired t-test, but it has an assumption of similar variability in both samples (as well as t-test)

# Tests for continuous data, more than 2 samples – examples

| Test | Parametric | Non-parametric |
|---|---|---|
| Paired | Repeated measures ANOVA (Analysis Of VAriance) – RMANOVA | Friedman test („ANOVA“) |
| Unpaired | One-way ANOVA (and its variants) | Kruskal-Wallis test („ANOVA“) |

- When ANOVA rejects $H_0$, it is necessary to find out which specific samples differ from each other – post hoc tests

# Choose the best test

In a clinical trial, patients take either a new drug to treat epilepsy or a placebo. The study is randomized (the study group is randomly drawn). Only patients, which have at least one and at most ten seizures in three months are included. The study evaluates a number of seizures during the first year of treatment

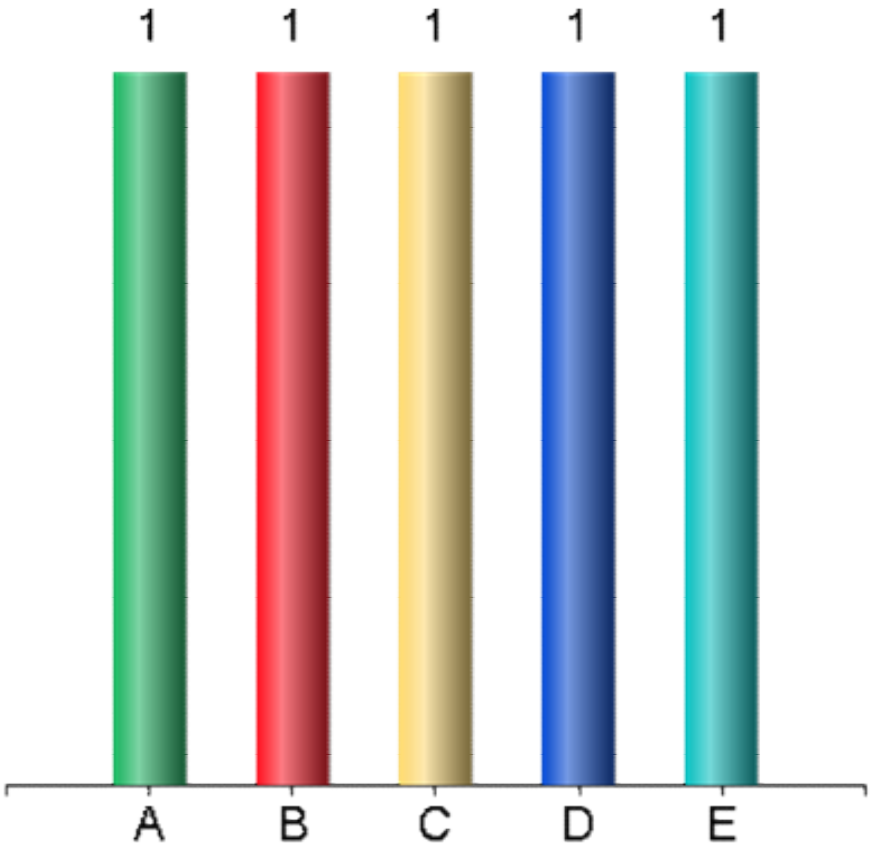
A. Paired t-test

B. Unpaired t-test

✔ C. Mann-Whitney U-test

D. Sign test

E. Repeated measures ANOVA



# ANOVA

- Analysis of variance
  - tests null hypothesis about more than two samples
  - requirements: Normal distribution, equal standard deviations
  - requires further analyses to find out which sample is different

# Nonparametric „ANOVA“

- Kruskal-Wallis test (unpaired)
- Friedman test (paired)

# Multiple comparisons problem

- When we perform more tests at once, the probability that some of them will give a statistically significant result only due to chance (i.e. type I error – $H_0$ is wrongly refuted) – increases (e.g. during post hoc tests following ANOVA
- For example, when performing 10 tests at $\alpha = 0.05$, the probability that **none** of them will give a significant result (given that $H_0$ is true in all of them) equals $(1-\alpha)^{10} = 60\%$, i.e. in 40% the $H_0$ is wrongly rejected.
- That is why the multiple comparisons corrections are applied (Bonferroni, Benjamini-Hochberg…) to further decrease α (and thus make the criteria for refuting $H_0$ stricter).
- Bonferroni correction: initial α is divided by the number of tests (or, alternatively, all p-values are multiplied by the number of tests with α left unchanged)
  - very “conservative“.

# Post hoc tests in ANOVA

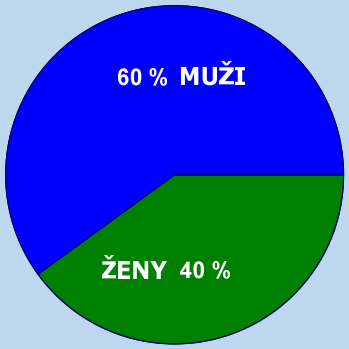
- Each group with each other (“football matches“)
  - Bonferroni correction $\alpha / [n (n - 1) / 2]$
  - Tukey, Scheffé (ANOVA)
  - Dunn (Kruskal-Wallis)
  - Neményi (Friedman)

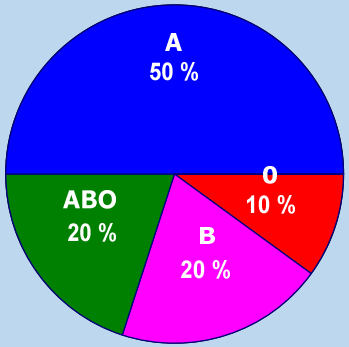
- Each group with control group
  - Bonferroni correction $\alpha / (n - 1)$
  - A priori, we are not interested in comparing the other groups between themselves
  - Dunnett (ANOVA)
  - Dunnett rank sum (nonparametric tests)

# “Manual“ multiple testing correction

- Useful in situations, where there are no standardized post hoc test as a part of statistical software
  - e.g. genetic tests – parameter in many candidate polymorphisms, comparing categorical data in more groups
- Bonferroni: α is divided by a number of tests (k)
- Bonferroni-Holm: each test has a different α-value. In a test with the lowest p-value, α(corr) equals α/k, in the second one α/(k-1), in the third one α/(k-2) … until the last one where it equals α
- Benjamini-Hochberg (FDR): each test has a different α-value. In a test with the lowest p-value, α(corr) equals α/k, in the second one α/(k/2), in the third one α/(k/3) … until the last one where it equals α
- When we find p > α(corr), results of following tests are not statisticaly significant either
- Alternatively, we may leave the α unchanged and create p(corr)-values by multiplying the p-values by denominators (dividing α in the examples above)

# Tests for categorical data




- From the contingency table, its probability under the assumption that $H_0$ is valid (i.e. the p-value) can be determined, as well as the effect size – e.g. the association between a mutation and a disease (expressed as RR – relative risk; OR – odds ratio)
- Sometimes, a reduction of larger tables into 2x2 table is advantageous [this is especially suitable in ordinal data – e.g. heart failure staging NYHA I-IV can be transformed into binary data as mild failure (NYHA I+II) and severe failure (NYHA III+IV)]

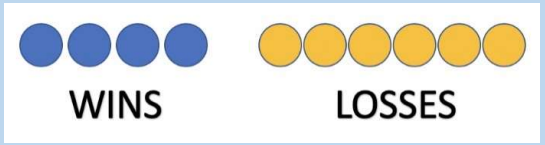
| | nemoc | zdraví |
|---|---|---|
| mutace | 50 | 2 |
| ne | 4 | 48 |

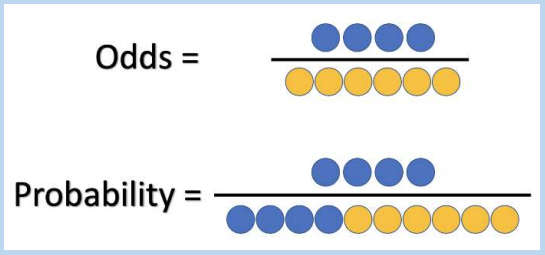
- Paired design can be used (typically presence/absence of the disease in time)

**Before & After**

| Before | After | |
|---|---|---|
| | Non-Smoker | Smoker |
| Non-Smoker | 20 | 5 |
| Smoker | 16 | 9 |

# Relative risk and odds ratio in 2x2 tables




| Exposure Status | Event Occurred | |
|---|---|---|
| | Yes | No |
| Exposed | a | b |
| Not Exposed | c | d |

$$\text{Relative Risk} = \frac{a/(a+b)}{c/(c+d)}$$

$$\text{Odds Ratio} = \frac{a/b}{c/d} = \frac{ad}{cb}$$

www.statpearls.com

- probability vs. odds
- RR is suitable for prospective studies, while the design is not important in OR
- If the dependent (modelled) variable is the same (e.g. event in the table), values of RR (a/(a+b)) and OR (a/b) are similar in when the occurrence is low
- RR is more intuitive, OR is more universal, commonly used in e.g. logistic regression
- It is always necessary to determine which variable will be independent and which one dependent

# Tests for categorical data - examples

| Test | 2 x 2 contingency table | More categories/measurements ‡ |
|---|---|---|
| Paired | McNemar test | Cochran Q test (Binary data, more measurements)<br>Sign test (ordinal data, two measurements |
| Unpaired | Chi-square ($\chi^2$) test*<br>Fisher exact test | Chi-square ($\chi^2$) test*<br>Cochran-Armitage test (table 3x2, ordinal data) |

‡ when $H_0$ is rejected, a serie of tests for 2 x 2 tables with appropriate multiple testing correction must follow

* under the assumption of certain minimal counts in each cell of the table (cca $n \geq 5$)

# Example

The study aimed to investigate an association between the blood group in AB0 system (A, B, AB and 0) and the presence of acute complications of the blood transfusion. How many fields does the respective contingency table have?

| Ranking | Response | Votes |
|---|---|---|
| Correct Answer | | |
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| Others | | |

# Example

In the previous case, $\chi^2$ test yielded $p < 0.05$ and a series of post hoc tests for 2x2 tables „each with each“ followed. One of the tests showed higher number of complications in the patients with AB group compared to the A group, $p = 0.05$ (5 %). How will the p-value change when Bonferroni correction is applied (p, not α-value is corrected here)? The result should be in percents (a natural number), eventually rounded to percents.

| Ranking | Response | Votes |
|---|---|---|
| Correct Answer | | |
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| Others | | |

# Regression models

- „Regression towards the mean“ (Francis Galton) – but methods already by Friedrich Gauss
- The goal is to estimate a value of modelled variable (dependent variable = regressand) using other known parameters (factors = regressors – categorical or continuous variables)
- The contribution of individual factors may be assessed separately (univariate models) or together in mutual interaction (multivariate models)
- For each factor, its effect size with **confidence intervals** may be determined (usually 95 %, i.e. where is the value with 95 % probability)
- Assumption: factors are **independent**
- Most often:
  - Linear regression (dependent variable is continuous – e.g. fasting plasma glucose)
  - Logistic regression (dependent variable is Binary – e.g. disease))
  - Cox regression (dependent variable is survival – survival time and endpoint)

# Assessing the contribution of factors

- Linear regression – regression coefficient β (standardized, unstandardized) and 95% confidence interval (CI)
  - Unlike in correlation, it is important which variable is independent and which one dependent
  - When the regressor is categorical, it is ANOVA in fact
- Logistic regression – OR and 95% CI
- Cox regression – hazard ratio (HR) and 95% CI

# Interpretation of regression models

- When β ± 95% CI includes 0, the contribution of the factor is not significant (under 0, the value of outcome is decreased, over 0 is increased)
- In OR and HR, same is true when 95% CI include 1 (under 1, the probability of an outcome is decreased, over 1 is increased)
- 95% CI can thus replace the p-value
- When the independent variable is categorical, one category has to be set as the reference one and regression coefficients / OR / HR are attributed to each other category
- When the independent variable is continuous, β / OR / HR corresponds to 1 unit (e.g. 1 year of age – assumes linear effect, otherwise it is better to categorize)

# Choose the right statement

In a cross-sectional study including 700 hospitalized patients between 80 – 90 years of age, signs of cognitive disability were found in 40 %. Association with potential risk factors (age, hypertension, diabetes) was assessed using univariate logistic regression. A presence of cognitive disability was consecutively associated with: age (OR = 1.20; 95 % CI = 1.12 – 1.40 per each year of age), hypertension (OR 1.40; 95 % CI 1.20 – 1.78) and diabetes (OR 2.80; 95 % CI 2.00 – 6.40).

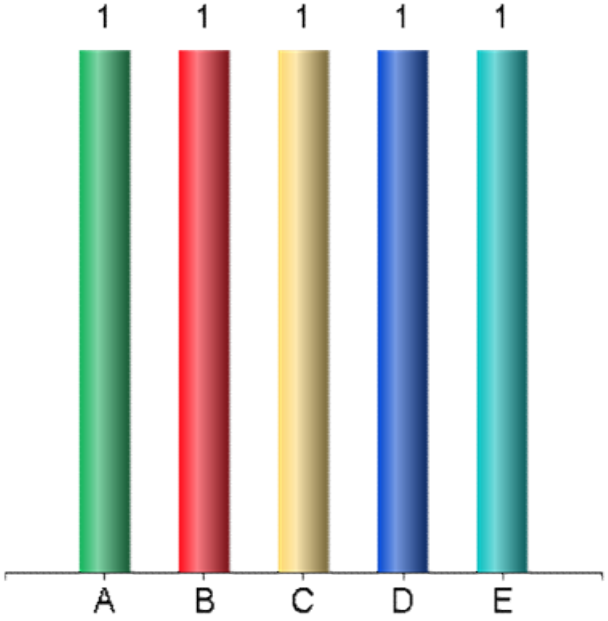
A. Age is not a statistically significant factor for cognitive disability

B. A probability of cognitive disability occurrence is two times higher in diabetics than in hypertensive patients

C. Age, diabetes and hypertension are mutually independent risk factors

D. When we test a statistical significance of associations, the p -value is < 0.05 in all cases

E. We may conclude that the factors lead to cognitive dysfunction.



# What to do with the ordinal data?

- Tests for categorical data, ANOVA (but: we ignore the order)
- Nonparametric tests (with many categories)
- Dichotomization and tests for binary data (often in medicine)
- Special tests – Cochran-Armitage (typically genetics), sign test

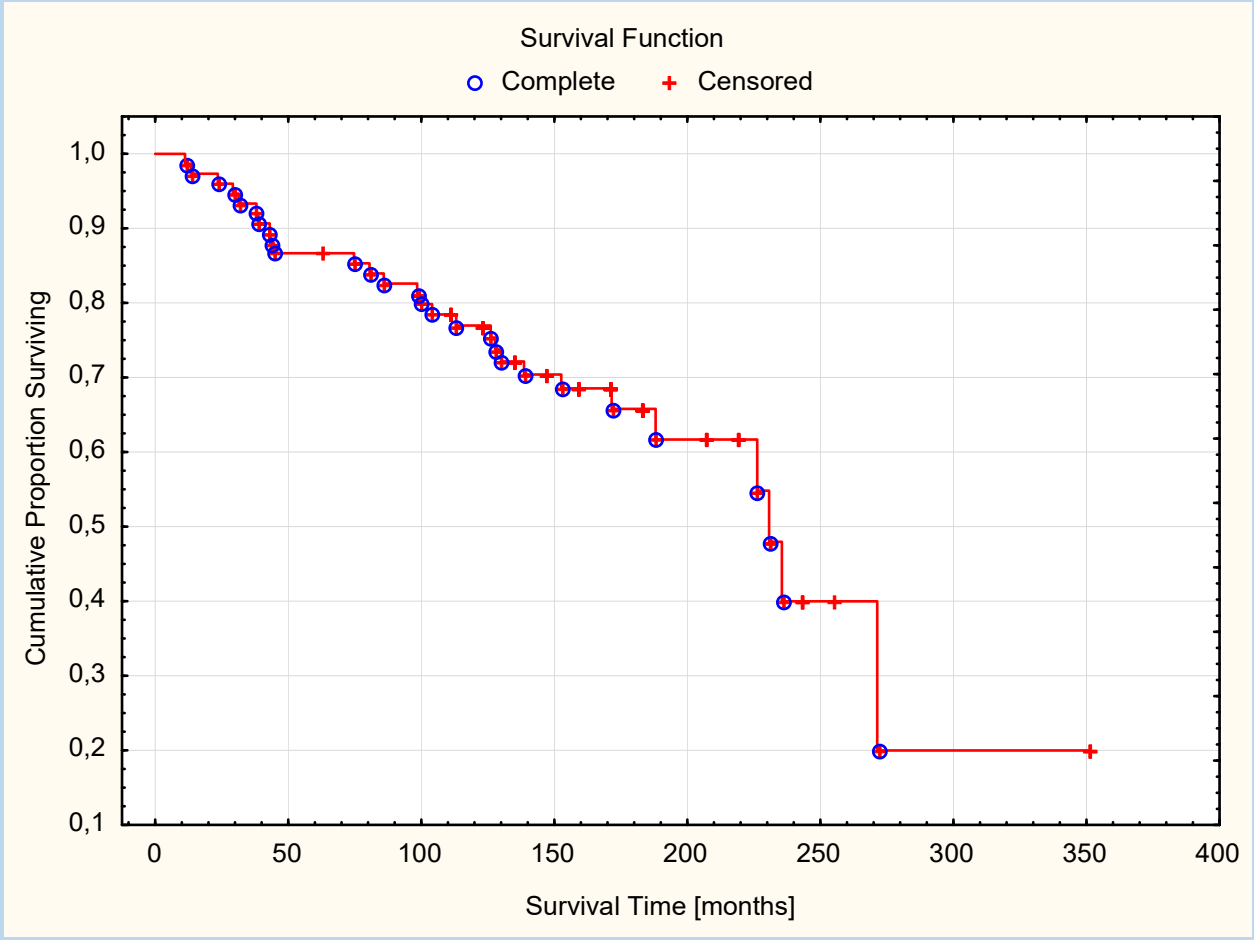
# Survival analysis

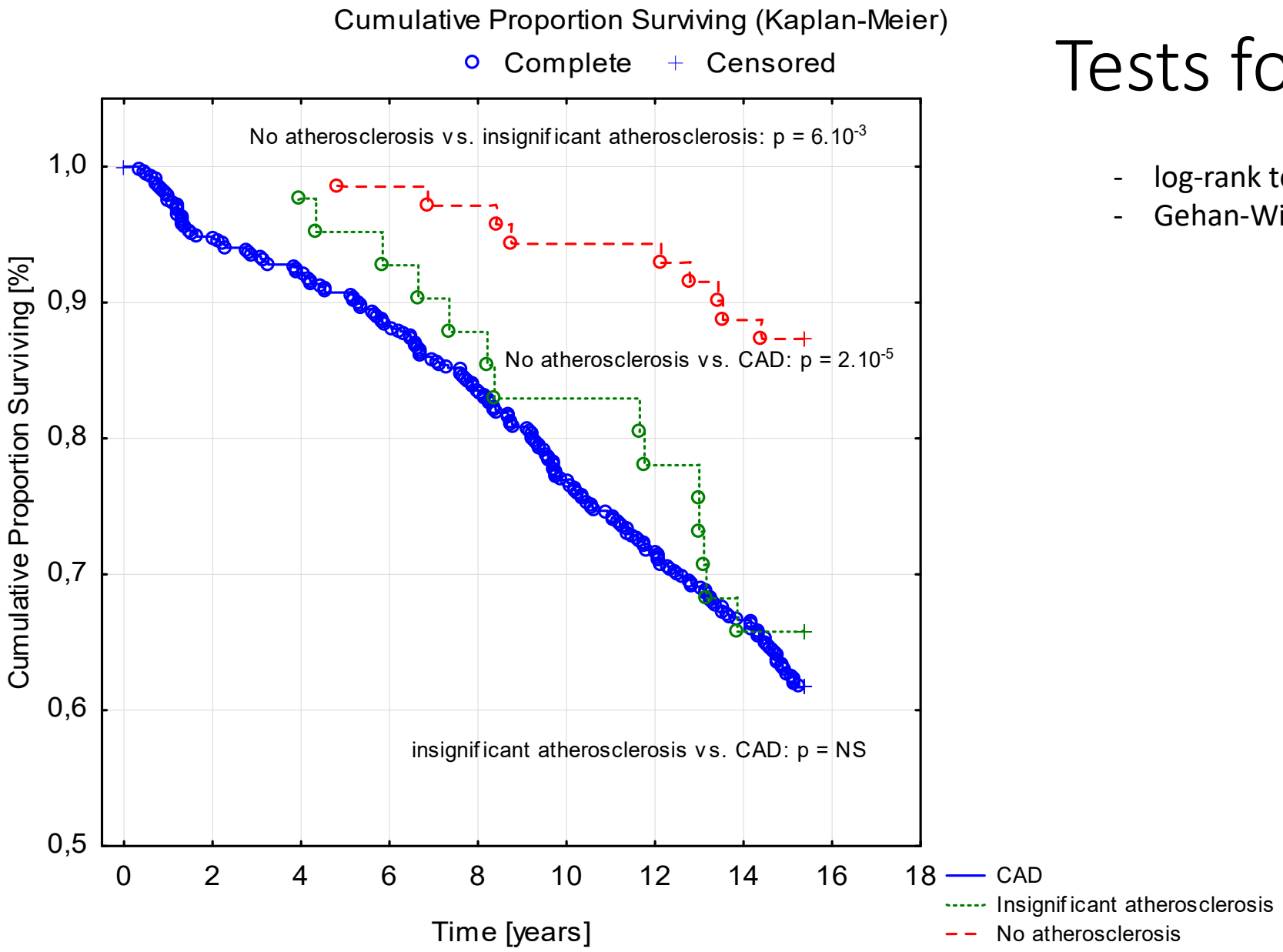
- Group of methods assessing an occurrence of given even (endpoint) in typically decreasing number of study group members „survivors“
- What is assessed?
- Endpoint
  - Occurs only once (if it can occur more times, the first occurrence is usually assessed).
- Censored data
  - still alive at the end of the study (event did not occur)
  - lost from the study
  - died for another cause
- Time of follow-up (survival time)

# Methods of survival analysis

- Life-tables
- Kaplan-Meier graphs
- Log rank test
- Gehan-Wilcoxon‘s test
- Cox regression

# Kaplan-Meier survival curve





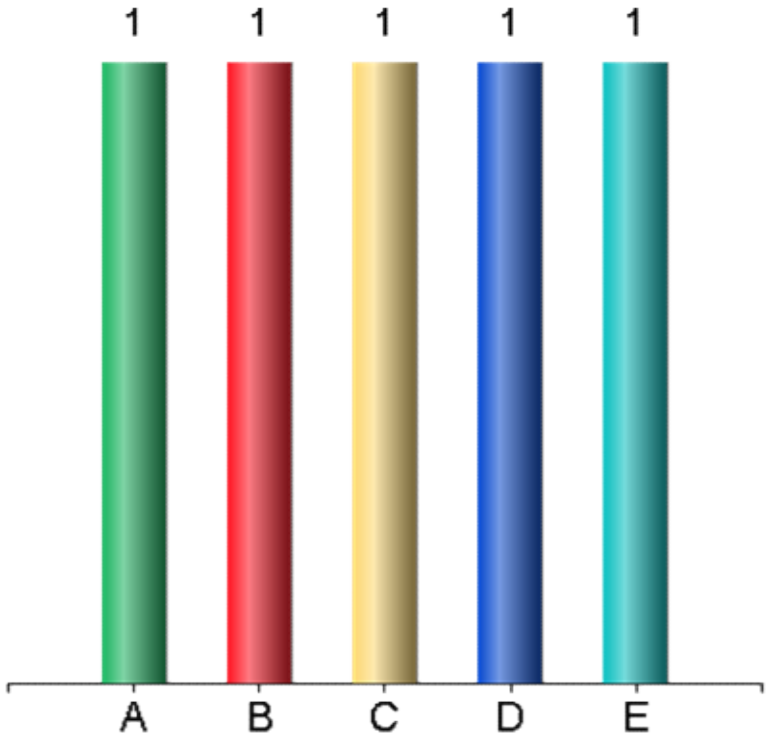

# Tests for survival

- log-rank test
- Gehan-Wilcoxon test

# Choose the right answer...

Four patients enrolled to the study investigating the re-occurrence of myocardial infarction (endpoint). In following years, subsequent events took place consecutively: one patients moved to Argentina and was thus lost from follow-up, one suffered the infarction and next month he died during a car accident, further one died of lung cancer and the last one lived until the end of the study in full health. The last point of Kaplan-Meier curve is at the value of:

- ✓ A. 66.6%
- B. 50%
- C. 33.3%
- D. 25%
- E. 0%

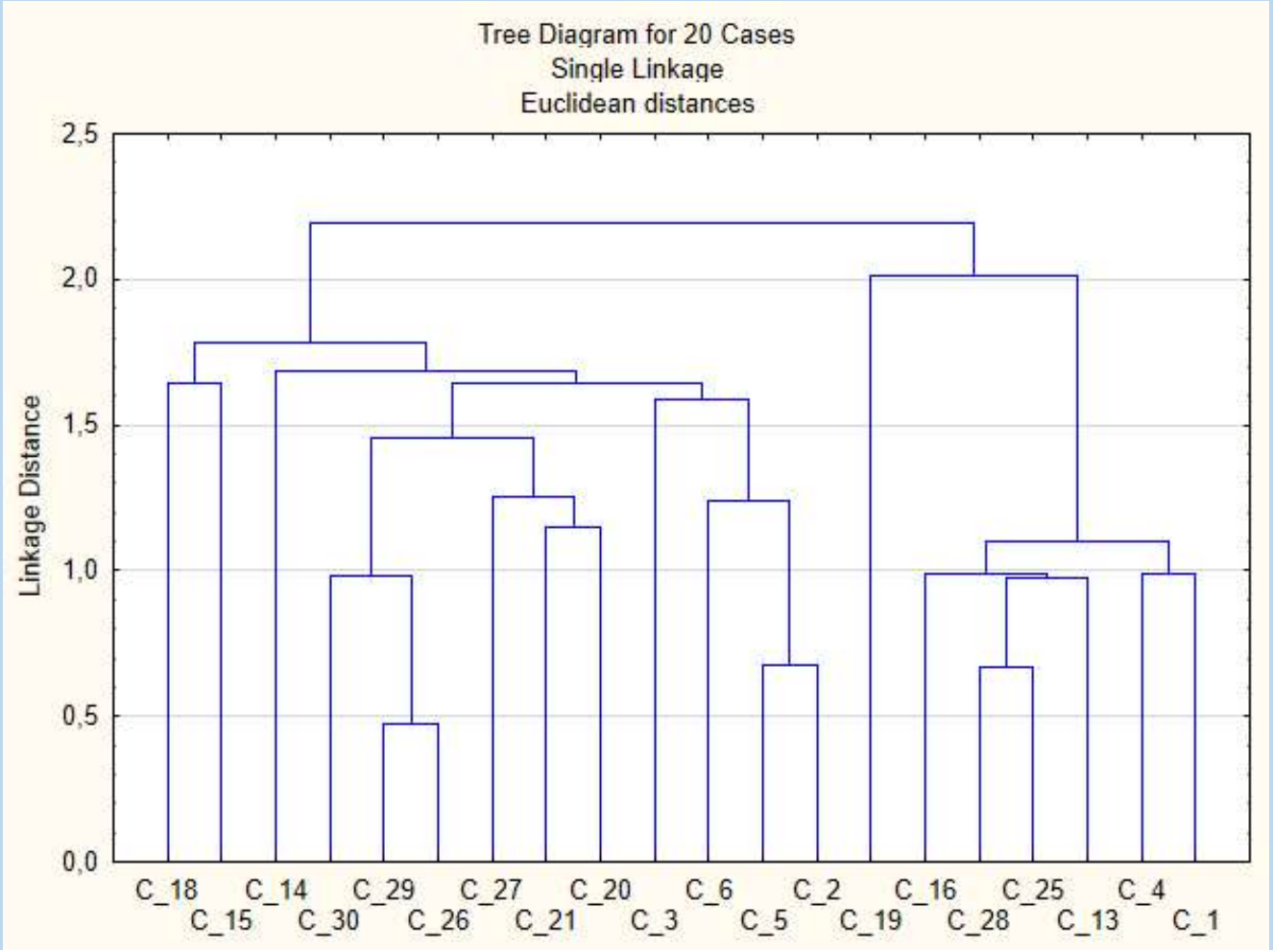


# Cluster analysis

- multidimensional analysis (1 parameter = 1 dimension)
- measure of distance
- amalgamation algorithm
- data standardization is necessary to assess different parameters together (to unify scales, all parameters are expressed in the same units: σ of their distribution – i.e. the z-score; mean = 0)

- k means clustering
- hierarchical tree (dendrogram)

# Dendrogram


Tree Diagram for 20 Cases
Single Linkage
Euclidean distances
Linkage Distance
2,5
2,0
1,5
1,0
0,5
0,0
C_18
C_15
C_14
C_30
C_29
C_26
C_27
C_21
C_20
C_3
C_6
C_5
C_2
C_19
C_16
C_28
C_25
C_13
C_4
C_1

# Choose the right answer...

A lonely island is visited by anthropologists, who discover human skulls of unknown origin there. They use the cluster analysis to assign them to some of the human populations nearby. Besides the genetic markers, they also measure the cranial index (in percents, mean = 85, SD = 10), facial index in percents, mean = 80, SD = 5) and the braincase volume (in $cm^3$, mean = 1500, SD = 200). What happens if the data are not standardized before the analysis:

A. Nothing, standardization is used for better visualization of the data.

B. The braincase volume will not be relevant for the analysis.

C. Cluster analysis will not be technically possible.

D. The assignment to a cluster will depend mainly on the braincase volume.

E. The mutual correlation of cranial and facial index will increase.