

Lecture 10

Correlation and regression

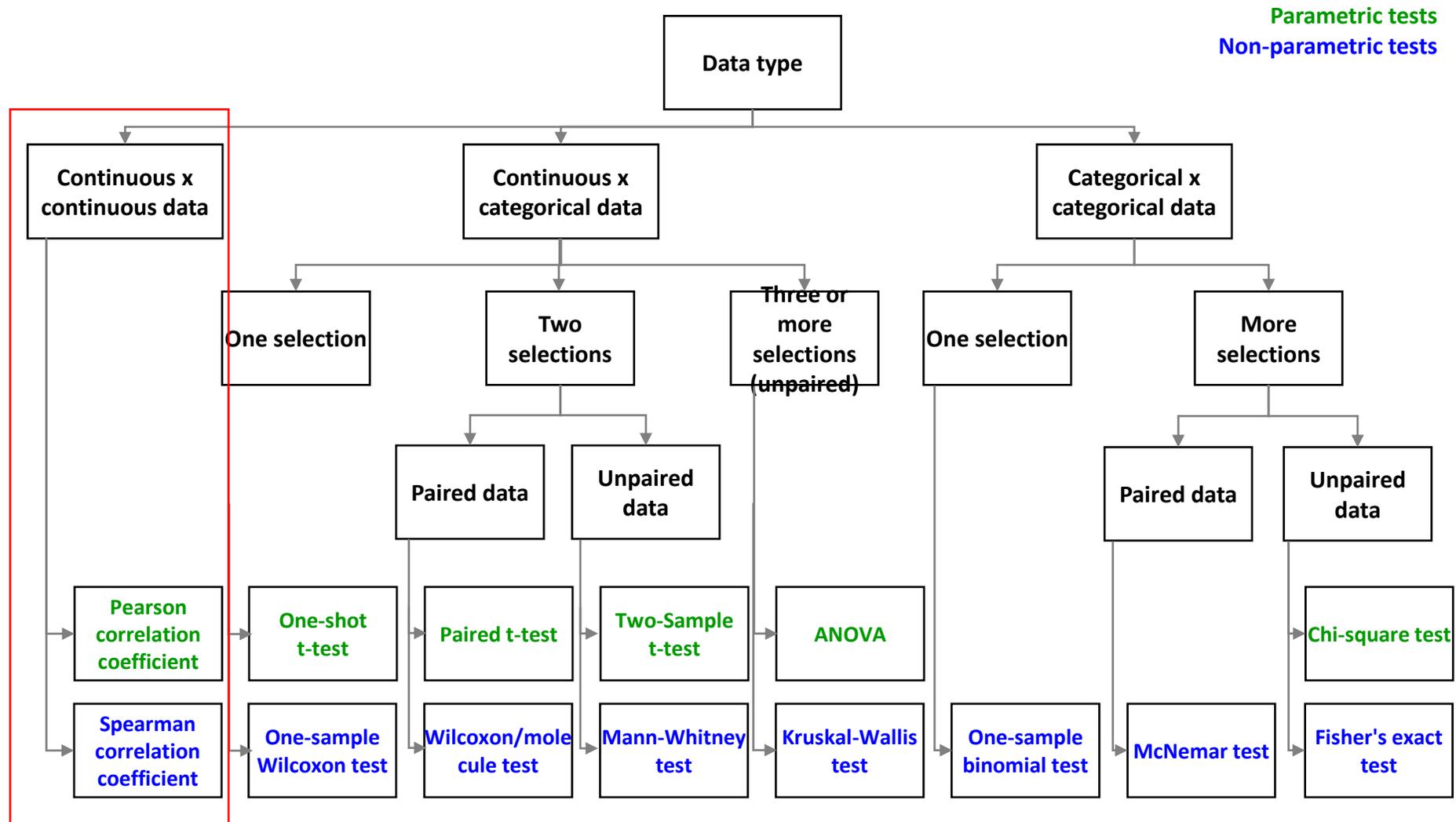
Parametric and non-parametric correlation

Linear regression

Annotation

- Correlation analysis is used to evaluate the degree of relationship between two continuous variables.
- Similar to other statistical methods, correlations can be parametric or non-parametric
- Regression analysis models the relationship between two or more variables, i.e. how one variable (the explained variable) depends on other variables (the predictors).
- Regression analysis, like ANOVA, is a tool for explaining the variability of the variable being evaluated

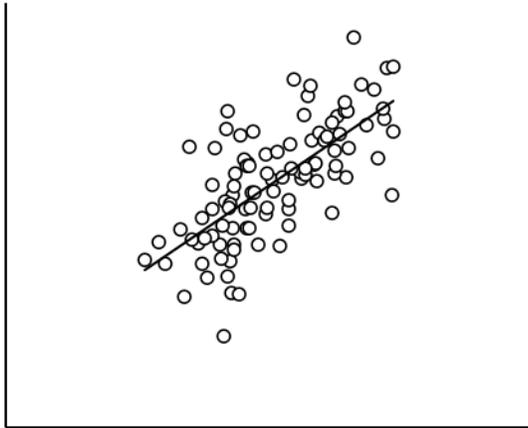
Basic decision making on the selection of statistical tests



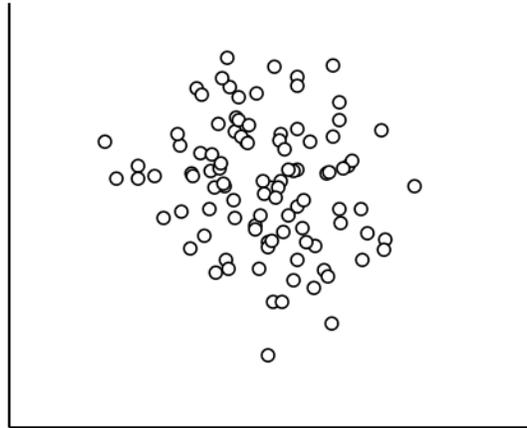
Description of the relationship of continuous variables

- The basic tool for describing the relationship of continuous variables is the XY graph, which allows to assess the type and strength of their relationship.

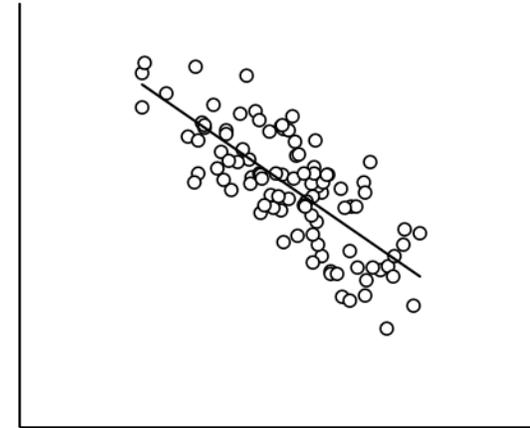
Positive linear relationship



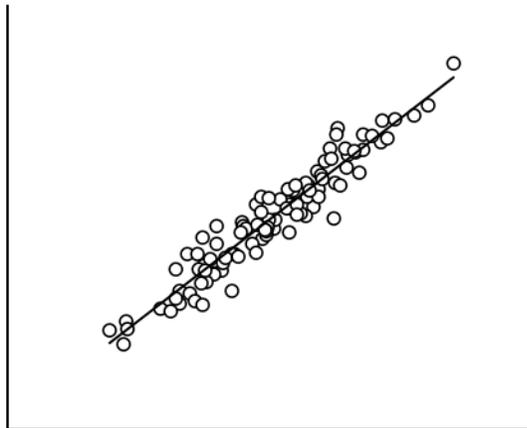
A casual relationship



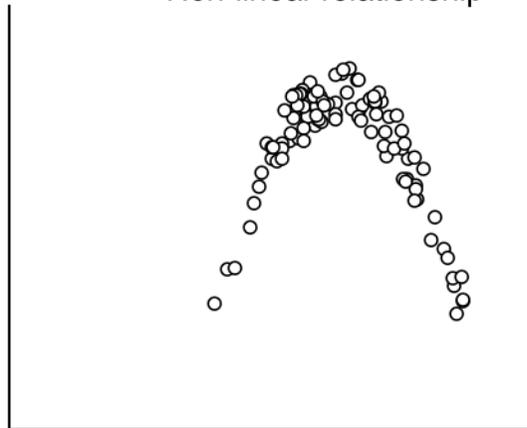
Negative linear relationship



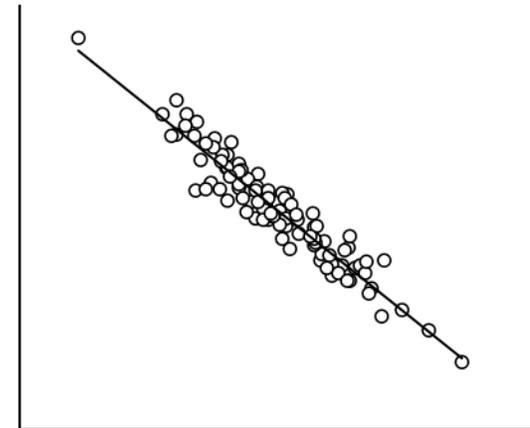
Strong positive linear relationship



Non-linear relationship

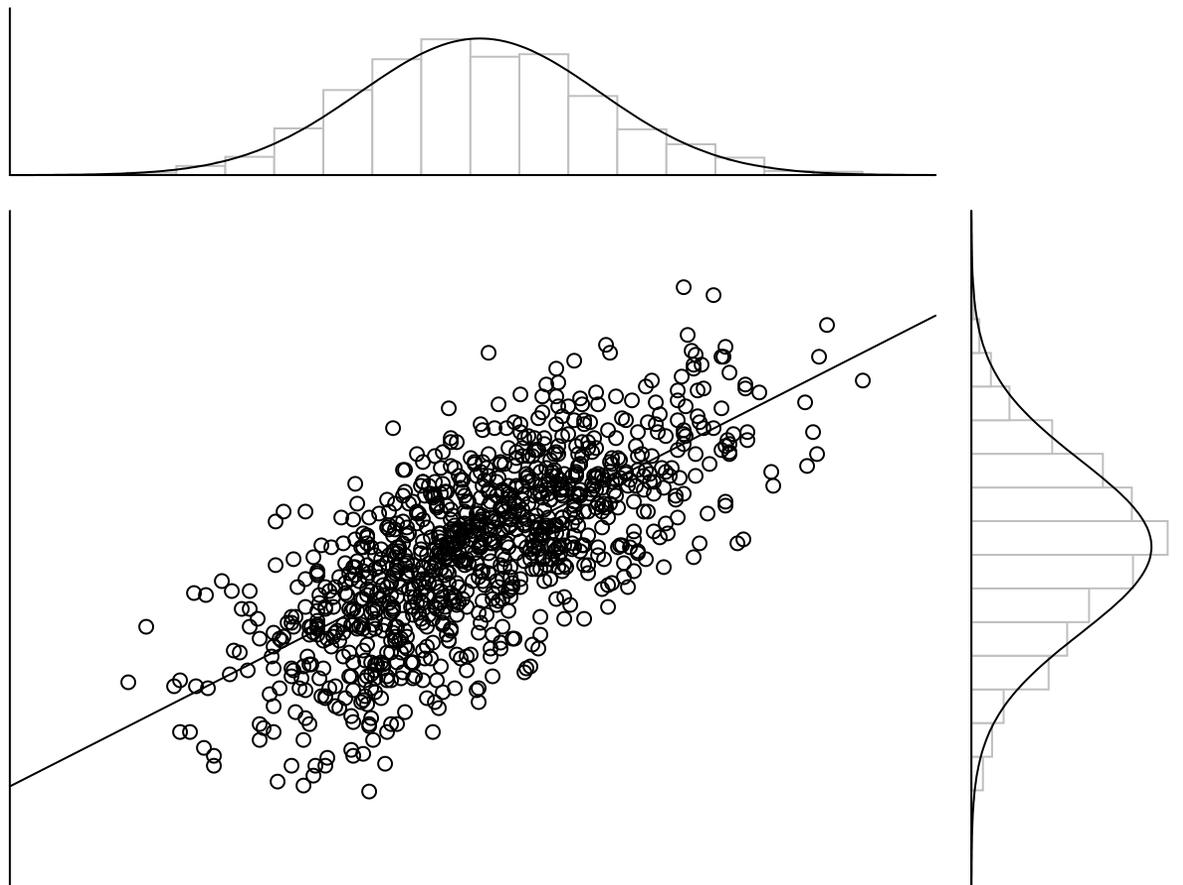


Strong linear negative relationship



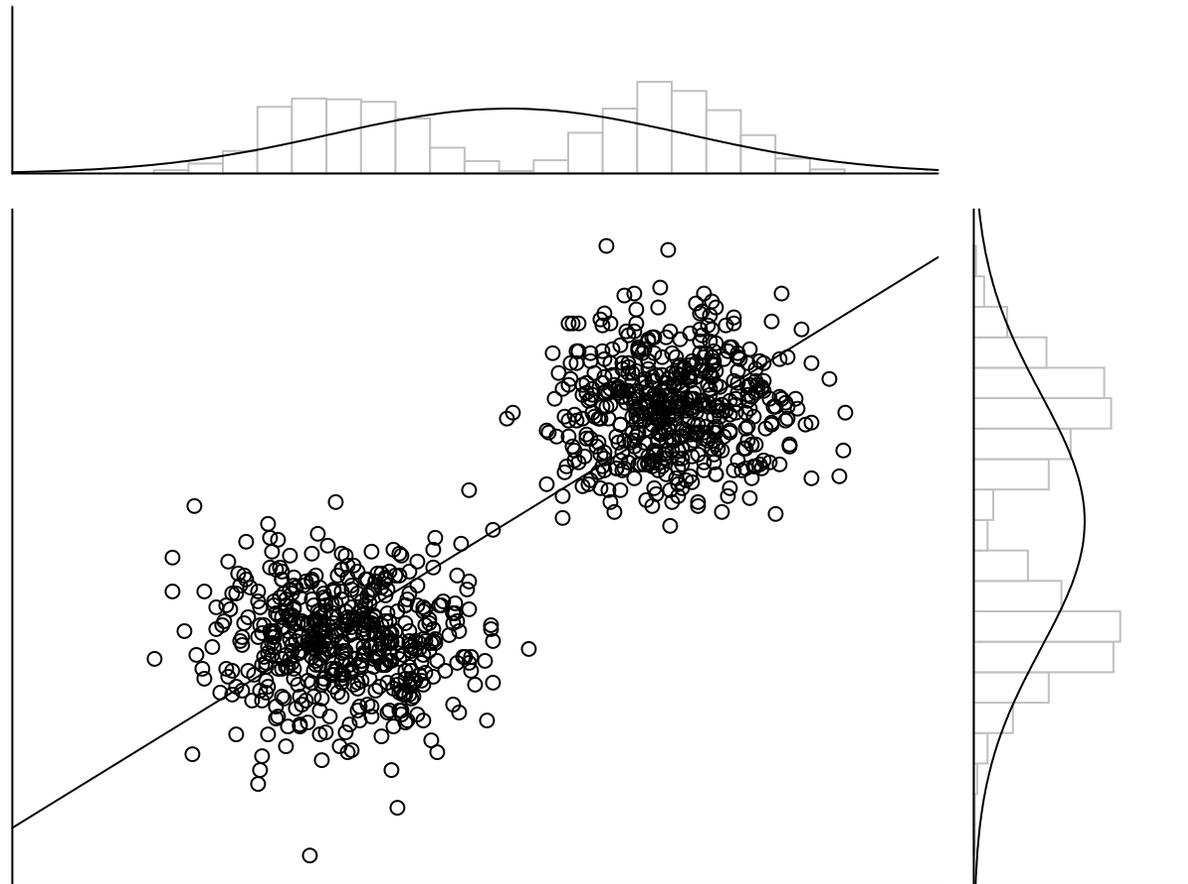
Assumptions of parametric correlation analysis

- A correct interpretation of parametric correlation analysis assumes a linear relationship between the variables and a normal distribution of the values of both variables.



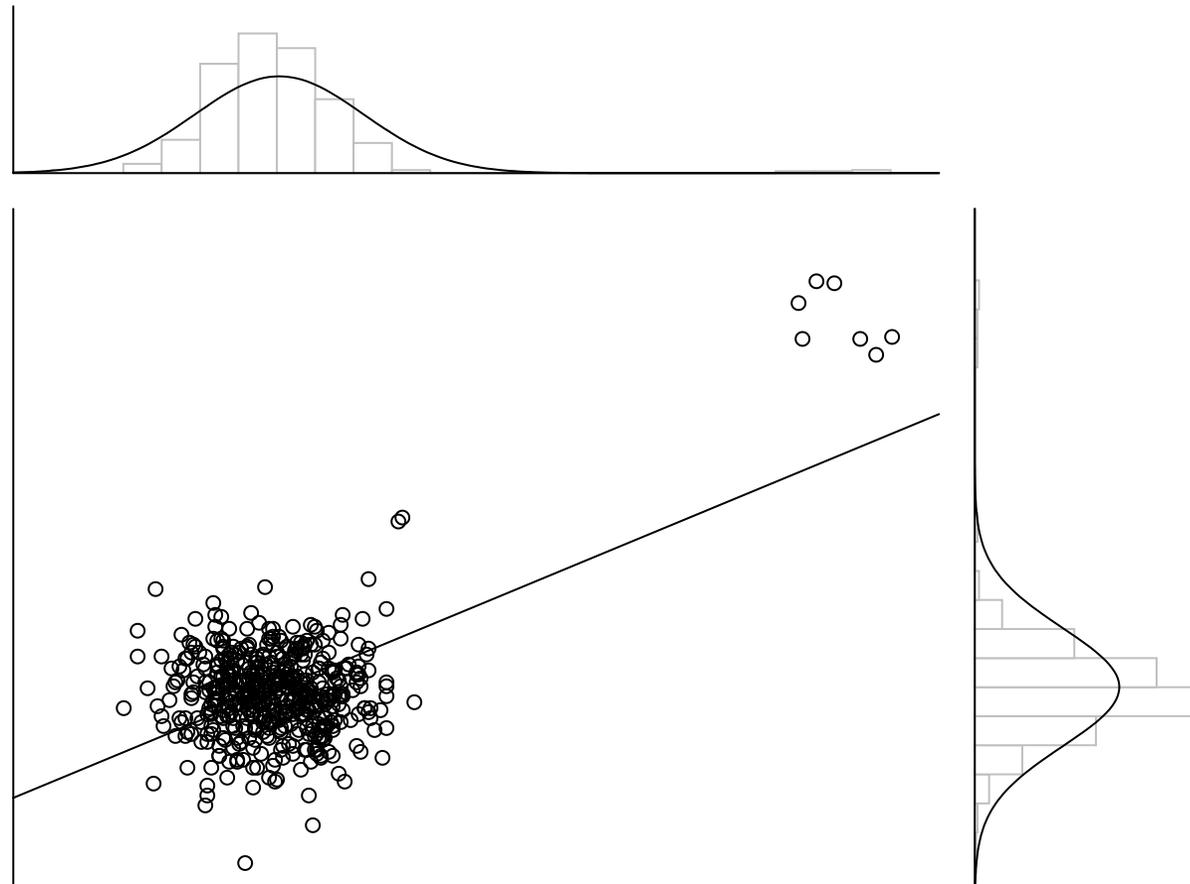
Bimodal distribution of values entering the correlation analysis

- In the case of a bimodal distribution of values entering the correlation analysis, it is not appropriate to calculate the correlation analysis; the result cannot be interpreted as a description of a linear relationship of continuous variables, but as a consequence of the existence of subsets of objects in the data.



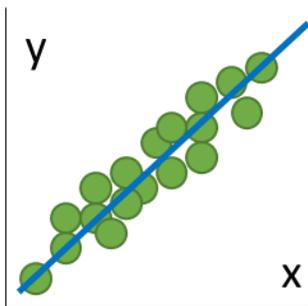
Presence of outliers in the data entering the correlation analysis

- If outliers are present in the data entering the correlation analysis, it is not appropriate to calculate the correlation analysis; the result cannot be interpreted as a description of a linear relationship of continuous variables, but as a consequence of the presence of outliers in the data.

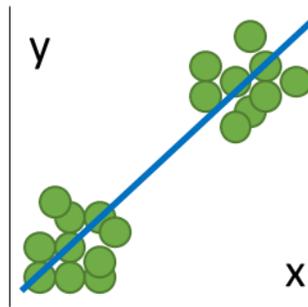


Correlation and covariance - parametric measures of the relationship between continuous variables

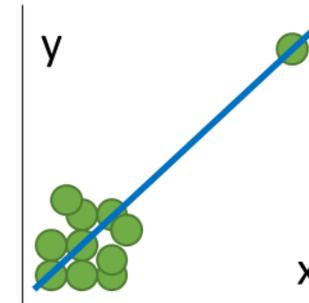
- Covariance and Pearson's correlation coefficient are basic methods for describing the linear relationship of continuous variables
- The assumption for the calculation of covariance and Pearson correlation coefficient is:
 - Data normality in both dimensions
 - Linearity of the relationship between variables



Linear relationship - seamless use of covariance or Pearson correlation coefficient

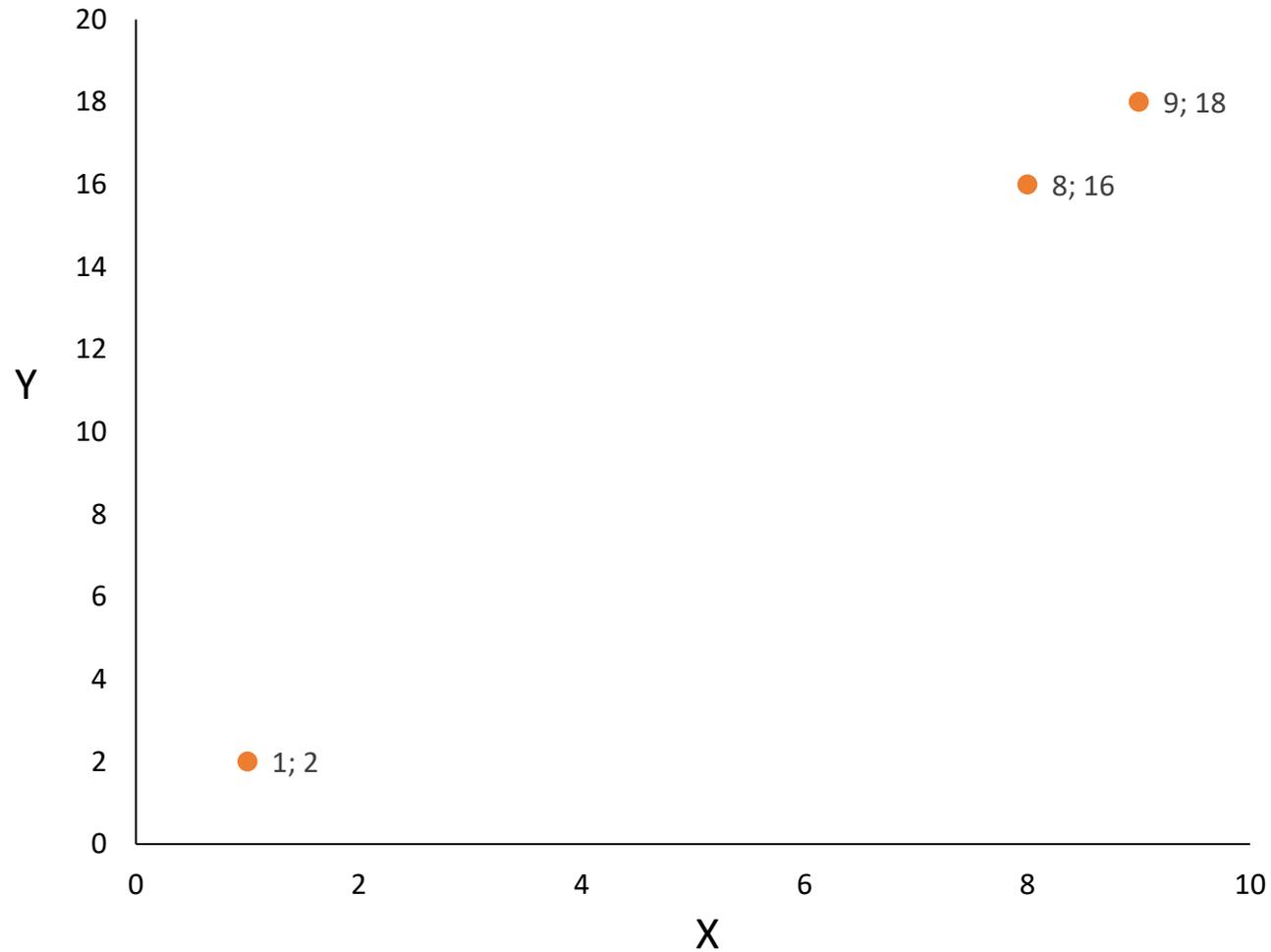


The correlation is given by two sets of values - it leads to the identification of groups of objects in the data

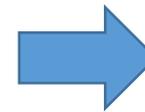


The correlation is given by the outlier - the analysis only describes the effect of the outlier

Calculation of covariance I

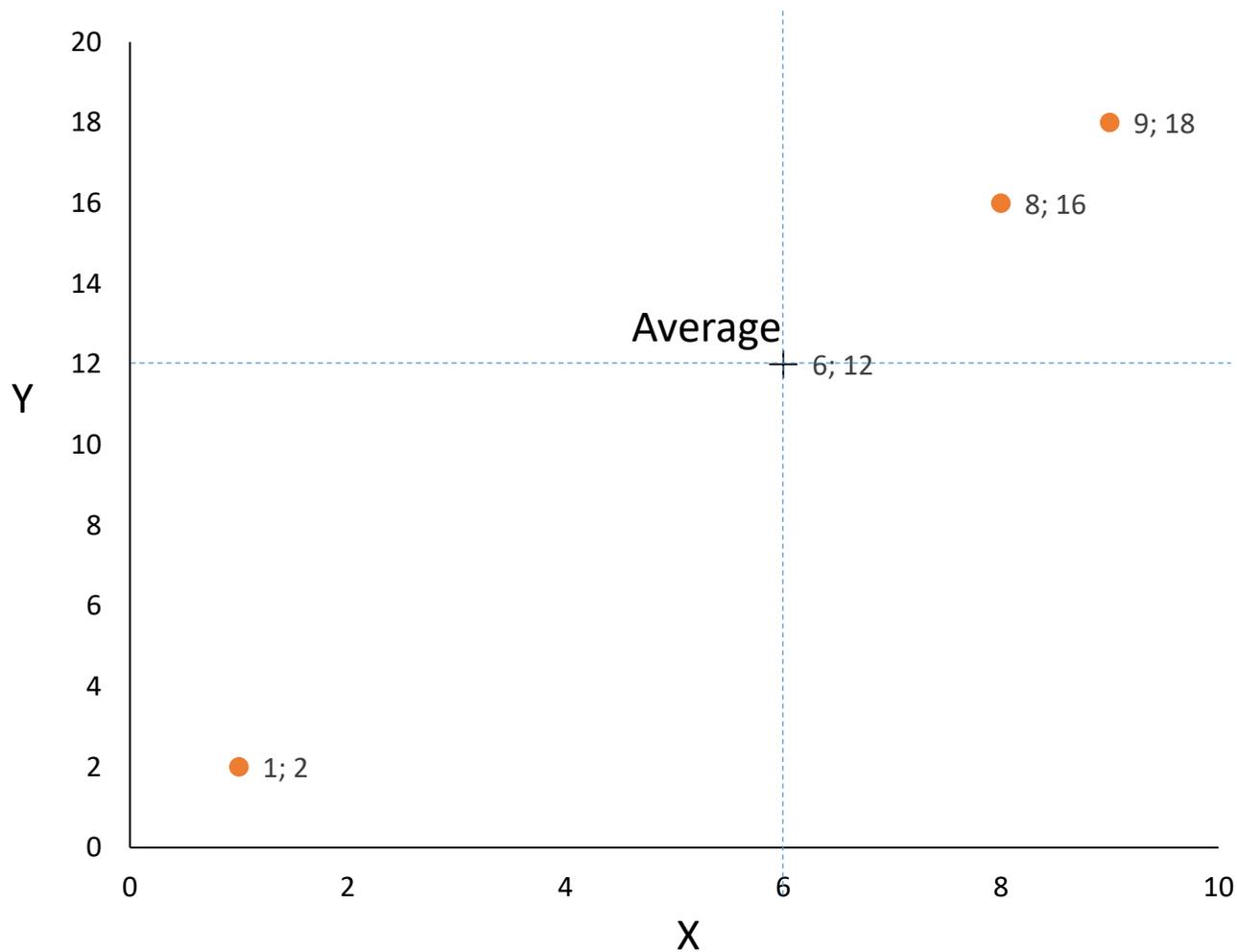


Covariance = shared variance

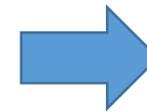


How to describe the relationship of variables num

Calculation of covariance II



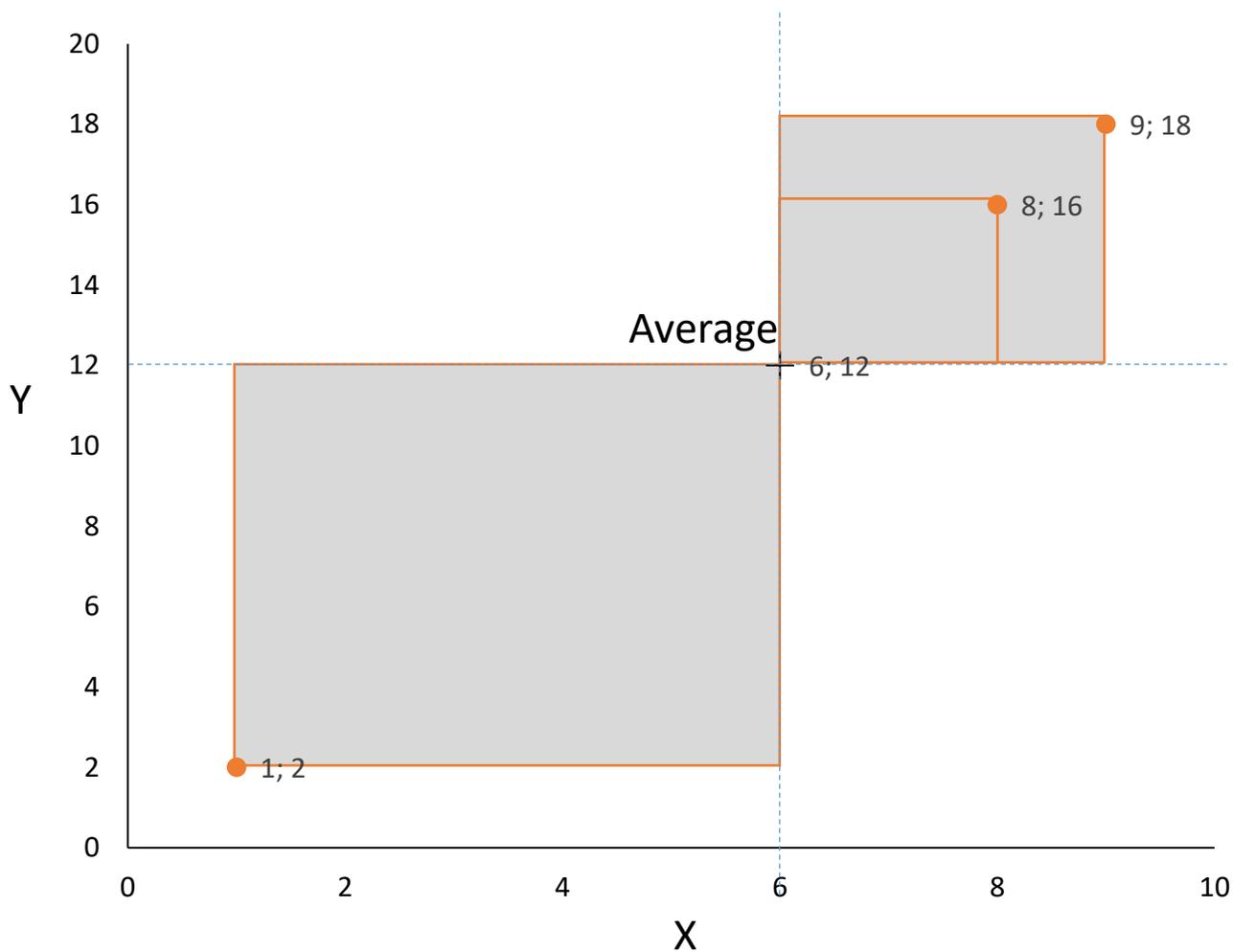
Covariance = shared variance



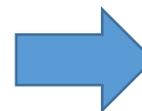
How to describe the relationship of variables numerically?

Data occur in different quadrants according to the mean !

Calculation of covariance III



Covariance = shared variance

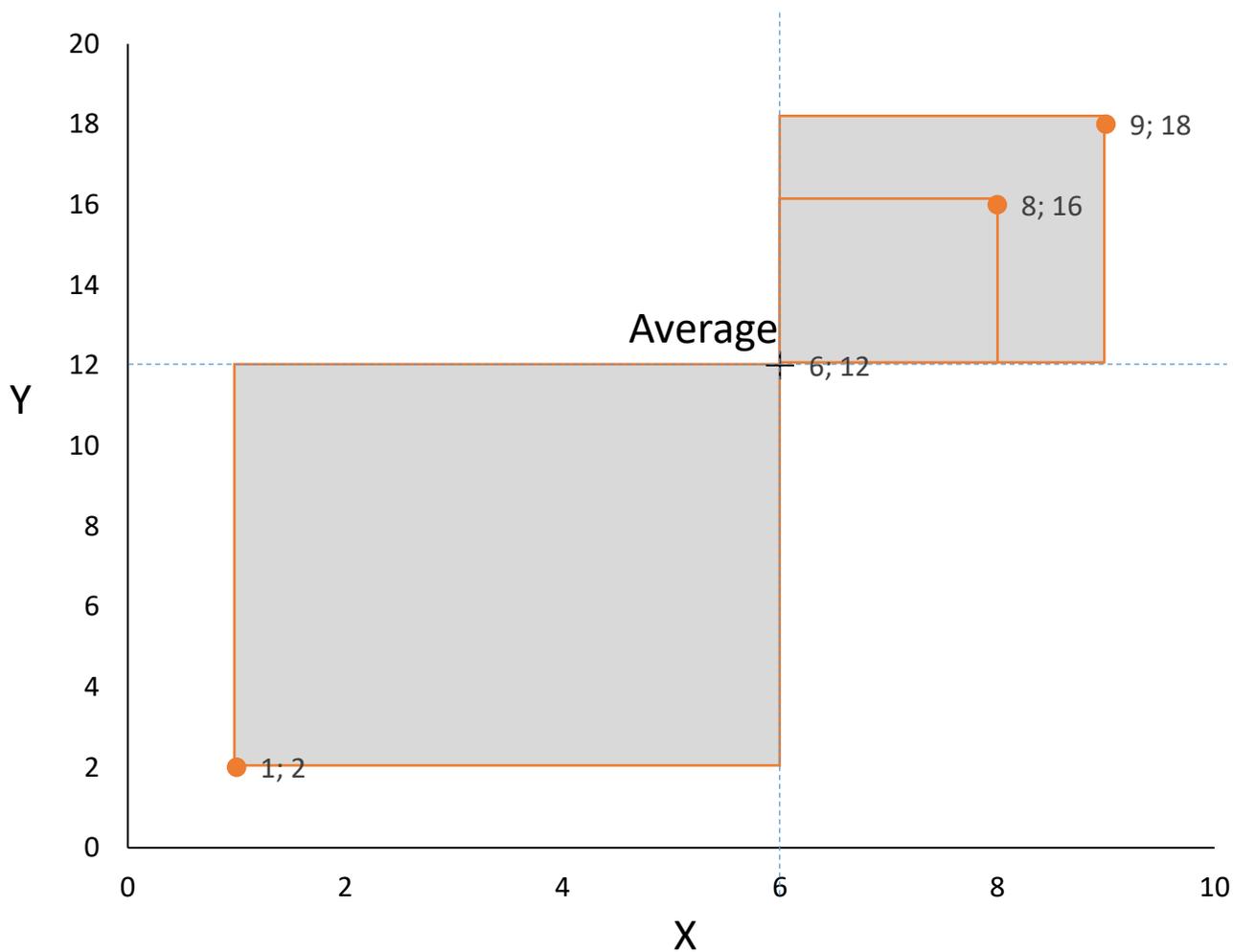


How to describe the relationship of variables numerically?

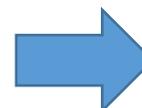
Data occur in different quadrants according to the mean !

Let us calculate the shared variance similarly to the variance of !!

Calculation of covariance IV



Covariance = shared variance



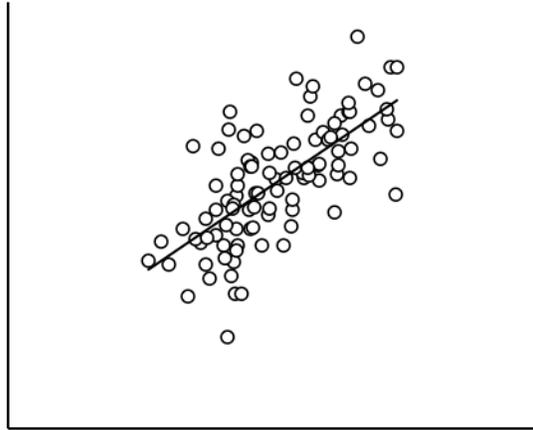
How to describe the relationship of variables numerically?

Data occur in different quadrants according to the mean !

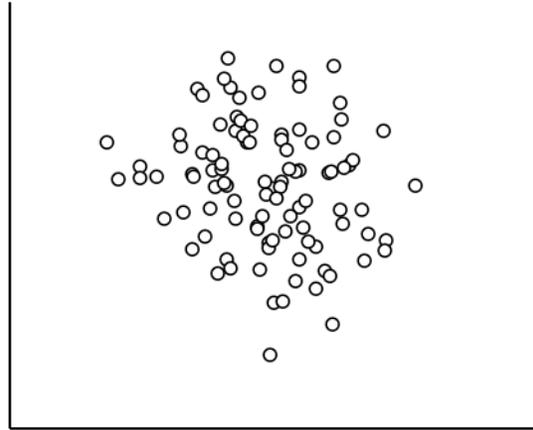
Let us calculate the shared variance similarly to the variance of !!

$$Cov(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x}) * (y_i - \bar{y})}{N - 1}$$

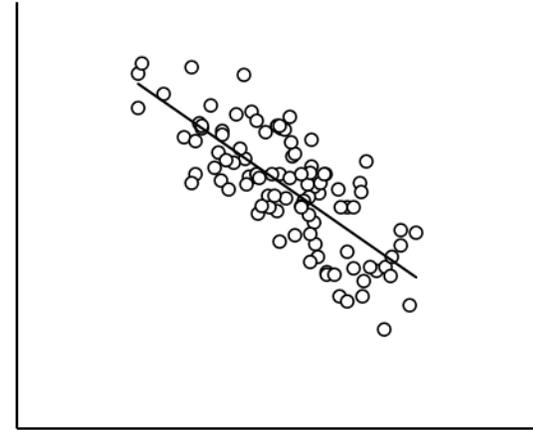
Calculation of covariance IV



Cov = ?

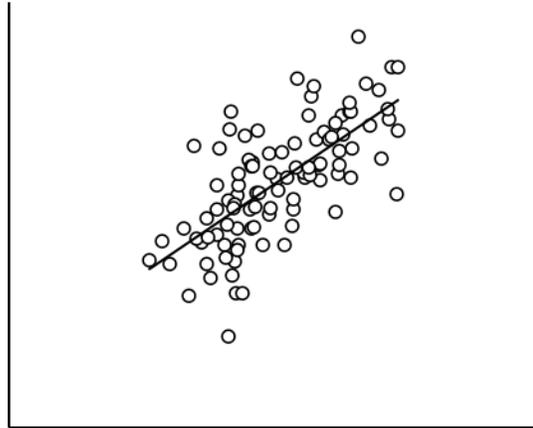


Cov = ?

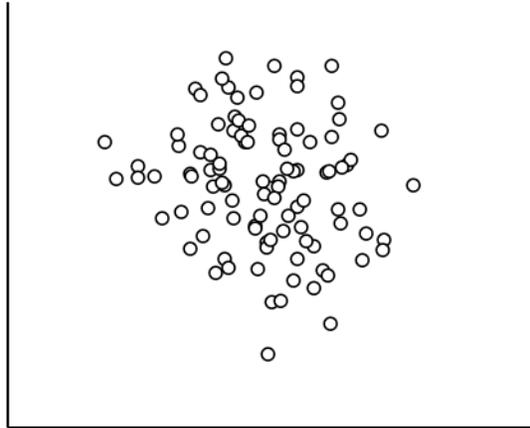


Cov = ?

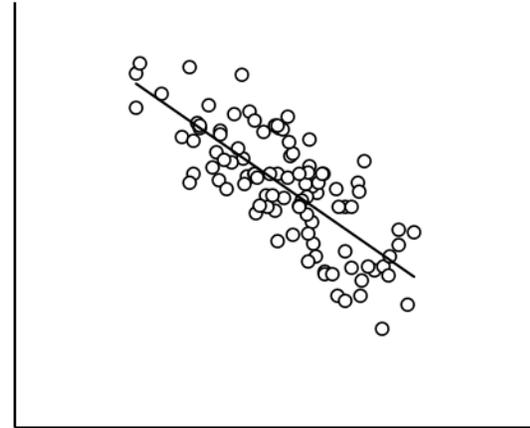
Calculation of covariance IV



Cov = positive number



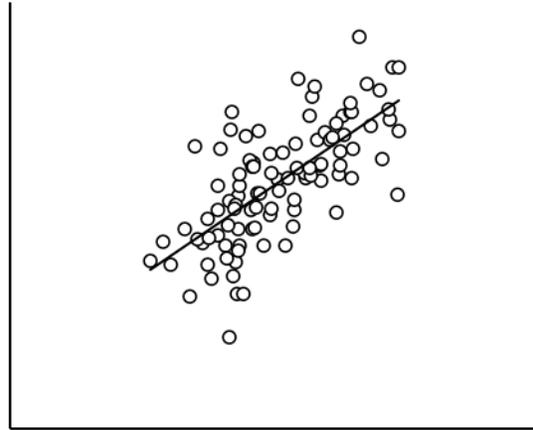
Cov = 0



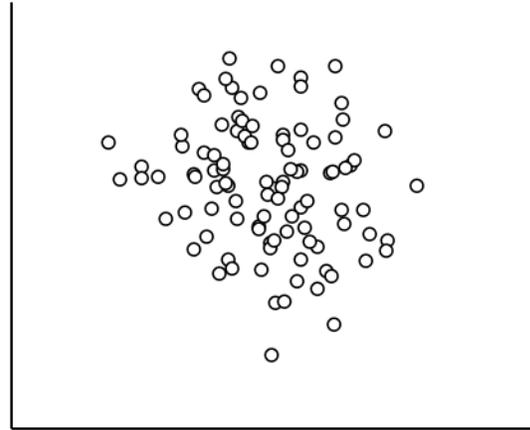
Cov = negative number

Is there a given minimum and maximum covariance?

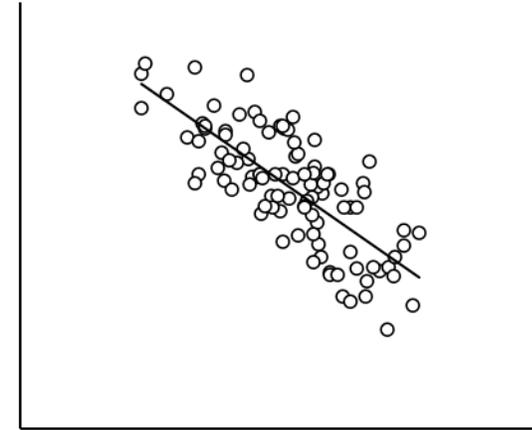
Calculation of covariance IV



Cov = positive number



Cov = 0



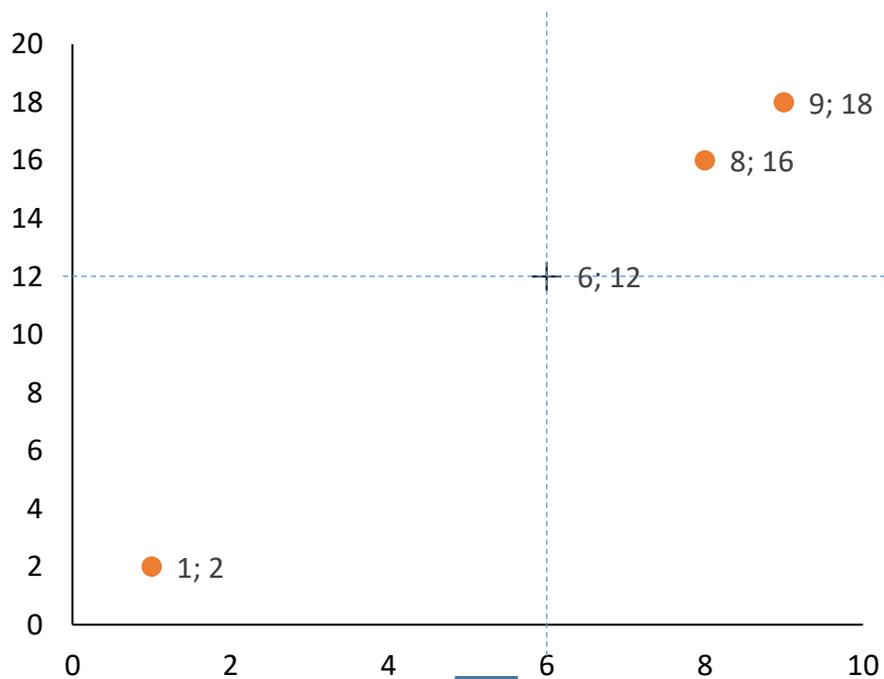
Cov = negative number

Is there a given minimum and maximum covariance?

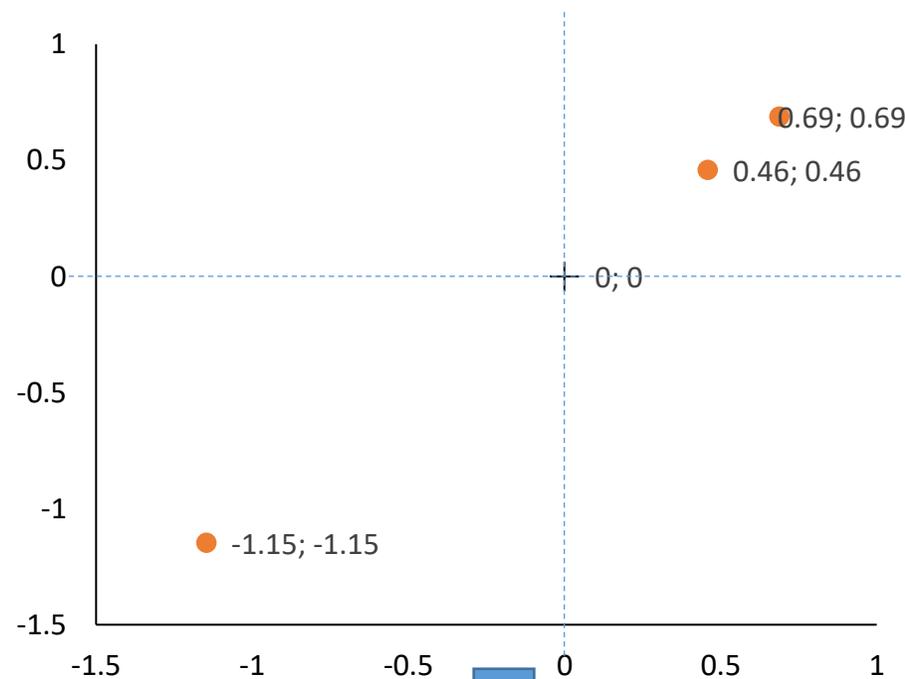
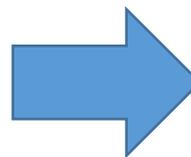
None, theoretically there can be covariance from $-\infty$ to $+\infty$; disadvantage in interpretation

Covariance of standardised data

- How does the covariance calculation work on data with a standard normal distribution (mean = 0, variance = 1)?



Cov = 38

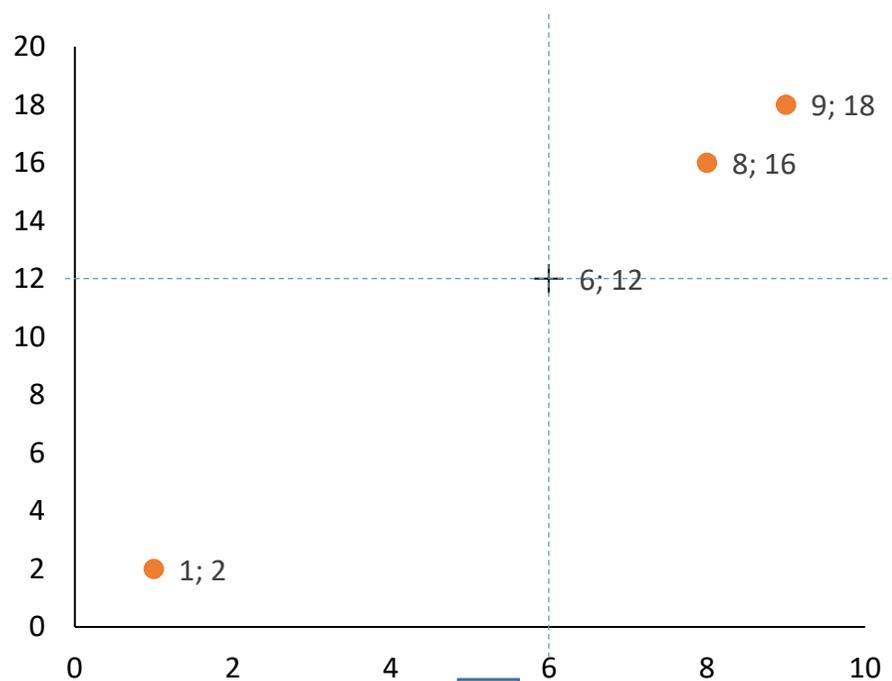


Cov = 1

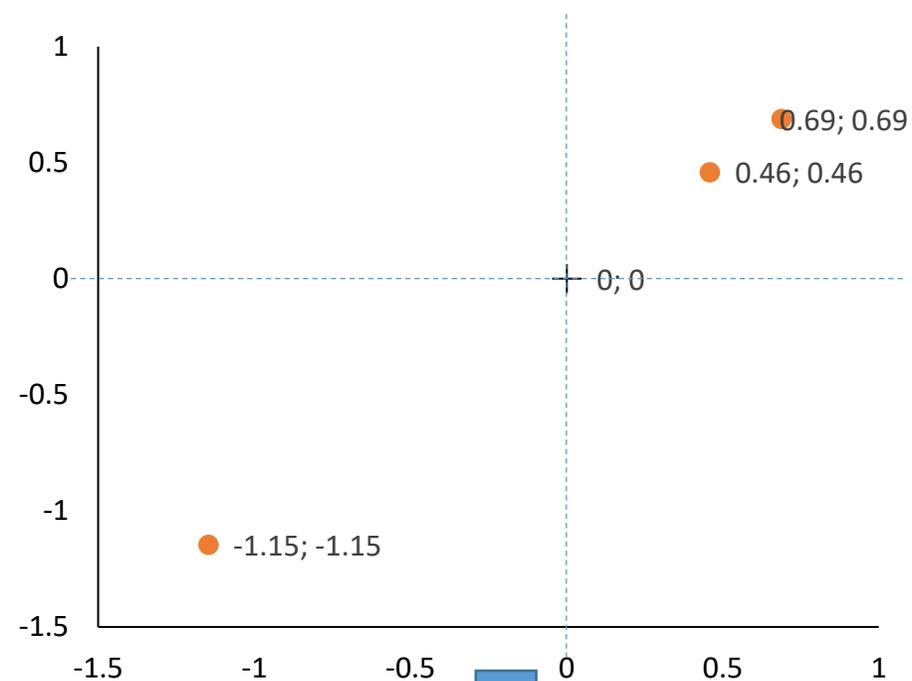
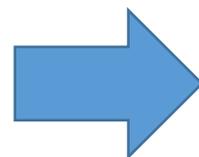
Calculation of Pearson correlation coefficient

- Pearson correlation coefficient represents a standardized form of covariance

$$r(x, y) = \frac{Cov(x, y)}{S_x S_y}$$

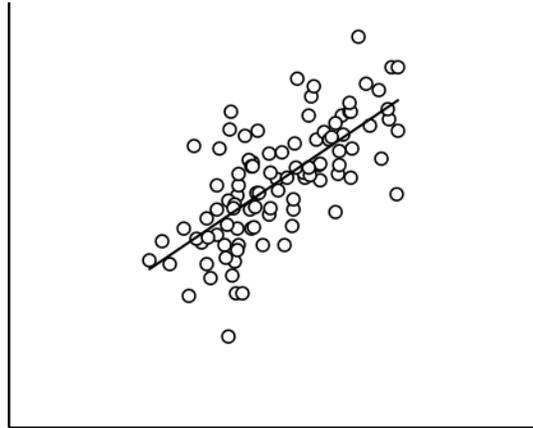


Cov = 38; r = 1

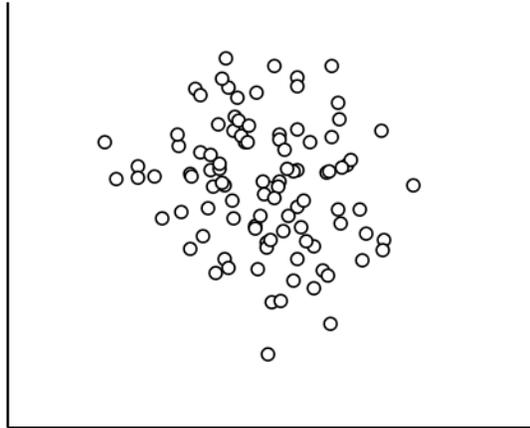


Cov = 1; r=1

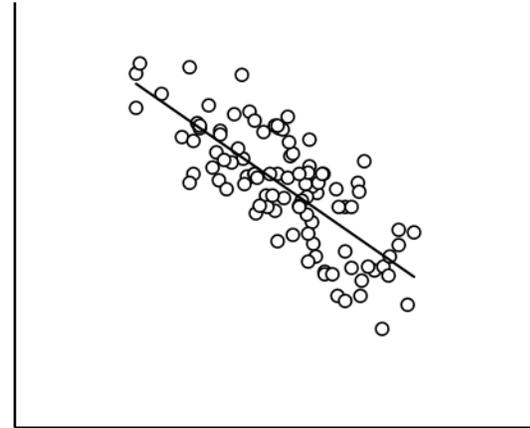
Calculation of Pearson correlation coefficient



$r = \text{positive number} \leq 1$



$r = 0$



$r = \text{negative number} \geq -1$

Is there a given minimum and maximum of the Pearson correlation coefficient?

Yes, the Pearson correlation coefficient is in the range $\langle -1; 1 \rangle$

Pearson correlation coefficient testing

P_i (ground)	10	14	15	32	40	20	16	50
P_i (rostl.)	19	22	26	41	35	32	25	40

$I = 1, \dots, n; n = 8; v = 6$

$$r = \frac{Cov(x, y)}{S_x \cdot S_y} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \right] \left[\sum y_i^2 - \frac{1}{n} (\sum y_i)^2 \right]}} = 0,7176$$

I. $H_0 : \rho = \phi : \alpha = 0,05$

tab : $r(v=6) = 0,7076$

II. $H_0 : \rho = \phi$

$$t = \left[\frac{r}{\sqrt{1-r^2}} \right] \cdot \sqrt{n-2} \quad v = n-2$$

$$t = \frac{0,7176}{0,6965} \cdot \sqrt{6} = 2,524 \left. \vphantom{t} \right\} P \leq 0,05$$

tab : $t_{0,975}^{(n-2)} = 2,447$

Comparison of two correlation coefficients (r)

1. $n_1 = 1258$
 $r_1 = 0,682$

2. $n_2 = 462$
 $r_2 = 0,402$

Blood pressure x oxygen radical

$$Z_i = 1.1513 \cdot \log \frac{(1 + r_i)}{(1 - r_i)}$$

$Z_1 = 0,833$



$Z_2 = 0,426$

Test : $H_0 : \rho_1 = \rho_2 ; \alpha = 0,05$

$$Z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0,407}{0,0545} = 7,461$$

tabulky : $Z_{0,975} = 1,96$

$7,461 \gg 1,96 \Rightarrow P \ll 0,01$

Non-parametric correlation (Spearman's correlation coefficient - r_s)

P_i in soil	1	2	3	6	7	5	4	8
P_i in rossl.	1	2	4	8	6	5	3	7
d_i	0	0	1	2	-1	0	-1	-1

$$i = 1, \dots, n; n = 8 \Rightarrow v = 6$$

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)} = 0,9048$$

$$\text{tab : } r_s(v = 6) = 0,89$$

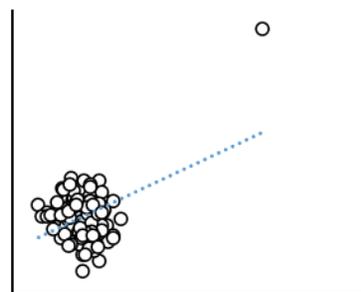
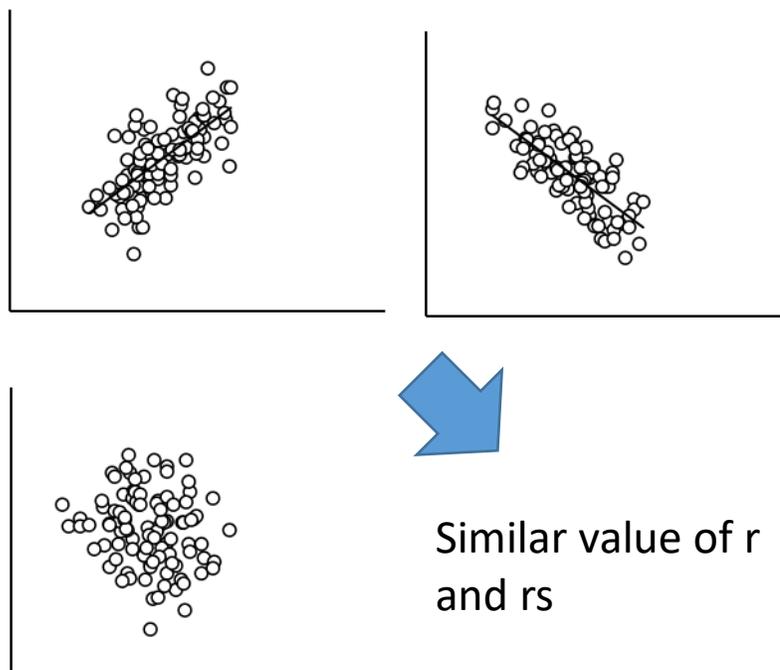
Patient no.	1	2	3	4	5	6	7
Doctor 1	4	1	6	5	3	2	7
Doctor 2	4	2	5	6	1	3	7
d_i	0	-1	1	-1	2	-1	0

$$r_s = 1 - \frac{6 \cdot 8}{7(49 - 1)} = 0,857$$

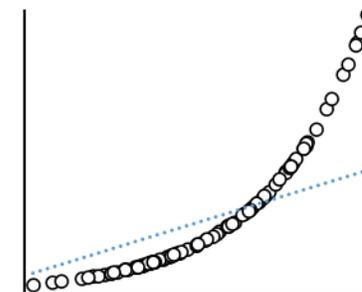
P = 0,358

Pearson and Spearman correlation coefficient

- Comparison of Pearson (r) and Spearman (r_s) correlation coefficient values allows to assess the type of relationship between variables

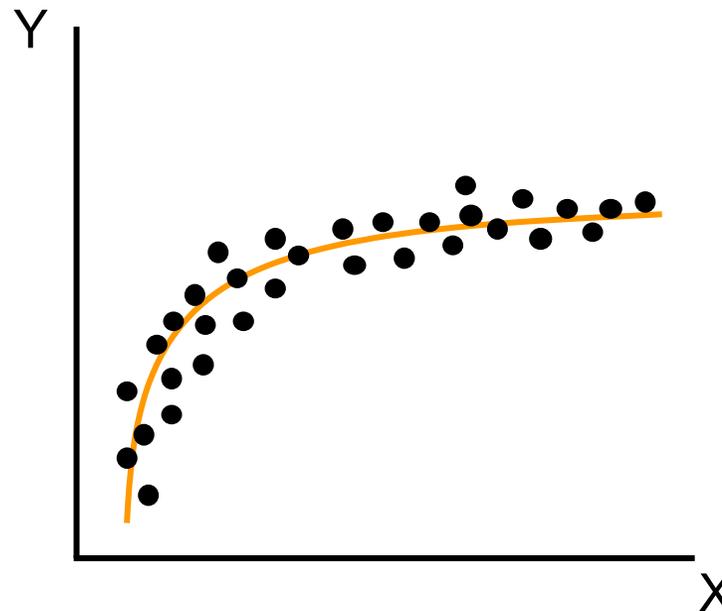
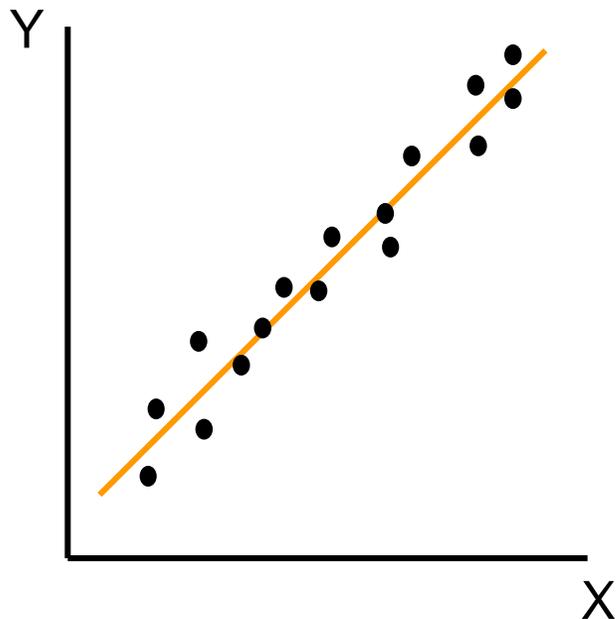


High r (due to outlier) and low r_s (outlier removed by transformation to order)



Low r (due to the non-linearity of the relationship) and high r_s (in the order of the strong relationship between the two variables)

Correlation in graphs I.



Relationships very often imply a functional relationship between Y and X.

$$Y = a + b \cdot X$$

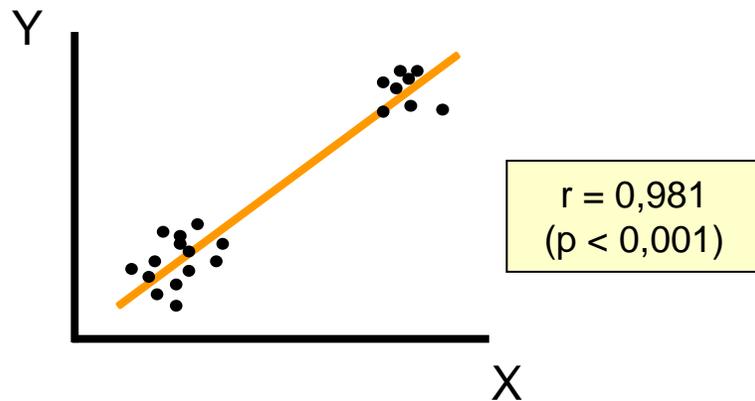
$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3$$

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2$$

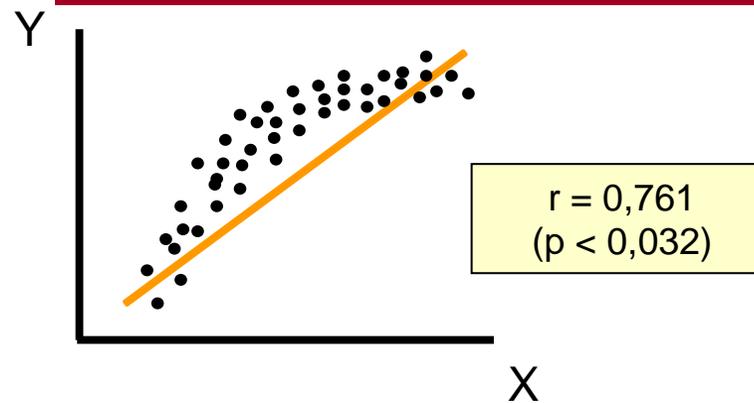
$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_1 \cdot X_2$$

Correlations in graphs II.

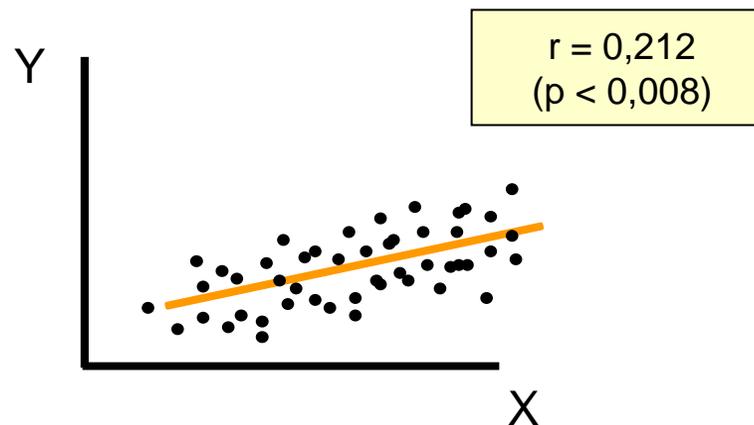
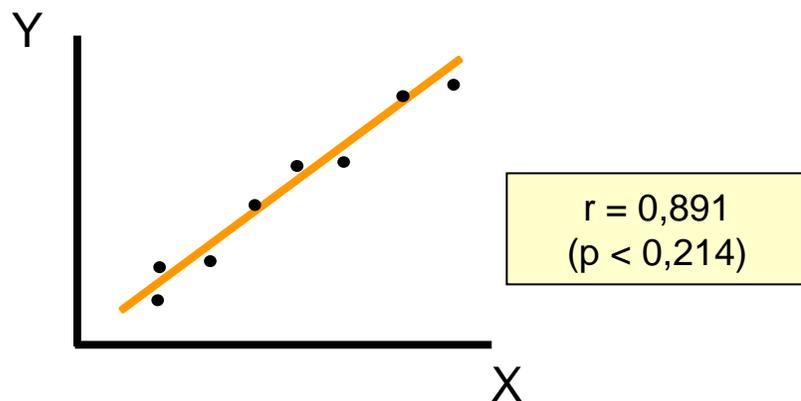
The problem of value distribution



Model type problem

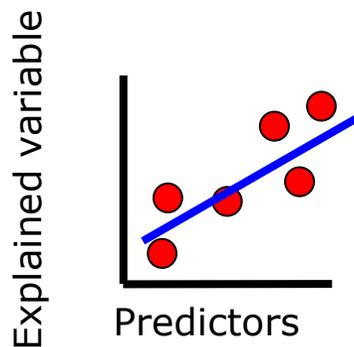
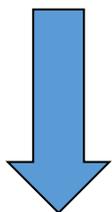


Sample size problem



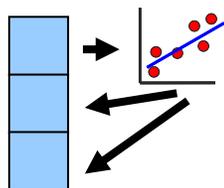
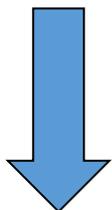
Creating models

1. Creation of the model



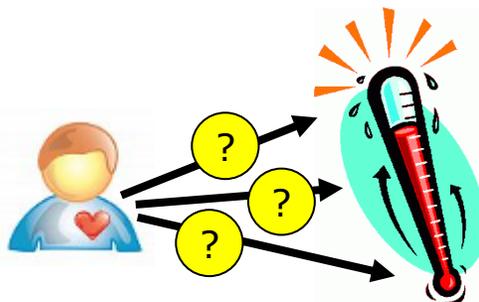
- Parameters influencing the explained patient characteristics
- Equations allowing prediction
- Validity of the model only in the range of predictors

2. Model validation



- Danger of "overlearning" the model
- Testing the model on known data
- Cross-validation

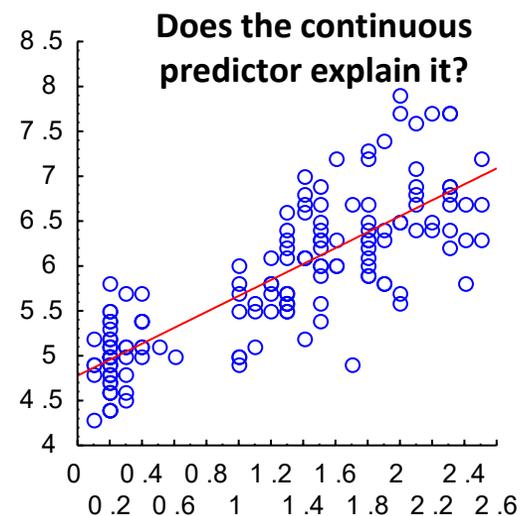
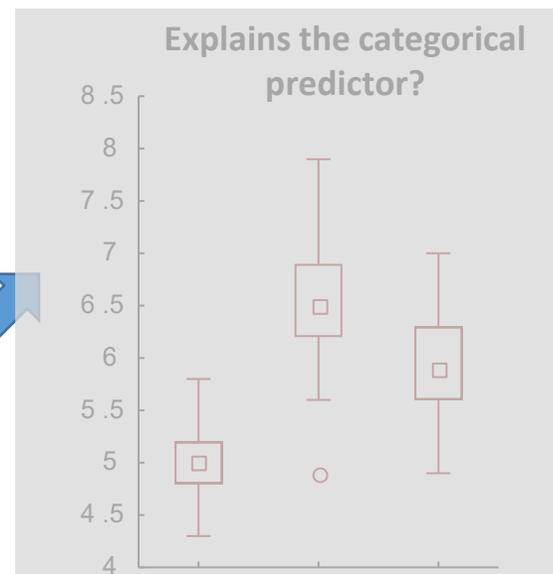
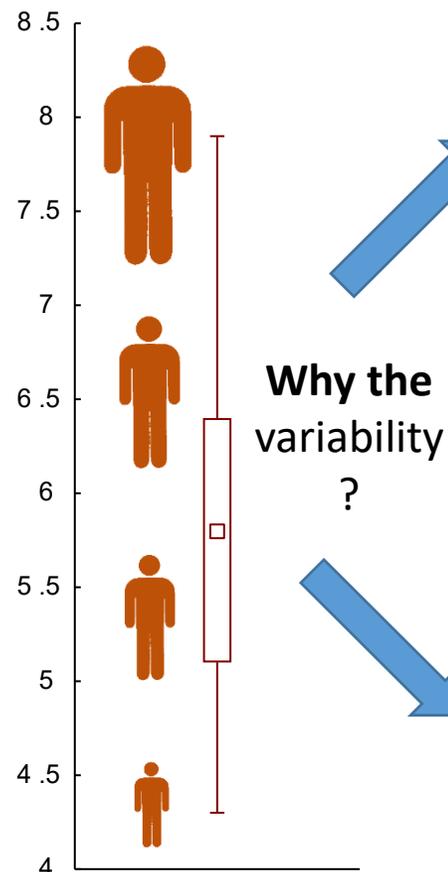
3. Application of the mode



- Individual prediction of the condition of non-small patients
- The model must be supported by correct statistics and extensive data

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



Basics of regression analysis

- Regression - a functional relationship between two or more variables

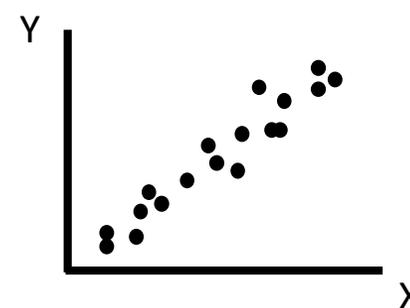
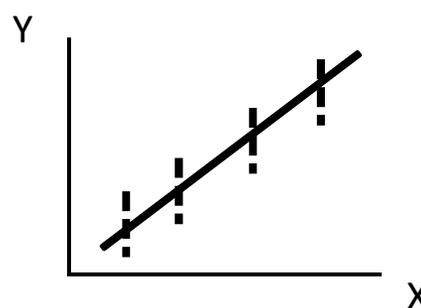
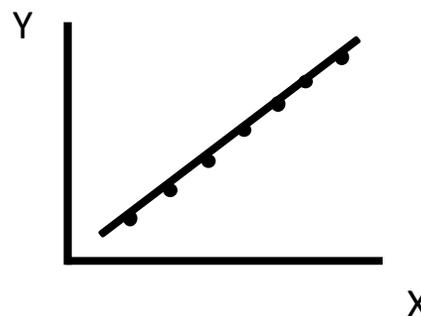
One-dimensional
 $y = f(x)$

Multidimensional
 $y = f(x_1, x_2, x_3, \dots, x_p)$

Relationship
between x , y

Deterministic

Regression, stochastic



For each x there is a probability distribution y

Linear regression I

$$Y = a + b \cdot x + e \quad \approx \quad \alpha + \beta \cdot X + \varepsilon$$

y — $\alpha \approx a$ (**intercept**): $a = \bar{y} - b \cdot \bar{x}$

y — $\beta \cdot X \approx b \cdot x$ (**sklon; slope**)

y — $\varepsilon \approx e$ - **náhodná složka**: $N(0; \sigma_e^2) = N(0; \sigma_y^2 x)$

The components forming y are added together

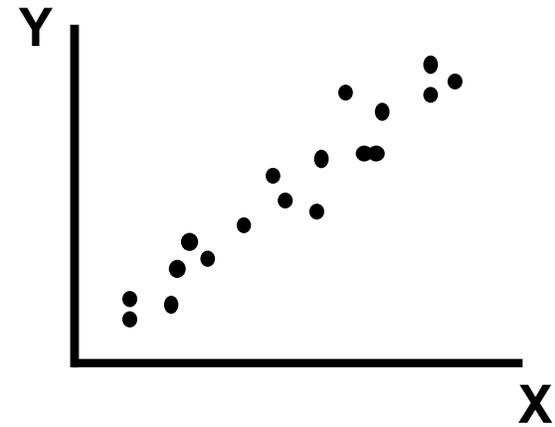
ε - random component of the line model = residuals of the line

$$\sigma_e^2 (\sigma_{y \cdot x}^2) \Rightarrow \text{rozptyl reziduí}$$

Linear Regression II

$$\begin{matrix} 1 \\ \vdots \\ n \end{matrix} \quad \mathbf{x} \quad \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

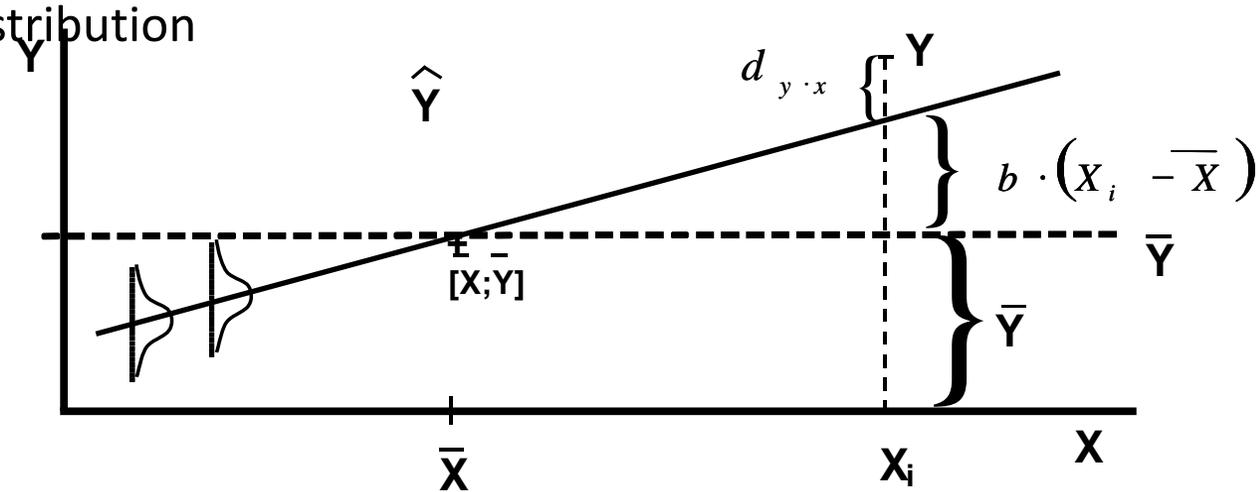
$$\begin{matrix} 1 \\ \vdots \\ n \end{matrix} \quad \mathbf{y} \quad \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$



$$\begin{matrix} 1 \\ \vdots \\ n \end{matrix} \quad \hat{\mathbf{y}} \quad \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} = \mathbf{a} + \mathbf{b} \cdot \begin{matrix} \mathbf{x} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix} \quad \longrightarrow \quad \begin{matrix} \mathbf{y} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix} - \begin{matrix} \hat{\mathbf{y}} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix} = \begin{matrix} \mathbf{e} \\ \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \end{matrix}$$

Linear regression IV

- Least squares method
 - X: Fixed, non-stochastic variable
 - The distribution of y values for each x is normal
 - The distribution of y values for each x has the same variance
 - The residues are independent of each other and have a normal distribution



$$d_{y \cdot x} = y - \hat{y} \quad \boxed{d_{y \cdot x} = y - \bar{y} - b(X_i - \bar{X})} \quad \hat{y} = \bar{y} + b(X_i - \bar{X})$$

The meaning of line interlacing
minimisation of deviations

$$d_{y \cdot x}^2 \rightarrow \sum [y - \hat{\alpha} - \hat{\beta}(X_i - \bar{X})]^2$$

Linear regression V

I. $b \sim \beta: b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$ $S_b^2 \sim \sigma_\beta^2: \frac{1}{\sum (X_i - \bar{X})^2} \cdot S_{y \cdot x}^2$

$S_{y \cdot x}^2 =$ **mean squared deviation from regression**

$S_{y \cdot x} =$ **sample standard deviation from regression**

$$S_{y \cdot x}^2 = \frac{\sum d_{y \cdot x}^2}{n-2} = \frac{\sum Y_i^2 - \frac{\sum Y_i^2}{n} - b^2 \cdot \sum (X_i - \bar{X})^2}{n-2}$$

II. $a \sim \alpha: a = \bar{Y} - b \cdot \bar{X}$ $S_a^2 \sim \sigma_\alpha^2$ $S_a^2 = \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum X^2} \right] \cdot S_{y \cdot x}^2$

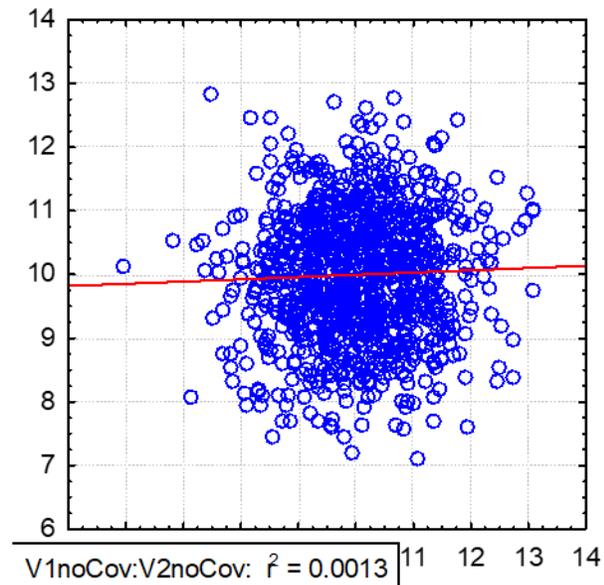
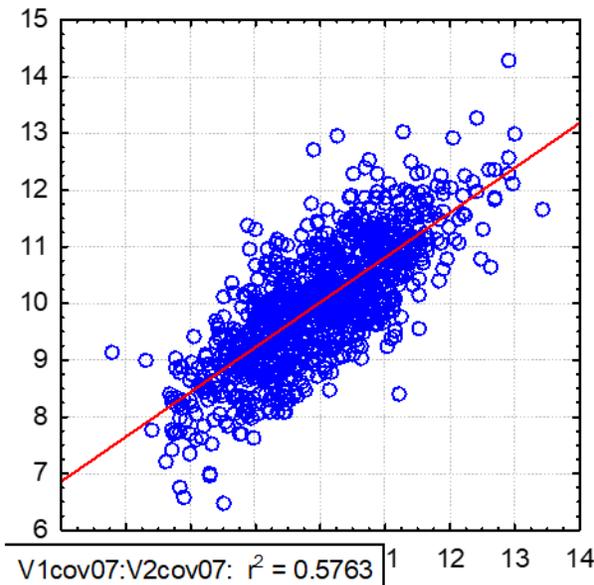
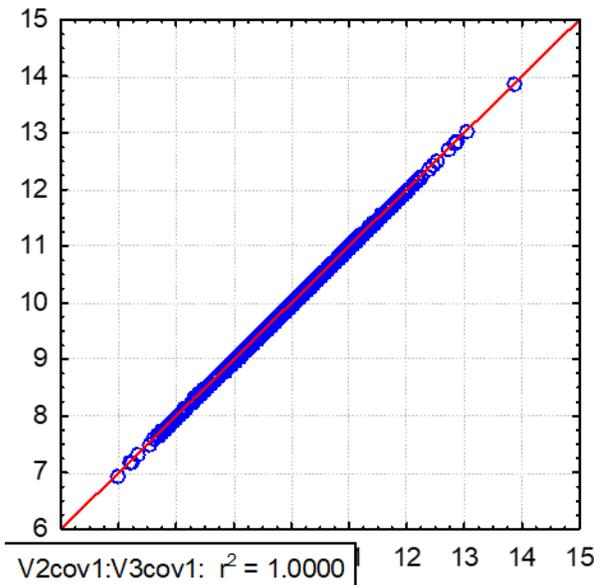
intercept

III. \hat{Y} : **model value**

$$\hat{Y}_i = a - b \cdot X_i$$
$$S_{\hat{y}_i} = (S_{y \cdot x}) \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum X^2}}$$

Exhausted variability and its statistical significance

- A basic indicator of model quality is the amount of variability explained by the model
- It is generally denoted by R^2 and is reported as a percentage or proportion of the total variability (in the case of linear regression, it is the Pearson correlation coefficient squared)
- Statistical significance of the calculated variability can be tested by analysis of variance



Analysis of variance in regression

- Calculation of the statistical significance of the variance extracted by the regression model

Overall ANOVA

$$SS / SS_{BT} \text{ (variance ratio)}$$

$$MS / MS_{BE} = F$$

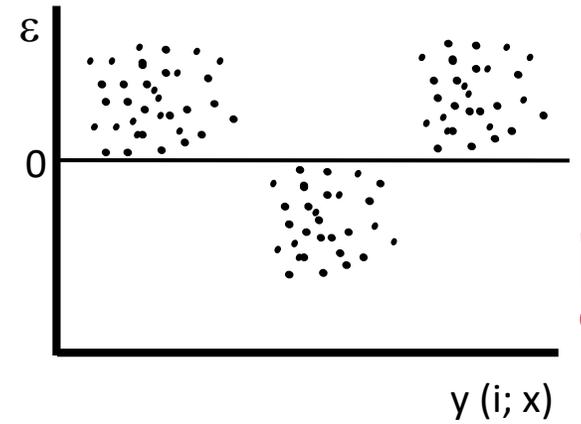
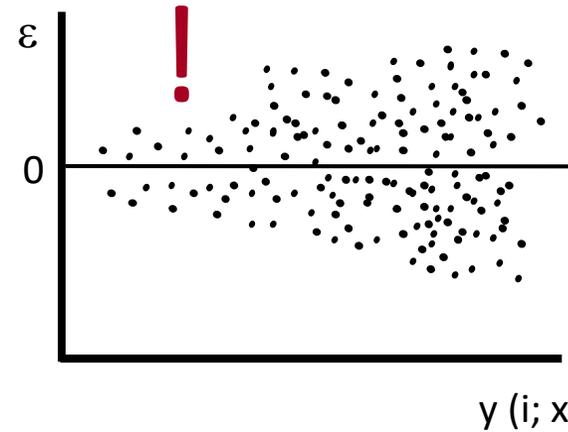
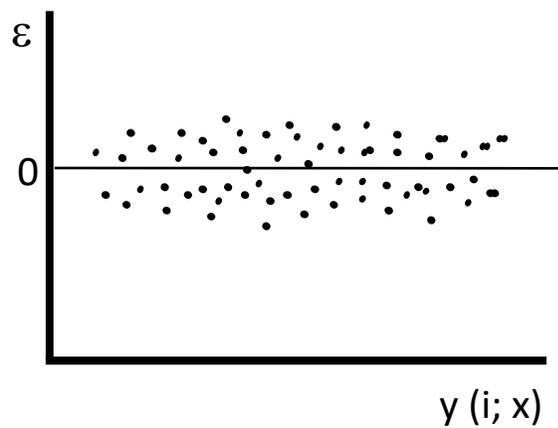
Analysis of variance of the regression model (straight lines here)

Source of dispersion	st.v.	SS	MS	F
Model (straight line)	1	SS_{MOD}	MS_{MOD}	MS_{MOD} / MS_R
Residuum	$on - 2$	SS_R	MS_R	
Total	$on - 1$	SS_T		

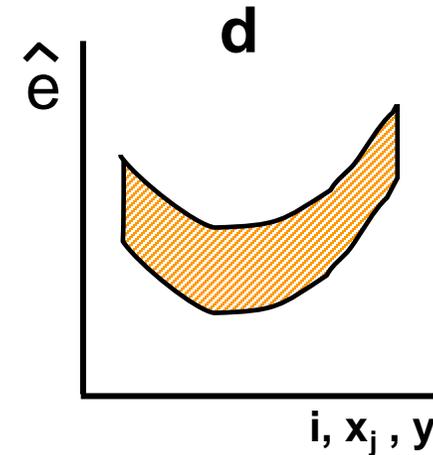
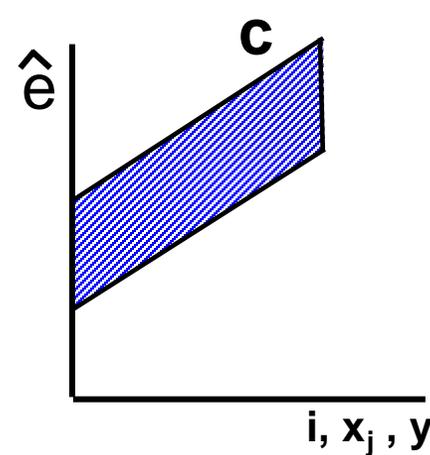
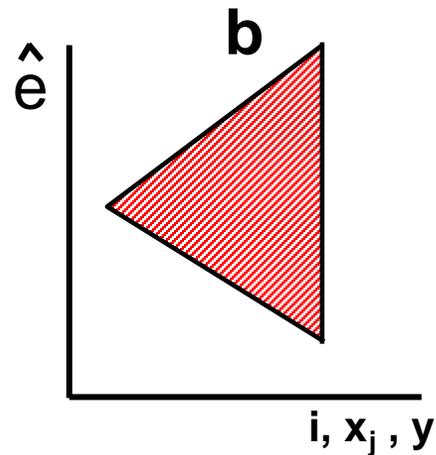
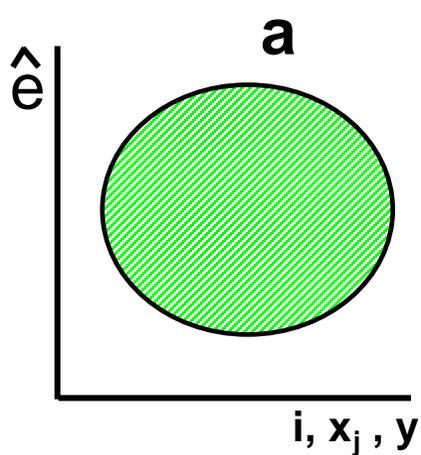
**$(SS / SS_{MODT}) \cdot 100 =$
% of the variance of
Y "drained" by the
line = coefficient of
determination $(R)^2$**

Linear regression: analysis of residuals

Graphs of model residuals (examples)



General shapes of model residues (diagram)

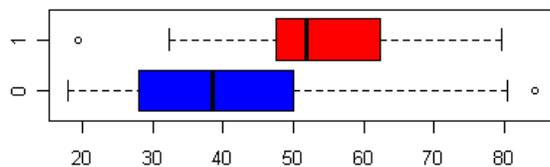


Adjustment of variables for the influence of other variables

1. In the first step, we define a regression model of the relationship between age and the adjusted parameter
2. For each patient, its residual from the regression line is calculated $\uparrow\downarrow$
3. The residual (representing the value of the parameter minus the effect of age, its average is 0) is added to the average value of the parameter
4. The resulting adjusted value has the effect of age subtracted, but at the same time the numerical value of the parameter is not changed

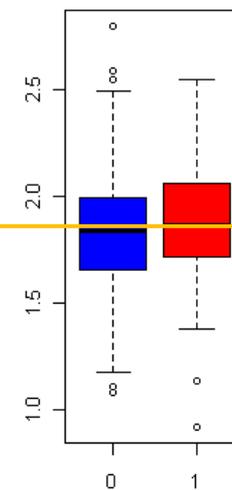
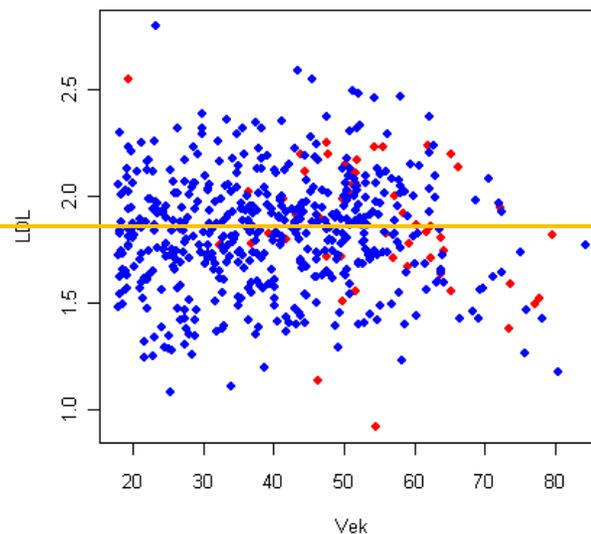
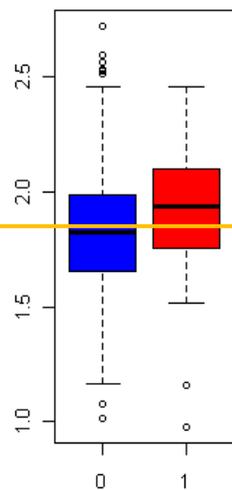
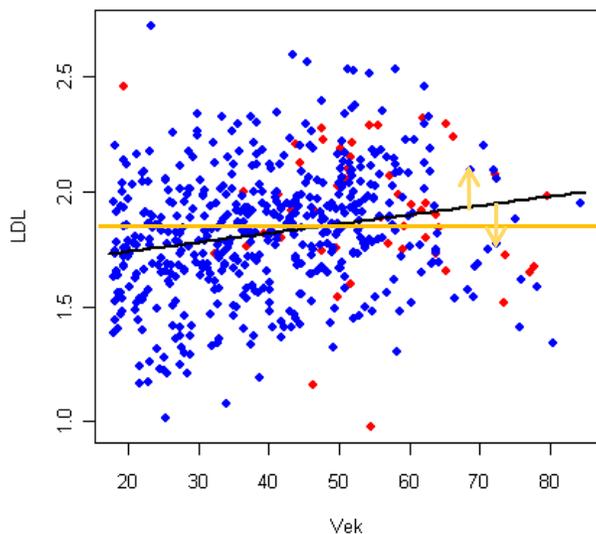
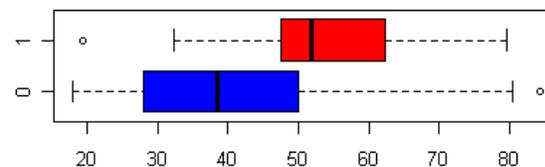
Original data

Vek



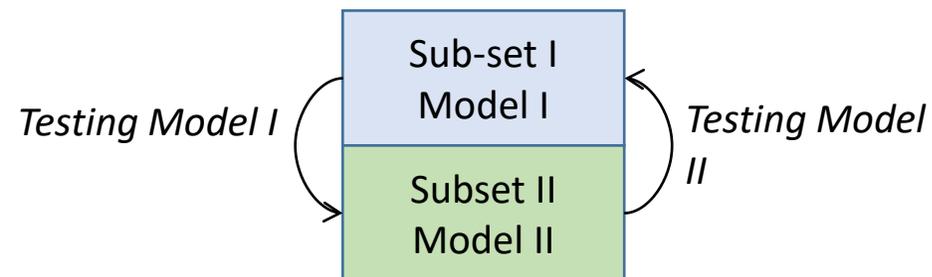
Adjusted data

Vek



Model validation on an independent set

- When creating models, a problem can occur when the created model is perfectly "trained" to solve a given problem on the data set on which it was created
- For this reason, it is problematic to test the results of the model on the same dataset on which it was created -> this is a proof by circle
- The solution is to test the results of the model on a set of known outcomes (here, known object groupings) that did not participate in the model definition
 - Cross-validation
 - the data file is randomly divided into several subfiles (2 or more)
 - A model is built on one subset and its results tested on the remaining subsets
 - The calculation is performed sequentially on all sub-files
 - One out leave out
 - The model is created on the whole file without one object
 - the model is tested on this object
 - the procedure is repeated for all objects
 - Permutation methods
 - Jackknife, bootstrap - the model is gradually created on random subsamples of the file and tested on the rest of the data

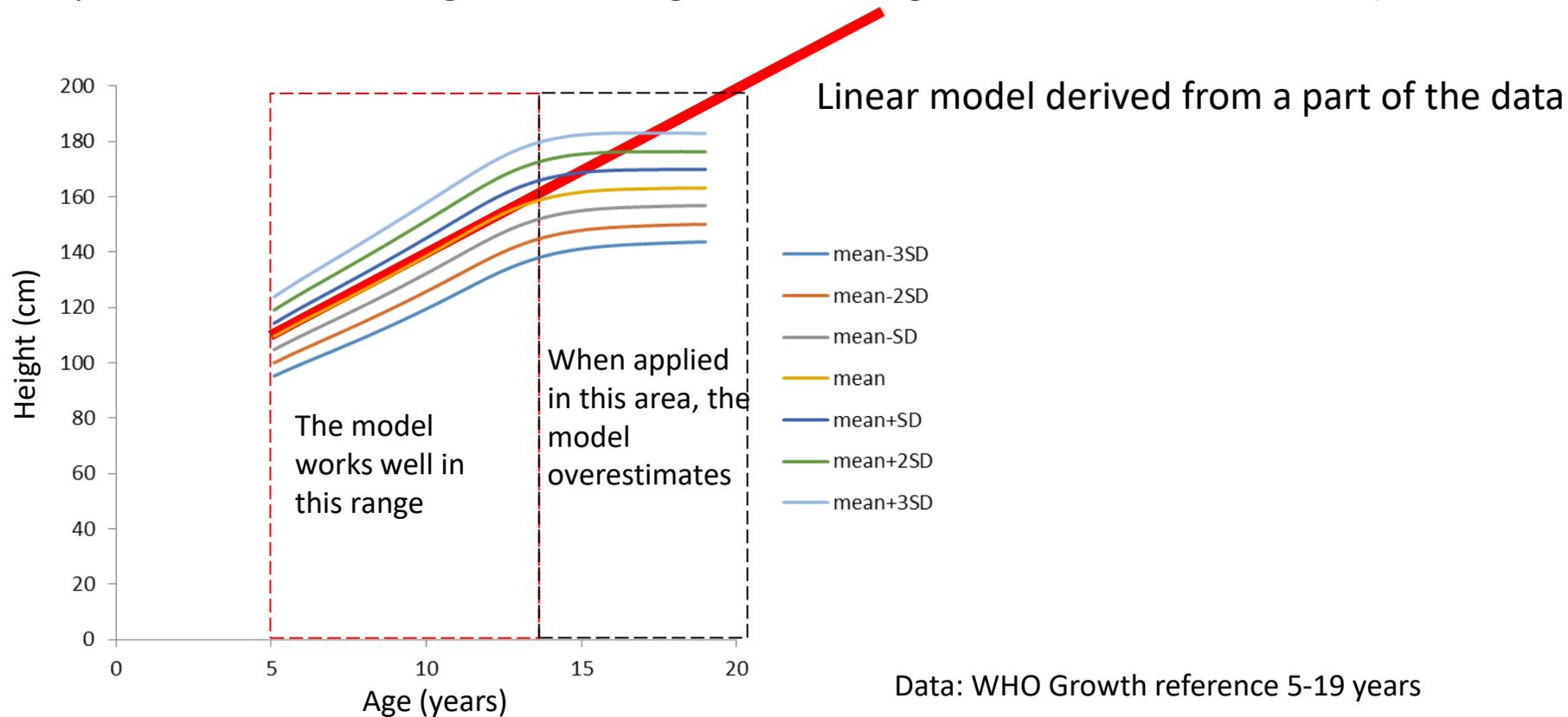


Statistical significance vs. practical use of the model

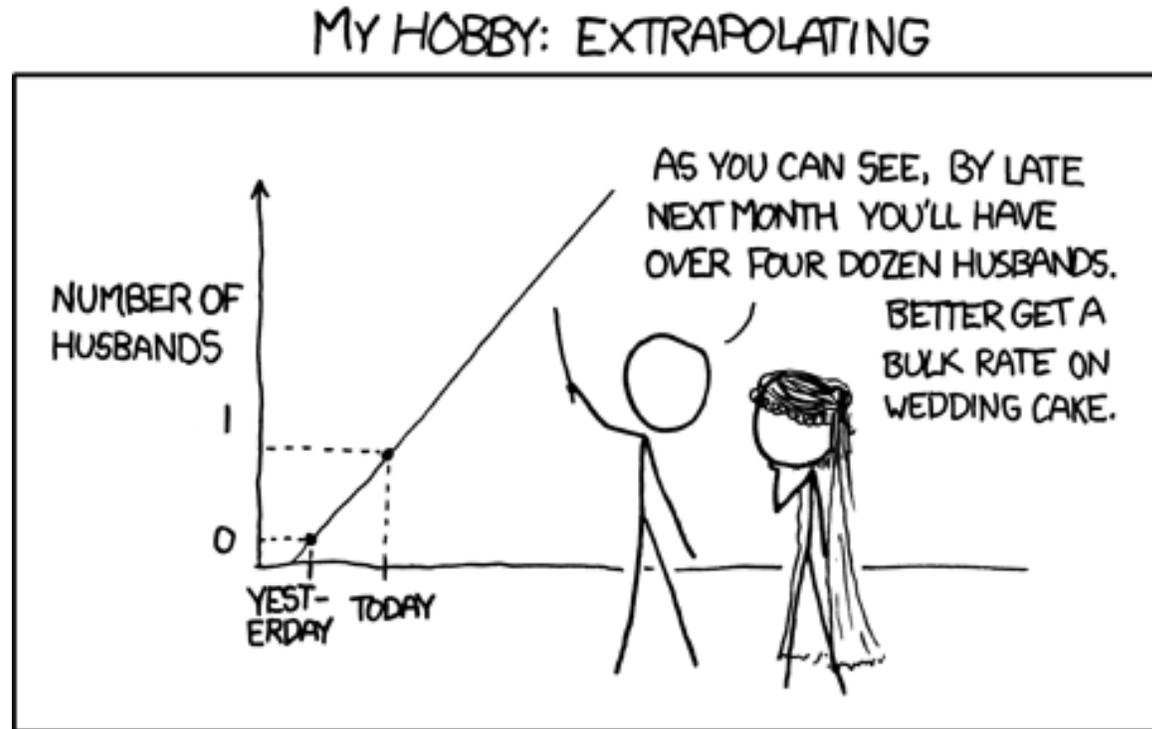
- When applying the model in practice, it is necessary to take into account both the statistical significance found and the practical significance of the model outputs
- This is an analogy to statistical vs. practical significance of differences in e.g. t-test
- Statistical significance = relationship between variables, the difference between groups is not a mere coincidence (or the probability that it is not a coincidence is low enough)
- Practical significance of the model
 - In terms of predictors: the change in the predicted value when the predictor changes is practically significant (e.g. the magnitude of the increase in blood pressure when age changes by 10 years)
 - In terms of objects: individual patient prediction is accurate enough to be practically usable (prediction of various events - hospitalization, death, development of complications, treatment outcome, etc.)

Scope of applicability of the model

- Models can only be applied within the range of predictors on which they were developed
- This is due to our ignorance of the behaviour of the relationships between predictors and the predicted variable outside the boundaries within which the model was defined (typical examples are e.g. dose-response curves, child growth vs. age, bacterial growth vs. substrate, etc.).



Scope of applicability of the model: example



General principles of prediction model development

- Requirements for a quality prediction model
 - Maximum predictive power
 - Maximum interpretability
 - Minimum complexity
- Creation of models
 - Does not contain redundant variables
 - It is tested on independent data
- Selection of variables
 - Forward and backward elimination algorithms are only an auxiliary indicator in the selection of the variables of the final model
 - Both classical statistical methods (ANOVA) and expert knowledge of the meaning of variables and their substitutability are applied in the selection of variables

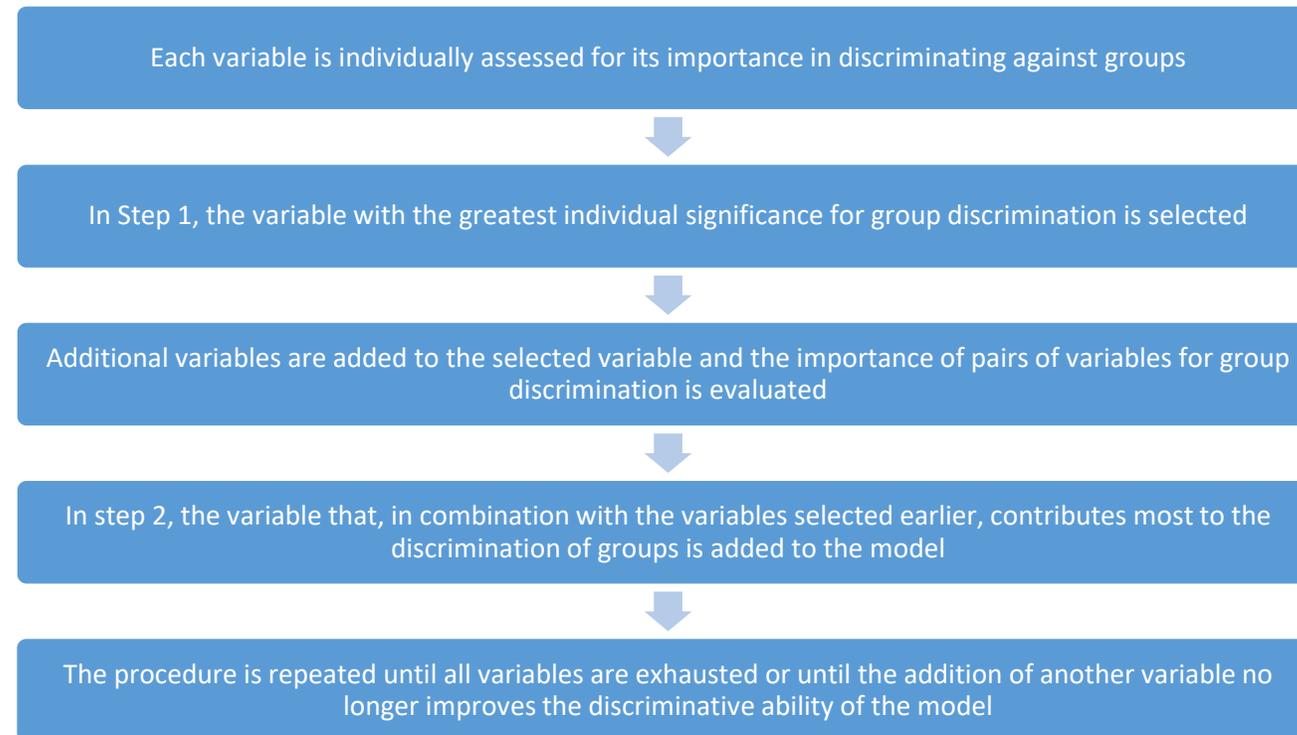
Forward and backward elimination

- Forward and backward stepwise elimination of variables from a model is a common technique used in regression, discriminant and other models
- Variables are gradually added (subtracted) to the model according to their importance in the model

Schematic of forward elimination of variables in the model

In the case of backward elimination, the process starts from a model with all variables and gradually eliminates the variables with the smallest contribution to the discriminatory power of the model

The process needs to be expertly controlled, e.g. the presence of redundant variables is risky



Steps of regression analysis

- Regression analysis (and other stochastic models in general) should proceed in the following steps
 1. Verification of general assumptions - normality of data, linearity of relationship
 2. Calculation of the model
 3. Analysis of model residuals to verify the suitability of applying a linear or other model
 4. Analysis of explained variability testing whether the model significantly explains the variability in the data
 5. Testing regression coefficients
 1. Assessment of the significance of model components
 2. Practical sensibility of the model
 6. Conclusion on the usability and meaningfulness of the model

Binary endpoint prediction

ROC analysis

Logistic regression

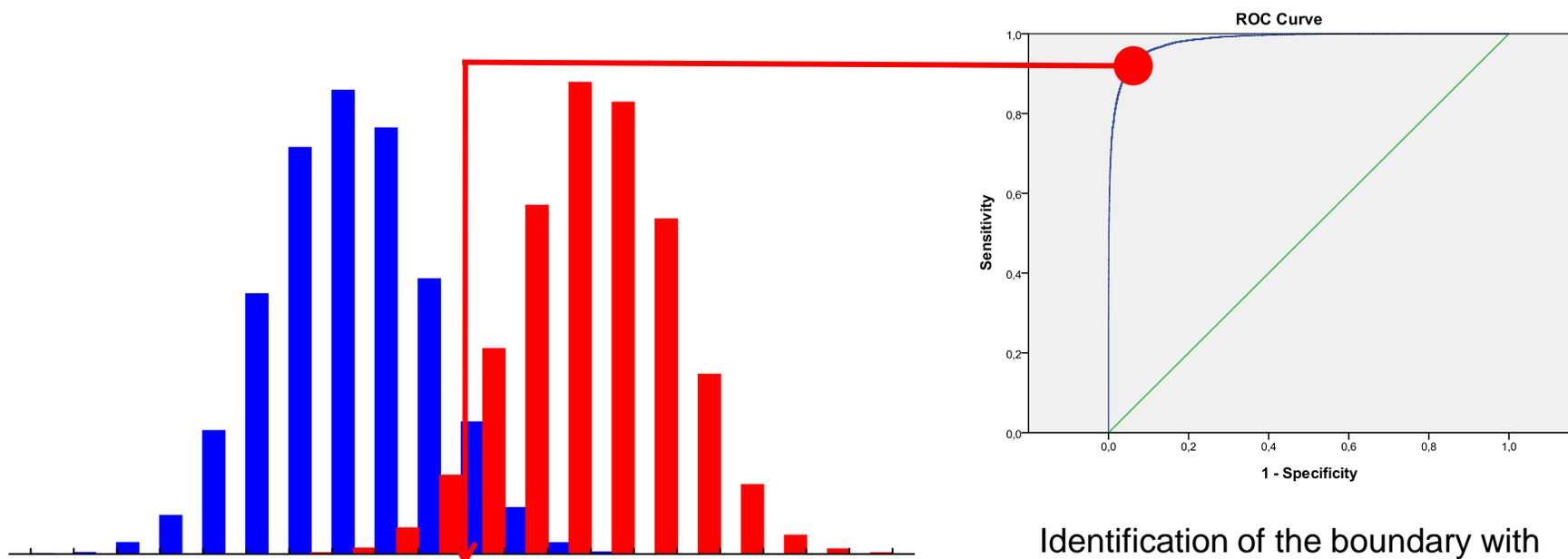
ROC analysis

- A tool for identifying the cut-off (boundary of the continuous data distribution) in continuous data with respect to the best possible binary endpoint distinction
- The result is a binarization of a continuous variable that is often more interpretable than results on continuous data
- The identification of a specific cut-off is related to a preference for either sensitivity or specificity for endpoint identification
- The preference for sensitivity or specificity is to some extent subjective to the real objective of the analysis
 - High sensitivity - a screening test where all possible patients need to be caught (e.g. severe disease that needs to be caught at an early stage)
 - High specificity - if it is necessary to catch only really sick patients (e.g. we don't want to expose patients to unnecessary treatment of a minor disease)

ROC analysis

- Identification of cutt offs for categorizing continuous variables to maximize their sensitivity and specificity when used in models

Where is the optimal boundary between the groups?



Identification of the boundary with the highest sensitivity and specificity for distinguishing groups

Sensitivity and specificity

- Key concepts in the description of the relationship between two binary variables = situation when we predict a binary endpoint with a binary predictor

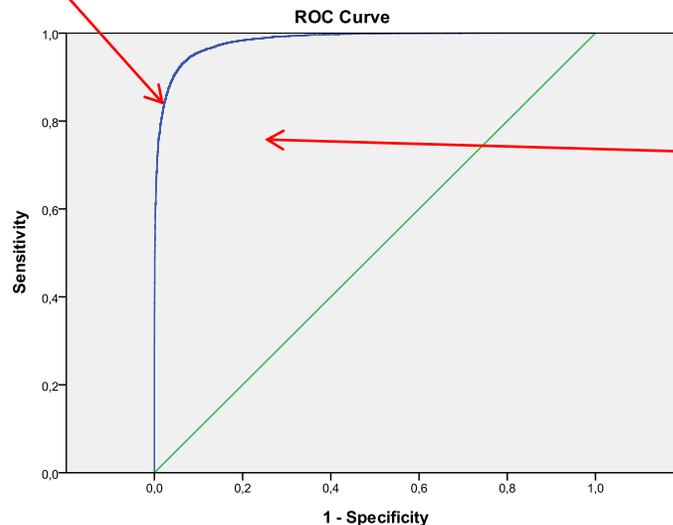
	1 - sick	0 - healthy
1 - risk group	Truly positive	False Positive
0 - non-risk group	False Negative	Truly negative

$$\textit{sensitivita} = \frac{\textit{skutečně pozitivní}}{\textit{skutečně pozitivní} + \textit{falešně negativní}}$$

$$\textit{specifická} = \frac{\textit{skutečně negativní}}{\textit{skutečně negativní} + \textit{falešně pozitivní}}$$

ROC outputs

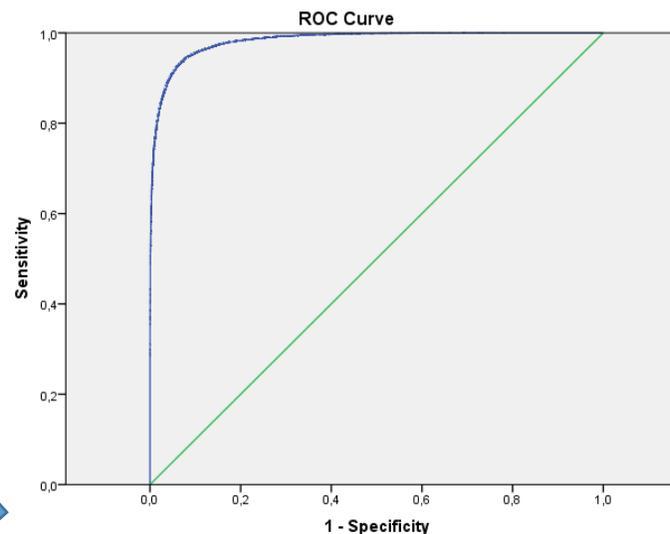
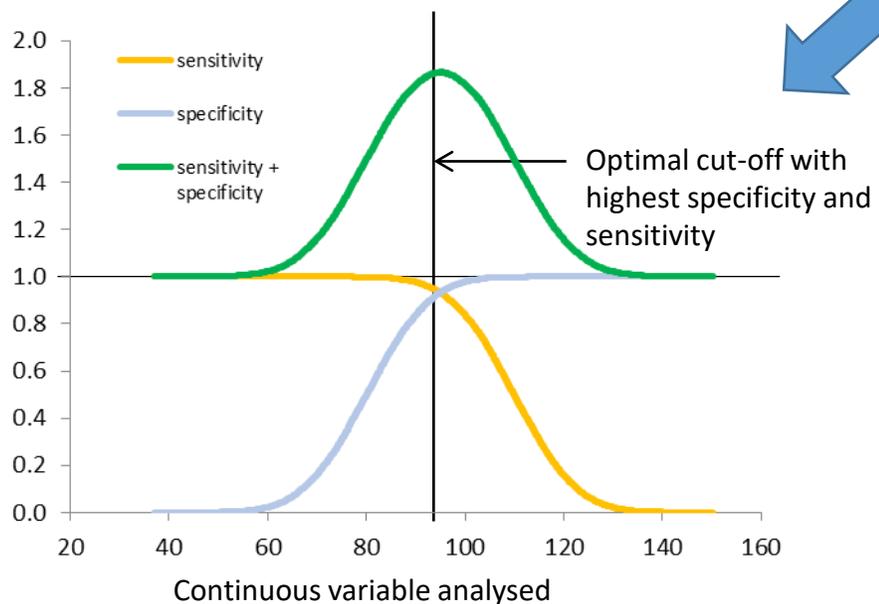
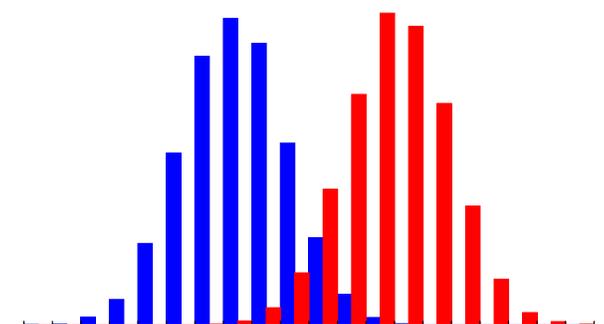
- **Sensitivity and specificity at each point on the curve - can be supplemented with IS**
- The best combination of sensitivity and specificity determines the appropriate split point of the continuous variable
- When identifying the cut-off, it is also necessary to check that the resulting risk group does not contain only the minimum values (a cut-off separating one patient is almost meaningless)



AUC (area under the curve) + IS
The different from 0.5, the better
the endpoint identification
AUC significance testing

ROC - example

Distinguishing two groups of patients
(blue=healthy; red=sick)



Area Under the Curve

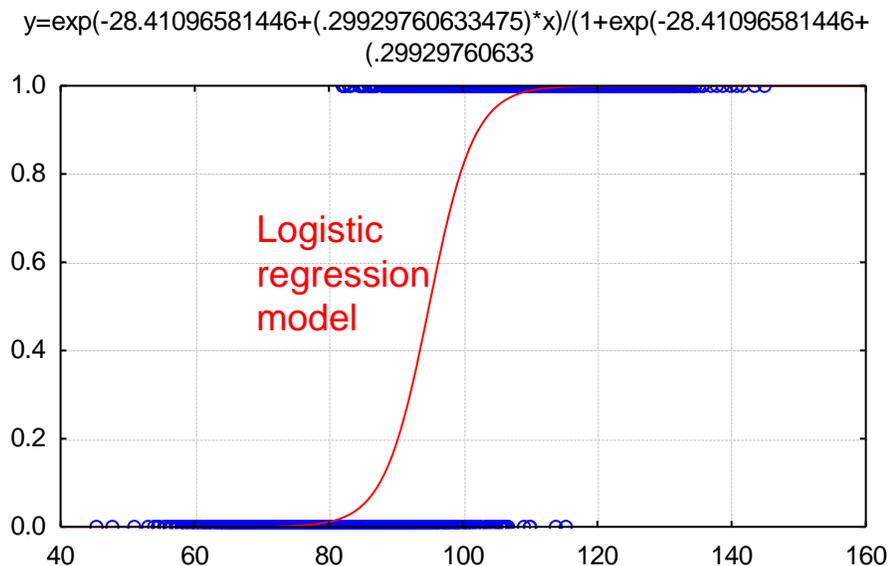
Test Result Variable(s):Var1

Area	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,983	,000	,000	,982	,984

a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5

Logistic regression

- Logistic regression is an essential tool for analyzing the dependence of a binary endpoint (death, complication, taxon occurrence, category membership, etc.) on continuous or binary predictors
- The aim of the analysis is:
 - Identification of relationships between predictors and endpoint and their description (odds ratio)
 - Creation of a prediction model allowing patients to be assigned to evaluation groups
- Logistic regression belongs to the group of generalized linear models (linear statistical models with link function)



Logistic regression example: prediction of a binary characteristic (y-axis) using a continuous variable (x-axis)

The principle of logistic regression

- In logistic regression, we model the effect of continuous or binary predictors on an endpoint with a binomial distribution - > classical linear regression cannot be used
- We predict the probability of occurrence of the phenomenon using the equation:

$$P(x) = \frac{\exp(a + b * x)}{1 + \exp(a + b * x)}$$

- Where is the $\frac{\exp(\text{rovnice})}{1 + \exp(\text{rovnice})}$ the link function for the logistic regression and the equation $a+b*x$ is the linear model used
- The concept of a link function is related to generalized linear models, where the link function converts the problem of nonlinear dependence of y on x into a linear model
- Simply put "non-linear relationship=link function(linear model)"
- Generalized linear model with link function "identity" = linear model

Odds ratio and logistic regression

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a SEPALLEN	5,140	1,007	26,080	1	,000	170,773	23,748	1228,028
Constant	-27,831	5,434	26,236	1	,000	,000		

a. Variable(s) entered on step 1: SEPALLEN.

- It describes the level of risk associated with:
 - **For continuous variables** with a change of 1 (for this reason, continuous variables are often converted into interpretable units - e.g. age by decades, concentration by hundreds of units)
 - **For binary variables** associated with the occurrence of a property (coded as 1)
 - For classical dummies, it is a risk to all other patients without a given trait
 - For binary variables coded against the reference category, this is an increase over patients in the reference category
- Odds ratio is the exponential value of the coefficient of the regression equation

Logistic regression: summary

- A basic tool for identifying factors influencing the occurrence of binary endpoints and predicting the individual probability of endpoint occurrence
- Applicable as an analogue of discriminant analysis for 2 groups
- Describes the riskiness of predictors for a binary endpoint in the form of odds ratios
- For multivariate models, it is important to analyze parameter redundancy and stability of multivariate models
- Cross-validation of the models, or other methods of testing the fit of the models on independent data, is necessary for practical deployment
- Cannot work with censored data (survival analysis)
- **Standard risk factor analysis methodology for binary endpoints (occurrence of something - death, taxon, etc.)**

Multivariate data analysis: introduction

Principles and applications of multivariate data analysis

Annotation

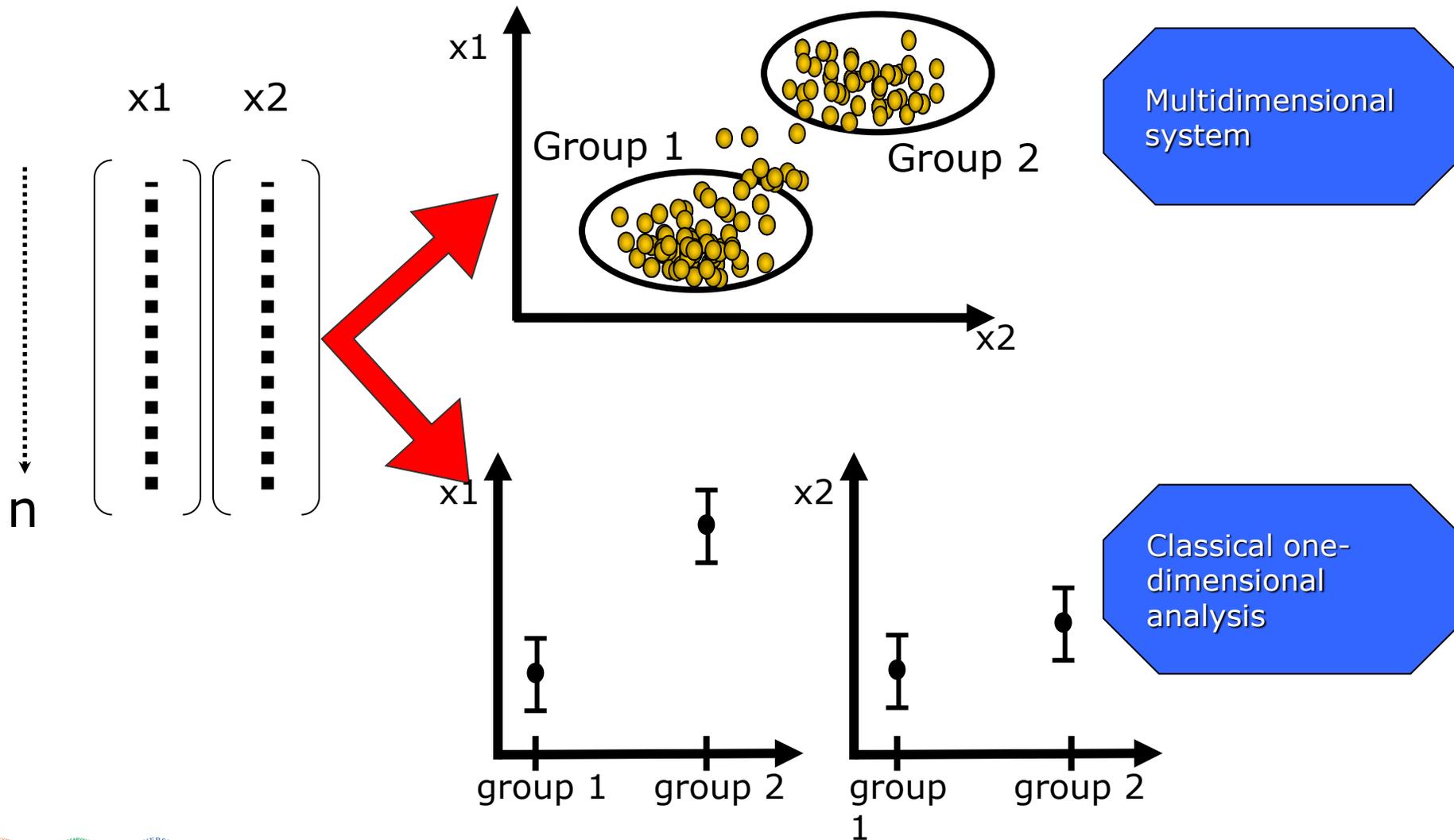
- Multivariate data analysis is a superstructure over classical, univariate statistics and is particularly suitable for biological and medical data, which are multivariate by nature
- However, in multivariate analysis it is necessary to remember that it is usually based on the same principles as univariate analysis and therefore it is necessary to observe the assumptions on which the calculation is based. This is important to note, especially given the relative availability of multivariate analyses in modern statistical software.

Relationship between classical and multivariate statistics

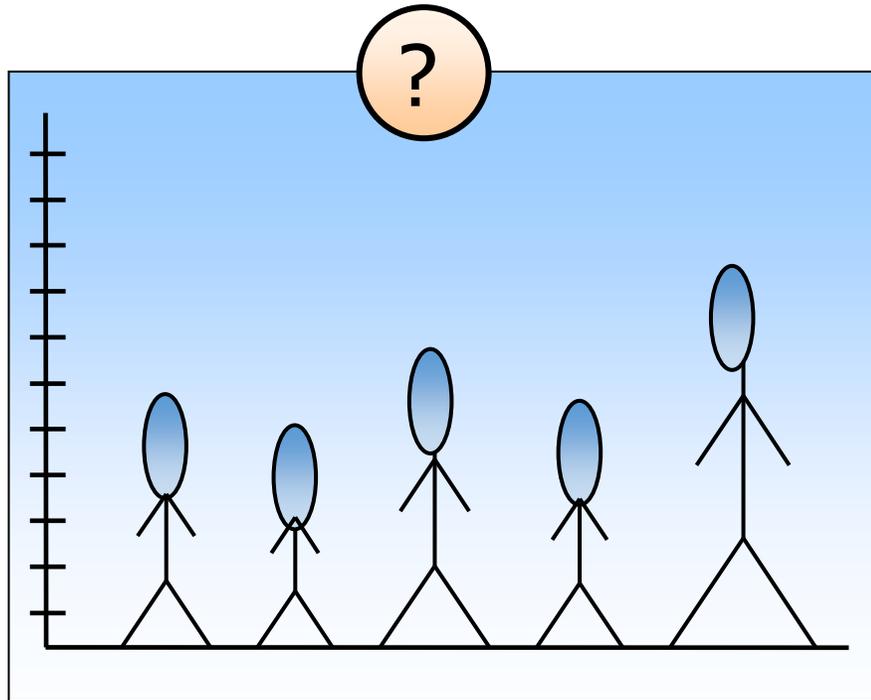
- Multivariate data analysis uses classical statistical approaches
- At the same time, she is also sensitive to their problems
- Data aggregation via summary statistics or contingency tables - correspondence analysis
- Correlation - principal component analysis, factor analysis, discriminant analysis



Multidimensional perception of reality - a new quality of data analysis



Routine aggregation of data "liquidates" the individuality of the individual

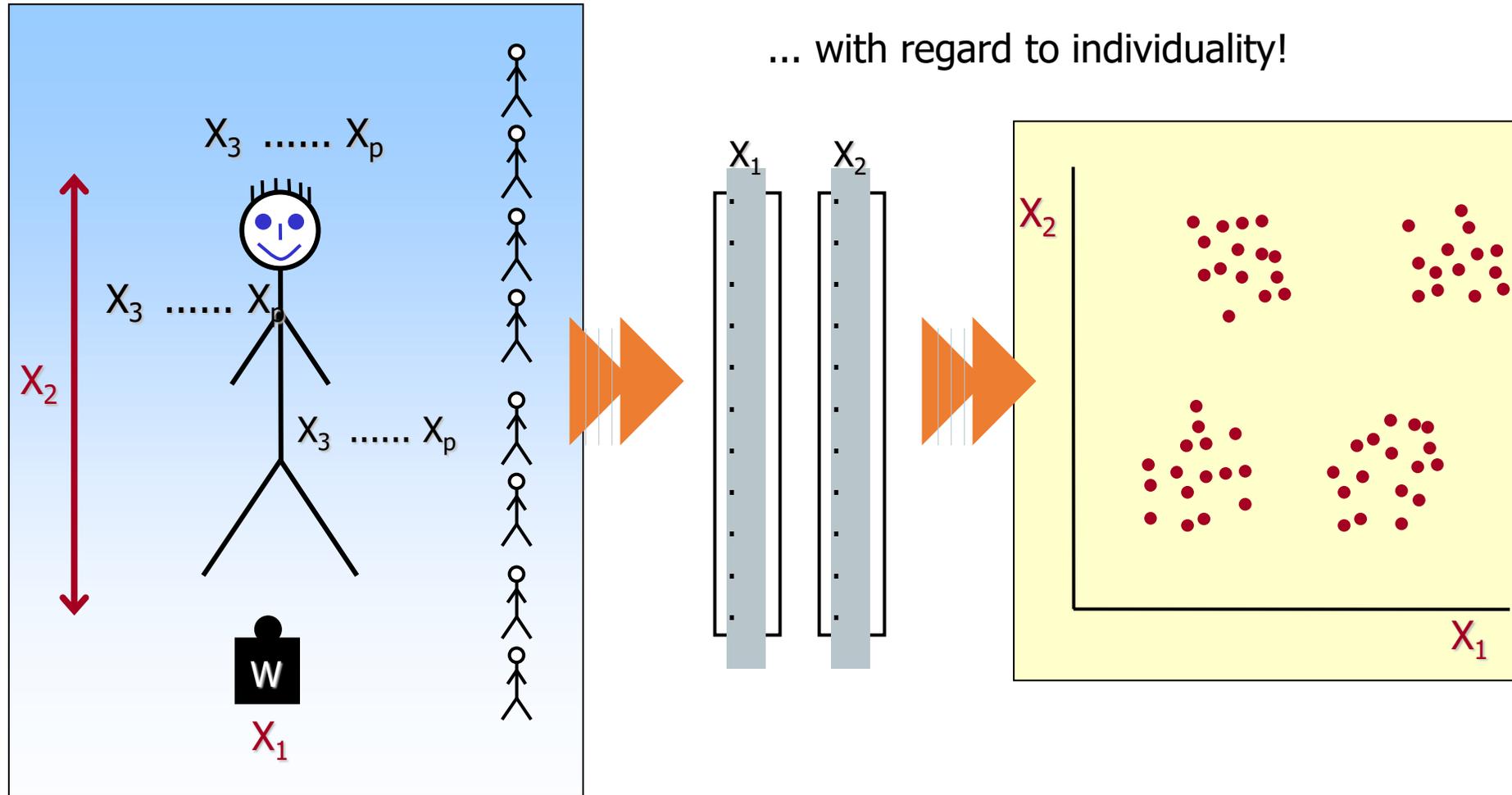


Diameter \pm SE

COMMON STATISTICAL SUMMARY

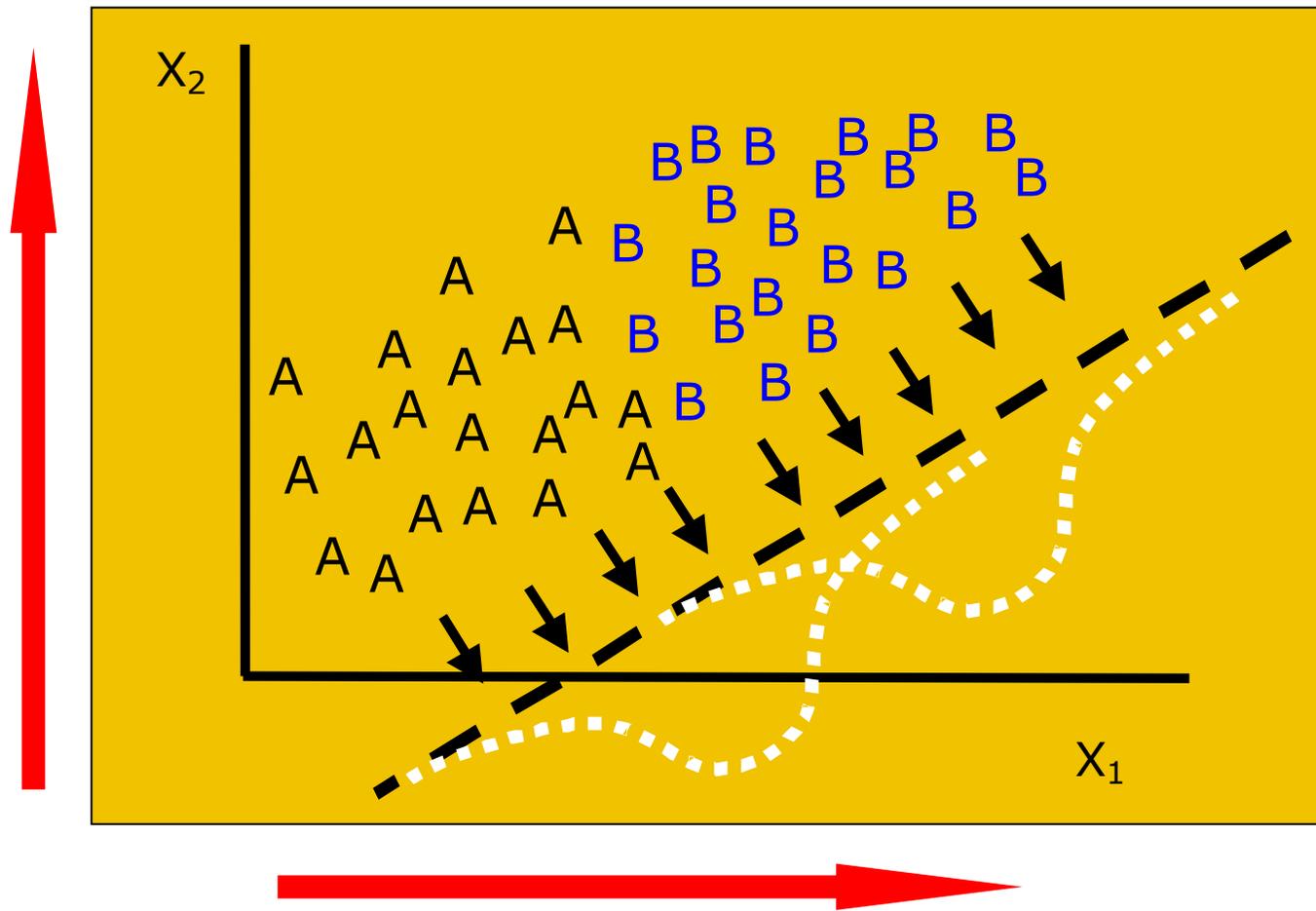
- ✓ Data transparency
- ✓ Does not distinguish the original measurement

Multidimensional evaluation



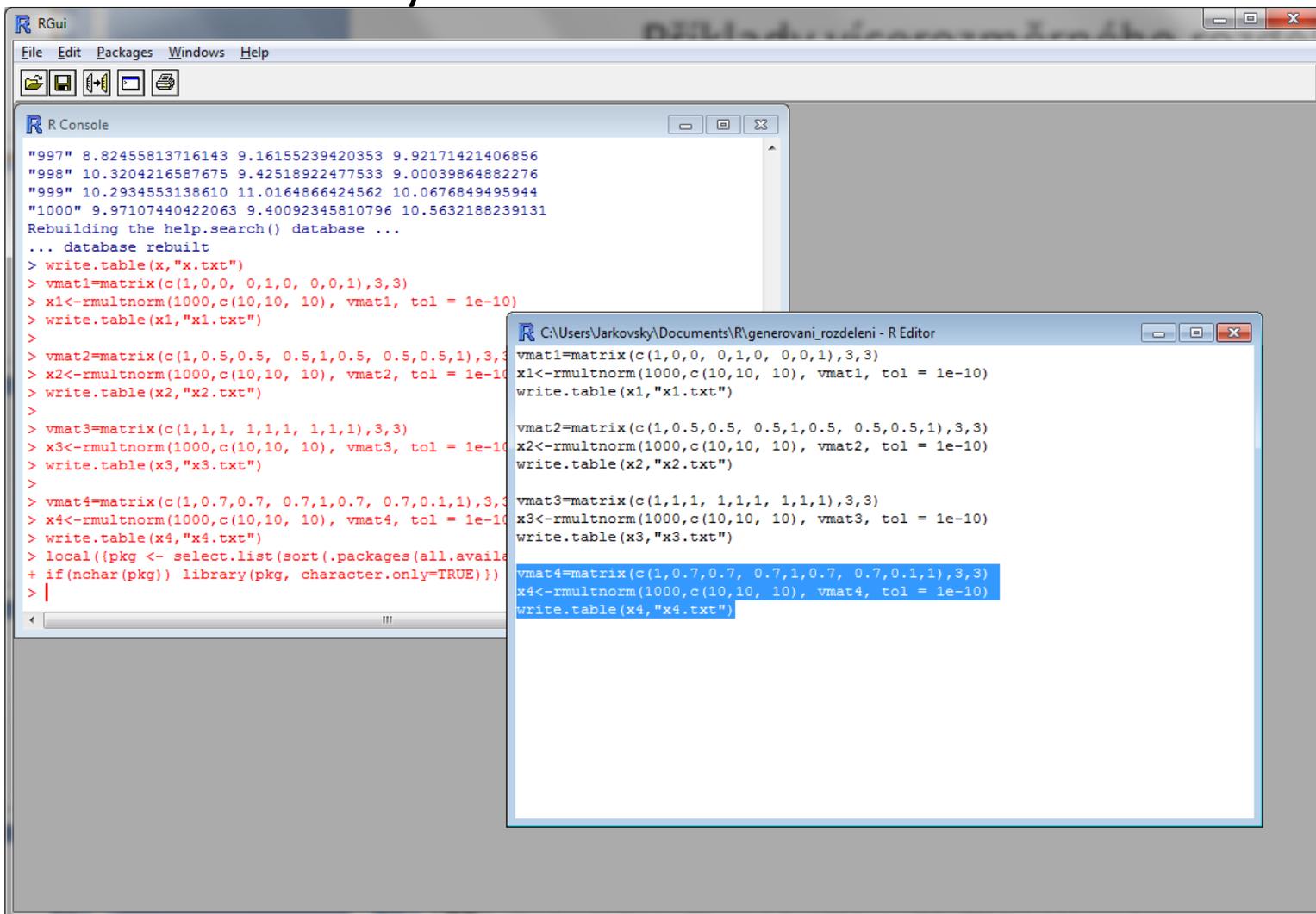
Multidimensional evaluation - new quality

Only the combined parameters have adequate information power



Examples of multivariate distribution

- R - MSBVAR library



```
RGui
File Edit Packages Windows Help
R Console
"997" 8.82455813716143 9.16155239420353 9.92171421406856
"998" 10.3204216587675 9.42518922477533 9.00039864882276
"999" 10.2934553138610 11.0164866424562 10.0676849495944
"1000" 9.97107440422063 9.40092345810796 10.5632188239131
Rebuilding the help.search() database ...
... database rebuilt
> write.table(x,"x.txt")
> vmat1=matrix(c(1,0,0, 0,1,0, 0,0,1),3,3)
> x1<-rmultnorm(1000,c(10,10, 10), vmat1, tol = 1e-10)
> write.table(x1,"x1.txt")
>
> vmat2=matrix(c(1,0.5,0.5, 0.5,1,0.5, 0.5,0.5,1),3,3)
> x2<-rmultnorm(1000,c(10,10, 10), vmat2, tol = 1e-10)
> write.table(x2,"x2.txt")
>
> vmat3=matrix(c(1,1,1, 1,1,1, 1,1,1),3,3)
> x3<-rmultnorm(1000,c(10,10, 10), vmat3, tol = 1e-10)
> write.table(x3,"x3.txt")
>
> vmat4=matrix(c(1,0.7,0.7, 0.7,1,0.7, 0.7,0.1,1),3,3)
> x4<-rmultnorm(1000,c(10,10, 10), vmat4, tol = 1e-10)
> write.table(x4,"x4.txt")
> local({pkg <- select.list(sort(.packages(all.available))
+ if(nchar(pkg)) library(pkg, character.only=TRUE))})
> |

R C:\Users\Uarkovsky\Documents\R\generovani_rozdeleni - R Editor
vmat1=matrix(c(1,0,0, 0,1,0, 0,0,1),3,3)
x1<-rmultnorm(1000,c(10,10, 10), vmat1, tol = 1e-10)
write.table(x1,"x1.txt")

vmat2=matrix(c(1,0.5,0.5, 0.5,1,0.5, 0.5,0.5,1),3,3)
x2<-rmultnorm(1000,c(10,10, 10), vmat2, tol = 1e-10)
write.table(x2,"x2.txt")

vmat3=matrix(c(1,1,1, 1,1,1, 1,1,1),3,3)
x3<-rmultnorm(1000,c(10,10, 10), vmat3, tol = 1e-10)
write.table(x3,"x3.txt")

vmat4=matrix(c(1,0.7,0.7, 0.7,1,0.7, 0.7,0.1,1),3,3)
x4<-rmultnorm(1000,c(10,10, 10), vmat4, tol = 1e-10)
write.table(x4,"x4.txt")
```

Multivariate distribution characteristics

- The basic characteristic of a multivariate distribution is a vector of means (vector of means)

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix}$$

- and the covariance matrix

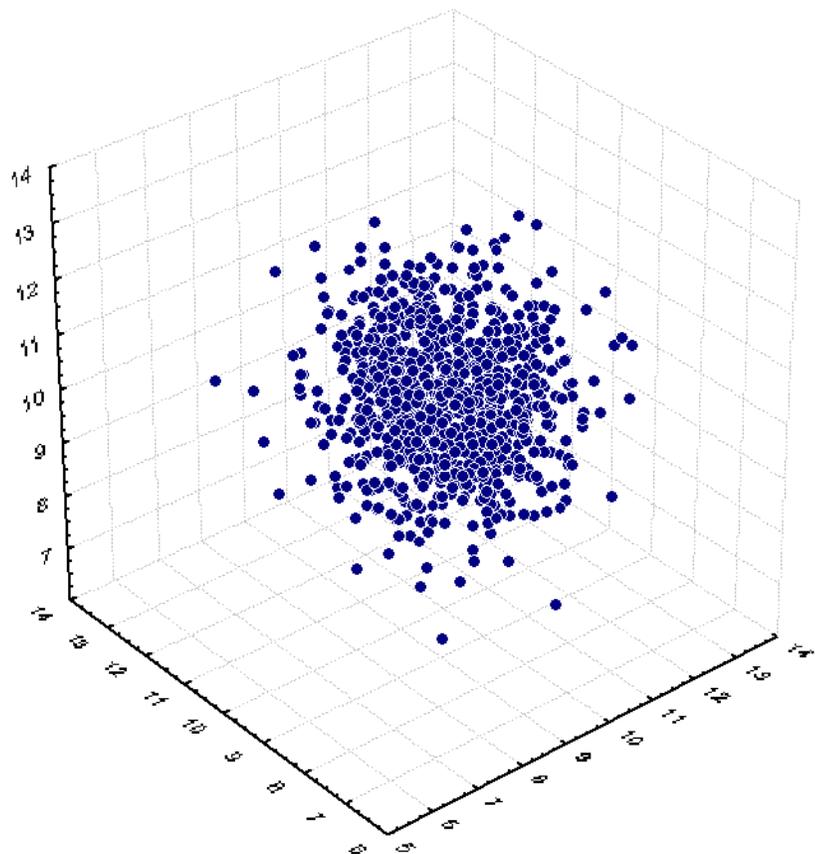
$$\Sigma = \text{var}(\mathbf{X}) = \text{cov}(\mathbf{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_p \\ \sigma_2\sigma_1 & \sigma_2^2 & \cdots & \sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p\sigma_1 & \sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{pmatrix}$$

- where σ_{ij} is the covariance of two random variables, i.e.

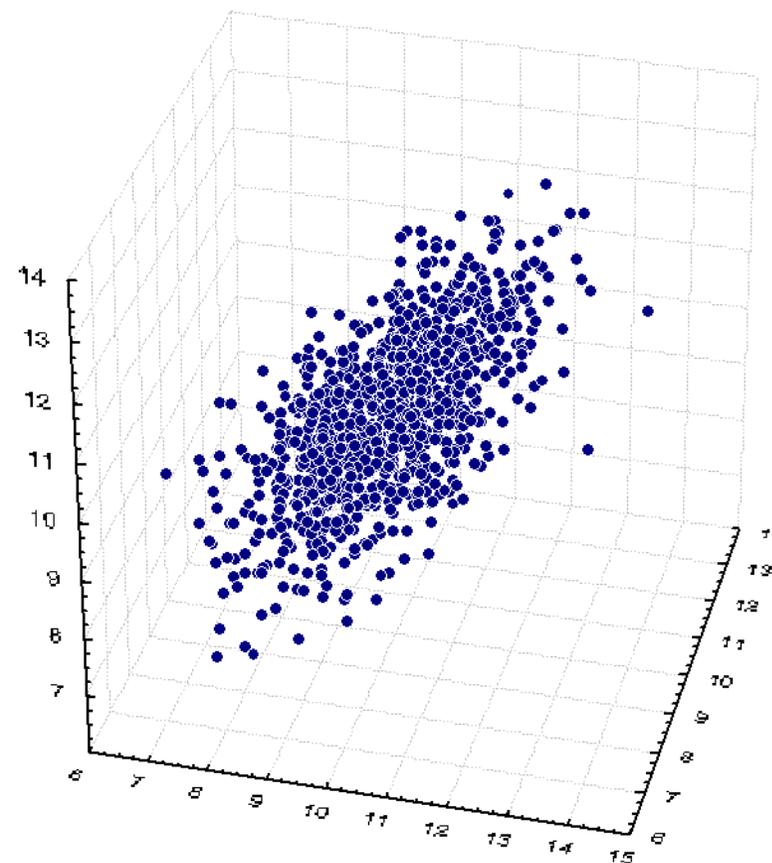
$$\sigma_{ij} = \text{cov}(X_i, X_j) = E((X_i - E(X_i))(X_j - E(X_j)))$$

Example of multivariate distribution I

```
vmat1=matrix(c(1,0,0, 0,1,0, 0,0,1),3,3)  
x1<-rmultnorm(1000,c(10,10, 10), vmat1, tol = 1e-10)  
write.table(x1, "x1.txt")
```

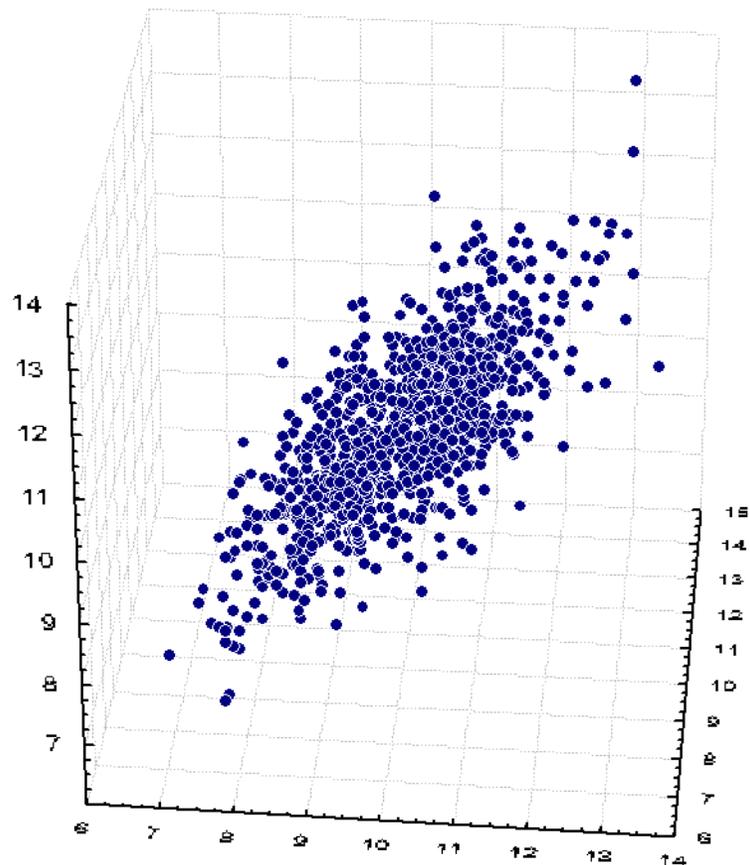


```
vmat2=matrix(c(1,0.5,0.5,1,0.5,0.5,0.5,1),3,3)  
x2<-rmultnorm(1000,c(10,10, 10), vmat2, tol = 1e-10)  
write.table(x2, "x2.txt")
```

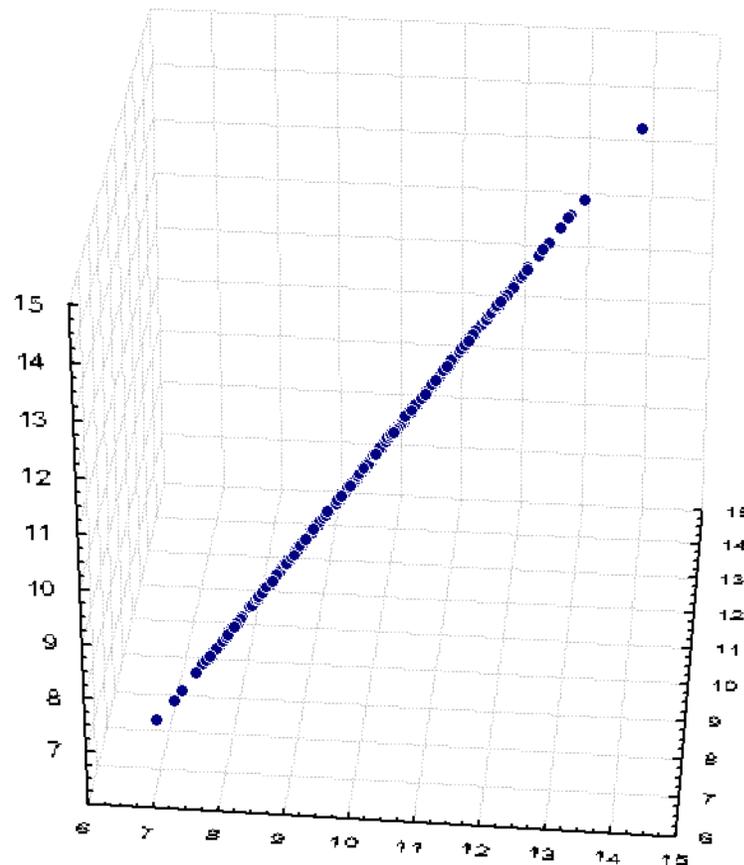


Example of multivariate distribution II

```
vmat4=matrix(c(1,0.7,0.7,1,0.7,0.7,0.1,1),3,3)  
x4<-rmultnorm(1000,c(10,10, 10), vmat4, tol = 1e-10)  
write.table(x4, "x4.txt")
```

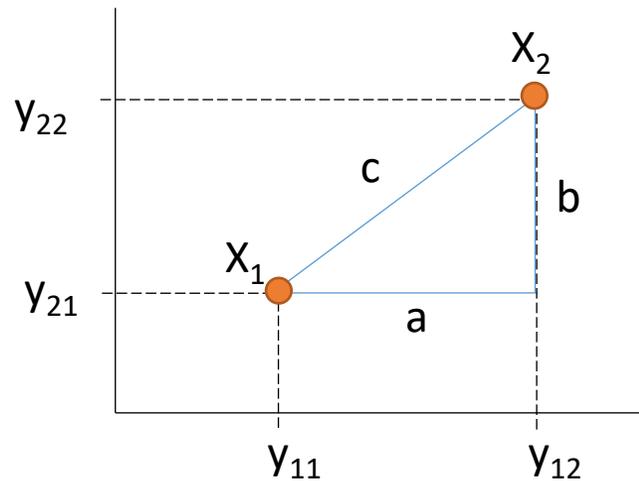


```
vmat3=matrix(c(1,1,1,1,1,1,1,1),3,3)  
x3<-rmultnorm(1000,c(10,10, 10), vmat3, tol = 1e-10)  
write.table(x3, "x3.txt")
```



Multidimensional evaluation is based on simple principles

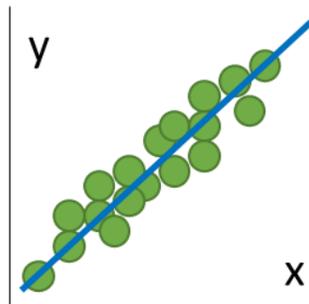
- The easiest measure of the relationship between two objects in multidimensional space is their distance
- The simplest type of this distance (unfortunately with limited application to community data) is the Euclidean distance based on the Pythagorean theorem



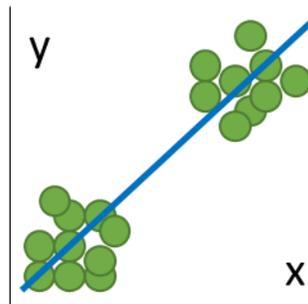
$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2}$$

Correlation as a principle for calculating multivariate analyses

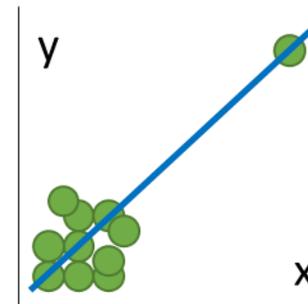
- Covariance and Pearson correlation coefficient is the basis of principal components analysis, factor analysis as well as other multivariate analyses working with linear dependence of variables
- The assumption for the calculation of covariance and Pearson correlation coefficient is:
 - Data normality in both dimensions
 - Linearity of the relationship between variables
- For multivariate analyses, the most serious problem is the presence of outliers



Linear relationship - seamless use of Pearson correlation coefficient



The correlation is given by two sets of values - it leads to the identification of groups of objects in the data



The correlation is given by the outlier - the analysis only describes the effect of the outlier

Analysis of contingency tables as a principle for calculating multivariate analyses

- The abundance of taxa (or number of any objects) at sites can be thought of as a contingency table, and the measure of the relationship between rows (sites) and columns (taxa) is the magnitude of the chi-square

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

Calculated for each cell of the table

	☠	😊
A	10	0
B	0	10

Observed table

	☠	😊
A	5	5
B	5	5

Expected table

The chi-squared value defines the degree of deviation of a given cell (in our context, taxon-locality relationship) from a situation where there is no relationship between rows and columns (taxon-locality)

Concepts of multivariate analyses

- **Multidimensional methods:** the name multidimensional is based on the type of input data, this data consists of individual objects and each of them is characterized by its parameters and each of these parameters can be considered as one dimension of the object.
- **Matrix algebra:** Matrix algebra is the basis for working with data and calculations of multidimensional methods, matrices form both input and output data and calculations are performed on them.
- **NxP matrix:** N objects with p parameters then form the so-called NxP matrix, which is the first type of data input to multivariate analyses.
- **Association matrices:** on the basis of these matrices, association matrices are calculated, on which further calculations are then performed. These are square matrices containing information about the similarity or dissimilarity (so-called metrics) of either objects (Q mode analysis) or parameters (R mode analysis). The scale of similarity varies according to the method used and the type of data, some methods allow the use of user metrics.

Input matrix of multivariate analyses

NxP MATRIX



Parameter values for individual objects

ASSOCIATION MATRIX

Calculation of similarity metrics distances



Correlation, covariance, distance, similarity

Basic types of multivariate analyses

CLUSTER ANALYSIS

- creating clusters of objects based on their similarity
- identification of object types

CLASSIFICATIONS

- Model for assigning unknown patients to predefined groups
- A series of algorithms

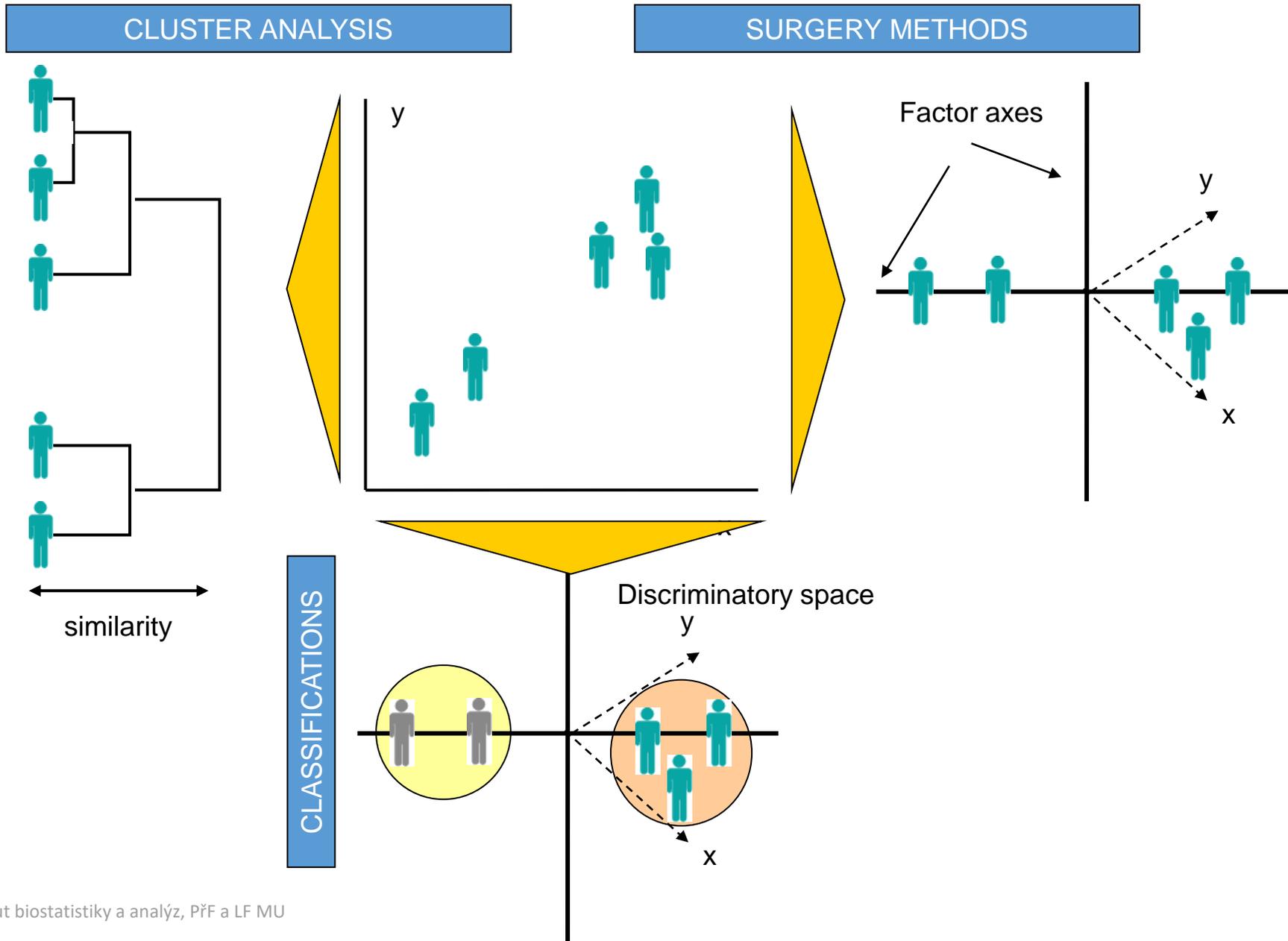
SURGERY METHODS

- simplification of a multidimensional problem into fewer dimensions
- the principle is to create new dimensions that better exploit the variability of the data

MODELLING

- Prediction models with multiple predictors
- Regression methods and other types of algorithms

Types of multivariate analyses



Thank you for your attention, I hope you took something away from the semester 😊

