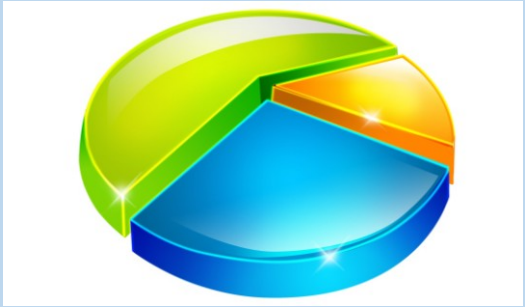
Statistical methods in biology and medicine II

Repetition

- descriptive statistics
  - data presentation
- statistical induction
  - conclusions about populations are based on samples
- statistical inference
  - hypotheses testing
  - multiple comparison

Repetition - kinds of data

- Continuous (always quantitative) – the parameter can theoretically be of any value in a given interval (e.g. glucose concentration: 0-∞; ejection fraction: 0-100%)
  - Ratio vs. 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 level of education (basic, high school, university) is an example of...

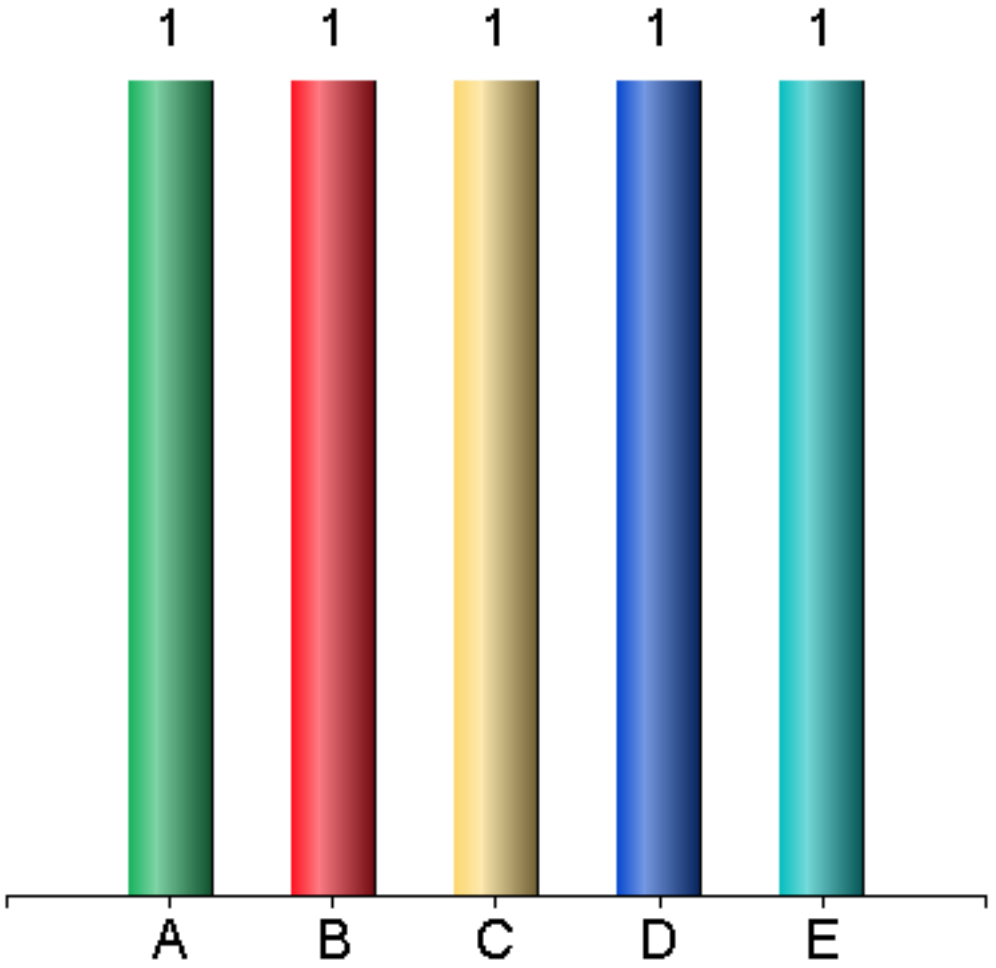
A. Ordinal variable

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)**	**correct (1-β)**
H_0 confirmed	**correct (1-α)**	**type II error (β)**

- 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-β – also **power of a test**
- Statistical significance p – 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)

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. Express its practical (not statistical) significance

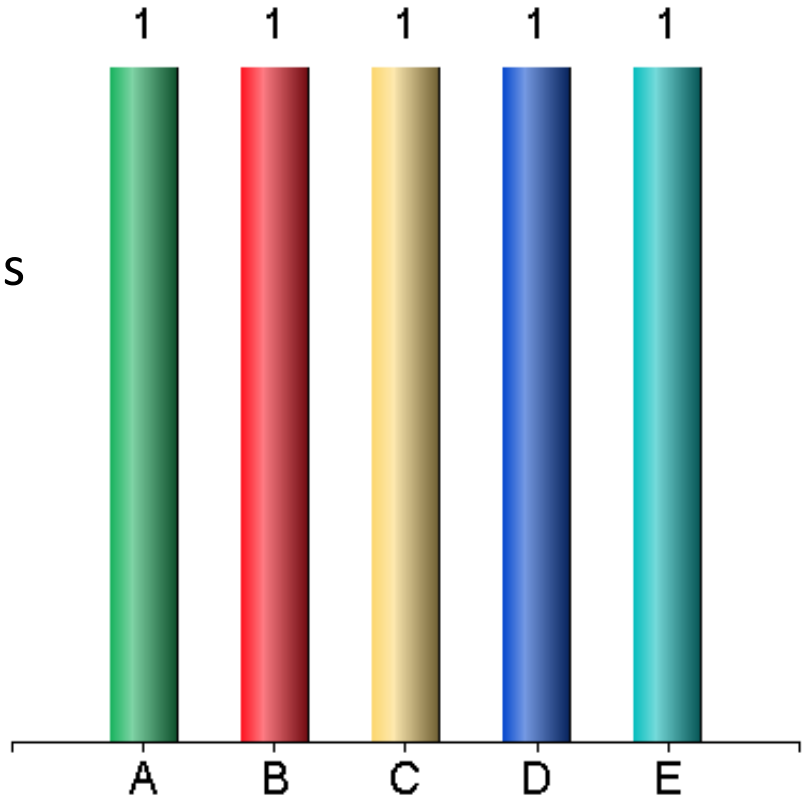
B. Increases with increasing variability of the data

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

D. Is expressed by a letter p

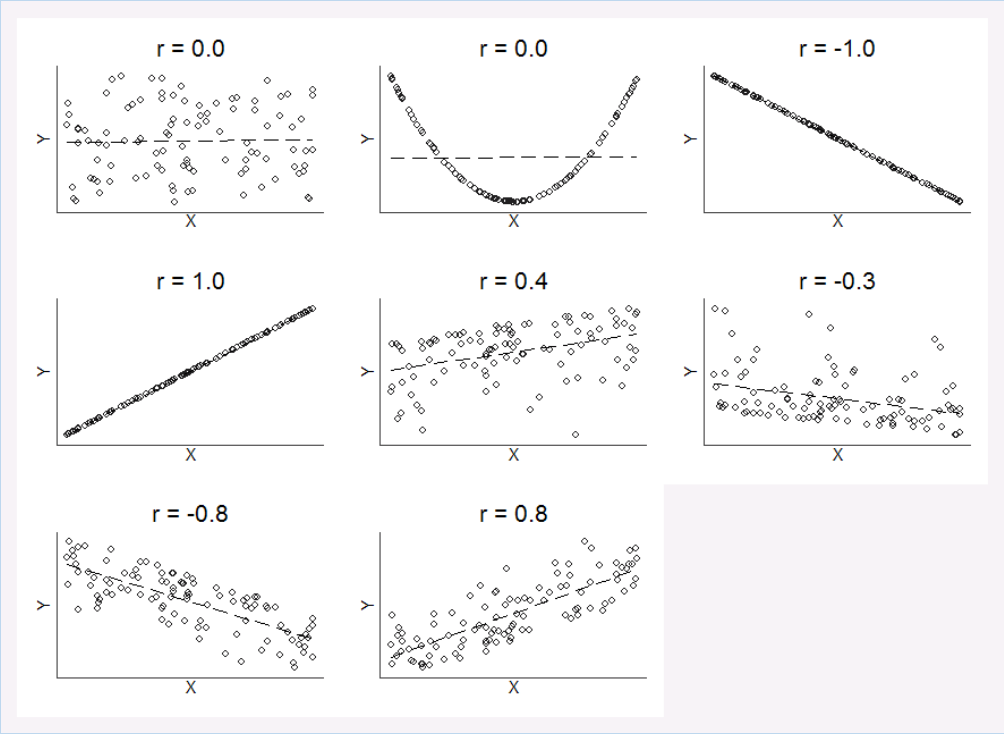
E. Is a probability that the alternative hypothesis is true in a case

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 define 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

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

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)

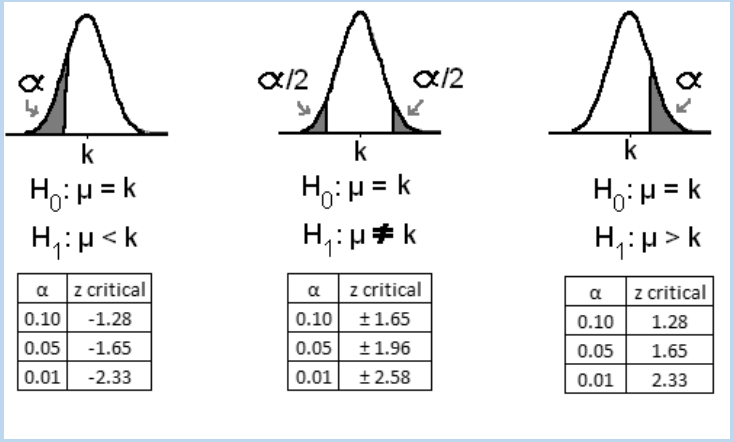
Unpaired (unmatched samples)

- 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) t-test	Wilcoxon paired test Sign test
Unpaired	Unpaired (independent) t-test	Mann-Whitney U-test * 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)

Advanced statistics

- Multiple samples comparison (ANOVA)
- Contingency tables (Fisher`s exact test, χ^2-test)
- Survival analysis
- Cluster analysis

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 treatmen

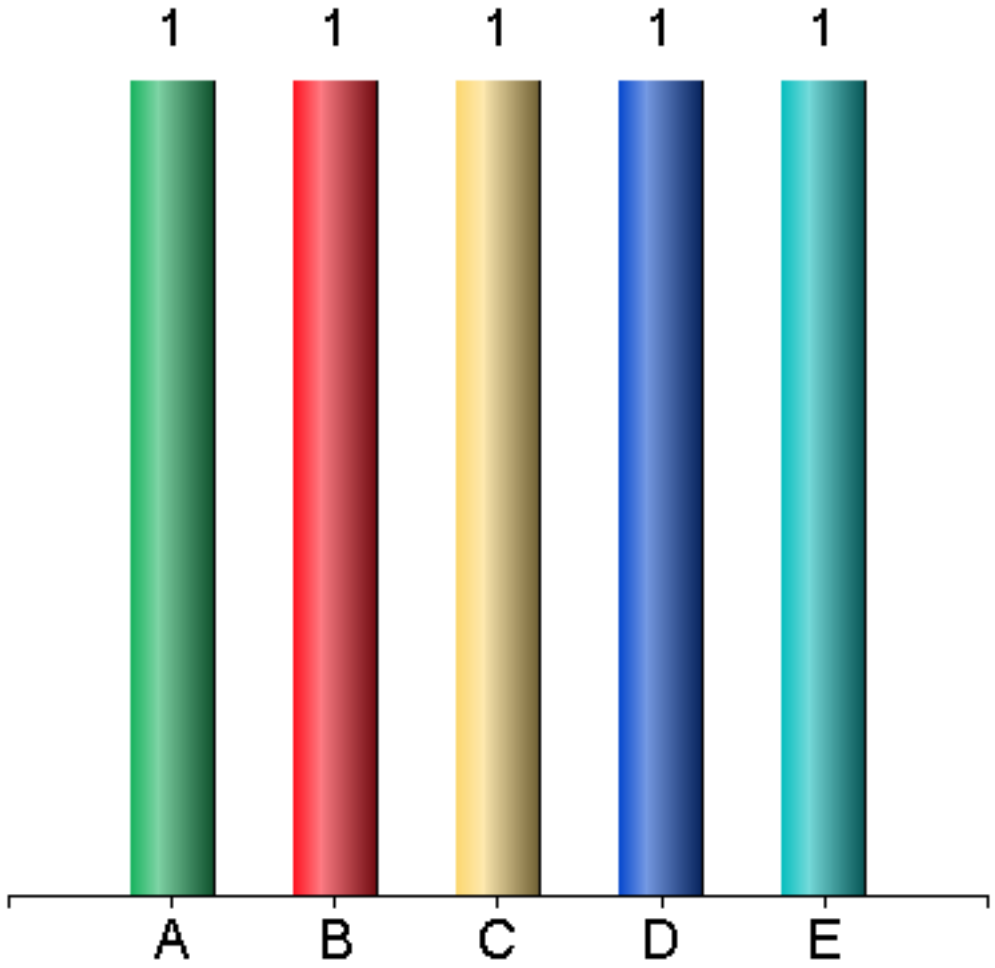
A. Paired t-test

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

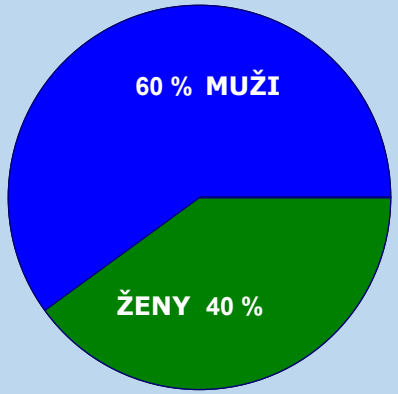
- When we perform more tests at once, the probability that some of them will give a statistically significant result only due to chance (i.e. type I error – H_0 is wrongly refuted) – increases (e.g. during post hoc tests following ANOVA
- For example, when performing 10 tests at $\alpha = 0.05$, the 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“.

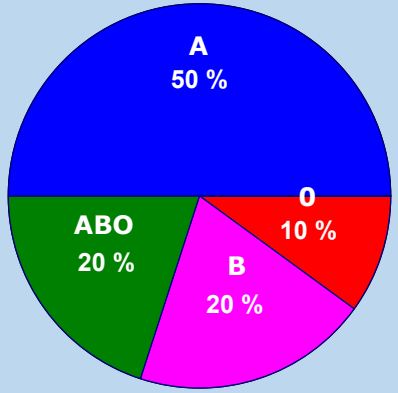
Contingency tables

- The association beween two categorical variables can be expressed in a contingency table (n x m, see below an example for 2x2 table

Počet z Cislo	Skupina		
Genotyp ADRB1	control	TTC	Celkový součet
CC	15	15	30
CG	12	10	22
GG		1	1
Celkový součet	27	26	53

Tests for categorical data

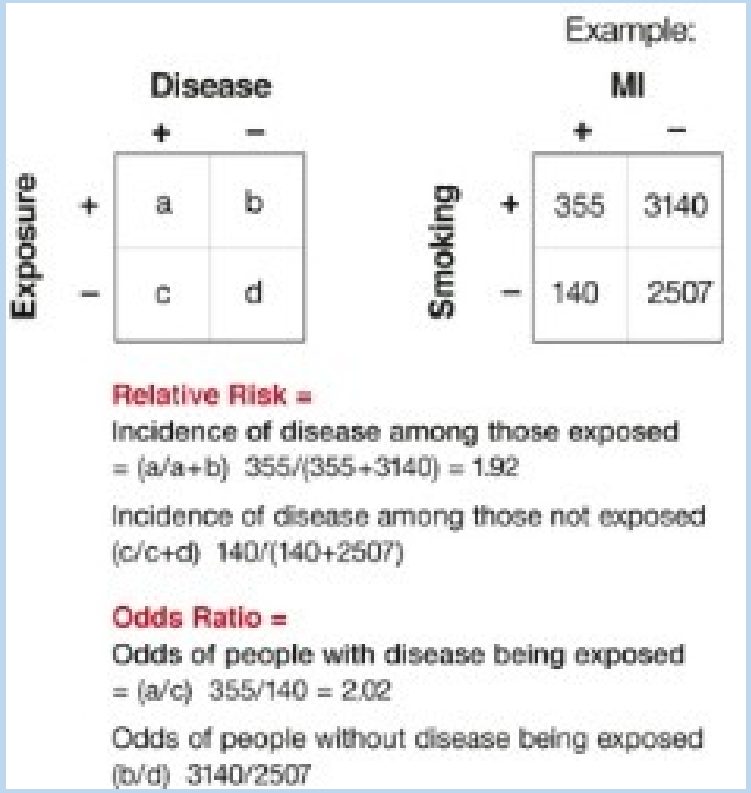




- From the contingency table, its probability under the assumption that H_0 is valid (i.e. the p-value) can be determined, as well as the effect size – e.g. the association between a mutation and a disease (expressed as RR – relative risk; OR – odds ratio)
- Sometimes, a reduction of larger tables into 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)]
- For binary variables, paired design can be used (typically presence/absence of the disease in time)

	nemoc	zdraví
mutace	50	2
ne	4	48

Relative risk and odds ratio in 2x2 tables



www.mdedge.com

- probabilities 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. disease 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

Tests for categorical data - examples

Test	2 x 2 contingency table	More categories/measurements ‡
Paired	McNemar test	Cochran Q test (Binary data, more measurements) Sign test (ordinal data, two measurements
Unpaired	Chi-square (χ^2) test* Fisher exact test	Chi-square (χ^2) test* 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, χ^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“ (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 (univariable models) or together in mutual interaction (multivariable models)
- Assumption: factors are **independent**
- Most often:
  - Linear regression (dependent variable is continuous)
  - Logistic regression (dependent variable is binary)
  - Cox regression (dependent variable is survival – survival time and endpoint)

Contribution of factors

- Linear regression – regression coefficient β (standardized, unstandardized) and 95% confidence interval (CI) – i.e. estimation, where the coefficient really is with 95% likelihood
  - Unlike in correlation, it is important which variable is independent and which one dependent
  - When the regressor is categorical, the regression model equals ANOVA in fact
- Logistic regression – OR and 95% CI
- Cox regression – hazard ratio (HR) and 95% CI
- When β ± 95% CI includes 0, the contribution of the factor is not significant (under 0, the value of outcome is decreased, over 0 it is increased)
- In OR and HR, the same is true when 95% CI include 1 (under 1, the probability of an outcome is decreased, over 1 it 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 comprising 700 hospitalized patients aged between 80 – 90 years, 40 % had signs of cognitive dysfunction. Association with candidate risk factors (age, hypertension, diabetes) was assessed using univariable logistic regression. The presence of cognitive dysfunction was subsequently associated to: age (OR = 1.20; 95 % CI = 1.12 – 1.40 per each subsequent year), hypertension (OR 1.40; 95 % CI 1.20 – 1.78), and diabetes as well (OR 2.80; 95 % CI 2.00 – 6.40).

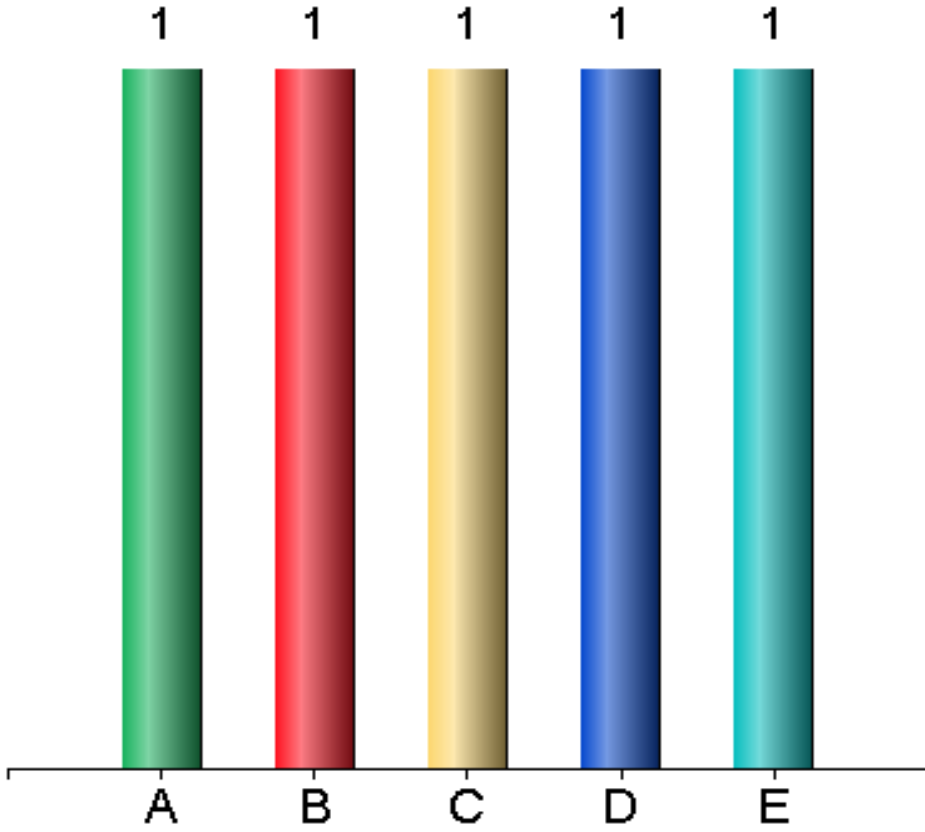
A. Factor of age is not statistically significant with respect to cognitive dysfunction

B. The probability of cognitive dysfunction is twice higher in diabetics than in hypertonics

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

D. P-value is < 0.05 in all cases

E. There is causal relationship between the factors and cognitive dysfunction



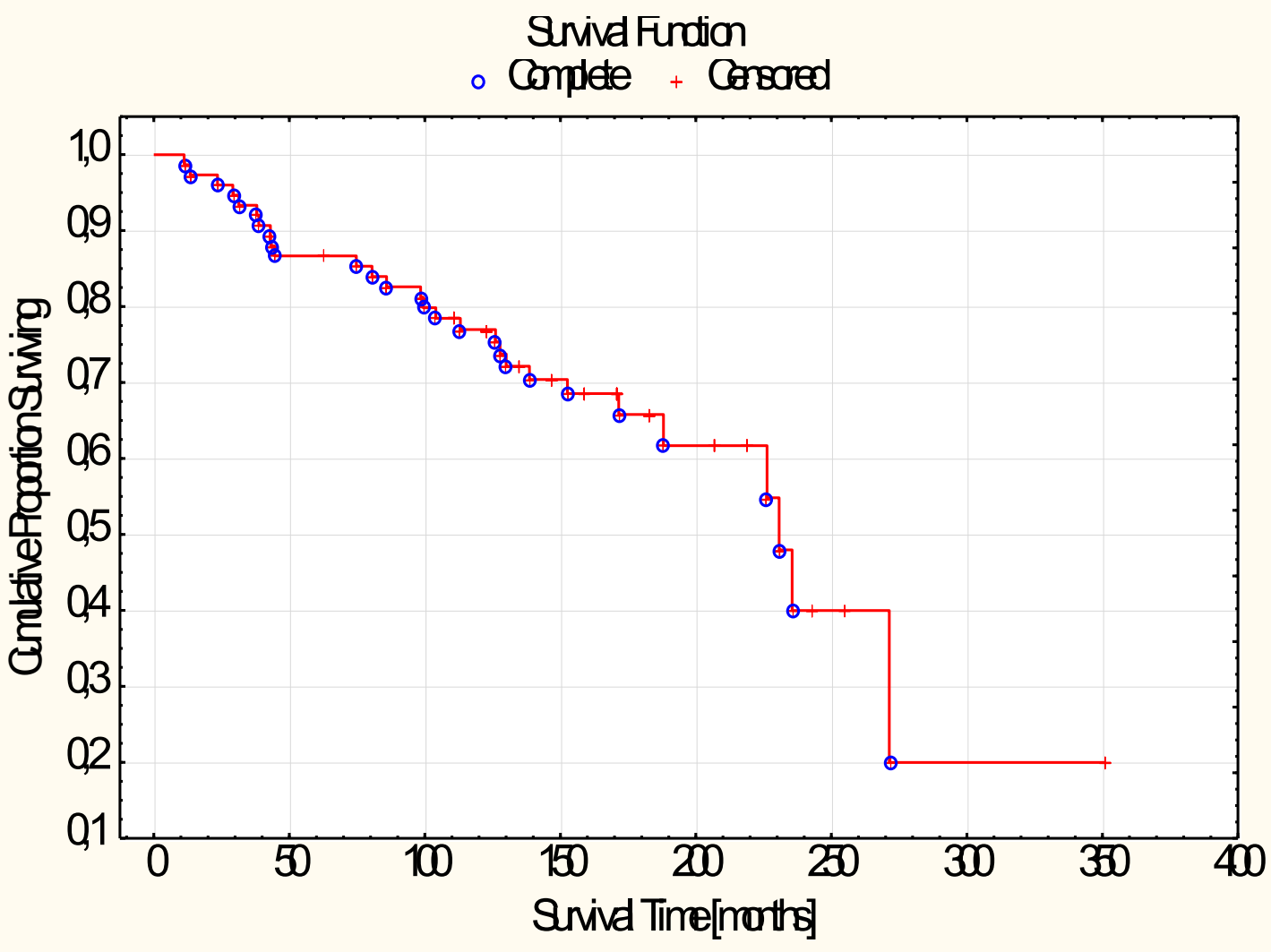
What to do with the ordinal data?

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

Survival analysis

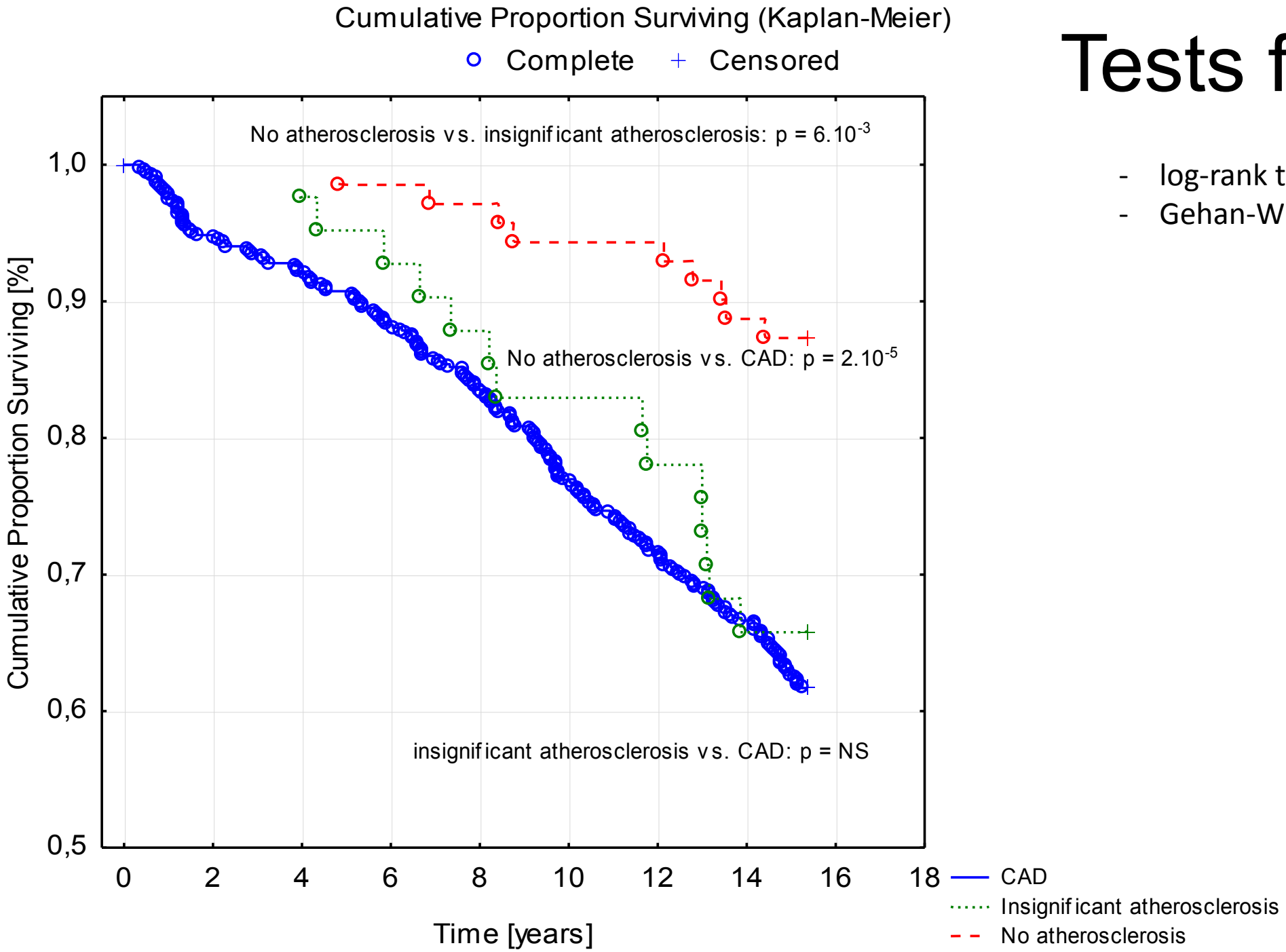
- probability of the given event (death) decreases with decreasing number of study group members „survivors“
- censored data
  - still alive at the end of the study (event did not occur)
  - lost from the study
  - died for another cause
- Kaplan-Meier graphs
- Log rank test

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:

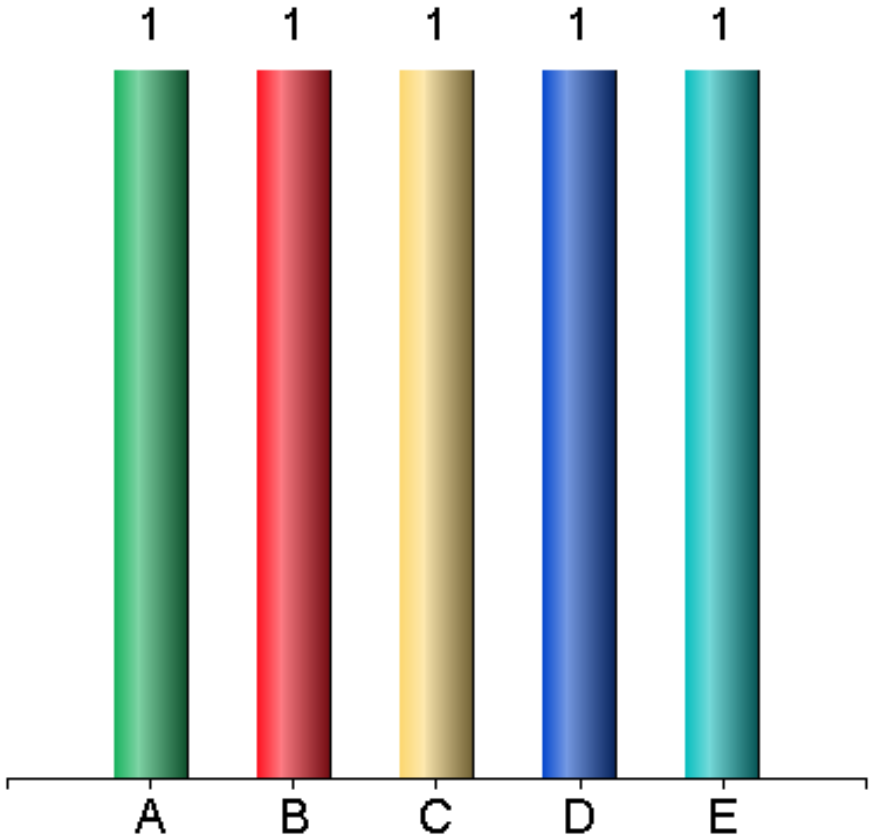
A. 66.6%

B. 50%

C. 33.3%

D. 25%

E. 0%



Cluster analysis

- multidimensional analysis
- measure of distance
- amalgamation algorithm
- data normalization

- k means clustering
- hierarchical tree (dendrogram)

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.