MUNI FI

Statistics II | 1

Introduction to statistics, Linear regression model

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

12 September 2022 (updated 16 September 2022)

Probability and statistics

- Probability theory *(teorie pravděpodobnosti)*:

  branch of mathematics, deals with the description of random phenomena and experiments; provides theoretical ideas, definitions, derivations, assertions and proofs for describing and working with random phenomena and experiments.
- (Mathematical) Statistics *(matematická statistika)*:

  deals with the collection, organization, analysis, interpretation and presentation of data; uses the tools of probability theory.
- Statistical / machine learning *(statistické / strojové učení)*

Methods of mathematical statistics

Descriptive statistics *(popisná statistika)*:

- summarizes the sample using summary statistics and indices;
- frequency tables, sample moments (mean, variance, standard deviation, skewness, kurtosis) and quantiles (median, quartiles, IQR), contingency tables, sample correlation.

Exploratory Data Analysis *(exploratorní analýza dat)*:

- analysis of the data, usually using visualization methods;
- frequency plot, boxplot, histogram, scatter plot, QQ plot.

Statistical inference *(statistická inference)*:

- derives probabilistic properties (parameters or probability distribution) of the population based on the analysis of the data sample;
- requires the so-called model – assumptions about the population and the sample;
- estimates of parameters – point and interval (confidence intervals), testing statistical hypothesis, prediction, classification, clustering.

Methods of statistical inference

Parametric methods:

- the model assumes a probability distribution or some class of them, parameters of these distributions are estimated;
- most of *classical* methods, e. g., *t*-test, linear regression model, multiple regression model, generalized linear models, analysis of variance, correlation analysis.

Nonparametric methods:

- millimalistic assumptions for the model are specified, no specific probability distribution is required;
- e. g., rank statistics and test and corresponding variants of ANOVA, correlation analysis; Functional Data Analysis.

Semiparametric methods:

- a combination of both approaches;
- e. g., Cox model of proportional hazards in survival analysis.

Statistical data types (1)

Stevens, Stanley Smith (1946). On the Theory of Scales of Measurement. Science 103 (2684), 677–680.

Nominal data *(nominální data)*:

- defined operations: $=$, $\neq$, classification, set membership;
- categorial data, discrete, in *R*: `factor`;
- values cannot be compared and ordered;
- e. g., blood type;
- for two categories: dichotomous data, often `TRUE/FALSE` or `1/0`;
- dummy variable $1/0$ encodes membership of a specific category.

Ordinal data *(ordinální data)*:

- additionally defined: order, rank;
- additional operations: $<$, $>$, comparison, sorting;
- categorial data, discrete, in *R*: `ordered factor`;
- distance between values cannot be quantified;
- e. g., the highest education attained, achieved grade in a course.

Statistical data types (2)

Interval data *(intervalová data)*:

- additionally defined: distance;
- additional operations: $+$, $-$;
- typically continuous numerical data, in *R*: `numeric`;
- ratio of values cannot be quantified, zero is not correctly defined;
- e. g., temperature in °C.

Ratio data *(poměrová data)*:

- additionally defined: ratio;
- additional operations: $*$, $/$;
- typically continuous numerical data, in *R*: `numeric`;
- all physical variables in accordance with SI, e. g., temperature in K.

Continuous data can be treated (with a certain loss of information) as well as discrete data: we divide the data into intervals, which further play the role of categories, so-called interval data.

Testing of statistical hypothesis (Neyman & Pearson, 1933)

- H_0: Null hypothesis *(nulová hypotéza)*, the statement being tested,
- H_1: Alternative hypothesis *(alternativní hypotéza)*, the statement being tested against H_0,
- statistical test **rejects *(zamítne)***, or **does not reject *(nezamítne)***, H_0 in favor of H_1.

	H_1 is true	H_0 is true
test rejects H_0 in favor of H_1	*True Positive* right decision $P = 1 - \beta$	*False Positive* Type I error $P = \alpha$
test does not reject H_0 in favor of H_1	*False Negative* Type II error $P = \beta$	*True Negative* right decision $P = 1 - \alpha$

- Level of significance of the test $= \alpha = P(H_0 \text{ rejected} \mid H_0 \text{ true})$
- Power of the test $= 1 - \beta = P(H_0 \text{ rejected} \mid H_1 \text{ true})$
- Cannot be ensured both $\alpha = 0$ and $\beta = 0$, even $\alpha, \beta \to \min$.
- General methodology – Neyman-Pearson lemma: Good criterion for the selection of hypotheses is a likelihood ratio.

Testing procedure

▶ *Classically*, using critical region *(kritický obor)*:

1. Set up H_0 a H_1,
2. choose a model corresponding to data annd hypotheses,
3. choose suitable test and test statistic T with known probability distribution under H_0,
4. decide about α, β and sample size, in our case $\alpha = 0.05$,
5. calculate the observed value t of the test statistic T,
6. calculate critical region W corresponding to t and H_1,
7. H_0 is rejected in favor of H_1, if and only if $t \in W$.

▶ Using p-value *(p-hodnota)*, a common method nowadays, especially in software:

6. calculate p-value p of the test,
7. H_0 is rejected in favor of H_1, if and only if $p < \alpha$.

▶ Using $100(1-\alpha)\%$ confidence interval *(interval spolehlivosti)* for suitable parameter.

Definition (p-value)

P-value p is the (highest) probability that, under validity of H_0, the test statistic T exhibits an equally or more extreme value than the value t observed on the test sample.

$$p = 2\,\min\{P(T \geq t), P(T \leq t)\};\ \text{or}\ p = P(T \geq t);\ \text{or}\ p = P(T \leq t);$$

according to the variant of H_1 (two-sided or one-sided).

- *P*-value is just a tool for deciding whether or not to reject H_0 in favor of H_1; it does not quantify the significance of the observed effect.
- If the test is performed correctly, it is ensured that $P(\text{Type I error}) \leq \alpha$.
- Neither $p = P(H_0)$ nor $p = P(\overline{H_1})$.
- The power of a test can usually be increased using larger sample.
- Non-rejection of H_0 does not imply H_0 is true.

Linear regression model

Regression model

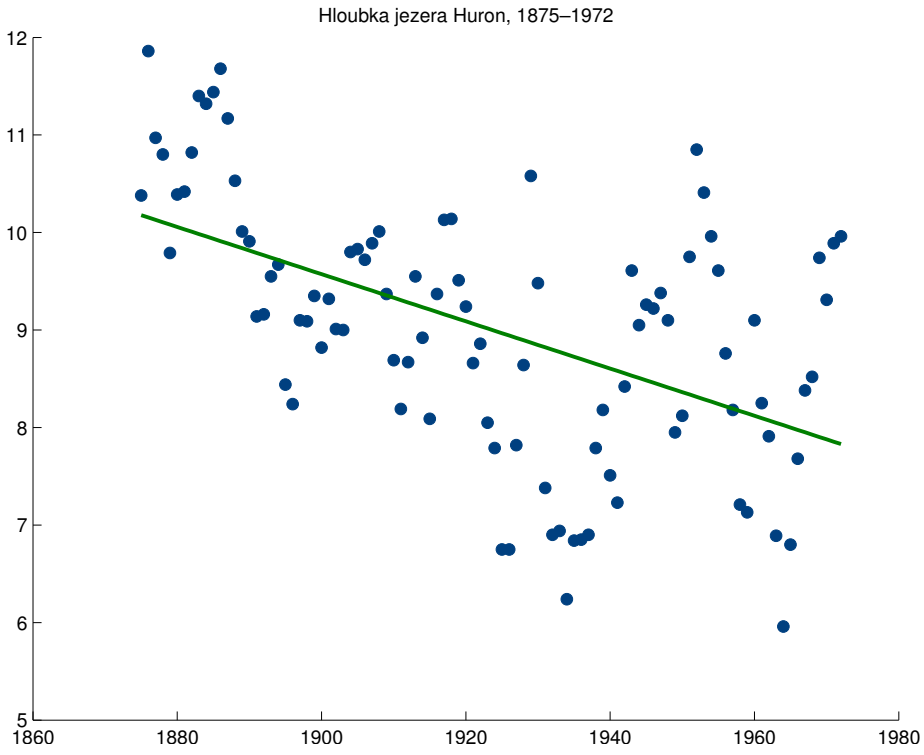
- Examination of the relationship between two numerical quantities, non-random independent variable x and observed random variable Y.
- Data: pairs (x_i, Y_i), $i = 1, \ldots, n$.
- Regression model *(regresní model)*:

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

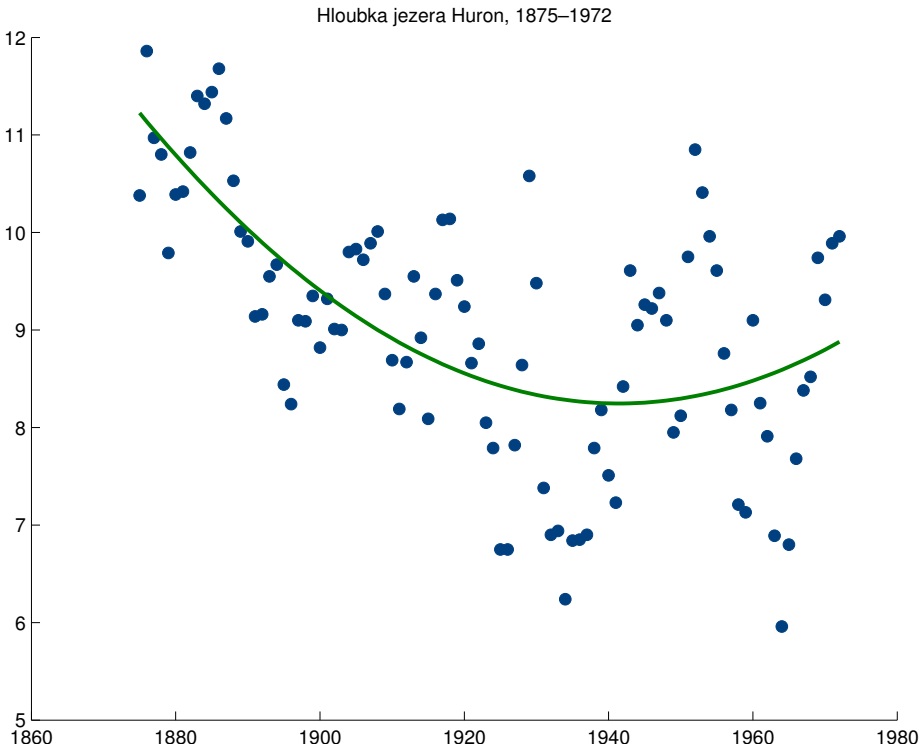
- x_i = known points (vectors) of fixed plan *(pevný plán)*,
- Y_i = observed (measured) values,
- $m(x)$ = regression function *(regresní funkce)* in the form of a function linear in parameters,
- ε_i = measurement errors, $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
- Task: given data (x_i, Y_i), find the *suitable* regression function $m(x)$.

Regression model

$$m(x) = \beta_0 + \beta_1 x$$



$$m(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$



Linear regression model

Assume the regression function

$$m(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{il} = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij}, \quad i = 1, \ldots, n,$$

which is a linear function of unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$.

Linear regression model

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1l} + \varepsilon_1, \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nl} + \varepsilon_n, \end{aligned}$$

written in matrix form as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1l} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nl} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}, \quad \text{i. e.,} \quad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_l)'$ = vector of $k = l + 1$ regression coefficients *(regresní koeficienty)*,
- $\boldsymbol{X}$ = regression / design matrix *(matice plánu)* consists of $(n \times k)$ nonrandom numbers x_{ij}, regressors / predictors *(regresory / prediktory)*,
- $n > k$,
- $r(\boldsymbol{X}) = k = l + 1$, i. e., the design matrix has full rank *(plná hodnost)*, its columns are linearly independent.

Random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$:

- are nonsystematic: $\mathrm{E}(\varepsilon_i) = 0$, i.e., $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$,
- have homogeneous variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$,
- are mutually uncorrelated: $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$;
- variance-covariace matrix *(kovarianční matice)* of the vector of observations is $\mathrm{Var}(\boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$.
- Hence, observations are uncorrelated and have homogeneous variance.

Least squares method

Optimization: find such $\boldsymbol{\beta}$ which minimizes the sum of quadratic deviations,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i - \beta_0 - \sum_{j=1}^{l} \beta_j x_{ij} \right]^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \longrightarrow \min .$$

- Ordinary Least Squares (OLS) estimate *(odhad metodou nejmenších čtverců)*

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_l) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

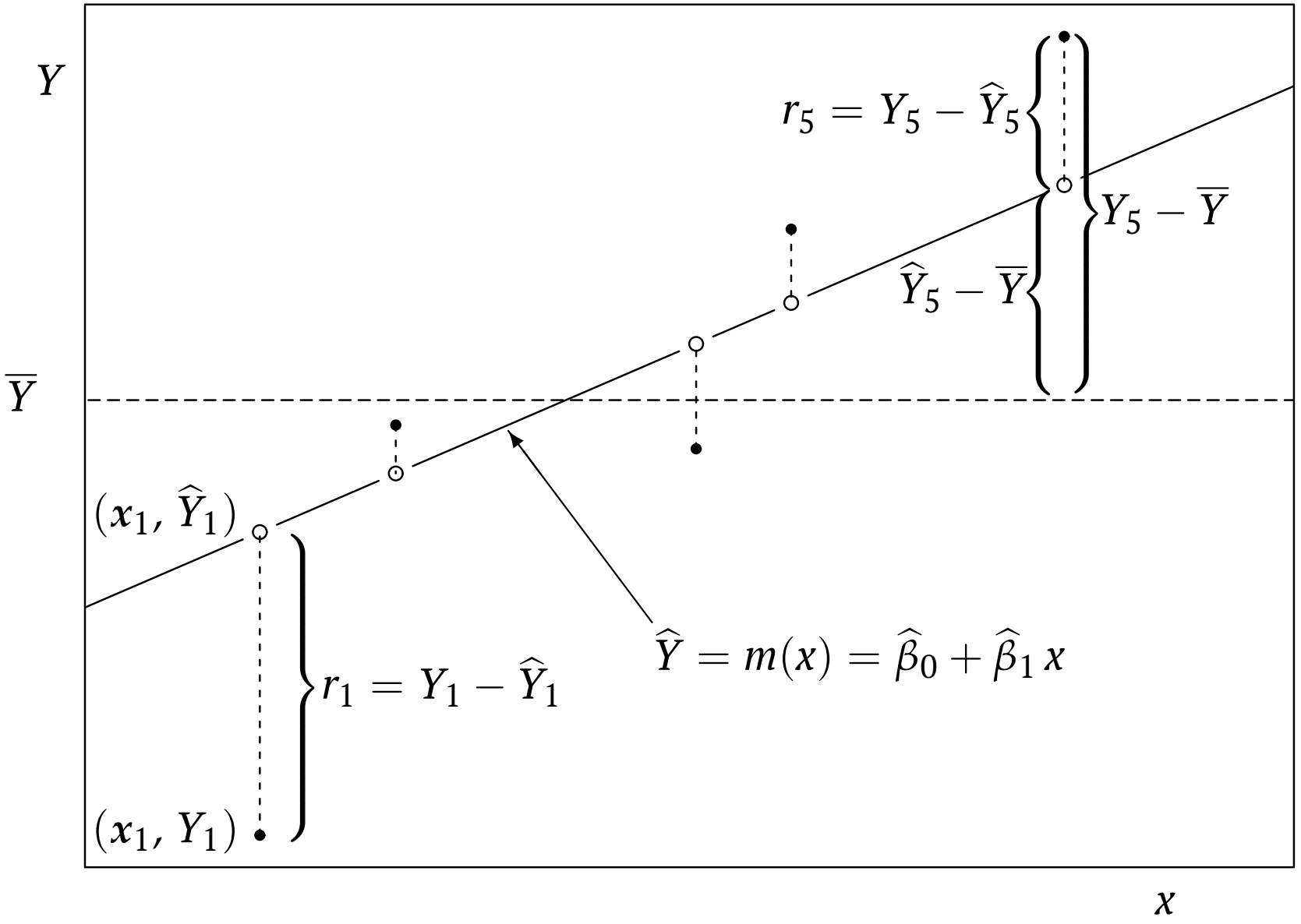
- predicted / fitted values

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}, \quad \text{i. e.,} \quad \widehat{Y}_i = \widehat{\beta}_0 + \sum_{j=1}^{l} \widehat{\beta}_j x_{ij},$$

- residuals *(rezidua)* $r_i = Y_i - \widehat{Y}_i$,
- residual sum of squares *(reziduální součet čtverců)*

$$S_e = S(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \sum_{i=1}^{n} \left[Y_i - \widehat{\beta}_0 - \sum_{j=1}^{l} \widehat{\beta}_j X_{ij} \right]^2 = \sum_{i=1}^{n} r_i^2 .$$

Linear regression model and least squares method



Theorems

Theorem (Gaussov-Markovov)

OLS estimate $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is BLUE = Best Linear Unbiased Estimate *(nejlepší nestranný lineární odhad)* of vector $\boldsymbol{\beta}$ and its variance-covariance matrix *(kovariančí matice)* is $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.

Theorem

Fitted values $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, residual sum of squares $S_e = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$,

where $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is so-called hat matrix.

Theorem

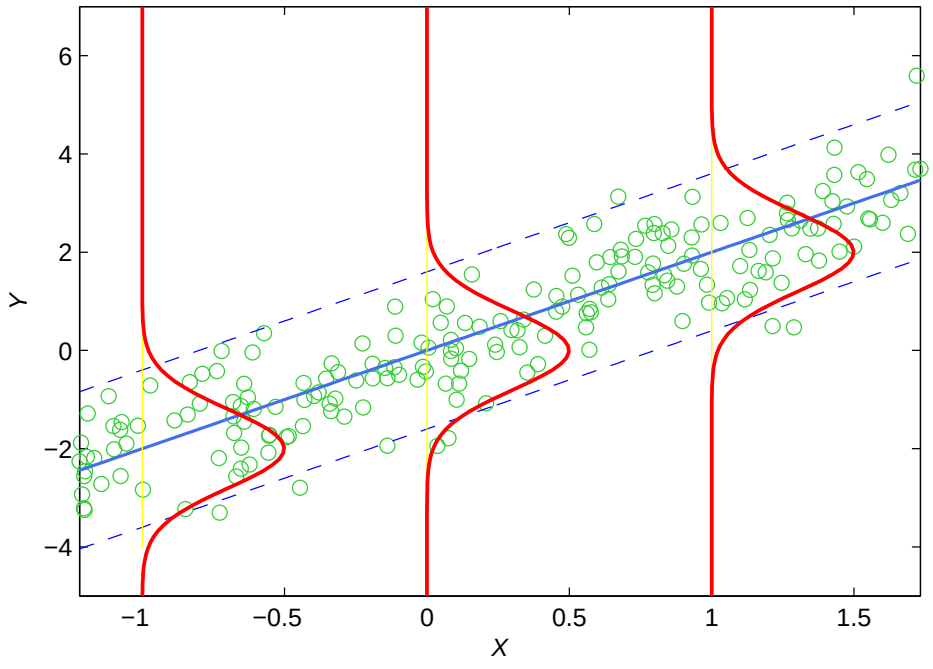
$$\widehat{\sigma^2}_{\text{OLS}} = \frac{S_e}{n-l-1} = \frac{S_e}{n-k}$$

is an unbiased estimate of the variance σ^2 of random errors.

Probability distributions in linear regression model

Additionaly, let us assume that the observations have n-dimensional gaussian (normal) disribution

$$\boldsymbol{Y} \sim \mathrm{N}_n\left(\boldsymbol{X}\boldsymbol{\beta},\, \sigma^2 \boldsymbol{I}_n\right).$$



Theorem

- OLS estimate has gaussian distribution, $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \sim \mathrm{N}_k\left(\boldsymbol{\beta},\, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$,
- statistic $K = (n-k)\dfrac{\widehat{\sigma^2}_{\mathrm{OLS}}}{\sigma^2} \sim \chi^2(n-k)$ has chi-square distribution,
- OLS estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and statistic K are independent.

Test of significance of particular regression coefficient

H_0: $\beta_j = 0$, i.e., regression coefficient β_j is not significant,
H_1: $\beta_j \neq 0$, i.e., regression coefficient β_j is significant

Under H_0, test statistic

$$T_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma^2}_{\text{OLS}} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{jj}}}$$

has Student $t_{1-\alpha/2}(n-k)$ probability distribution.

H_0 is rejected at the level of significance α, if $|T_j| \geq t_{1-\alpha/2}(n-k)$.

$100(1-\alpha)$% confidence interval for regression coefficient β_j is

$$\left(-\sqrt{\widehat{\sigma^2}_{\text{OLS}} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{jj}}\, t_{1-\alpha/2}(n-k), \quad \sqrt{\widehat{\sigma^2}_{\text{OLS}} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{jj}}\, t_{1-\alpha/2}(n-k) \right).$$

Omnibus F-test in linear regression model

H_0: $\beta_1 = \cdots = \beta_l$,
H_1: $\exists j \in \{1, \ldots, l\} : \beta_j \neq 0$, i.e., at least one β_j is significant

Note that the intercept β_0 is not included in these hypotheses.

Under H_0, test statistic

$$F = \frac{1}{k-1} \cdot \frac{S_{\widehat{Y}}}{\widehat{\sigma^2}_{\text{OLS}}} = \frac{n-k}{k-1} \cdot \frac{S_{\widehat{Y}}}{S_e}$$

where $S_{\widehat{Y}} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$ is regression sum of squares,

has Fisher-Snedecor $F(k-1, n-k)$ probability distribution.

H_0 is rejected at the level of significance α, if $F \geq F_{1-\alpha}(k-1, n-k)$.

Coefficient od determination = R squared

Definition

Coefficient of determination *(index determinace)* R squared:

$$R^2 = \frac{S_{\widehat{Y}}}{S_T} = 1 - \frac{S_e}{S_T},$$

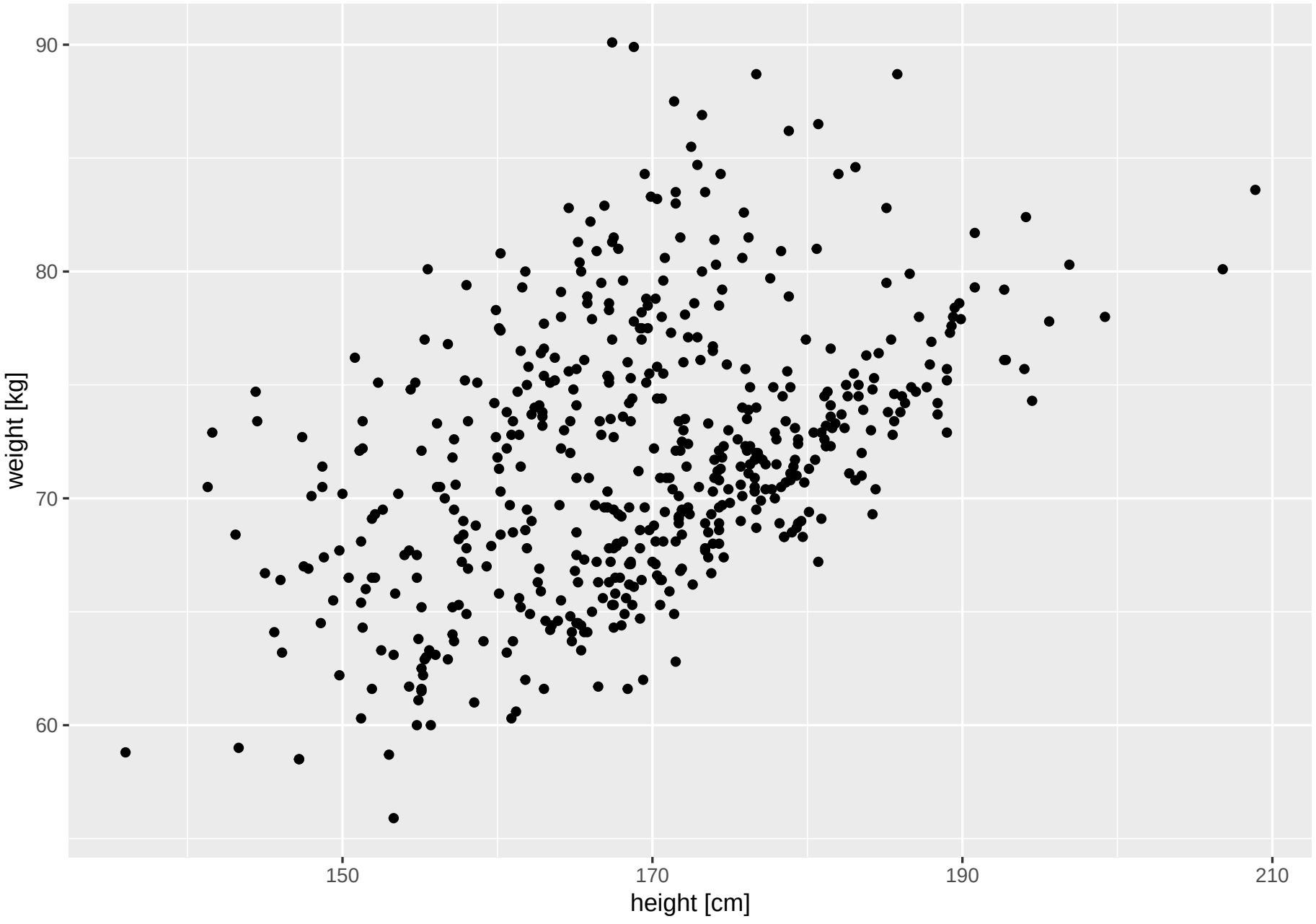
where $S_{\widehat{Y}} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$ and $S_T = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ is total sum of squares.

Adjusted *(korigovaný)* coefficient of determination R bar squared:

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2).$$

- If the predicted values exactly match the observed values, then $R^2 = 1$.
- Linear regression model with only the intercept β_0 has $R^2 = 0$.
- R^2 quantifies the fraction of the variance in the data which is explained by the linear regression model.

Example: weight vs. height of people



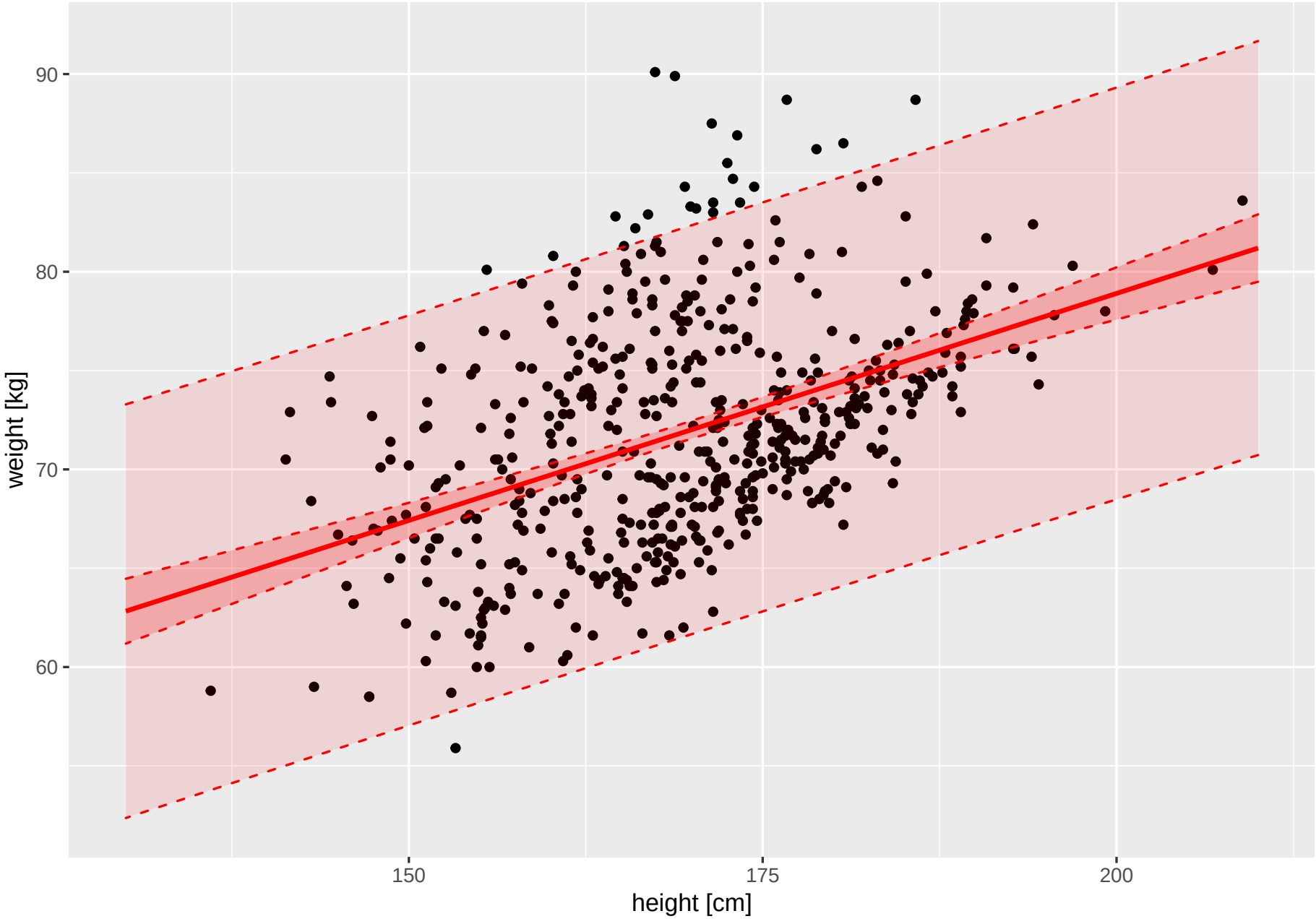
Data

```
'data.frame': 528 obs. of  21 variables:
 $ ID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Sex           : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
 $ Height        : num  173 160 165 164 161 ...
 $ Weight        : num  67.7 71.3 66.3 65.5 60.3 69 75 69.6 74.7 76.1 ...
```

```
m1 <- lm(Weight ~ Height, data = dt) # or
m1 <- lm(Weight ~ 1 + Height, data = dt)
summary(m1)
```

```
Call:
lm(formula = Weight ~ Height, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-12.277  -3.832  -1.121   3.487  18.685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.9708     3.4759   9.485   <2e-16 ***
Height        0.2296     0.0205  11.201   <2e-16 ***

Residual standard error: 5.26 on 526 degrees of freedom
Multiple R-squared:  0.1926,  Adjusted R-squared:  0.1911
F-statistic: 125.5 on 1 and 526 DF,  p-value: < 2.2e-16
```

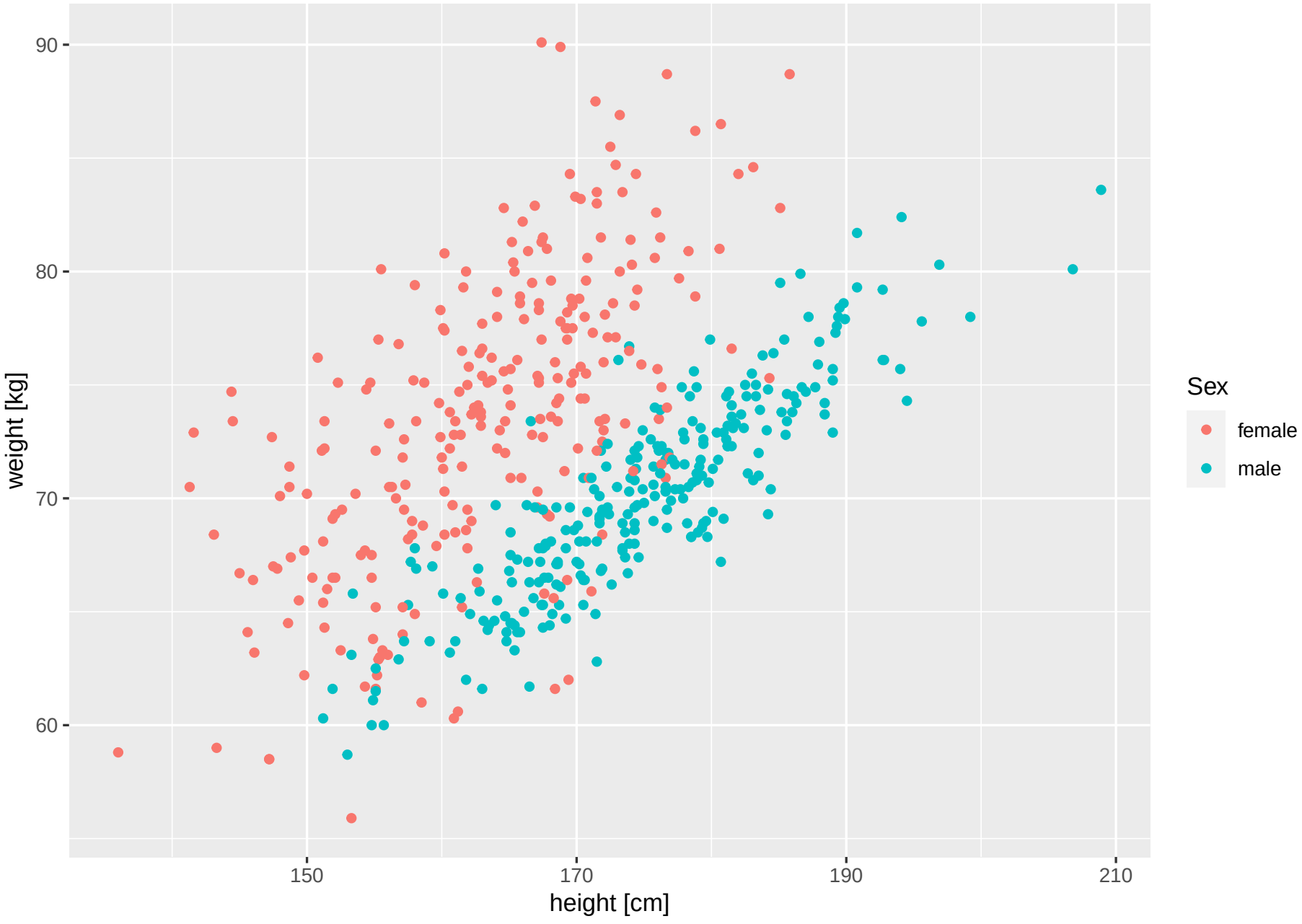
OLS estimate with 95% confidence and prediction intervals



Confidence intervals and predictions

- value of the regression function – point estimate of the value of the regression function $m(x)$ for given x;
  `predict(..., type = "none")`
- confidence interval *(interval spolehlivost pro hodnoty regresní funkce)* = interval estimate of the value of the regression function $m(x)$ for given x;
  `predict(..., type = "confidence")`
- confidence band *(pás spolehlivosti kolem regresní funkce)* – band estimate for the whole regression function $m(x)$
- prediction *(predikce pozorování)* – point estimate $\widehat{Y}(x)$ for given x;
  `predict(..., type = "none")`
- prediction interval *(predikční interval)* – for the predicted observation $\widehat{Y}(x)$ for given x;
  `predict(..., type = "prediction")`

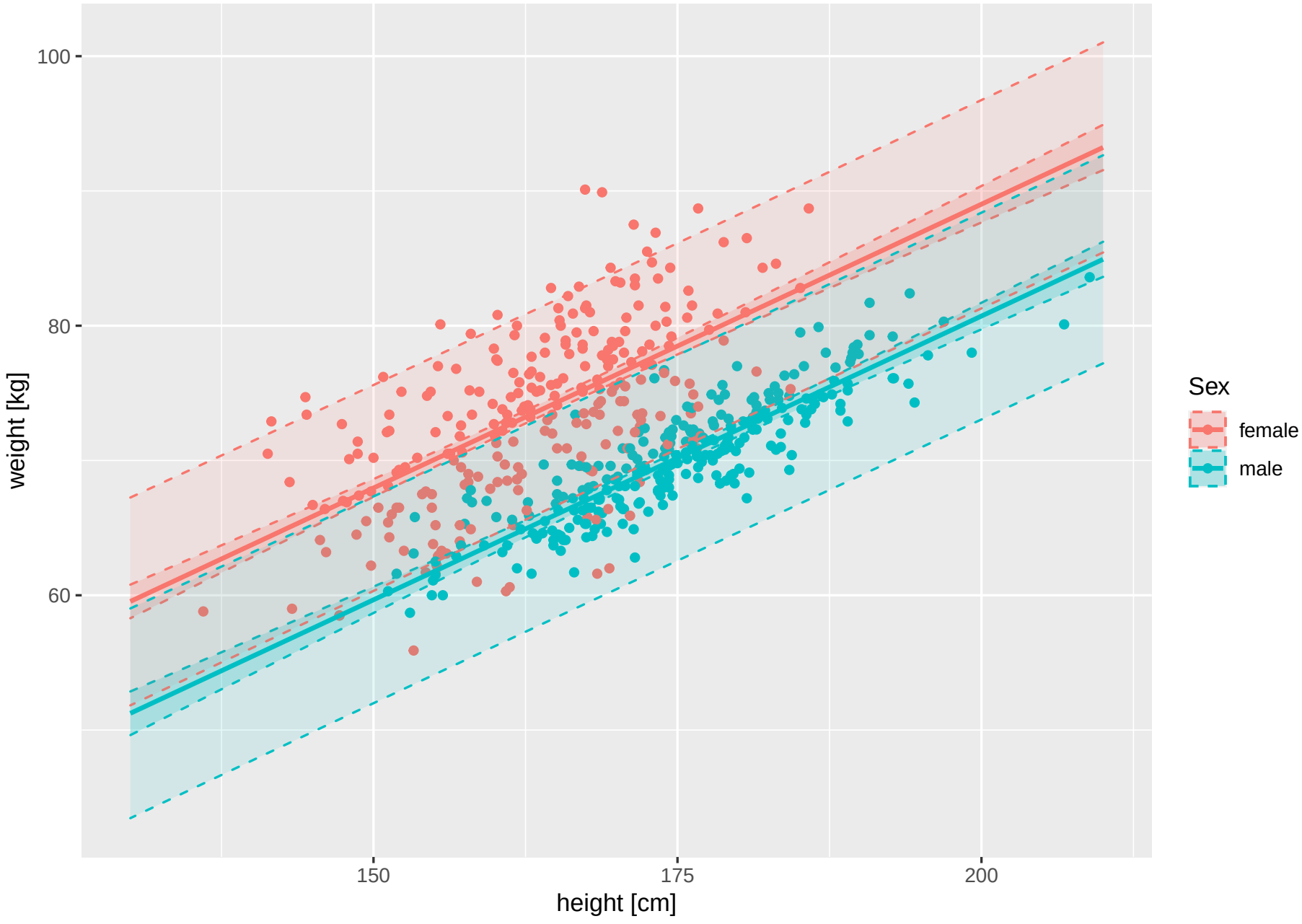
Example: weight vs. height of people



```
m3 <- lm(Weight ~ 1 + Height + Sex, data = dt)
```

```
Call:
lm(formula = Weight ~ 1 + Height + Sex, data = dt)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1307  -2.0173   0.0736   1.9999  14.8114

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.79944    2.88700   1.662    0.097 .
Height       0.42108    0.01761  23.905   <2e-16 ***
Sexmale     -8.29905    0.39340 -21.096   <2e-16 ***

Residual standard error: 3.873 on 525 degrees of freedom
Multiple R-squared:  0.563, Adjusted R-squared:  0.5614
F-statistic: 338.2 on 2 and 525 DF,  p-value: < 2.2e-16
```

OLS estimate with 95% confidence and prediction intervals



Statistical vocabulary

random variable	*náhodná veličina*	X
random sample	*máhodný výběr*	$\boldsymbol{X}$
mean	*střední hodnota*	$\mathsf{E}(X);\ \mu_X$
variance	*rozptyl*	$\mathsf{Var}(X);\ \sigma_X^2$
standard deviation	*směrodatná odchylka*	σ_X
variance-covariance matrix	*kovarianční matice*	$\mathsf{Var}(\boldsymbol{X})$
sample mean	*výběrový průměr*	$\overline{X}$
sample variance	*výběrový rozptyl*	S_X^2
sample standard deviation	*výběrová sm. odchylka*	S_X
statistic	*statistika*	$T(\boldsymbol{X})$
probability distribution	*rozdělení pravděpodobnosti*	$\mathsf{N}; \mathsf{t}; \mathsf{F}; \chi^2; \ldots$
quantile	*kvantil*	
median	*medián*	
estimate	*odhad*	$\widehat{\mu}; \widehat{\sigma}^2; \ldots$
confidence interval	*interval spolehlivosti*	
prediction interval	*predikční interval*	

MUNI FI

Statistics II | 2

Analysis of variance (ANOVA)

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

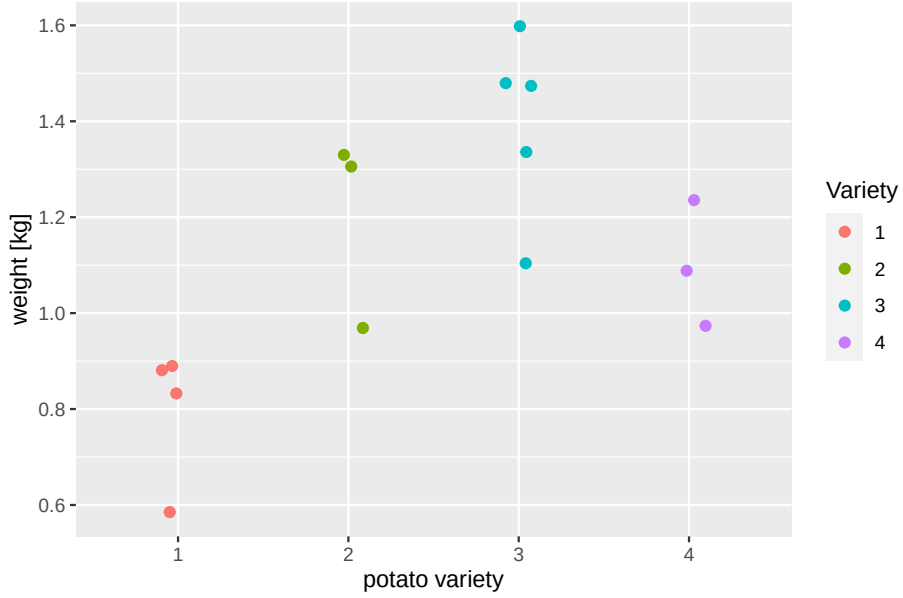
19 September 2022

One-way (single-factor) ANOVA

Example

Analysis of 4 varieties of potatoes based on the weights of the clusters of potato tubers.

variety	weight [kg]
1	0.9, 0.8, 0.6, 0.9
2	1.3, 1.0, 1.3
3	1.3, 1.5, 1.6, 1.1, 1.5
4	1.1, 1.2, 1.0



At a 5% significance level, test the null hypothesis that the mean weight of a cluster of potatoe tubers does not depend on the variety. If you reject the null hypothesis, find which pairs of varieties significantly differ.

- variety: grouping factor – categorical, nominal or ordinal type
- weight: observed random variable – numerical, interval or ratio type

Notation and assumptions

- The factor A has $a \geq 3$ levels.
- The ith level has n_i observations $Y_{i1}, \ldots, Y_{i\,n_i}$, which form a random sample from $\mathrm{N}(\mu_i, \sigma^2)$ probability distribution, $i = 1, \ldots, a$.
- Y_{ij}: first index – group by the level of the factor, second index – order in the group.
- The particular random samples are stochastically independent.
- Model of one-way ANOVA *(jednofaktorová analýza rozptylu / jednoduché třídění)*:

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where ε_{ij} are stochastically independent random variables with $\mathrm{N}(0, \sigma^2)$ probability distribution, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$.

Notation

level	count	observations	sum	average	distribution
1	n_1	$\boldsymbol{Y}_1 = (Y_{11}, \ldots, Y_{1n_1})'$	$Y_{1\cdot} = \sum_{j=1}^{n_1} Y_{1j}$	$\overline{Y}_{1\cdot} = \frac{1}{n_1} Y_{1\cdot}$	$Y_{1j} \sim N(\mu_1, \sigma^2)$
2	n_2	$\boldsymbol{Y}_2 = (Y_{21}, \ldots, Y_{2n_2})'$	$Y_{2\cdot} = \sum_{j=1}^{n_2} Y_{2j}$	$\overline{Y}_{2\cdot} = \frac{1}{n_2} Y_{2\cdot}$	$Y_{2j} \sim N(\mu_2, \sigma^2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	n_a	$\boldsymbol{Y}_a = (Y_{a1}, \ldots, Y_{an_a})'$	$Y_{a\cdot} = \sum_{j=1}^{n_a} Y_{aj}$	$\overline{Y}_{a\cdot} = \frac{1}{n_a} Y_{a\cdot}$	$Y_{aj} \sim N(\mu_a, \sigma^2)$
	n		$Y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} Y_{ij}$	$\overline{Y}_{\cdot\cdot} = \frac{1}{n} Y_{\cdot\cdot}$	

Dot – summing over the index, overline – averaging.

Statement of the problem

At a significance level of α, we test H_0 against H_1,

H_0: all levels of the factor have equal means,
H_1: at least one pair of levels has different means.

- It is a generalization of the two-sample t-test.
- Note that it is not the same as to apply the two-sample t-test to each of $a(a-1)/2$ pairs of levels. This so-called multiple testing problem) does not guarantee that $P(\text{Type I error}) \leq \alpha$.
- Significance level corrections (e.g., Bonferroni correction) are not feasible for a large number of levels.
- In the 1930s, R. A. Fisher introduced the analysis of variance (ANOVA) *(analýza rozptylu)*, which guarantees $P(\text{Type I error}) = \alpha$.

If the null hypothesis H_0 is rejected, we are further interested in finding which pairs of levels have significantly different means. The so-called multiple comparison *(mnohonásobné porovnávání)* methods are used for this:

- Tukey's method – preferred if all groups have similar sample sizes;
- Scheffé's method – preferred if sample sizes are considerably different.

Model of one-way (single-factor) ANOVA

Definition (M_A, one-way ANOVA model)

Observations Y_{ij} follow model M_A, if

$$Y_{ij} = \underbrace{\mu + \alpha_i}_{\mu_i} + \varepsilon_{ij},$$

for $i = 1, \ldots, a$, $j = 1, \ldots, n_i$,
where ε_{ij} are *i.i.d.* random variables with $\mathrm{N}(0, \sigma^2)$ probability distribution,

- μ = overall / grand mean *(střední hodnota)* of random variable Y,
- α_i = the effect *(efekt)* of the ith level of factor A,
- $\mu_i = \mu + \alpha_i$ = mean of Y by the ith level of factor A,
- ε_{ij} = random errors.

Equivalent expressions of the hypotheses

$$H_0\colon\ \alpha_1 = \cdots = \alpha_a = 0,$$
$$H_1\colon\ \exists i \in \{1, \ldots, a\}\colon\ \alpha_i \neq 0$$

$$H_0\colon\ \mu_1 = \cdots = \mu_a,$$
$$H_1\colon\ \exists i, j \in \{1, \ldots, a\}\colon\ \mu_i \neq \mu_j$$

Definition (M_0, minimal / null model)

Under H_0, observations Y_{ij} follow model M_0, a submodel of M_A,

$$Y_{ij} = \mu + \varepsilon_{ij}$$

Derivation – using linear regression model

Model M:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{1}_{n_{a-1}} & \vdots & & \ddots & \mathbf{1}_{n_{a-1}} & \mathbf{0} \\ \mathbf{1}_{n_a} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}_{n_a} \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Solving the system of *normal equations*: $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & n_1 & n_2 & \cdots & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & \cdots & 0 \\ n_2 & 0 & n_2 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ n_{a-1} & \vdots & & \ddots & n_{a-1} & 0 \\ n_a & 0 & \cdots & \cdots & 0 & n_a \end{pmatrix}, \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \cdots & \mathbf{1}'_{n_{a-1}} & \mathbf{1}'_{n_a} \\ \mathbf{1}'_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_{n_2} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathbf{1}'_{n_{a-1}} & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}'_{n_a} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_{a-1} \\ Y_a \end{pmatrix} = \begin{pmatrix} Y_{\cdot\cdot} \\ Y_{1\cdot} \\ \vdots \\ \vdots \\ Y_{a-1\cdot} \\ Y_{a\cdot} \end{pmatrix}.$$

The design matrix $\boldsymbol{X}$ is not of full rank *(plné hodnosti)*. (Calculate its rank.)

Estimates of parameters

Least-squares estimate of the vector of parameters in linear regression model:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

But the design matrix $\boldsymbol{X}$ is not of full rank, thus $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist.

Any pseudoinverse matrix *(psudoinverzní matice)* $(\boldsymbol{X}'\boldsymbol{X})^{-}$ can be used instead, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$, e.g., $(\boldsymbol{X}'\boldsymbol{X})^{-} = \mathrm{diag}\left(0, \frac{1}{n_1}, \ldots, \frac{1}{n_a}\right)$; or one additional equation is necessary.

Usually, the additional equation is

$$\sum_{i=1}^{a} n_i \alpha_i = 0,$$

leading to following estimators:

overall (grand) mean:	$\widehat{\mu} = \overline{Y}_{\cdot\cdot}$
effects (A):	$\widehat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$
mean of group $(A = i)$:	$\widehat{\mu}_i = \widehat{\mu} + \widehat{\alpha}_i = \overline{Y}_{i\cdot}$
- in model M_0:	$\widehat{\mu}_i = \widehat{\mu} = \overline{Y}_{\cdot\cdot}$

Sum of squares quantifies the variability

- Total sum of squares *(celkový součet čtverců)*
  = variability of observations around the overall mean:

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2 \qquad \sim \chi^2(df_T = n - 1),$$

- Between-groups sum of squares *(regresní součet čtverců)*
  = variability of group means around the overall mean, i.e., explained by factor A:

$$S_A = \sum_{i=1}^{a} n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2 \qquad \sim \chi^2(df_A = a - 1),$$

- Within-groups / error / residual sum of squares *(reziduální součet čtverců)*
  = variability of observations within each group around the group mean, i.e., unexplained by factor A:

$$S_e = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2 \qquad \sim \chi^2(df_e = n - a).$$

The df quantities are degrees of freedom *(stupně volnosti)* of the statistics.

Theorem

$$S_T = S_A + S_e$$

The testing in one-way ANOVA relies on comparison of models M and M_0.

Theorem (Omnibus ANOVA F-test)

$$F_A = \frac{MS_A}{MS_e} = \frac{\frac{S_A}{df_A}}{\frac{S_e}{df_e}} = \frac{\frac{S_A}{a-1}}{\frac{S_e}{n-a}} = \frac{\frac{S_T - S_e}{df_T - df_e}}{\frac{S_e}{df_e}} = \left(\frac{S_T}{S_e} - 1\right)\frac{n-a}{a-1}.$$

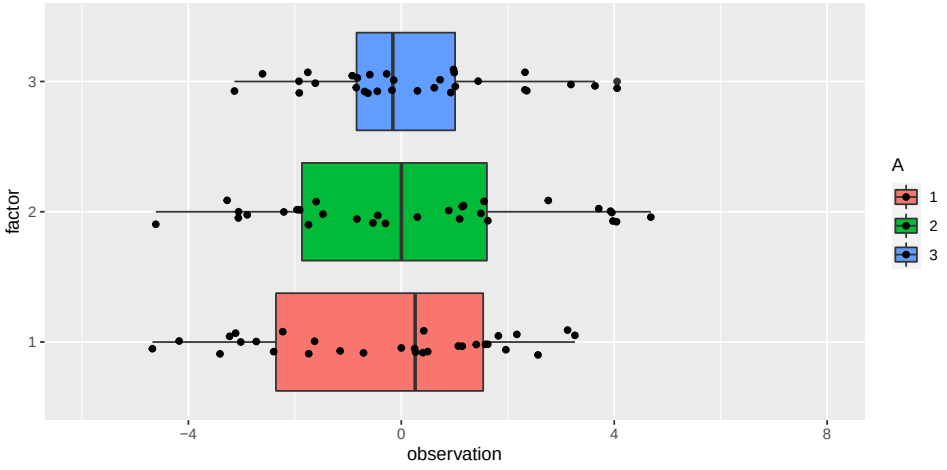
Under the null hypothesis H_0, i.e, if M_0 is correct,
the statistic F_A has Fisher-Snedecor $\mathrm{F}(a-1, n-a)$ probability distribution with $(a-1)$ and $(n-a)$ degrees of freedom.

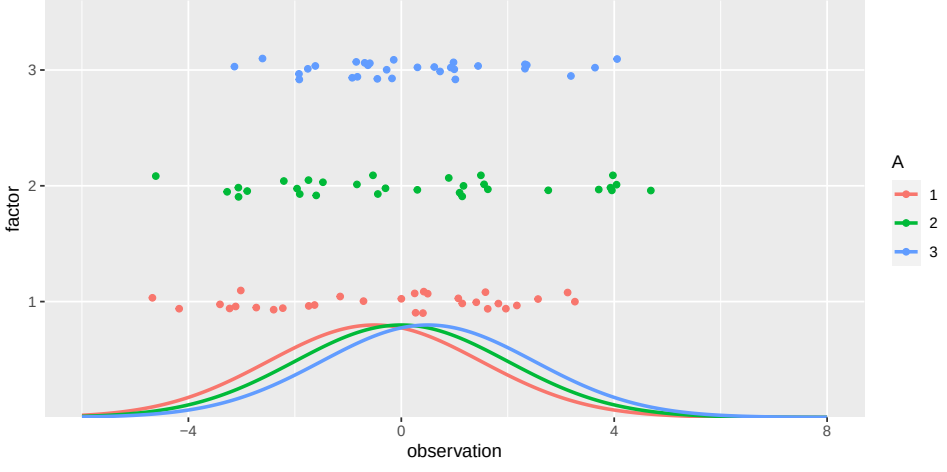
The null hypothesis H_0 is rejected, i.e., factor A is not significant, if

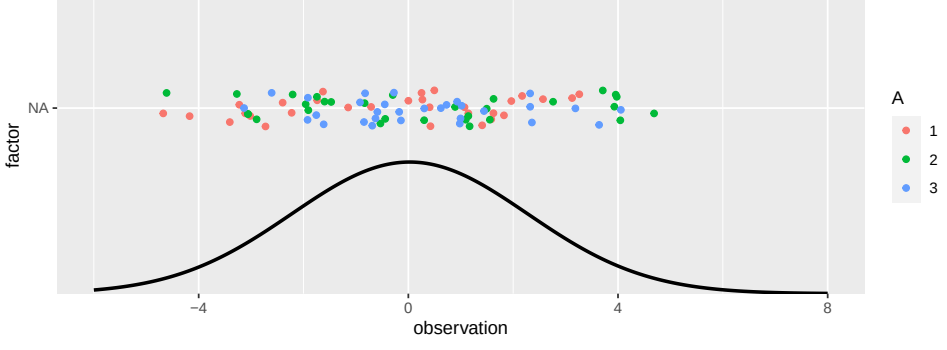
$$F_A \geq F_{1-\alpha}(a-1, n-a).$$

Why the test statistic F_A has Fisher-Snedecor probability distribution?

Illustration: similar means





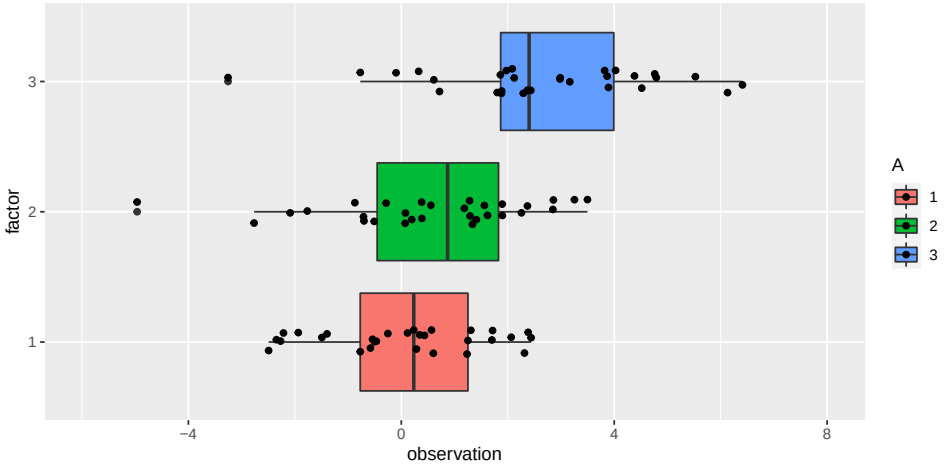


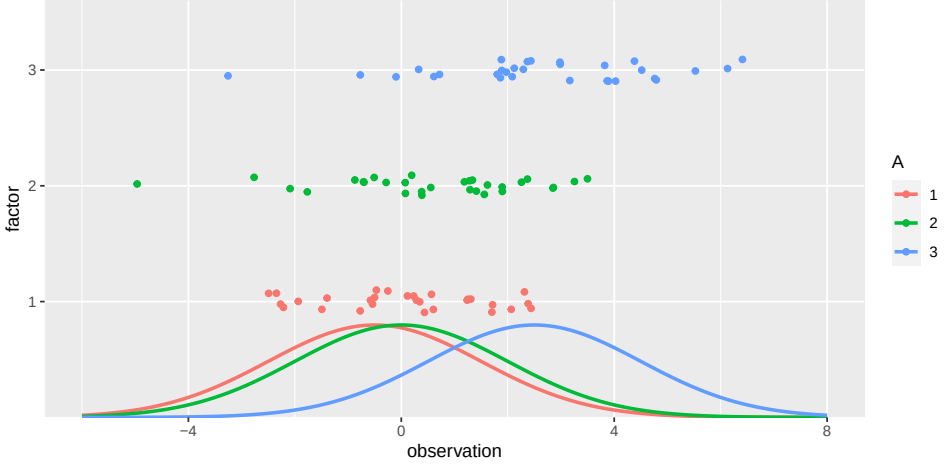
$$n_1 = n_2 = n_3 = 30,$$

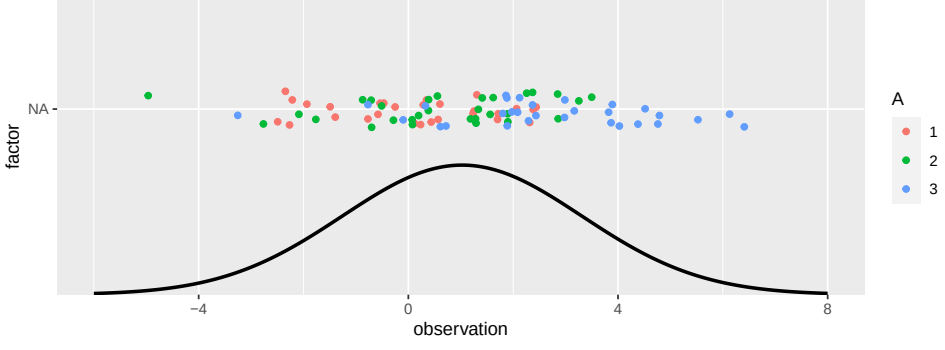
$$\mu_1 = -0.5, \mu_2 = 0, \mu_3 = 0.5, \sigma^2 = 4$$

```
      S.T       S.A       S.e
 450.6129 6.443252 444.1697

            Df Sum Sq Mean Sq F value
A            2    6.4   3.222   0.631
Residuals   87  444.2   5.105
```

Illustration: significantly different means







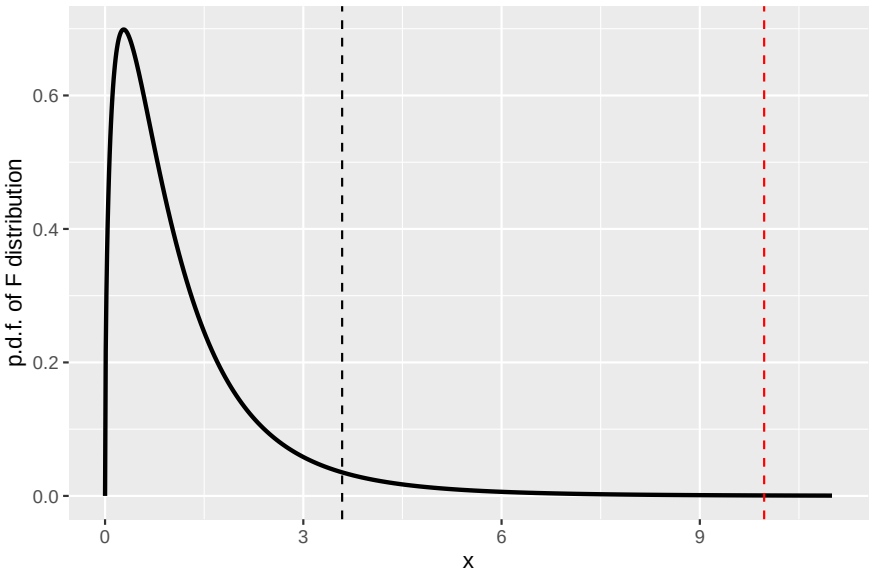
$$n_1 = n_2 = n_3 = 30,$$

$$\mu_1 = -0.5, \mu_2 = 0, \mu_3 = 2.5, \sigma^2 = 4$$

```
      S.T       S.A       S.e
560.0406 218.7568 341.2838

            Df Sum Sq Mean Sq F value
A            2  218.8  109.38   27.88
Residuals   87  341.3    3.92
```

ANOVA table

source of variability	degrees of freedom	sum of squares	mean squares	value of the test statistic	p-value
group	$df_A = a - 1$	S_A	$MS_A = \frac{S_A}{df_A}$	$F_A = \frac{MS_A}{MS_e}$	p_A
residual	$df_e = n - a$	S_e	$MS_e = \frac{S_e}{df_e}$		
total	$df_T = n - 1$	S_T			



Example – potatoes:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Variety	3	0.816	0.27200	9.973	0.0018
Residuals	11	0.300	0.02727		

Multiple comparison

For chosen $k \neq l$, we test the equality of means in the kth and lth level:

H_0: $\mu_k = \mu_l$,
H_1: $\mu_k \neq \mu_l$

Theorem (Tukey's method)

H_0 is rejected at the level of significance α, if

$$|\overline{Y}_{k\cdot} - \overline{Y}_{l\cdot}| \geq \sqrt{\frac{S_e}{(n-a)n_k}}\, q_{1-\alpha}(a, n-a),$$

where $q_{1-\alpha}(a, n-a)$ are quantiles (numerically computed) of the studentized range.

Theorem (Scheffé's method)

H_0 is rejected at the level of significance α, if

$$|\overline{Y}_{k\cdot} - \overline{Y}_{l\cdot}| \geq \sqrt{S_e \frac{a-1}{n-a}\left(\frac{1}{n_k} + \frac{1}{n_l}\right) F_{1-\alpha}(a-1, n-a)}.$$

Coefficients of the linear model of ANOVA in software

Different parametrization of the one-way ANOVA is used by most of the statistical software (including *R*).

Definition

Random variables Y_{ij} follow the model

$$Y_{ij} = \mu^* + \alpha_i^* + \varepsilon_{ij},$$

for $i = 1, \ldots, a$, $j = 1, \ldots, n_i$,
where ε_{ij} are *i.i.d.* random variables with $\mathrm{N}(0, \sigma^2)$ probability distribution,

- $\mu^* = \mu_1$ = mean of the first level of factor A, i.e., $\alpha_1^* = 0$,
- α_i^* = the effect of the ith level of factor A, $\alpha_1^* = 0$ is fixed,
- $\mu_i = \mu^* + \alpha_i^*$ = mean of Y by the ith level of factor A.

Equivalent expressions of the hypotheses

$$H_0\colon\ \alpha_2^* = \cdots = \alpha_a^* = 0,$$
$$H_1\colon\ \exists i \in \{2, \ldots, a\}\colon\ \alpha_i^* \neq 0$$

Testing the homogeneity of variances and normality

To verify the homogeneity of variances, i.e. to verify the consistency of variances in individual levels of the factor, we use

- Levene's test,
- Bartlett's test.

To verify the normality of the observations in each group, we use

- normal QQ-plot,
- Lilliefors test,
- Shapiro–Wilk test.

Levene's and Bartlett's test of homogeneity of variances

Theorem (Levene's test)

Denote $Z_{ij} = |Y_{ij} - \overline{Y}^*_{i\cdot}|$, where $\overline{Y}^*_{i\cdot}$ is sample mean / median / 10% trimmed mean. Under the hypothesis of homogeneity of variances, test statistic

$$L = \frac{\frac{1}{a-1}\sum_{i=1}^{a} n_i \left(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot}\right)^2}{\frac{1}{n-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_{i\cdot}\right)^2}$$ has $\mathrm{F}(a-1, n-a)$ probability distribution.

The homogeneity of variances is rejected at the level of significance α, if $L \geq F_{1-\alpha}(a-1, n-a)$.

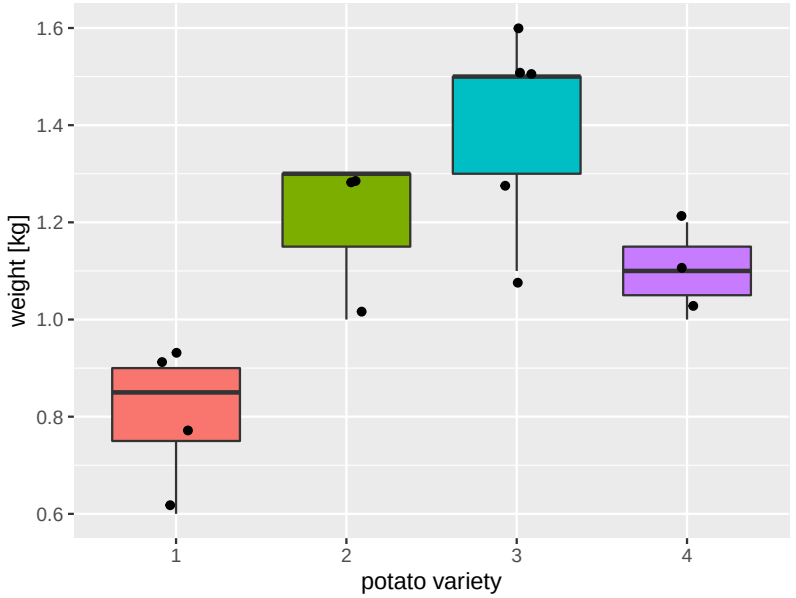
Theorem (Bartlett's test)

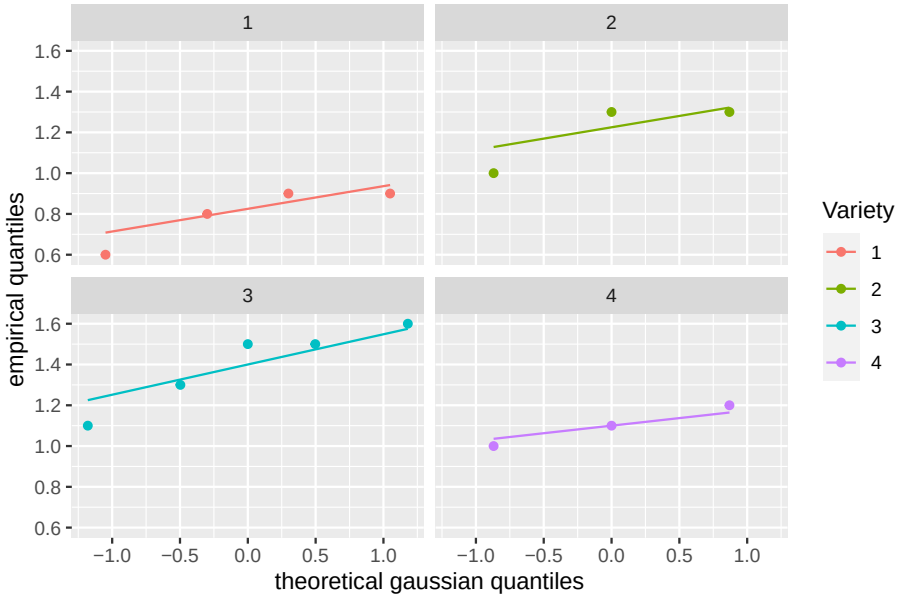
Under the hypothesis of homogeneity of variances, test statistic

$$B = \frac{1}{C}\left[(n-a)\ln\frac{S_e}{n-a} - \sum_{i=1}^{a}(n_i - 1)\ln S_i^2\right]$$ has asymptotically $\chi^2(a-1)$ probability distribution. Here, S_j^2 denotes the sample variance in the ith level of the factor and $C = 1 + \frac{1}{3(a-1)}\left(\sum_{j=1}^{a}\frac{1}{n_j - 1} - \frac{1}{n-a}\right)$. The homogeneity of variances is rejected at the level of significance α, if $B \geq \chi^2_{1-\alpha}(a-1)$.

Example: ANOVA in *R*


```
'data.frame': 15 obs. of  2 variables:
 $ Weight : num  0.9 0.8 0.6 0.9 1.3 1 1.3 1.3 1.5 1.6 ...
 $ Variety: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 2 2 2 3 3 3 ...
```






```
  Bartlett test of homogeneity of variances
data:  Weight by Variety
Bartlett's K-squared=1.0417, df=3, p-value=0.7912

Levene's Test for Homogeneity of Variance (center=median)
      Df F value Pr(>F)
group  3  0.1874 0.9027

  Variety Shapiro.p.value
  <fct>             <dbl>
1 1                 0.161
2 2                 0
```

ANOVA

```
# using "aov"
aov.model <- aov(Weight~Variety, data=dt)
# or using "lm" and "anova"
M.A <- lm(Weight~Variety, data=dt)
anova.model <- anova(M.A)
```

ANOVA table

```
            Df Sum Sq Mean Sq F value Pr(>F)
Variety      3  0.816 0.27200   9.973 0.0018 **
Residuals   11  0.300 0.02727
```

Estimates of coefficients of the linear regression model

```
(Intercept)    Variety2    Variety3    Variety4
        0.8         0.4         0.6         0.3
```

Estimates of effects and means

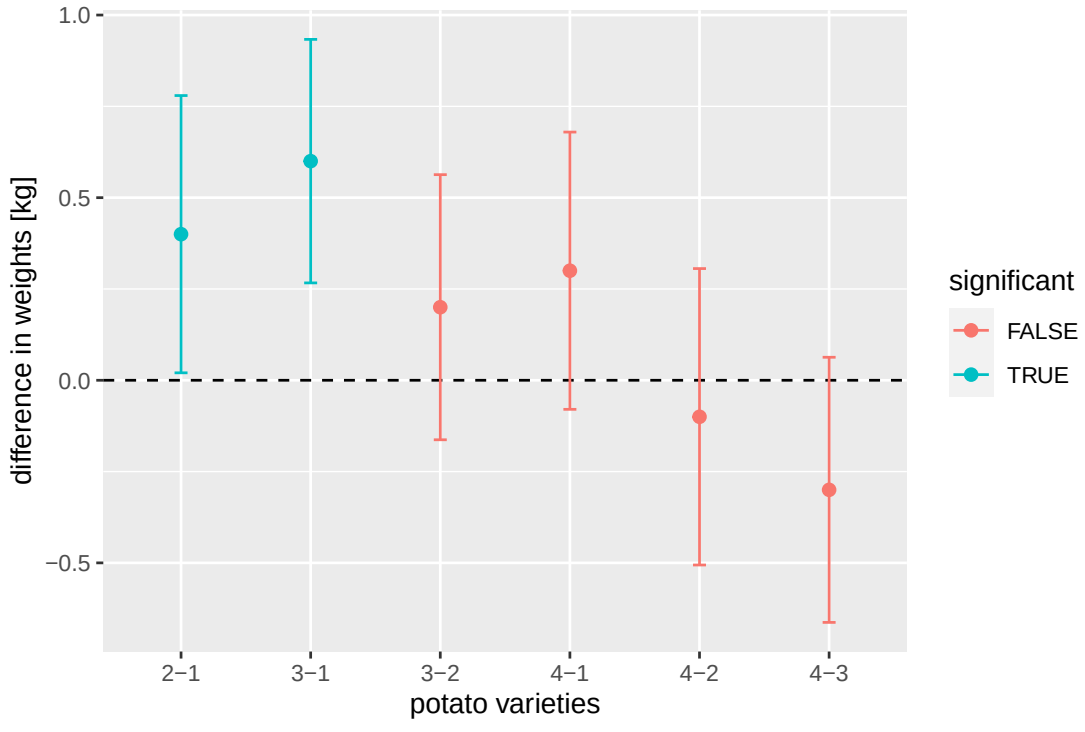
```
Tables of effects
 Variety
        1    2    3     4
    -0.34 0.06 0.26 -0.04
rep  4.00 3.00 5.00  3.00
```

```
Tables of means
Grand mean
1.14
 Variety
      1   2   3   4
    0.8 1.2 1.4 1.1
rep 4.0 3.0 5.0 3.0
```

Example: multiple comparison

Tukey's method:

	diff	lwr	upr	p adj
2-1	0.4	0.02040199	0.77959801	0.0381806
3-1	0.6	0.26659524	0.93340476	0.0010299
4-1	0.3	-0.07959801	0.67959801	0.1391459
3-2	0.2	-0.16296512	0.56296512	0.3885221
4-2	-0.1	-0.50580735	0.30580735	0.8783019
4-3	-0.3	-0.66296512	0.06296512	0.1172041



Significantly different: 1–2 and 1–3

Scheffé's method:

	Weight	groups
3	1.4	a
2	1.2	ab
4	1.1	ab
1	0.8	b

Significantly different: 1–3

Using ANOVA to compare nested linear regression models

ANOVA – comparison of submodels

Assume two linear regression models for the **same data** of size n:

1. model M_1 with design matrix $\boldsymbol{X}_1$ with rank $r_1 = r(\boldsymbol{X}_1)$ and residual sum of squares $S_{e\,1}$;
2. **submodel** M_2 of model M_1 with design matrix $\boldsymbol{X}_2$ with rank $r_2 = r(\boldsymbol{X}_2)$ which is formed by omitting some columns of $\boldsymbol{X}_1$, and with residual sum of squares $S_{e\,2}$.

Asumming the validity of model M_1, we further test

H_0: model M_2 is valid, too, i.e., M_1 can be simplified to M_2,
H_1: model M_2 is not valid.

Under the null hypothesis H_0, test statistic

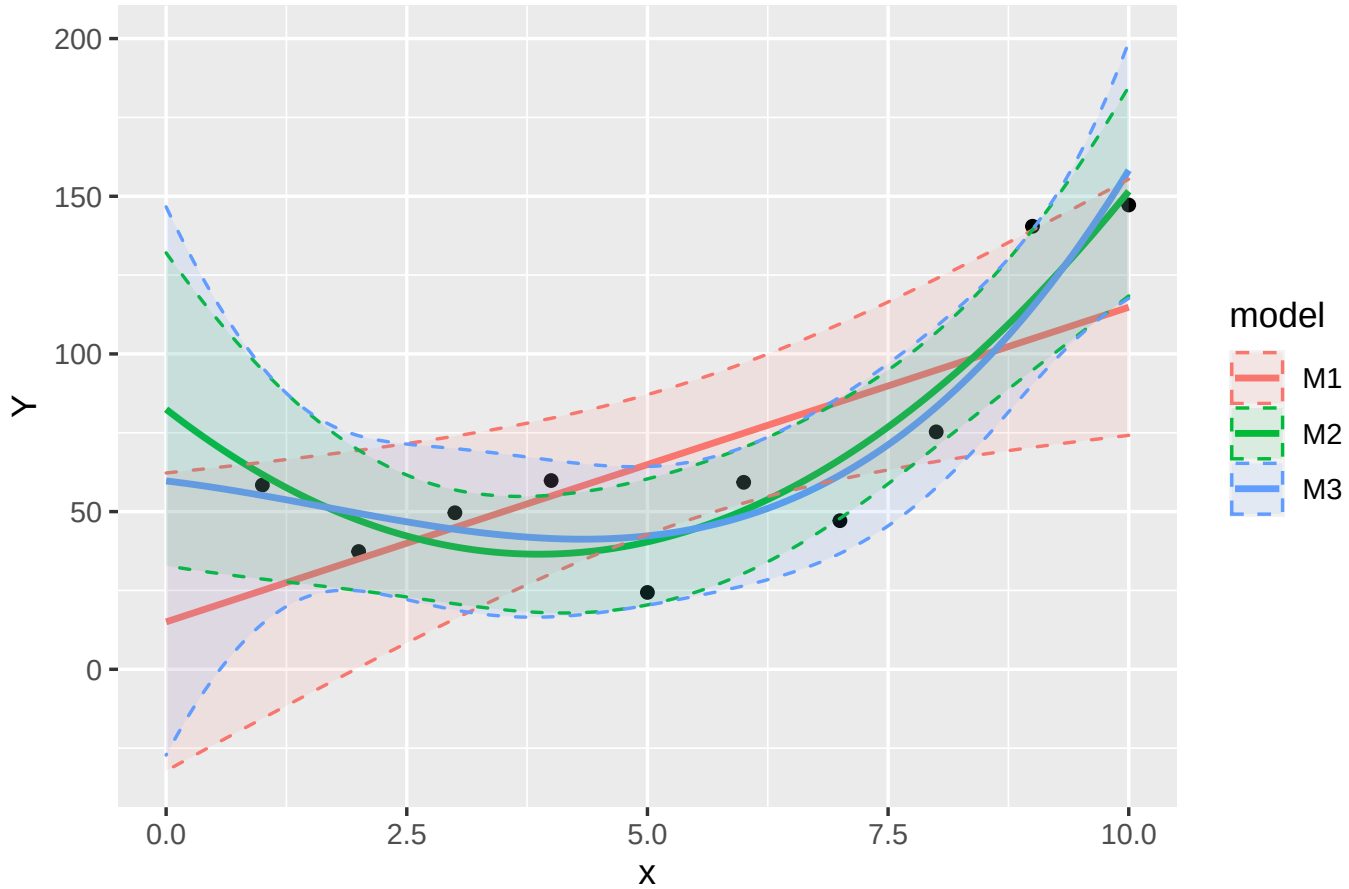
$$F = \frac{\frac{S_{e2}-S_{e1}}{r_1-r_2}}{\frac{S_{e1}}{n-r_1}} = \frac{S_{e2}-S_{e1}}{S_{e1}} \cdot \frac{n-r_1}{r_1-r_2} = \left(\frac{S_{e2}}{S_{e1}} - 1\right) \frac{n-r_1}{r_1-r_2}$$

has Fisher-Snedecor $\mathrm{F}(r_1 - r_2, n - r_1)$ probability distribution.

H_0 is rejected at the level of significance α, if $F \geq F_{1-\alpha}(r_1 - r_2, n - r_1)$.

Obviously $r_2 < r_1 < n$, $S_{e\,2} \geq S_{e\,1}$. Compare F statistic with one-way ANOVA.

Example: Which of these linear models is the best?




```
'data.frame': 10 obs. of  2 variables:
 $ x: int  1 2 3 4 5 6 7 8 9 10
 $ Y: num  58.4 37.3 49.6 59.9 24.4 ...
```

- $M_3 : \widehat{Y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
- $M_2 : \widehat{Y} = \beta_0 + \beta_1 x + \beta_2 x^2$
- $M_1 : \widehat{Y} = \beta_0 + \beta_1 x$

M_3

```
lm(formula=Y~1 + x + I(x^2) + I(x^3), data=dt)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.7207    35.5280   1.681    0.144
x            -3.5743    26.6299  -0.134    0.898
I(x^2)       -1.3063     5.4929  -0.238    0.820
I(x^3)        0.2649     0.3294   0.804    0.452
Residual standard error: 18.31 on 6 degrees of freedom
Multiple R-squared:  0.8694,  Adjusted R-squared:  0.8042
F-statistic: 13.32 on 3 and 6 DF,  p-value: 0.004624
```

M_2

```
lm(formula=Y~1 + x + I(x^2), data=dt)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.4500    20.9805   3.930  0.00568 **
x           -23.7340     8.7623  -2.709  0.03026 *
I(x^2)        3.0647     0.7763   3.948  0.00555 **
Residual standard error: 17.84 on 7 degrees of freedom
Multiple R-squared:  0.8554,  Adjusted R-squared:  0.814
F-statistic:  20.7 on 2 and 7 DF,  p-value: 0.001151
```

M_1

```
lm(formula=Y~1 + x, data=dt)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.027     20.475   0.734   0.4840
x              9.978      3.300   3.024   0.0165 *
Residual standard error: 29.97 on 8 degrees of freedom
Multiple R-squared:  0.5333,  Adjusted R-squared:  0.475
F-statistic: 9.143 on 1 and 8 DF,  p-value: 0.01647
```

Compare M_3 and M_2

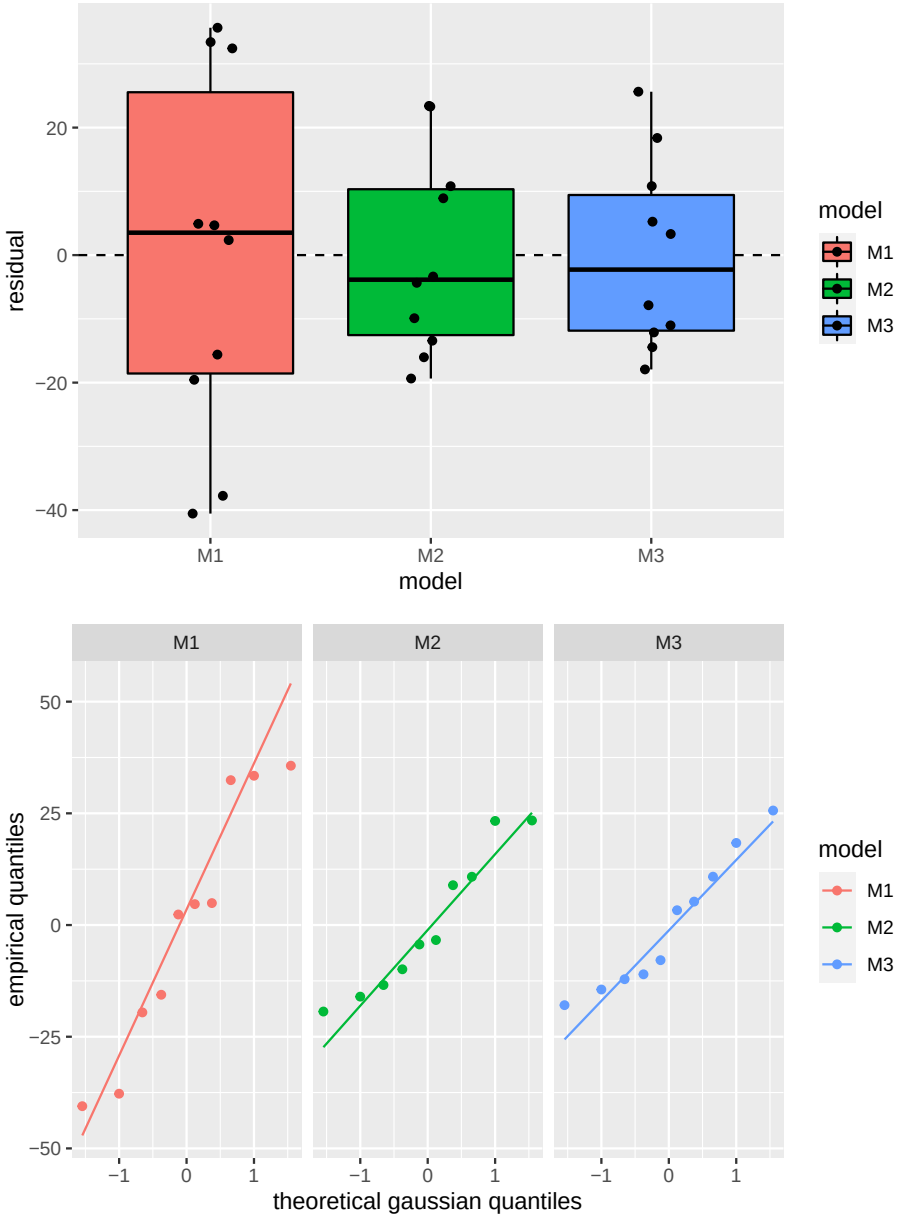
```
anova(M3, M2)

Analysis of Variance Table
Model 1: Y~1 + x + I(x^2) + I(x^3)
Model 2: Y~1 + x + I(x^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      6 2010.7
2      7 2227.4 -1   -216.76 0.6469 0.4519
```

Compare M_2 and M_1

```
anova(M2, M1)

Analysis of Variance Table
Model 1: Y~1 + x + I(x^2)
Model 2: Y~1 + x
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      7 2227.4
2      8 7186.6 -1   -4959.2 15.585 0.0055
```

And the winner is… M_2

Check residuals…



Two-way ANOVA

Example

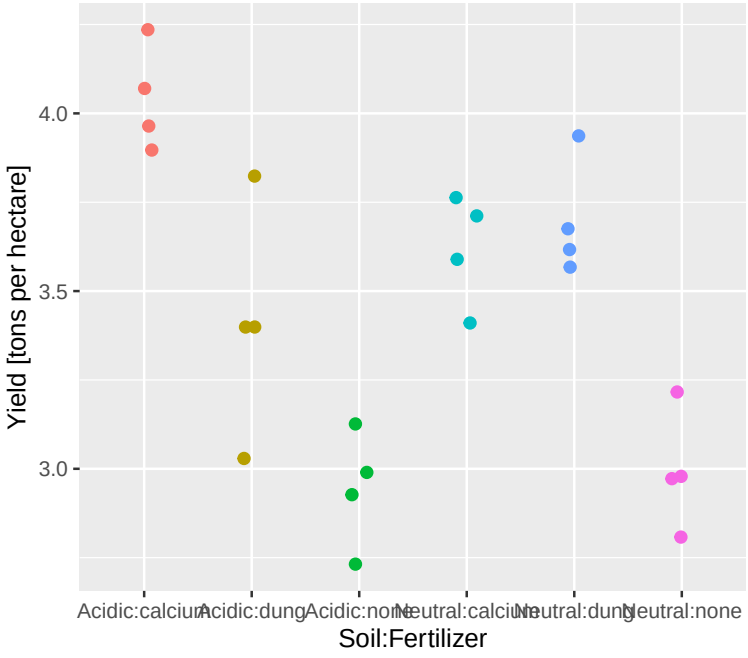
Examination of hay yields (tons per hectare) based on the type of soil (normal; sour) and fertilizer (none; dung; calcium).

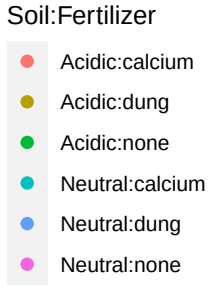
soil type (A)	fertilizer (B) none	dung	calcium
normal	2.8, 3.2, 3.0, 3.0	3.7, 3.6, 3.9, 3.6	3.4, 3.8, 3.7, 3.6
sour	3.1, 2.7, 3.0, 2.9	3.4, 3.4, 3.0, 3.8	4.2, 4.0, 4.1, 3.9

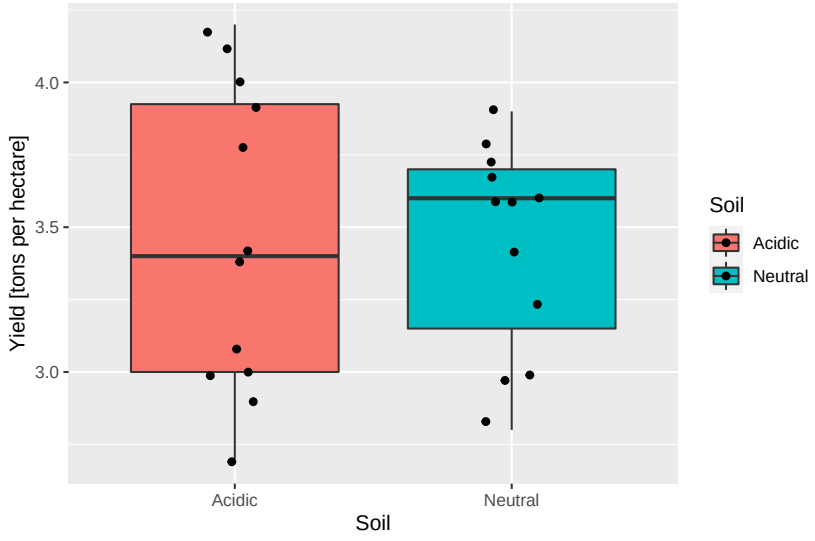
At a 5% significance level, test following hypotheses:

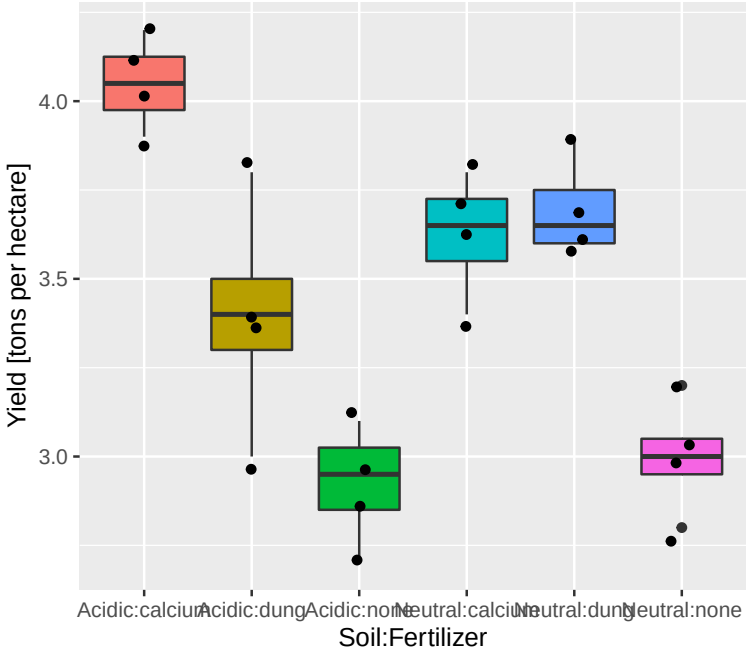
- Soil type does not significantly affect hay yields.
- Method of fertilization does not significantly affect hay yields.
- Soil type and method of fertilization do not interact with respect to hay yields.

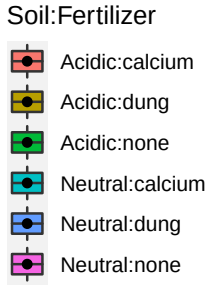
If you reject the null hypothesis, find which pairs differ.

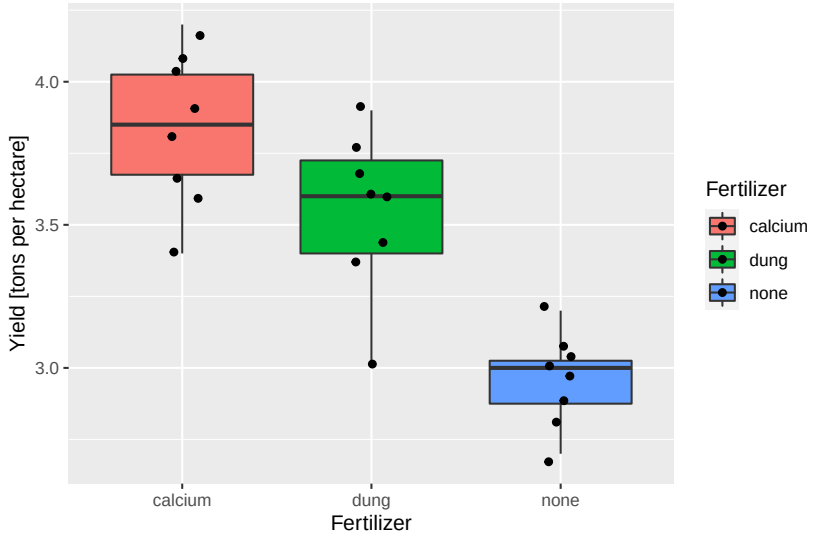

Yield [tons per hectare]
4.0
3.5
3.0
Acidic:calcium
Acidic:dung
Acidic:none
Neutral:calcium
Neutral:dung
Neutral:none
Soil:Fertilizer


Soil:Fertilizer
Acidic:calcium
Acidic:dung
Acidic:none
Neutral:calcium
Neutral:dung
Neutral:none


Yield [tons per hectare]
4.0
3.5
3.0
Acidic
Neutral
Soil
Soil
Acidic
Neutral


Yield [tons per hectare]
4.0
3.5
3.0
Acidic:calcium
Acidic:dung
Acidic:none
Neutral:calcium
Neutral:dung
Neutral:none
Soil:Fertilizer


Soil:Fertilizer
Acidic:calcium
Acidic:dung
Acidic:none
Neutral:calcium
Neutral:dung
Neutral:none


Yield [tons per hectare]
4.0
3.5
3.0
calcium
dung
none
Fertilizer
Fertilizer
calcium
dung
none

Notation and assumptions

- Two factors, A has $a \geq 2$ levels, B has $b \geq 2$ levels
- The combination of i th level of factor A and j th level of factor B has n_{ij} observations $Y_{ij1}, \ldots, Y_{ijn_{ij}}$, which form a random sample from $\mathrm{N}(\mu_{ij}, \sigma^2)$ probability distribution.
- Y_{ijk}: first index – group by the level of factor A, second index – group by the level of factor B, third index – order in the group.
- The particular random samples are stochastically independent.
- Model of two-way ANOVA *(dvoufaktorová analýza rozptylu / dvojné třídění)*:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

where $\varepsilon_{ij}\,k$ are stochastically independent random variables with $\mathrm{N}(0, \sigma^2)$ probability distribution, $i = 1, \ldots, a,\ j = 1, \ldots, b,\ j = k, \ldots, n_{ij}$.

Models and parameters in two-way ANOVA

$$
\begin{aligned}
M_+\colon &\quad Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \varepsilon_{ijk} \\
M_2\colon &\quad Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\
M_B\colon &\quad Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \\
M_A\colon &\quad Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \\
M_0\colon &\quad Y_{ijk} = \mu + \varepsilon_{ijk}
\end{aligned}
$$

for $i = 1, \ldots, a$, for $i = j, \ldots, b$, $k = 1, \ldots, n_{ij}$,
where ε_{ijk} are *i.i.d.* random variables with $\mathrm{N}(0, \sigma^2)$ probability distribution,

- μ = overall mean (grand mean) of the random variable Y,
- α_i = the (row) effect of the i th level of factor A,
- β_j = the (column) effect of the j th level of factor B,
- λ_{ij} = the interaction of the i th level of factor A and j th level of factor B,
- $\mu_{ij} = \mu + \alpha_i + \beta_j + \lambda_{ij}$ = mean of Y by the i th level of factor A and j th level of factor B,
- ε_{ij} = random errors.

Hypotheses ans submodels in two-way ANOVA

$$
\begin{aligned}
M_+\colon &\quad Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \varepsilon_{ijk} \\
M_2\colon &\quad Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\
M_B\colon &\quad Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \\
M_A\colon &\quad Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \\
M_0\colon &\quad Y_{ijk} = \mu + \varepsilon_{ijk}
\end{aligned}
$$

Interactions

$H_{0\,AB}$: all $\lambda_{ij} = 0$, i.e., the interaction is not significant,
$H_{1\,AB}$: $\exists\, i, j : \lambda_{ij} \neq 0$, i.e., the interaction is significant

Factor B

$H_{0\,B}$: all $\beta_j = 0$, i.e., factor B is not significant,
$H_{1\,B}$: $\exists\, j : \beta_j \neq 0$, i.e., factor B is significant

Factor A

$H_{0\,A}$: all $\alpha_i = 0$, i.e., factor A is not significant,
$H_{1\,A}$: $\exists\, i : \alpha_i \neq 0$, i.e., factor A is significant

Possible sequences of submodels

$$
M_+ \xrightarrow{H_{0\,AB}} M_2 \xrightarrow{H_{0\,B}} M_B \xrightarrow{H_{0\,A}} M_0, \quad \text{or} \quad M_+ \xrightarrow{H_{0\,AB}} M_2 \xrightarrow{H_{0\,A}} M_A \xrightarrow{H_{0\,B}} M_0
$$

Estimates of parameters

Model M_+ (and its submodels similarly) is written as linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\,(\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b, \lambda_{11}, \ldots, \lambda_{ab})'$$

with design matrix $\boldsymbol{X}$ of size $n \times (1 + a + b + ab)$ and rank $r(\boldsymbol{X}) = ab$.

Additional equations $\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \lambda_{ij} = 0, \quad \sum_{j=1}^{b} \lambda_{ij} = 0$

leads to model of full rank with following estimators:

overall (grand) mean:	$\widehat{\mu} = \overline{Y}_{\cdot\cdot\cdot}$
interactions:	$\widehat{\lambda}_{ij} = \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot}$
row (A) effects:	$\widehat{\alpha}_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}$
column (B) effects:	$\widehat{\beta}_j = \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot}$
mean of group $(A = i, B = j)$:	$\widehat{\mu}_{ij} = \overline{Y}_{ij\cdot}$
- in model M_2:	$\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j = \overline{Y}_{i\cdot\cdot} + \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot}$
- in model M_B:	$\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\alpha}_i = \overline{Y}_{i\cdot\cdot}$
- in model M_A:	$\widehat{\mu}_{ij} = \widehat{\mu} + \widehat{\beta}_j = \overline{Y}_{\cdot j\cdot}$
- in model M_0:	$\widehat{\mu}_{ij} = \widehat{\mu} = \overline{Y}_{\cdot\cdot}$

Sum of squares quantifies the variability

- $S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y}_{...})^2$, $\sim \chi^2(df_T = n-1)$,
- $S_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2 \sim \chi^2(df_{AB} = (a-1)(b-1))$,
- $S_B = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\,(\overline{Y}_{.j.} - \overline{Y}_{...})^2$, $\sim \chi^2(df_B = b-1)$,
- $S_A = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\,(\overline{Y}_{i..} - \overline{Y}_{...})^2$, $\sim \chi^2(df_A = a-1)$,
- $S_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y}_{ij})^2$, $\sim \chi^2(df_e = n - a\,b)$.

Theorem

$$S_T = S_{AB} + S_A + S_B + S_e$$

Table of two-way ANOVA with interactions

source of variability	degrees of freedom	sum of squares	mean squares	value of the test statistic	p-value
row A	$df_A = a - 1$	S_A	$MS_A = \frac{S_A}{df_A}$	$F_A = \frac{MS_A}{MS_e}$	p_A
column B	$df_B = b - 1$	S_B	$MS_B = \frac{S_B}{df_B}$	$F_B = \frac{MS_B}{MS_e}$	p_B
interaction	$df_{AB} = (a-1)(b-1)$	S_{AB}	$MS_{AB} = \frac{S_{AB}}{df_{AB}}$	$F_{AB} = \frac{MS_{AB}}{MS_e}$	p_{AB}
residual	$df_e = n - a\,b$	S_e	$MS_e = \frac{S_e}{df_e}$		
total	$df_T = n - 1$	S_T			

When any null hypothes is rejected, the multiple comparison usually follows.

Testing in two-way ANOVA with interactions

1. Test the significance of the interactions using F_{AB}:
$$\text{if } F_{AB} \geq F_{1-\alpha}((a-1)(b-1), n-ab), \text{ reject } H_{0\,AB}.$$
2. Test the significance of the B (column) factor using F_B,
at the same time, take into account effects of (row) factor A:
$$\text{if } F_B \geq F_{1-\alpha}(b-1, n-ab), \text{ reject } H_{0\,B}.$$
3. Test the significance of the A (row) factor using F_A,
do not take into account effects of (row) factor B:
$$\text{if } F_A \geq F_{1-\alpha}(a-1, n-ab), \text{ reject } H_{0\,A}.$$

Or use the corresponding p-values from ANOVA table, **from the bottom up**.

Two-way ANOVA without interactions

The interactions can be omitted. (Advantage: fewer parameters.)

Then, $S_{AB} = 0$, $df_e = n - a - b + 1$, the testing procedure starts with $H_{0\,B}$.

Coefficients of the linear model of two-way ANOVA in software

Different parametrization of the one-way ANOVA is used by most of the statistical software (including *R*).

Definition

Random variables Y_{ijk} follow model

$$Y_{ijk} = \mu^* + \alpha_i^* + \alpha_j^* + \lambda_{ij}^* + \varepsilon_{ij},$$

for $i = 1, \ldots, a$, $i = j, \ldots, b$, $k = 1, \ldots, n_{ij}$,

where ε_{ijk} are *i.i.d.* random variables with $\mathrm{N}(0, \sigma^2)$ probability distribution,

- $\mu^* = \mu_{11}$ = mean of the *top left* category $(A = 1, B = 1)$,
- $\alpha_1^* = 0$, $\beta_1^* = 0$, $\lambda_{ij}^* = 0$ are fixed,
- other effects $(i \geq 2$ or $j \geq 2)$ and interactions express the deviations in mean from the category $(A = 1, B = 1)$,
- $\mu_i = \mu^* + \alpha_i^* + \beta_j^* + \lambda_{ij}^*$ = mean of Y by the category $(A = i, B = j)$.

Example: Two-way ANOVA without interactions in *R*


```
'data.frame': 24 obs. of  4 variables:
 $ Soil      : Factor w/ 2 levels "Acidic","Neutral": 2 2 2 2 2 2 2 2 2 2 ...
 $ Fertilizer: Factor w/ 3 levels "calcium","dung",..: 3 3 3 3 2 2 2 2 1 1 ...
 $ Yield     : num  2.8 3.2 3 3 3.7 3.6 3.9 3.6 3.4 3.8 ...
```

```
aov.model <- aov(Yield~Soil+Fertilizer, data=dt)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Soil          1  0.002  0.0017   0.027    0.871
Fertilizer    2  3.182  1.5912  25.752 2.93e-06 ***
Residuals    20  1.236  0.0618
```

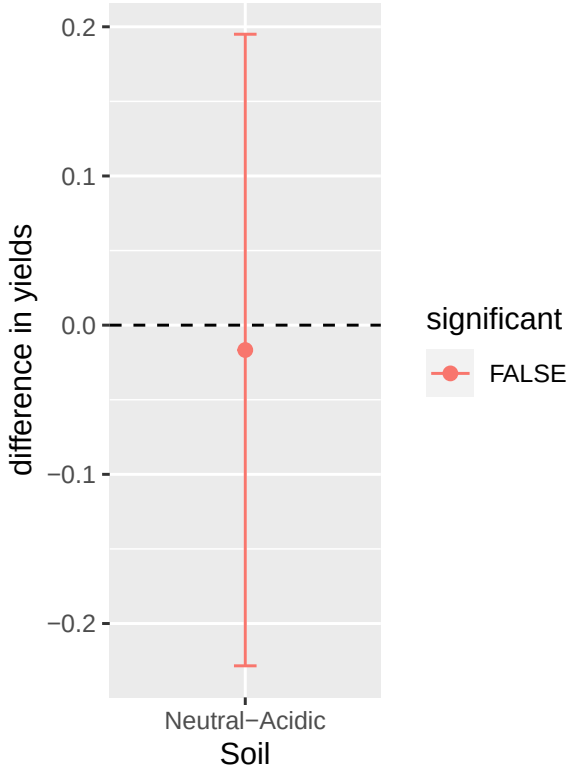
```
   (Intercept)    SoilNeutral Fertilizerdung Fertilizernone
    3.84583333    -0.01666667    -0.28750000    -0.87500000
```

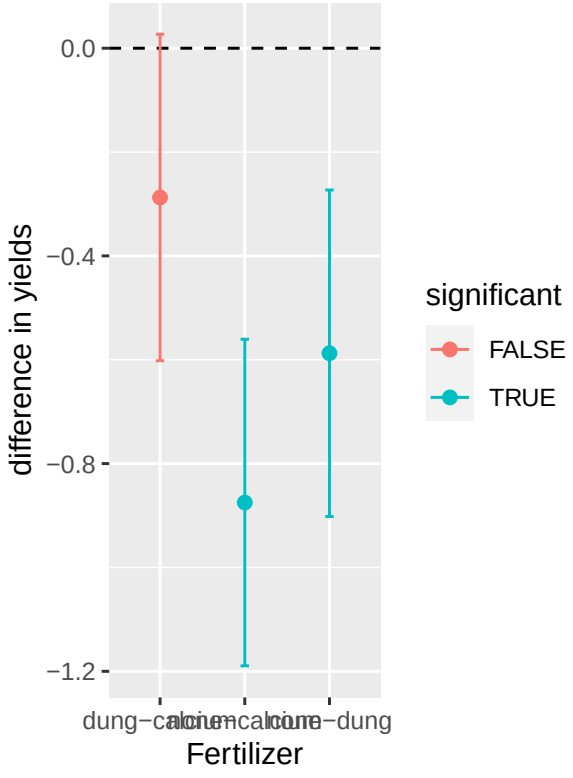
```
        Tables of effects
 Soil
Soil
   Acidic    Neutral
 0.008333 -0.008333
 Fertilizer
Fertilizer
calcium    dung    none
 0.3875  0.1000 -0.4875
```

```
   Tables of means
Grand mean
3.45
 Soil
Soil
 Acidic Neutral
  3.458   3.442
 Fertilizer
Fertilizer
calcium    dung    none
  3.838   3.550   2.963
```

Example: multiple comparison in two-way ANOVA without interactions

Tukey's method:





Scheffé's method:

	Yield	groups
Acidic	3.458333	a
Neutral	3.441667	a

	Yield	groups
calcium	3.8375	a
dung	3.5500	a
none	2.9625	b

	Yield	groups
Acidic:calcium	4.050	a
Neutral:dung	3.700	ab
Neutral:calcium	3.625	abc
Acidic:dung	3.400	bcd
Neutral:none	3.000	cd
Acidic:none	2.925	d

Example: Two-way ANOVA with interactions in *R*


```
aov.model <- aov(Yield~Soil+Fertilizer+Soil:Fertilizer, data=dt) # or simply
aov.model <- aov(Yield~Soil*Fertilizer, data=dt)
```

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
Soil              1  0.002  0.0017   0.044  0.83658
Fertilizer        2  3.182  1.5912  41.814 1.72e-07 ***
Soil:Fertilizer   2  0.551  0.2754   7.237  0.00494 **
Residuals        18  0.685  0.0381
```

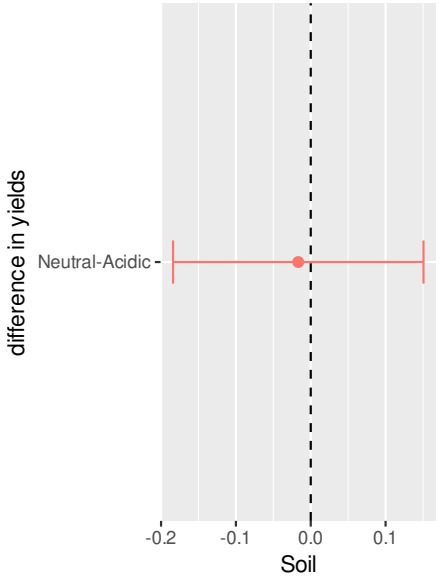
```
   (Intercept)                SoilNeutral             Fertilizerdung
         4.050                     -0.425                     -0.650
Fertilizernone SoilNeutral:Fertilizerdung SoilNeutral:Fertilizernone
        -1.125                      0.725                      0.500
```

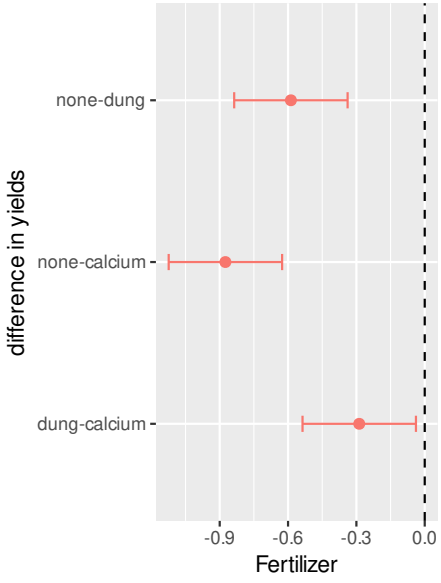
```
Tables of effects
 Soil
Soil
   Acidic    Neutral
 0.008333 -0.008333
 Fertilizer
Fertilizer
calcium    dung    none
 0.3875  0.1000 -0.4875
 Soil:Fertilizer
         Fertilizer
Soil      calcium  dung     none
  Acidic   0.20417 -0.15833 -0.04583
  Neutral -0.20417  0.15833  0.04583
```

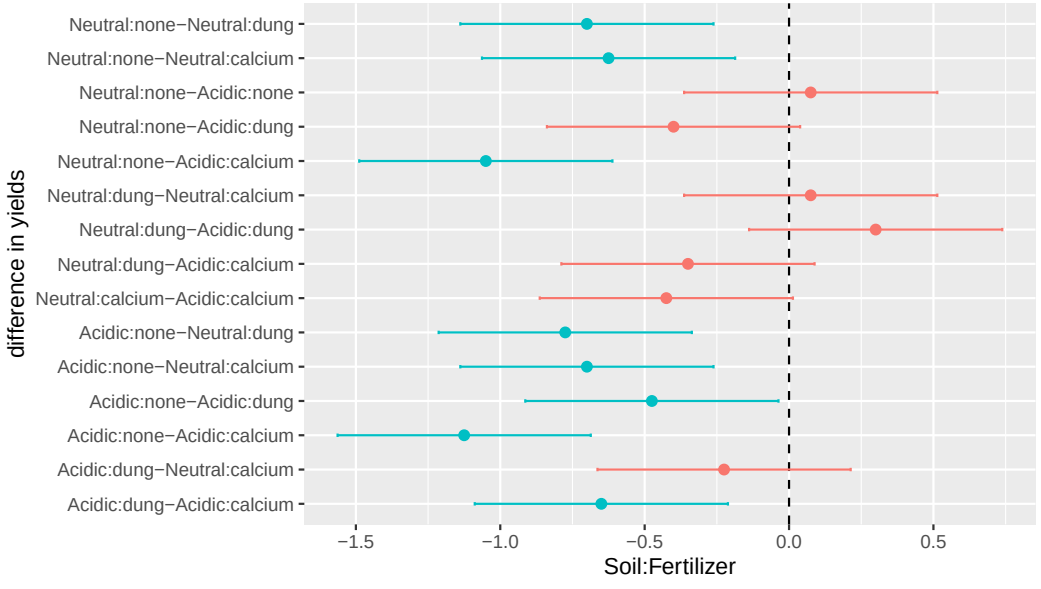
```
Tables of means
Grand mean
3.45
 Soil
Soil
 Acidic Neutral
  3.458   3.442
 Fertilizer
Fertilizer
calcium    dung    none
  3.838   3.550   2.963
 Soil:Fertilizer
         Fertilizer
Soil      calcium dung  none
  Acidic  4.050   3.400 2.925
  Neutral 3.625   3.700 3.000
```

Example: multiple comparison in two-way ANOVA with interactions

Tukey's method:







Scheffé's method:

	Yield	groups
Acidic	3.458333	a
Neutral	3.441667	a

	Yield	groups
calcium	3.8375	a
dung	3.5500	b
none	2.9625	c

	Yield	groups
Acidic:calcium	4.050	a
Neutral:dung	3.700	ab
Neutral:calcium	3.625	ab
Acidic:dung	3.400	bc
Neutral:none	3.000	c
Acidic:none	2.925	c

Assumptions of ANOVA

- Independence of particular random samples: very important assumption. Violation of the independence leads to change in the probability distribution of the test statistic and p-values.
- Normality of the data: verification by normal QQ-plot and statistical tests (Shapiro-Wilk, Lillieforse). ANOVA is not very sensitive to violation of the normality. Does not matter when when each group has at least 20 observations and the distributions are not very skewed. When strongly violated, use Kruskal-Wallis test.
- Homogeneity of the variances: verification by Levene and/or Bartlett test. Does not matter when slightly violated and all groups have similar number of observations. When strongly violated, use Kruskal-Wallis test.

Summary

- ANOVA – basic idea, typical examples
- assumptions – data types, normality and homogeneity of variances
- definition of ANOVA, model equation, parameters and their interpretation
- hypothesis and equivalent formulations
- test statistic in ANOVA, sums of squares
- ANOVA table – intepretation
- effects, group means, overall mean – calculation and interpretation
- methods of multiple comparison
- using ANOVA to compare nested linear regression models

MUNI FI

Statistics II | 3

Rank-based methods and tests

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

26 September 2022

Parametric methods

- Most of the statistical test use a parametric family, e.g., the normal distribution $\mathrm{N}(\mu, \sigma^2)$, for modeling random data.
- Based on the observed data $(X_1, \ldots, X_n)'$, we are able to calculate parameter estimates of the **parameters** (point estimates), e.g. $\widehat{\mu}, \widehat{\sigma}^2$, confidence intervals (interval estimates), and perform statistical tests on the parameters.
- All calculations are based on the assumption that the observed data comes from the specified parametric family.
- Typical assumptions of most parametric methods: interval or ratio type of data; normality of the sample; homogeneity of variances of random samples.
- A **probability model** describes ideas about the possible outcomes of a random event and the corresponding probabilities.
- **The model is essential** for determining the statistical uncertainty of an estimate (point as well as interval) or for deriving the critical region (and calculation of the p-value) of a test.

Nonparametric methods

If these assumptions are not met, we use **nonparametric** methods. Nonparametric statistics is about methods that do not make parametric assumptions about the data generating process.

- Nonparametric regression models,
- nonparametric (distribution-free) tests,
- nonparametric density estimation (e.g., kernel estimates),
- ...

Rank tests are nonparametric tests based on **ranks** of random variables in random sample.

Ordered random sample and ranks

Definition (Random sample)

Vector of random variables $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is a random sample *(náhodný výběr)* of sample size *(rozsah)* n, if the random variables are i.i.d. = independent identically distributed, i.e., have the same probability distribution and are mutually independent.

Definition (Ordered random sample)

Ordered random sample *(uspořádaný náhodný výběr)* is random vector

$$(X_{(1)}, X_{(2)}, \ldots, X_{(n)}), \qquad \text{where } X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

$X_{(i)}$ is the i-th order statistic *(pořádková statistika)*.

Definition (Rank)

Rank / rank statistic *(pořadí)* R_i of the random variable X_i is the order of X_i in the ordered random sample $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$.

If there are no ties in the sample,

$$R_i = \left|\left\{k : X_k \leq X_i\right\}\right|.$$

Otherwise, e.g., average ranks, are used,

$$R_i = \left|\left\{k : X_k < X_i\right\}\right| + 1 + \tfrac{1}{2}\left|\left\{k \neq i : X_k = X_i\right\}\right|.$$

Ordered random sample and ranks

Example

Consider random sample $\boldsymbol{X} = (2.0, 1.8, 2.1, 2.4, 1.9, 2.1, 2.0, 1.8, 2.3, 2.1)'$.

i	1	2	3	4	5	6	7	8	9	10
X_i	2.0	1.8	2.1	2.4	1.9	2.1	2.0	1.8	2.3	2.1
$X_{(i)}$	1.8	1.8	1.9	2.0	2.0	2.1	2.1	2.1	2.3	2.4
R_i	4	1	6	10	3	7	5	2	9	8
average R_i	4.5	1.5	7	10	3	7	4.5	1.5	9	7

R

- Permutation: `order(X)`
- Ordered sample: `sort(X)`, `X[order(X)]`
- Ranks: `rank(x)`, `rank(x, ties.method = "average")`

Sign test *(Znaménkový test)*

Let $(X_1, \ldots, X_n)$ be random sample from a continouos probability distribution with median $\widetilde{x}$, i.e.,

$$P(X_i < \widetilde{x}) = P(X_i > \widetilde{x}) = \tfrac{1}{2}, \qquad i = 1, \ldots, n.$$

Is the median equal to chosen number $x_0 \in \mathbb{R}$?

$$H_0 : \widetilde{x} = x_0 \qquad H_1 : \widetilde{x} \neq x_0.$$

Calculate diferences $X_i - x_0$ of the observations from the tested value, and denote T the number of positive differences, $T^+ = \big|\{i : X_i > x_0\}\big|$.

Let us define indicator random variables *(indikátorové náhodné veličiny)*,

$$Z_i = \begin{cases} 1, & X_i > x_0, \\ 0, & X_i \leq x_0. \end{cases}$$

- Verify, that $T^+ = Z_1 + \cdots + Z_n$.
- Specify the probability distribution of T^+ under H_0.
- Calculate $\mathrm{E}(Z_i)$, $\mathrm{Var}(Z_i)$, $\mathrm{E}(T^+)$ and $\mathrm{Var}(T^+)$ under H_0.

Sign test

Theorem (Sign test – for small sample size)

H_0 is rejected at the level of significance $\leq \alpha$, if

$$T^+ \leq k_\alpha \qquad \text{or} \qquad T^+ \geq n - k_\alpha.$$

Number k_α is the largest number from $\{0, \ldots, n\}$, for which

$$P(T^+ \leq k_\alpha) = \frac{1}{2^n} \sum_{i=0}^{k_\alpha} \binom{n}{i} \leq \frac{\alpha}{2} \quad \text{and} \quad P(T^+ \geq n - k_\alpha) = \frac{1}{2^n} \sum_{i=n-k_\alpha}^{n} \binom{n}{i} \leq \frac{\alpha}{2}.$$

Moivre-Laplace theorem implies that $U = \dfrac{T^+ - \mathsf{E}(T^+)}{\sqrt{\mathsf{Var}(T^+)}} \overset{as.}{\sim} \mathsf{N}(0;1)$ as $n \to \infty$.

Theorem (Sign test – asymptotic variant)

Under H_0, test statistic $U = \dfrac{2T^+ - n}{\sqrt{n}}$ has asymptotically standard normal distribution $\mathsf{N}(0;1)$ as $n \to \infty$.

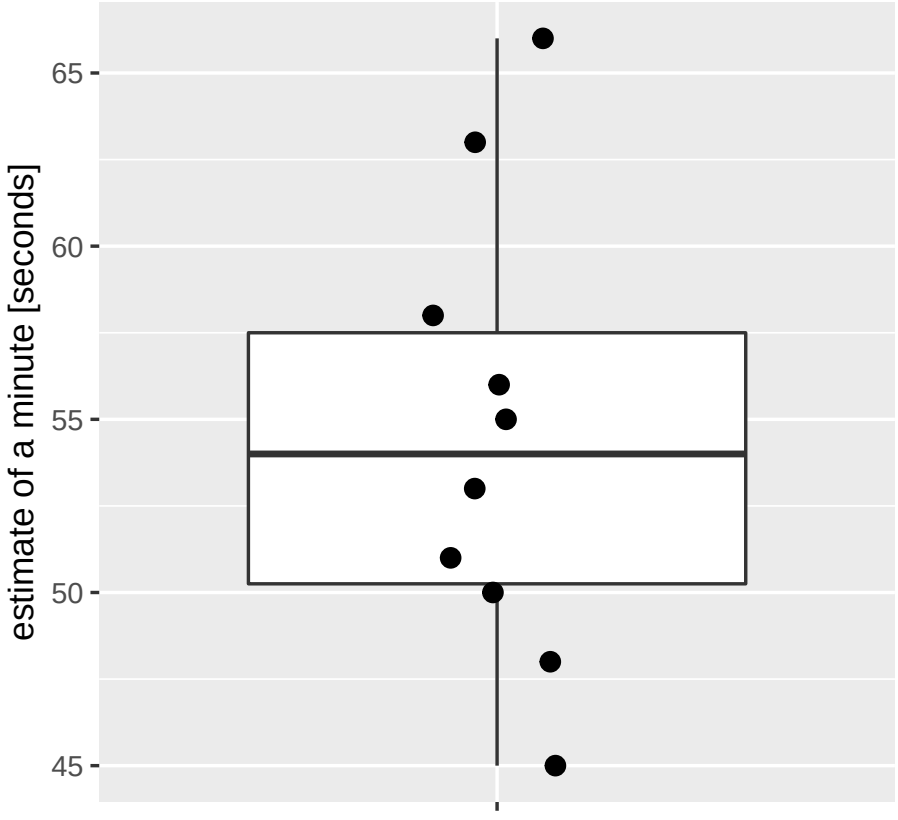
H_0 is rejected at the asymptoticlevel of significance α, if $|U| \geq u_{1-\alpha/2}$.

Example

Ten research participants had to guess independently of each other and without prior training when a minute has passed after the sound signal.

Observations in seconds: $X = (53, 48, 45, 55, 63, 51, 66, 56, 50, 58)'$.

Test hypothesis that half of the participants had a period of one minute underestimated and the second half had it overestimated.

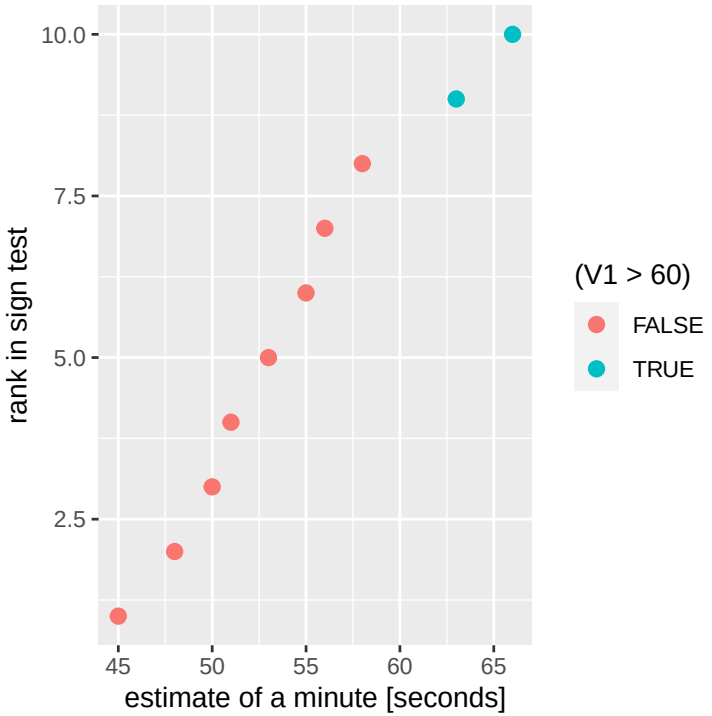


Example

$$H_0 : \tilde{x} = 60,\ H_1 : \tilde{x} \neq 60$$

i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58
$(X_i - 60)$	-7	-12	-15	-5	3	-9	6	-4	-10	-2

$n = 10, T^+ = 2, U = \frac{4-10}{\sqrt{10}} = -1.897, k_{0.05} = 1, u_{0.975} = 1.96, H_0$ is not rejected




```
SIGN.test(dt$V1, md = 60)
  One-sample Sign-Test
data:  X
s = 2, p-value = 0.1094
alternative hypothesis: true median is not equal t
95 percent confidence interval:
 48.64889 61.37778
sample estimates:
median of x
        54
Achieved and Interpolated Confidence Intervals:
                  Conf.Level  L.E.pt  U.E.pt
Lower Achieved CI     0.8906 50.0000 58.0000
Interpolated CI       0.9500 48.6489 61.3778
Upper Achieved CI     0.9785 48.0000 63.0000
```

Sign test

- Used especially in the case when the probability distribution of the observations X_i is significantly skewed, hence surely not gaussian. The t-test would be biased (incorrect p-values) in such a case.
- The test has low power. It is desirable to have a large sample.
- The asymptotical variant is sufficiently accurate when $n \geq 20$.
- Differences $X_i - x_0$ which are equal to zero are omitted, and the test is performed only for the remaining differences with reduced n.

Paired sign test *(párový znaménkový test)*

Let us assume **i.i.d.** pairs of (possibly dependent) observations $((Y_1, Z_1), \ldots, (Y_n, Z_n))$ from a bivariate continuous probability distribution.

$$H_0 : \tilde{z} - \tilde{y} = x_0 \qquad H_1 : \tilde{z} - \tilde{y} \neq x_0.$$

Calculate differences $X_i = Z_i - Y_i$ and perform the sign test on the new sample $(X_1, \ldots, X_n)'$.

Signed-rank Wilcoxon test *(Jednovýběrový Wilcoxonův test)*

Let $(X_1, \ldots, X_n)$ be random sample from a continouos probability distribution with symmetrical probability density function $f(x)$,

$$P(X_i < \tilde{x}) = \int_{-\infty}^{\tilde{x}} f(x)\,\mathrm{d}x = \int_{\tilde{x}}^{\infty} f(x)\,\mathrm{d}x = P(X_i > \tilde{x}) = \tfrac{1}{2}, \qquad i = 1, \ldots, n.$$

Is the median equal to chosen number $x_0 \in \mathbb{R}$?

$$H_0 : \tilde{x} = x_0, \qquad H_1 : \tilde{x} \neq x_0.$$

1. Calculate diferences $Y_i = X_i - x_0$
2. and sort them in nondecreasing order according to their absolute value,
$$|Y|_{(1)} \leq |Y|_{(2)} \leq \cdots \leq |Y|_{(n)}.$$
3. Denote R_i^+ the rank of $|Y_i|$ in this nondecreasing sequence.
4. Calculate the sum of R_i^+ ranks separately for positive and negative Y_i,
$$T^+ = \sum_{Y_i > 0} R_i^+, \qquad T^- = \sum_{Y_i < 0} R_i^+.$$

What is the sum $(T^+ + T^-)$ equal to?

Signed-rank Wilcoxon test

Alternative calculation, signed ranks

$$T^+ = \frac{n(n+1)}{4} + \frac{T}{2}, \quad T^- = T^+ - T, \quad \text{where} \quad T = \sum_{i=1}^{n} R_i^+ \operatorname{sgn}(Y_i).$$

Theorem (Signed-rank Wilcoxon test)

H_0 is rejected at the level of significance α, if

$$\min\{T^+, T^-\} \leq w_\alpha(n),$$

where $w_\alpha(n)$ is *critical calue* of the Wilcoxon test.

Theorem (Signed-rank Wilcoxon test – asymptotic variant)

Under H_0, test statistic $U = \dfrac{T^+ - \mathsf{E}(T^+)}{\sqrt{\mathsf{Var}(T^+}}$,

where $\mathsf{E}(T^+) = \frac{1}{4}n(n+1)$ and $\mathsf{Var}(T^+) = \frac{1}{24}n(n+1)(2n+1)$,

has asymptotically standard normal distribution $\mathsf{N}(0;1)$.

H_0 is rejected at the asymptotic level of significance α, if $|U| \geq u_{1-\alpha/2}$.

Example

$$H_0 : \widetilde{x} = 60,\ H_1 : \widetilde{x} \neq 60$$

i	1	2	3	4	5	6	7	8	9	10
X_i	53	48	45	55	63	51	66	56	50	58
$Y_i = (X_i - 60)$	−7	−12	−15	−5	3	−9	6	−4	−10	−2
R_i^+	6	9	10	4	2	7	5	3	8	1
$\operatorname{sgn} Y_i$	−1	−1	−1	−1	1	−1	1	−1	−1	−1

$$T = -41,\ T^+ = 7,\ T^- = 48,\ n = 10,\ w_{0.05}(10) = 8,$$
$$\mathrm{E}(T^+) = 27.5,\ \mathrm{Var}(T^+) = 96.25,\ U = -2.09,\ u_{0.975} = 1.96,\ H_0 \text{ is rejected}$$


```
wilcox.test(dt$V1, mu = 60)
  Wilcoxon signed rank exact test
data:  X
V = 7, p-value = 0.03711
alternative hypothesis: true location is not equal
```

Signed-rank Wilcoxon test

- Used especially to test whether data comes from a symmetric population with a specified median.
- T-test is parametric analogy of Wilcoxon signed-rank test for case of testing the mean of a sample from gaussian probability distribution.
- The Wilcoxon test assumes symmetry of the probability density of the observed variable around the median. In case of asymmetry of the probability density of the data, H_0 can be rejected even if $\tilde{x} = x_0$ holds. In the case of asymmetry of the probability density, e.g., the sign test is used.
- Differences $X_i - x_0$ which are equal to zero are omitted, and the test is performed only for the remaining differences with reduced n.
- The asymptotical variant is sufficiently accurate when $n \geq 30$.

Paired Wilcoxon test *(párový Wilcoxonův test)*

Let us assume **i.i.d.** pairs of (possibly dependent) observations $((Y_1, Z_1), \ldots, (Y_n, Z_n))$ from a bivariate continuous probability distribution.

$$H_0 : \tilde{z} - \tilde{y} = x_0 \qquad H_1 : \tilde{z} - \tilde{y} \neq x_0.$$

Calculate differences $X_i = Z_i - Y_i$ and perform the Wilcoxon signed-rank test on the new sample $(X_1, \ldots, X_n)'$.

Wilcoxon rank-sum test *(Dvouvýběrový Wilcoxonův test)*

Example

Two methods of fertilization were tested on a total of 13 experimental fields of the same soil quality: method A on 8 fields, method B on 5 fields.

fertilization	wheat yields in tons per hectare
A	5.7, 5.5, 4.3, 5.9, 5.2, 5.6, 5.8, 5.1
B	5.0, 4.5, 4.2, 5.4, 4.4

Does the fertilization method have an effect on wheat yields?

Comparison of two independent random samples:

- $(X_1, \ldots, X_m)$ coming from cumulative distribution function (c.d.f.) $F_X(x)$,
- $(Y_1, \ldots, Y_n)$ coming from c.d.f. $F_Y(y)$.

Test of the equality of c.d.f.s against an alternative of a **location shift**,

$$H_0 : F_X(x) = F_Y(x), \qquad H_1 : F_X(x) = F_Y(x - \Delta) \text{ for } \Delta > 0$$

Wilcoxon rank-sum test

1. Join both samples,
$$(Z_1, \ldots, Z_{m+n}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n),$$
2. and sort the combined sample in nondecreasing order,
$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(m+n)}.$$
3. Denote $(R_1, \ldots, R_m)$ the ranks of $(X_1, \ldots, X_m)$ and $(S_1, \ldots, S_n)$ the ranks of $(Y_1, \ldots, Y_n)$ in the combined ordered sample.
4. Calculate the sums of ranks of X- and Y-sample,
$$T_1 = \sum_{i=1}^{m} R_i, \qquad T_2 = \sum_{j=1}^{n} S_j,$$
5. and corresponding Mann-Whitney's statistics
$$U_1 = T_1 - \frac{m(m+1)}{2}, \qquad U_2 = T_2 - \frac{n(n+1)}{2}.$$
6. The Mann-Whitney's test statistic is $U_{\mathrm{MW}} = \min\{U_1, U_2\}$.

- U_1 = number of cases $X_i > Y_j$ out of all pairwise comparisons.
- $U_1 + U_2 = m\,n$.

Wilcoxon rank-sum test

Theorem (Mann-Whitney-Wilcoxon test)

H_0 is rejected at the level of significance α, if

$$U_{\mathrm{MW}} \leq w_\alpha(m, n),$$

where $w_\alpha(m, n)$ is *critical calue* of the Mann-Whitney-Wilcoxon test.

Theorem (Mann-Whitney-Wilcoxon test – asymptotic variant)

Under H_0, test statistic $U = \dfrac{U_{\mathrm{MW}} - \mathrm{E}(U_{\mathrm{MW}})}{\sqrt{\mathrm{Var}(U_{\mathrm{MW}})}}$,

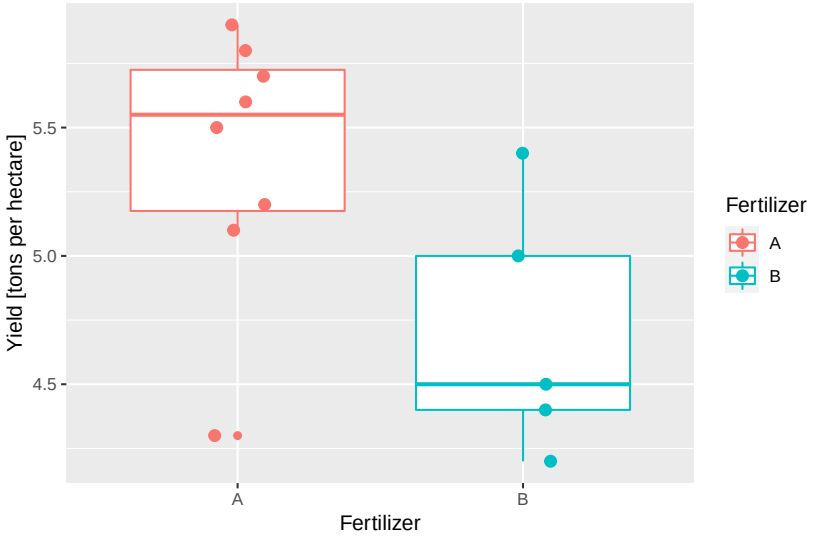
where $\mathrm{E}(U_{\mathrm{MW}}) = \frac{1}{2}\, m\, n$ and $\mathrm{Var}(U_{\mathrm{MW}}) = \frac{1}{12}\, m\, n(m + n + 1)$,

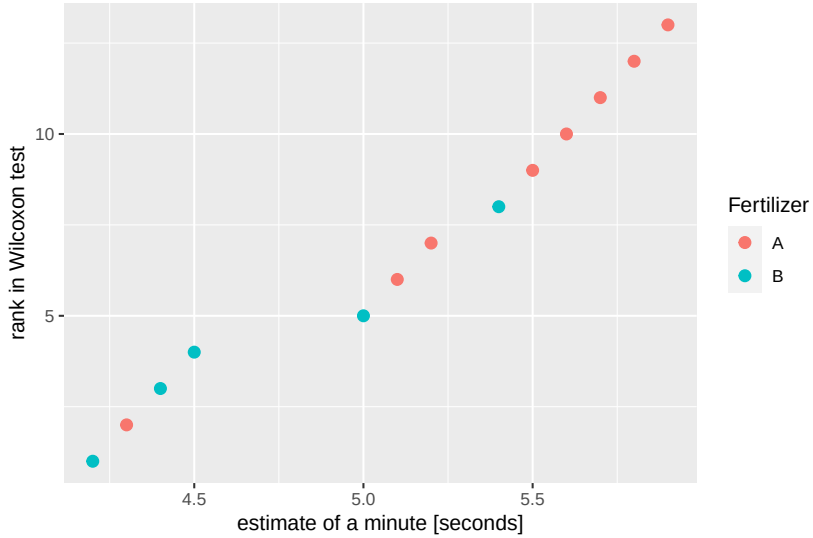
has asymptotically standard normal distribution $\mathrm{N}(0; 1)$.

H_0 is rejected at the asymptotic level of significance α, if $|U| \geq u_{1-\alpha/2}$.

Effect size $= \dfrac{U_1}{m\, n}$.

Example





$$H_0 : F_A(x) = F_B(x), \qquad H_1 : F_A(x) = F_B(x - \Delta)$$

Z_k	4.2	4.3	4.4	4.5	5.0	5.1	5.2	5.4	5.5	5.6	5.7	5.8	5.9	rank sum
A: R_i		2				6	7		9	10	11	12	13	$T_1 = 70$
B: S_j	1		3	4	5			8						$T_2 = 21$

$$m = 8, n = 5, T_1 = 70, U_1 = 34, T_2 = 21, U_2 = 6, U_{\text{MW}} = 6,$$

$$w_{0.05}(8, 5) = 6, \; U = \frac{6 - 20}{\sqrt{140/3}} = -2.049, \; u_{0.975} = 1.96, \; H_0 \text{ is rejected}$$

```
wilcox.test (dt|>filter(Fertilizer=="A")$Yield, dt|>filter(Fertilizer=="B")$Yield)
  Wilcoxon rank sum exact test
data:  pull(filter(dt, Fertilizer == "A"), Yield) and pull(filter(dt, Fertilizer ==
W = 34, p-value = 0.04507
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon rank-sum test

- Mann-Whitney U test = Mann–Whitney–Wilcoxon test = Wilcoxon rank-sum test = two-sample Wilcoxon test.
- The test assumes that the random samples are stochastically independent, and the data have of at least ordinal type.
- More general alternative is $H_1 : F_X(x) \neq F_Y(x)$.
- The more strict alternative $H_1 : F_X(x) = F_Y(x - \Delta)$ requires that the data comes from continuous probability distributions and is restricted to a shift in location. Rejection of H_0 then leads to a difference in medians.
- Mann–Whitney U test is preferable to the t-test when the data are ordinal but not of interval type.
- Mann–Whitney U test is more robust with respect to the presence of outliers.
- Mann-Whitney U test may have worse Type I error control when data are both heteroscedastic and non gaussian.
- The asymptotical variant is sufficiently accurate when $m, n > 10$.
- See also two-sample Kolmogorov-Smirnov test.

Median test for two independent samples

The test statistic

$$T = \frac{m}{2} + \frac{1}{2}\sum_{i=1}^{m} \operatorname{sgn}\left(R_i - \frac{m+n+1}{2}\right)$$

is equal to the number of X_i observations that are greater the the median of the combined sample. If the total sample size $(m+1)$ is odd, $\frac{1}{2}$ is added.

Theorem (Median test)

Under H_0, test statistic $U = \dfrac{T - \mathrm{E}(T)}{\sqrt{\mathrm{Var}(T}}$,

where $\mathrm{E}(T) = \dfrac{m}{2}$ and $\mathrm{Var}(T) = \begin{cases} \frac{m\,n}{4(m+n-1)}, & \text{for } m+n \text{ odd}, \\ \frac{m\,n}{4(m+n)}, & \text{for } m+n \text{ even}, \end{cases}$ has

asymptotically standard normal distribution $\mathrm{N}(0;1)$.

H_0 is rejected at the asymptotic level of significance α, if $|U| \geq u_{1-\alpha/2}$.

The test is particularly suitable in the case of so-called censored observations, when for some extreme values we only know that that they are smaller or larger than some limit, but we do not know their exact values.

Nonparametric one-way ANOVA

Kruskal-Wallis test = nonparametric analogy of one-way ANOVA and generalization of Mann-Whitney-Wilcoxon test.

Assumptions

- The factor A has $a \geq 3$ levels.
- The ith level has n_i observations $(Y_{i1}, \ldots, Y_{in_i})$, which form a random sample of at least ordinal type, coming from cumulative distribution function $F_i(x)$.
- Y_{ij}: first index – group by the level of the factor, second index – order in the group.
- The particular random samples are stochastically independent.

Hypothesis

A hypothesis that the factor A has no influence on the probability distribution of the observed variable Y, i.e., so-called *non-dominance* of cumulative distribution functions:

$H_0 : F_1(x) = F_2(x) = \cdots = F_a(x),$

$H_1 : \exists\, i \neq j : F_i(x) > F_j(x), \quad \text{or} \quad F_i(x) < F_j(x).$

Kruskal-Wallis test

1. Join all observations, $(Y_{1\,1},\ \ldots,\ Y_{a\,n_a})$,
2. and sort the combined sample in nondecreasing order,

$$Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}, \qquad n = \sum_{i=1}^{a} n_i.$$

3. Denote R_{ij} the (average) rank of Y_{ij} in the combined sample.
4. Calculate the sums of ranks in each category,

$$T_i = R_{i\cdot} = \sum_{j=1}^{n_i} R_{ij}, \qquad i = 1,\ \ldots,\ a.$$

A	observations	ranks	size	sum of ranks
1	$(Y_{1\,1},\ \ldots,\ Y_{1\,n_1})$	$(R_{1\,1},\ \ldots,\ R_{1\,n_1})$	n_1	T_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$(Y_{i\,1},\ \ldots,\ Y_{i\,n_i})$	$(R_{i\,1},\ \ldots,\ R_{i\,n_i})$	n_i	T_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	$(Y_{a\,1},\ \ldots,\ Y_{a\,n_a})$	$(R_{a\,1},\ \ldots,\ R_{a\,n_a})$	n_a	T_a
total			n	$\frac{n(n+1)}{2}$

Theorem (Kruskal-Wallis test)

Denote $$Q = \frac{12}{n(n+1)} \sum_{i=1}^{a} \frac{T_i^2}{n_i} - 3(n+1).$$

H_0 is rejected at the level of significance α, if $Q \geq h_\alpha(a-1)$, where $h_\alpha(a-1)$ is *critical value* of the test.

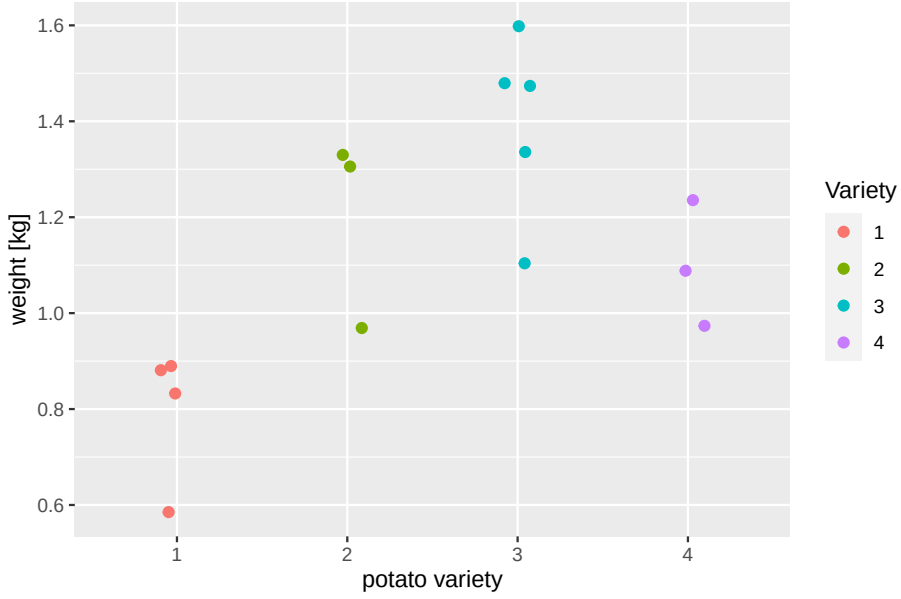
Under H_0, test statistic Q has asymptotically chi-squared probability distribution $\chi^2(a-1)$ with $(a-1)$ degrees of freedom, $\mathrm{E}(Q) = a-1$, and $h_\alpha(a-1) \approx \chi^2_{1-\alpha}(a-1)$.

H_0 is rejected at the asymptotic level of significance α, if $Q \geq \chi^2_{1-\alpha}(a-1)$.

For more than ca. 25 % of ties in data, following correction is used, $Q_K = \frac{Q}{K}$, where $K = 1 - \frac{\sum_k m_k(m_k^2-1)}{n(n^2-1)}$, and m_k denotes the number of ties.

Example

Analysis of 4 varieties of potatoes based on the weights of the clusters of potato tubers.



A	weight Y_{ij}	rank R_{ij}	n_i	T_i
1	0.9, 0.8, 0.6, 0.9	3.5, 2.0, 1.0, 3.5	4	10
2	1.3, 1.0, 1.3	11, 5.5, 11	3	27,5
3	1.3, 1.5, 1.6, 1.1, 1.5	11, 13.5, 15, 7.5, 13.5	5	60,5
4	1.1, 1.2, 1.0	7.5, 9.0, 5.5	3	22
total			15	120

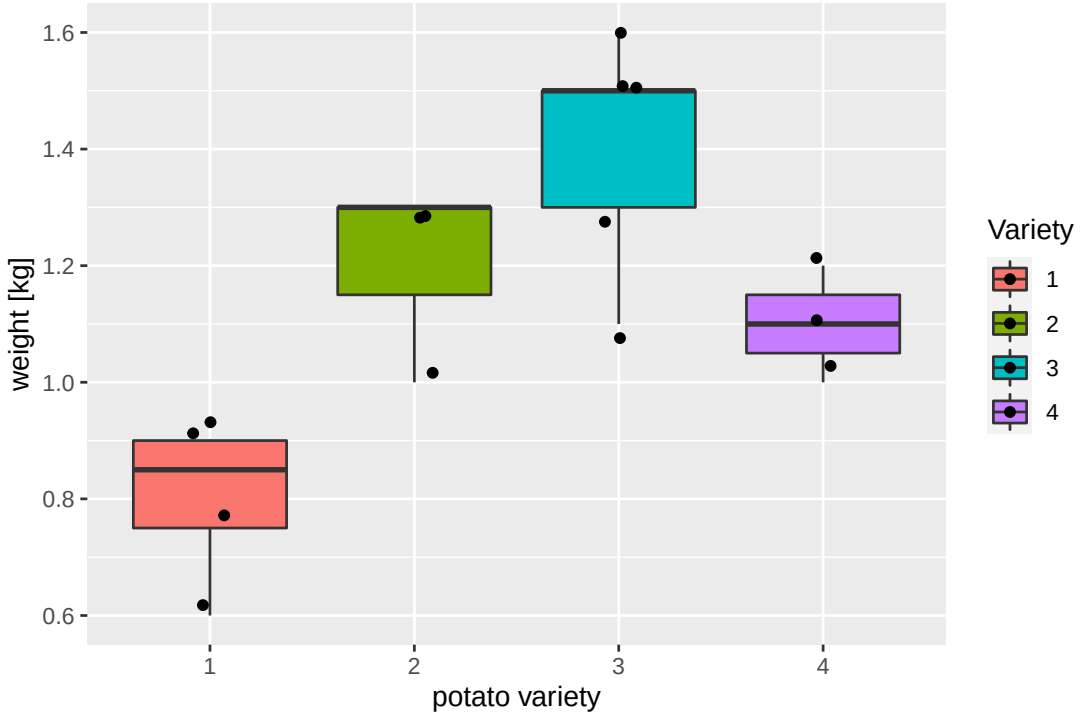
$Q = 10.523,\ K = 1 - \frac{48}{3360},\ Q_K = 10.676 > \chi^2_{0,95}(3) = 7.815,\ H_0$ is rejected

Example

Kruskal-Wallis test

```
KWtest <- with (dat, kruskal (Weight, Variety))
KWtest
```

```
$statistics
     Chisq    p.chisq
  10.67585 0.01361427
$parameters
  Df ntr   t.value
   3   4 2.200985
$rankMeans
 Variety    Weight r
1       1  2.500000 4
2       2  9.166667 3
3       3 12.100000 5
4       4  7.333333 3
$groups
  trt      means  M
1   3 12.100000  a
2   2  9.166667 ab
3   4  7.333333  b
4   1  2.500000  c
```



Median test

Median test uses the test statistic Denote A_i denotes the number of observations $(Y_{i1}, \ldots, Y_{in_i})$ from the i-th category that are greater than the median $\widetilde{Y}$ of the joined sample,

$$A_i = \left|\left\{j : Y_{ij} > \widetilde{Y}\right\}\right|, \qquad i = 1, \ldots, a.$$

If the total sample size n is odd, the A_i for which the median $\widetilde{Y}$ of the joined sample belongs to the corresponding category i, is increased by $\frac{1}{2}$.

Theorem (Median test)

Under H_0, test statistic

$$Q_{\mathrm{M}} = 4 \sum_{i=1}^{a} \frac{A_i^2}{n_i} - n$$

has asymptotically ($\min\{n_1, \ldots, n_a\} \to \infty$) chi-squared probability distribution $\chi^2(a-1)$ with $(a-1)$ degrees of freedom.

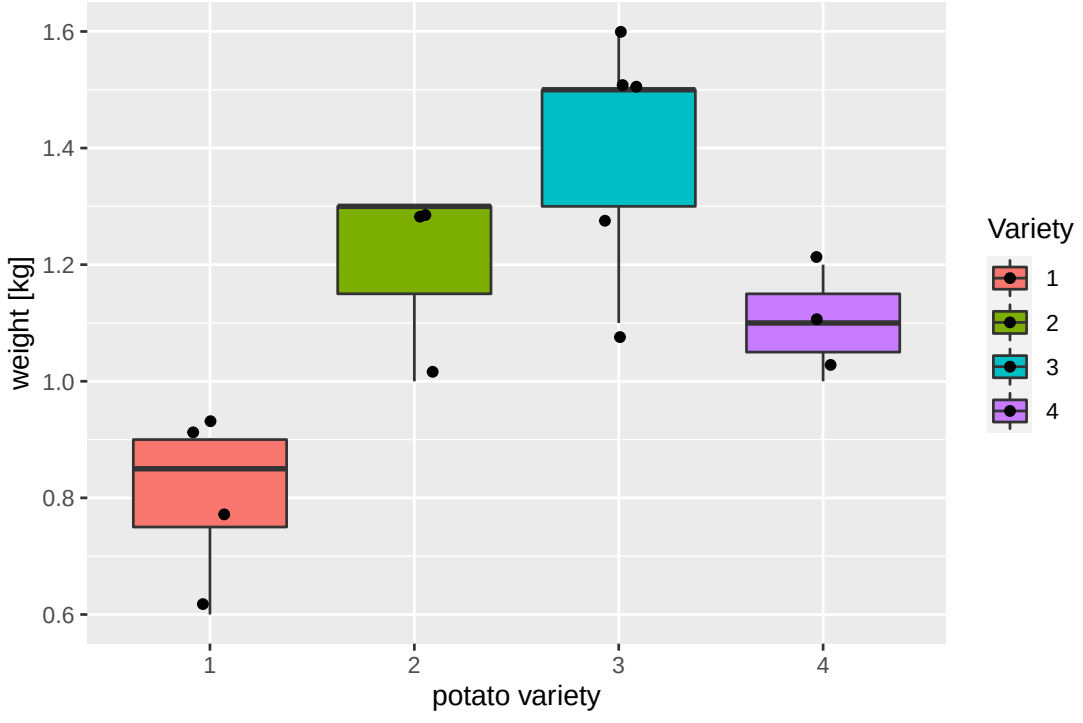
H_0 is rejected at the asymptotic level of significance α, if $Q_{\mathrm{M}} \geq \chi^2_{1-\alpha}(a-1)$.

Example

Median test

```
Mtest <- with (dat, Median.test (Weight, Variety))
Mtest
```

```
$statistics
     Chisq    p.chisq Median
  6.428571 0.09252244    1.1
$parameters
  Df ntr
   3   4
$Medians
  trt Median grather lessEqual
1   1   0.85       0         4
2   2   1.30       2         1
3   3   1.50       4         1
4   4   1.10       1         2
$comparison
        Median     Chisq      pvalu
1 and 2   0.90 7.0000000 0.0081509
1 and 3   1.10 5.7600000 0.0163950
1 and 4   0.90 7.0000000 0.0081509
2 and 3   1.30 2.8800000 0.0896860
2 and 4   1.15 0.6666667 0.414216178
3 and 4   1.25 4.8000000 0.028459737    *
```



Friedman test

- Block design of the data,
- a levels of factor A,
- b blocks for each $i = 1, \ldots, a$,
- Y_{ij} stands for observation of j-th block at the i-th level of factor A,
- Y_{ij} comes from a distribution with cumulative distribution function $F_{ij}(x)$.
- Friedman test is nonparametric analogy of block-design two-way ANOVA with one observation Y_{ij} in each group.

Hypothesis

$H_0 : F_{1j}(x) = \cdots = F_{aj}(x)$, i.e., the c.d.f. is identical in each block, but does not need to be identical across the factor levels,

H_1 : c.d.f.s differ also across factor levels.

Friedman test

1. Calculate ranks R_{ij} of Y_{ij} **separately in each block**,
2. sum the ranks for each level of factor A, $R_{i\cdot} = \sum_{j=1}^{b} R_{ij}$, $i = 1, \ldots, a$.
3. Test statistic is $Q_{\mathrm{F}} = \dfrac{12\,b}{a(a+1)} \sum_{i=1}^{a} \left(\dfrac{R_{i\cdot}}{b} - \dfrac{a+1}{2} \right)^2$.

Theorem (Friedman test)

Under H_0, test statistic Q_{F} has asymptotically $\chi^2(a-1)$ probability distribution .

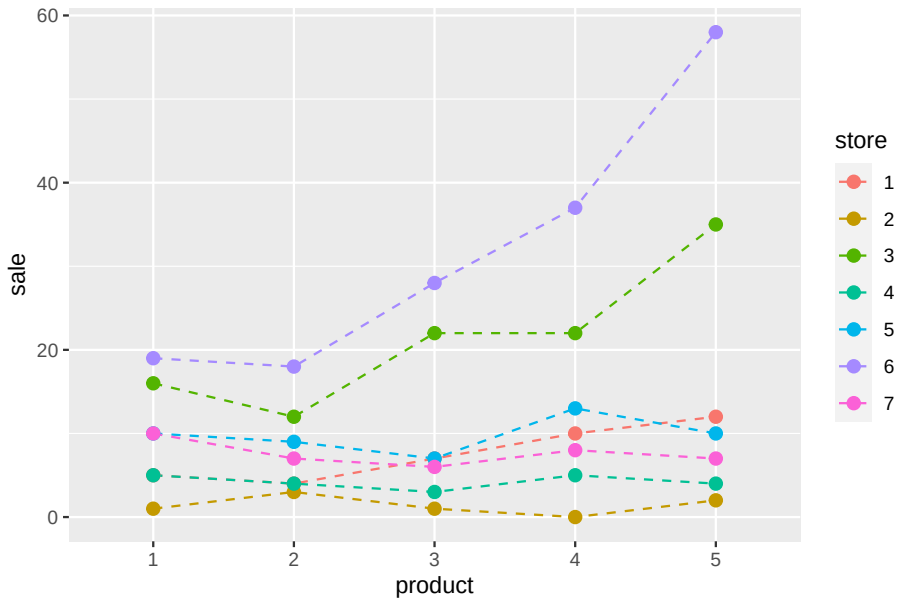
H_0 is rejected at the asymptotic level of significance α, if $Q_{\mathrm{F}} \geq \chi^2_{1-\alpha}(a-1)$.

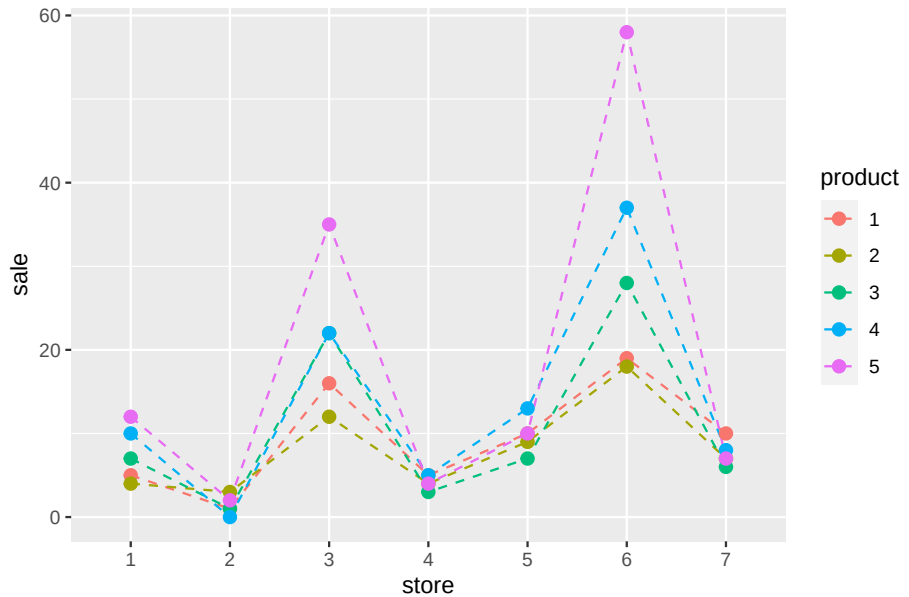
The $\chi^2(a-1)$ approximation is accurate enough when $n > 15$ or $k > 4$.

Example

Numbers of sold pieces of products: $a = 5$ products, $b = 7$ blocks.

B	1	2	3	4	5	6	7
$A = 1$	5	1	16	5	10	19	10
$A = 2$	4	3	12	14	9	18	7
$A = 3$	7	1	22	3	7	28	6
$A = 4$	10	6	22	5	13	37	8
$A = 5$	12	2	35	4	10	58	7






```
friedman.test(X$sale, X$product, X$store)
  Friedman rank sum test
data:  X$Y, X$A and X$B
Friedman chi-squared = 8.3284, df = 4, p-value = 0.08026
```

Van der Waerden test

Van der Waerden test test the same hypothesis as in the Kruskal-Wallis test but converts the ranks R_{ij} to quantiles of the standard normal distribution $\mathrm{N}(0;1)$.

1. $A_{ij} = \Phi^{-1}\left(\frac{R_{ij}}{n+1}\right)$,
2. $A_{i\cdot} = \sum_{j=1}^{n_i} A_{ij}, \qquad i = 1, \ldots, a,$
3. Test statistic is $$Q_{\mathrm{W}} = \frac{\sum_{i=1}^{a} \frac{A_{i\cdot}^2}{n_i}}{\frac{1}{n-1}\sum_{i=1}^{a}\sum_{j=1}^{n_i} A_{ij}^2}.$$

Theorem (Van der Waerden test)

Under H_0, test statistic Q_{W} has asymptotically $\chi^2(a-1)$ probability distribution.

H_0 is rejected at the asymptotic level of significance α, if $Q_{\mathrm{W}} \geq \chi^2_{1-\alpha}(a-1)$.

The test is robust as Kruskal-Wallis test, but efficient also for normal data.

Multiple comparison

When the null hypothesis H_0 is rejected, multiple comparison usually follows. Factor levels $A = k$ and $A = l$ are significantly different in their probability distributions (particularly in the location shift), if

$$|T_k - T_l| > \sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) h_\alpha (a-1)}\,.$$

In the case of balanced design, i.e. when $n_i = b$ for all $i = 1, \ldots, a$, so called Nemény test is preferred. It is based on the Tukeys's idea from ANOVA. Factor levels $A = k$ and $A = l$ are significantly different in their probability distributions (particularly in the location shift), if

$$\sqrt{2\,b}\;\left|\overline{Z}_{k\cdot} - \overline{Z}_{l\cdot}\right| > q_\alpha,$$

where $$Z_{ij} = \begin{cases} 1, & Y_{ij} > \tilde{Y}, \\ 0, & Y_{ij} \leq \tilde{Y}, \end{cases}$$ with group means $\overline{Z}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} Z_{ij}$.

Rank tests in *R*

- Sign test

```
SIGN.test(X, md = x0)                                    library("BSDA")
```

- Wilcoxon signed-rank test

```
wilcox.test(X, mu = x0)
```

- Wilcoxon rank-sum test

```
wilcox.test(X, Y)
```

- Paired Wilcoxon test

```
wilcox.test(X, Y, paired = TRUE)
```

- Kruskal-Wallis test

```
kruskal(Y, group)                                  library("agricolae")
```

- Median test

```
Median.test(Y, group)                              library("agricolae")
```

- Van der Waerden test

```
waerden.test(Y, group)                             library("agricolae")
```

- Friedman test

```
friedman.test(Y, group, block)
```

Summary

- parametric and nonparamwtric methods, assumptions, comparison
- ordered random sample, order statistic, rank
- rank tests (sign test, Wilcoxon's tests): hypotheses, principles of tests, calculation of ranks and test statistics
- Kruskall-Wallis test (median test, Van der Waerden test, Friedman test): model, hypothesis, test statistic, multiple comparison, comprarison with standard ANOVA

MUNI FI

Statistics II | 4

Goodness-of-fit test, testing probability distribution

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

3 October 2022

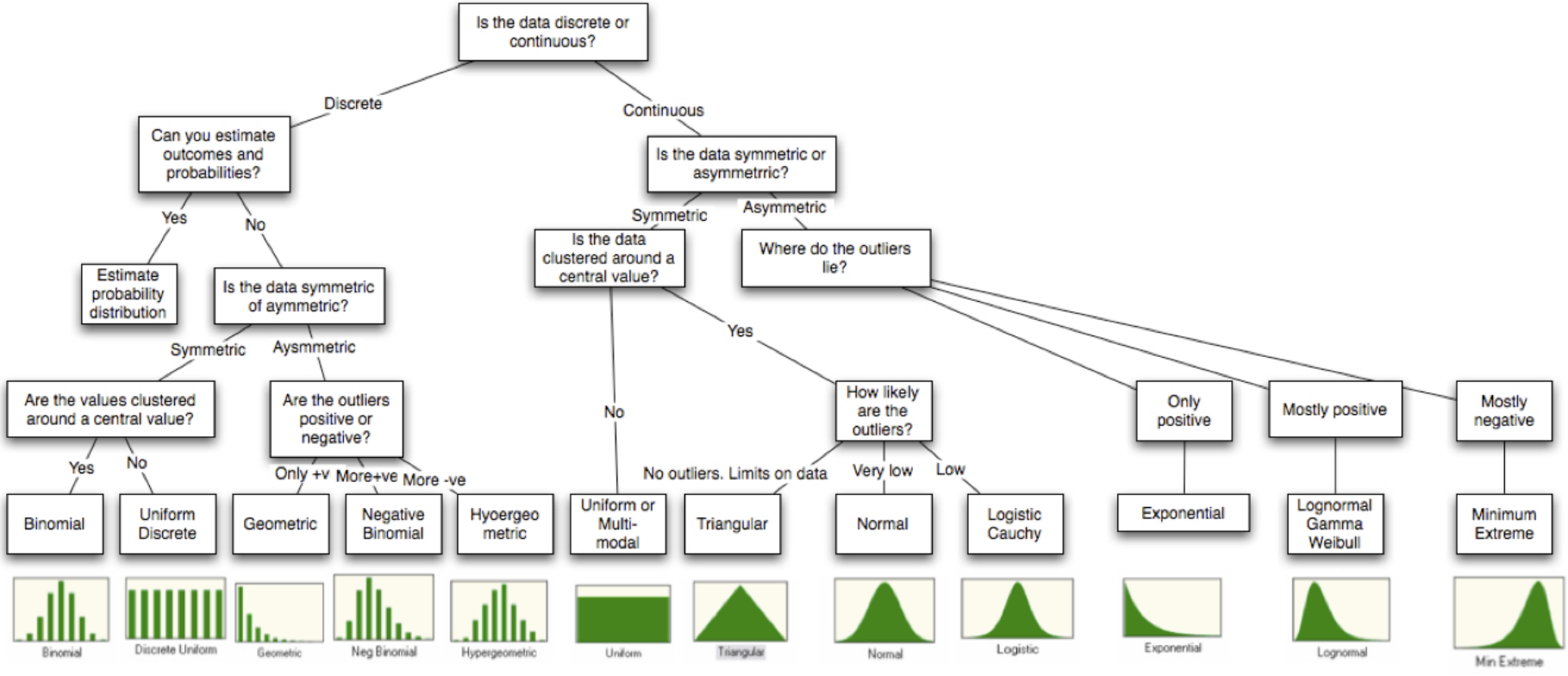
Motivation

Often, we need to test whether the random sample $(X_1, \ldots, X_n)$ comes from a specific probability distribution of a family of probability distributions:

- with given parameters, e.g., $\mathrm{N}(10, 4)$, $\mathrm{Ex}(3.5)$, $\mathrm{Po}(2)$, $\mathrm{Bi}(10, 0.6)$;
- with unkown parameters, e.g., gaussian, exponential, Poisson, binomial;
- with given probability mass function $p(x)$ or probability density function $f(x)$.

Some examples:

- Parametric tests require a specific probability distribution of the random sample, e.g., t-test requires normality of the data.
- ANOVA requires normality of the data.
- Quality control and compliance with the prescribed probability distribution.
- Random number generators.



(Aswath Damodaran: *Probabilistic approaches to risk*)

- frequency barplot compared with the probability mass function
- histogram compared with the probability density function
- quantile-quantile (QQ) plot for given probability distribution
- empirical cumulative distribution function compared with the theoretical cumulative distribution function

Some well-known probability distributions

Discrete:

- Bernoulli (alternative)
- binomial
- hypergeometric
- Poisson
- geometric
- negative binomial
- uniform discrete

Aboslutely continuous:

- gaussian (normal)
- chi-squared
- Student
- Fisher-Snedecor
- lognormal
- exponential
- gamma
- Weibull
- logistic
- beta
- uniform continuous

Examples

Example (1)

84 families was chosen randomly from a set of families of 5 children and the number of boys was detected for each family.

number of boys	0	1	2	3	4	5
number of familes	3	10	22	31	14	4

At a significance level of 0.05, test hypothesis that number of boys in families of 5 children has binomial distribution Bi $(5, 0.5)$.

Example (2)

Waiting time (in minutes) was observed for 70 clients of a certain company that they spent waiting for service (from the moment of taking their ticket).

waiting time	$(0, 3]$	$(3, 6]$	$(6, 9]$	$(9, 12]$	$(12, 15]$	$(15, 18]$	$(18, 21]$	$(21, 24]$
# of clients	14	16	10	9	8	5	3	5

At a significance level of 0.05, test hypothesis that waiting time has exponential distribution.

Pearson’s chi-squared test

Empirical and theoretical frequencies

Assume that n objects $(X_1, \ldots, X_n)$ are distributed into k disjoint categories, $A_1, \ldots, A_k$, while each subject corresponds to exactly one category.

Theoretical probability distribution P assigns the probability p_j that randomly chosen object X is a member of the category A_j, $p_j = P(X \in A_j)$.

Definition

- Empirical / observed frequencies *(empirické četnosti)* are the numbers $N_1, \ldots, N_k$ of objects $(X_1, \ldots, X_n)$ in individual categories.
- Theoretical / expected frequencies *(teoretické četnosti)* are the expected numbers $n_1, \ldots, n_k$ of objects in individual categories,

$$n_j = n\, p_j.$$

category	A_1	A_2	$\cdots$	A_k	sum
empirical freqs.	N_1	N_2	$\cdots$	N_k	n
probabilities	p_1	p_2	$\cdots$	p_k	1
theoretical freqs.	n_1	n_2	$\cdots$	n_k	n

Multinomial probability distribution

Definition

The joint probability distribution of empirical frequencies $(N_1, \ldots, N_k)$ is

$$P(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

for $n_j = 0, 1, \ldots, n$ and $n_1 + \cdots + n_k = n$.

This probability distribution is called *k*-variate multinomial *(multinomické)*,

$$(N_1, \ldots, N_k) \sim \mathsf{M}_k(n; p_1, \ldots, p_k).$$

Theorem

For $(N_1, \ldots, N_k) \sim \mathsf{M}_k(n; p_1, \ldots, p_k)$, it holds:

$$N_j \sim \mathsf{Bi}(n, p_j), \quad \mathsf{E}(X_j) = n p_j, \quad \mathsf{Var}(X_j) = n p_j (1 - p_j), \qquad j = 1, \ldots, k.$$

Moivre-Laplace theorem: For large n, large k and none *large* category,

$$\frac{N_j - n p_j}{\sqrt{n p_j (1 - p_j)}} \approx \frac{N_j - n p_j}{\sqrt{n p_j}} = \frac{N_j - n_j}{\sqrt{n_j}} \overset{as.}{\sim} \mathsf{N}(0, 1)$$

H_0: empirical distribution = theoretical distribution, i.e., all $N_j = n_j$,
H_1: empirické and theoretical distribution differ

Idea behind the test statistic K:

1. $N_1 - n_1, \ldots, N_k - n_k \longrightarrow 0$, better $\sum_{j=1}^{k}(N_j - n_j) \longrightarrow 0$,
2. $\sum_{j=1}^{k}|N_j - n_j| \longrightarrow 0$, better $\sum_{j=1}^{k}(N_j - n_j)^2 \longrightarrow 0$,
3. $\sum_{j=1}^{k}\frac{(N_j-n_j)^2}{p_j}$, better $\sum_{j=1}^{k}\frac{(N_j-n_j)^2}{n_j} \longrightarrow 0$.

$$K = \sum_{j=1}^{k}\frac{(N_j - n_j)^2}{n_j} = \sum_{j=1}^{k}\frac{N_j^2}{n_j} - 2\underbrace{\sum_{j=1}^{k}\frac{N_j\, n_j}{n_j}}_{=n} + \underbrace{\sum_{j=1}^{k}\frac{n_j^2}{n_j}}_{=n} = \sum_{j=1}^{k}\frac{N_j^2}{n_j} - n = \frac{1}{n}\sum_{j=1}^{k}\frac{N_j^2}{p_j} - n$$

The test statistic K is the sum of squares of k independent random variables with standard normal distribution with one binding condition $\sum_{j=1}^{k} N_j = n$.
What is the probability distribution of K and its expectation $\mathrm{E}(K)$?

Pearson's chi-squared test

Theorem (Pearson's chi-squared test *(Pearsonův test dobré shody)*)

Under H_0, the test statistic K has asymptotically chi-squared distribution with $(k-1)$ degrees of freedom,

$$K = \sum_{j=1}^{k} \frac{(N_j - n\,p_j)^2}{n\,p_j} = \frac{1}{n} \sum_{j=1}^{k} \frac{N_j^2}{p_j} - n \overset{as.}{\sim} \chi^2(k-1).$$

H_0 is rejected at the level of significance α, if $K \geq \chi^2_{1-\alpha}(k-1)$.

Assumptions *(Podmínka dobré aproximace)*

- $n_j \geq 5, \qquad j = 1, \ldots, k, \qquad$ **or**
- $n_j \geq 5\,q, \qquad$ where $q = \frac{1}{k} \cdot \left|\{j : n_j < 5\}\right| \qquad$ (so called Yarnold's criterion).

When these conditions are violated, it is necessary to appropriately merge some adjacent categories.

Alternatives

- $K = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$,

  where $O_j = N_j$ = **O**bserved frequencies, $E_j = n_j$ = **E**xpected frequencies.
- Deviation statistic / likelihood-ratio test *(deviační statistika, test poměrem věrohodnostní)*:

$$G = 2 \sum_{j=1}^{k} N_j \ln \frac{N_j}{n\, p_j} = 2 \sum_{j=1}^{k} O_j \ln \frac{O_j}{E_j}$$

  has, under H_0, asymptotically chi-squared distribution with $(k-1)$ degrees of freedom, $G \sim \chi^2(k-1)$.

  H_0 is rejected at the level of significance α, if $G \geq \chi^2_{1-\alpha}(k-1)$.

When the parameters are unknown

When testing the conformity of the probability distribution of random sample $(X_1, \ldots, X_n)$ with a theoretical probability distribution P with **unknown parameter(s)**, so-called modified minimum χ^2 method is used.

Modified minimum χ^2 method

Let us denote the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. The system of following equations is solved,

$$\sum_{j=1}^{k} \frac{N_j}{p_j(\boldsymbol{\theta})} \frac{\partial p_j(\boldsymbol{\theta})}{\partial \theta_i} = 0, \qquad i = 1, \ldots, m.$$

Theorem (Pearson's chi-squared test with unknown parameters)

When m parameters $\boldsymbol{\theta}$ estimated by the modified minimum χ^2 method are substitued into p_j and theoretical frequencies n_j, the degrees of freedom of the test statistic K is equal to $(k-1-m)$,

$$K \overset{as.}{\sim} \chi^2(k-1-m).$$

Degrees of freedom are reduced by the number of estimated parameters.

Algorithm for discrete probability distributions

1. The categories $A_1, \ldots, A_k$ must cover all possible outcomes $t_1, \ldots, t_k$ of the considered discrete probability distribution.
2. Calculate empirical frequencies,
$$N_j = \left|\{X_i = t_j\}\right|,$$
3. and theoretical frequencies using the probability function of the theoretical distribution,
$$n_j = n\,p_j = n\,P(X = t_j);$$
the theoretical cumulative distribution function $F(x)$ can also be used,
$$n_j = n\,p_j = n\left[F(t_j) - \lim_{t \to t_j^-} F(t)\right].$$
4. Verify the assumptions (e.g., Yarnold's criterion) and modify (merge) the categories unless the conditions are met.
5. Calculate the value of the test statistic K and decide to reject or not to reject H_0.

Algorithm for absolutely continuous probability distributions

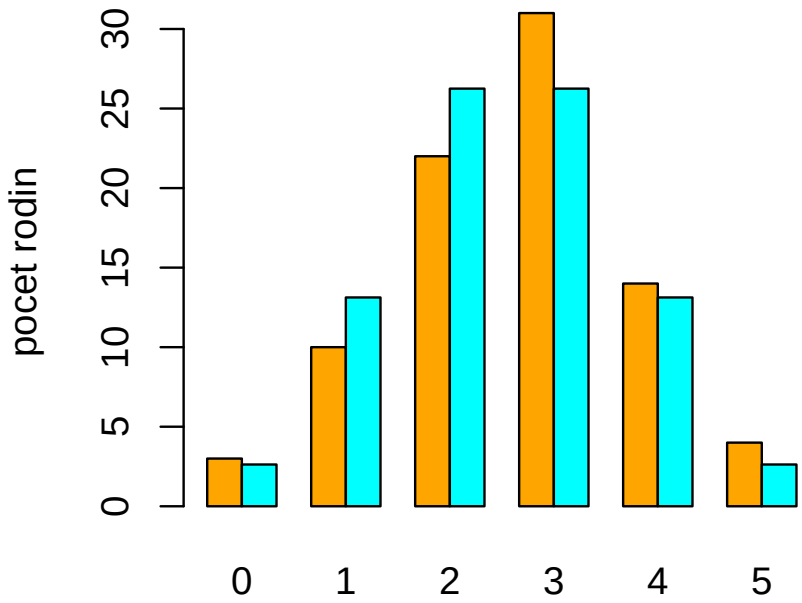
1. The categories $A_1, \ldots, A_k$ are defined as intervals
$$A_j = (t_{j-1}, t_j], \qquad j = 1, \ldots, k,$$
covering the entire range of possible outcomes of the considered absolutely continouso probability distribution. Sufficient number of values of the random sample $(X_1, \ldots, X_n)$ has to be in each interval. Recommended number of categories, i.e., intervals (as in the histogram construction): $k \approx \sqrt{n}$ for small n; sometimes $k \approx 1 + \log_2 n$; $k \approx 15 \left(\frac{n}{100}\right)^{2/5}$ for large n.
2. Calculate empirical frequencies,
$$N_j = \left|\{t_{j-1} < X_i \leq t_j\}\right|$$
3. and theoretical frequencies using the theoretical cumulative distribution function $F(x)$,
$$n_j = n\, p_j = n\, P(t_{j-1} < X \leq t_j) = n\left[F(t_j) - F(t_{j-1})\right].$$
4. Verify the assumptions (e.g., Yarnold's criterion) and modify (merge) the categories unless the conditions are met.
5. Calculate the value of the test statistic K and decide to reject or not to reject H_0.

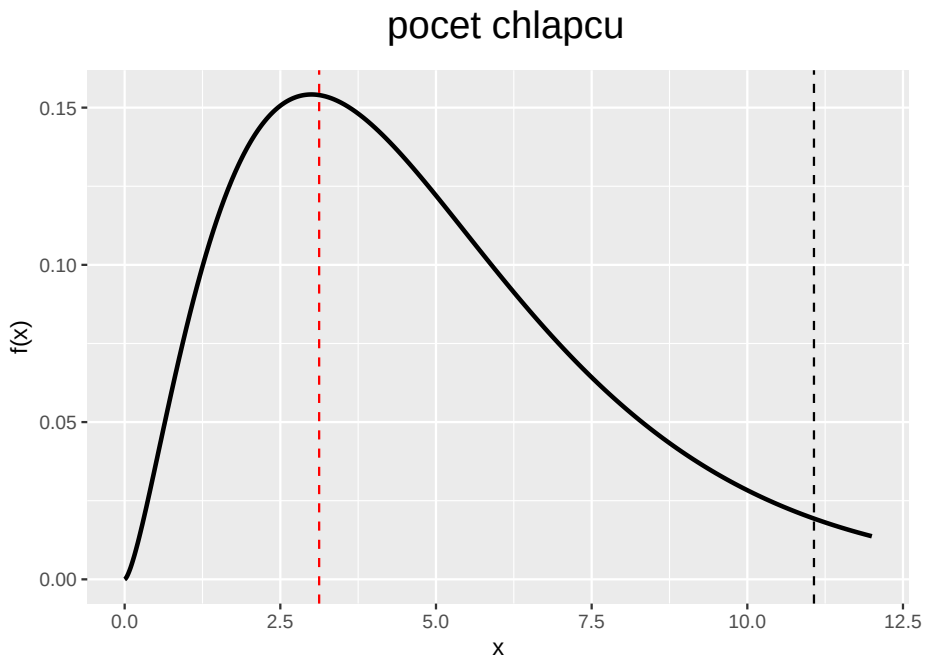
Example 1


```
  t.j N.j      p.j     n.j
1   0   3 0.03125   2.625
2   1  10 0.15625  13.125
3   2  22 0.31250  26.250
4   3  31 0.31250  26.250
5   4  14 0.15625  13.125
6   5   4 0.03125   2.625
```

```
q <- sum (n * p.j < 5) / k
[1] 0.3333333
n * p.j >= 5 * q
[1] TRUE TRUE TRUE TRUE TRUE TRUE
```

```
K <- sum (N.j^2 / (n * p.j)) - n
[1] 3.12381

qchisq (0.95, df = k - 1)
[1] 11.0705

K >= qchisq (0.95, df = k - 1)
[1] FALSE

1 - pchisq (K, df = k - 1)
[1] 0.6809048

chisq.test (N.j, p = p.j)
  Chi-squared test for given probabilities
data:  N.j
X-squared = 3.1238, df = 5, p-value = 0.68
```

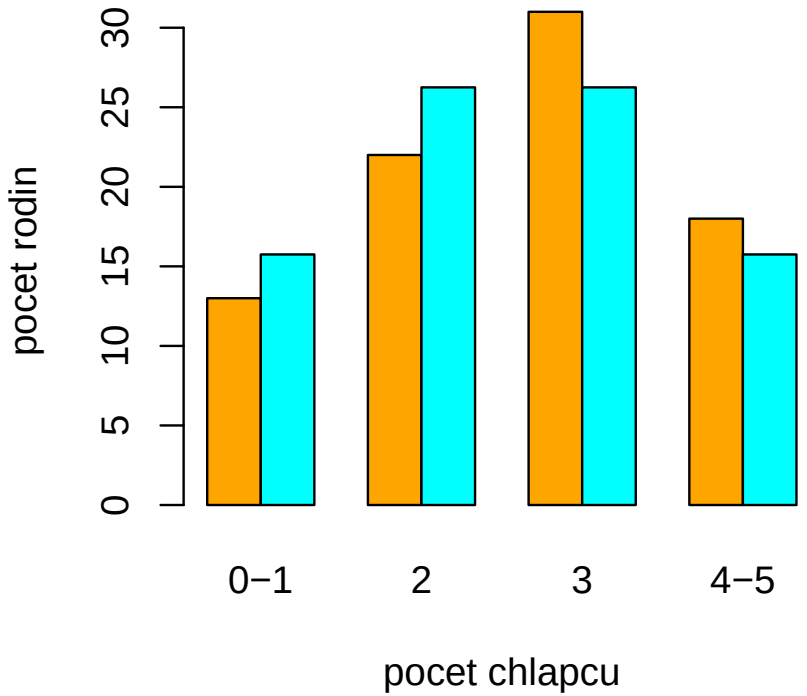


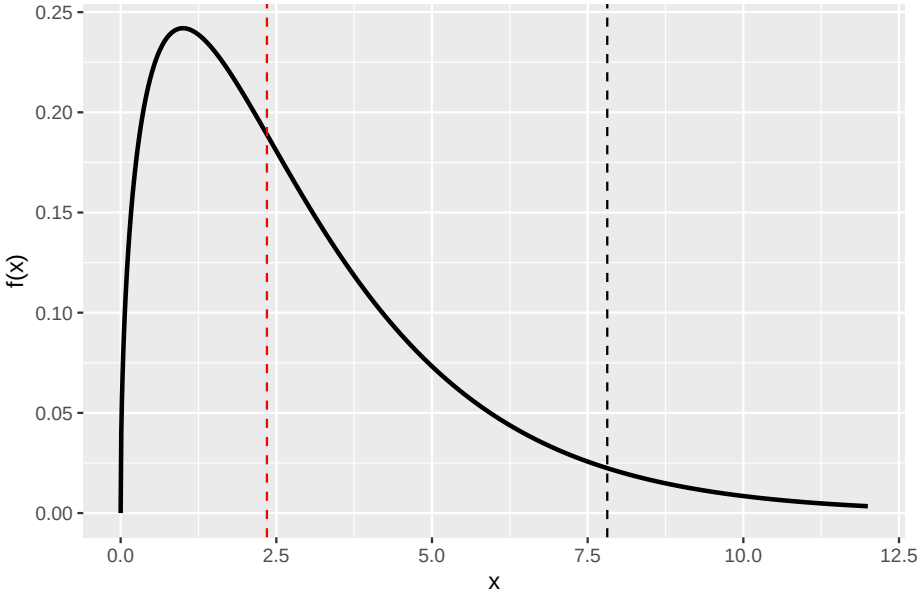


Example 1


```
n * p.j >= 5
[1] FALSE  TRUE  TRUE  TRUE  TRUE FALSE

  t.j2 N.j2   p.j2  n.j2
1  0-1   13 0.1875 15.75
2    2   22 0.3125 26.25
3    3   31 0.3125 26.25
4  4-5   18 0.1875 15.75
n * p.j2 >= 5
[1] TRUE TRUE TRUE TRUE

K <- sum (N.j2^2 / (n * p.j2)) - n
[1] 2.349206

qchisq (0.95, df = k2 - 1)
[1] 7.814728

K >= qchisq (0.95, df = k2 - 1)
[1] FALSE
```

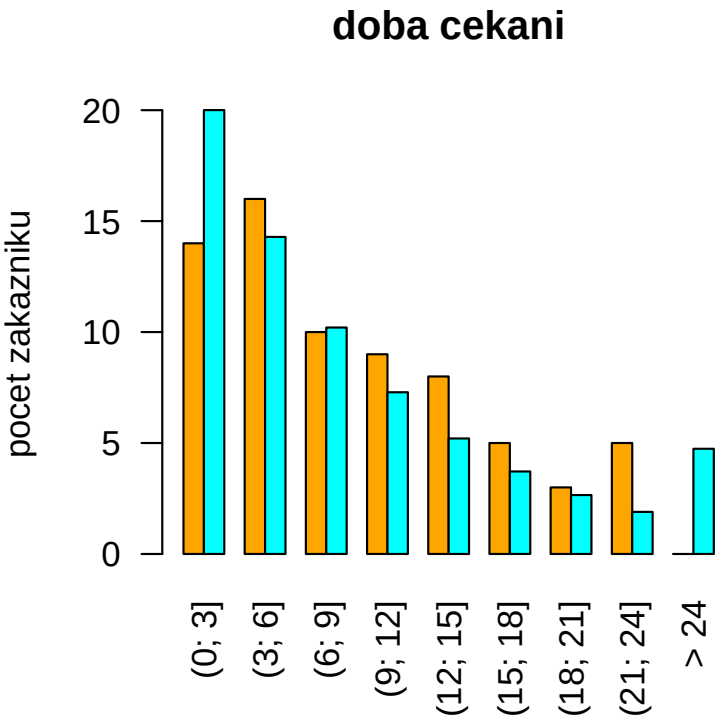
At asymptotic level of significance of 0.05, we do not reject the null hypothesis, that the number of boys in families with 5 children has binomial distribution $\mathrm{Bi}(5, 0.5)$.





Example 2

The *intensity* λ of the exponential distribution is estimated by the maximum likelihood method as $\widehat{\lambda} = \dfrac{1}{\overline{X}}$.



	A.j	N.j	p.j	n.j
1	(0; 3]	14	0.28576159	20.003311
2	(3; 6]	16	0.20410190	14.287133
3	(6; 9]	10	0.14577742	10.204419
4	(9; 12]	9	0.10411983	7.288388
5	(12; 15]	8	0.07436638	5.205647
6	(15; 18]	5	0.05311533	3.718073
7	(18; 21]	3	0.03793701	2.655591
8	(21; 24]	5	0.02709607	1.896725
9	> 24	0	0.06772447	4.740713

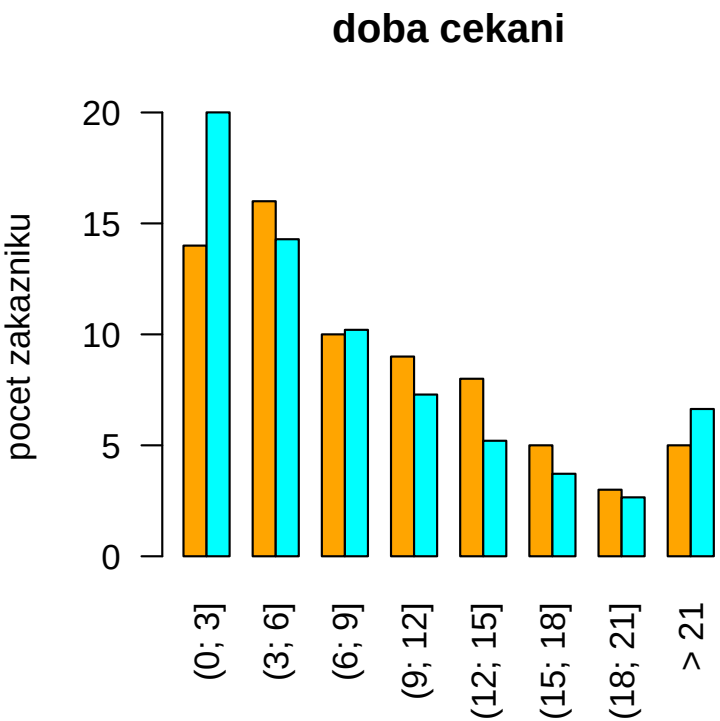
Note the last category, i.e., interval $(24, \infty]$, with no observation.

Example 2

assumptions, Yarnold's criterion

```
n * p.j >= 5
[1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE

q <- sum (n * p.j < 5) / k
[1] 0.4444444

n * p.j >= 5 * q
[1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
```

It is necessary to merge at least the last 2 categories. Then ...



	A.j2	N.j2	p.j2	n.j2
1	(0; 3]	14	0.28576159	20.003311
2	(3; 6]	16	0.20410190	14.287133
3	(6; 9]	10	0.14577742	10.204419
4	(9; 12]	9	0.10411983	7.288388
5	(12; 15]	8	0.07436638	5.205647
6	(15; 18]	5	0.05311533	3.718073
7	(18; 21]	3	0.03793701	2.655591
8	> 21	5	0.09482054	6.637438

Example 2

Yarnold's criterion

```
q <- sum (n * p.j2 < 5) / k2
[1] 0.25

n * p.j2 >= 5 * q
[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Pearson's chi-squared test, remeber $m = 1$ estimated parameter

```
K <- sum (N.j2^2 / (n * p.j2)) - n
[1] 4.803687

qchisq (0.95, df = k2 - 1 - 1)
[1] 12.59159

K >= qchisq (0.95, df = k2 - 1 - 1)
[1] FALSE
```

At the asymptotic level of significance of 0.05, we do not reject the null hypothesis, that the waiting times follow the exponential distribution.

Kolmogorov-Smirnov test and Lillieforse test

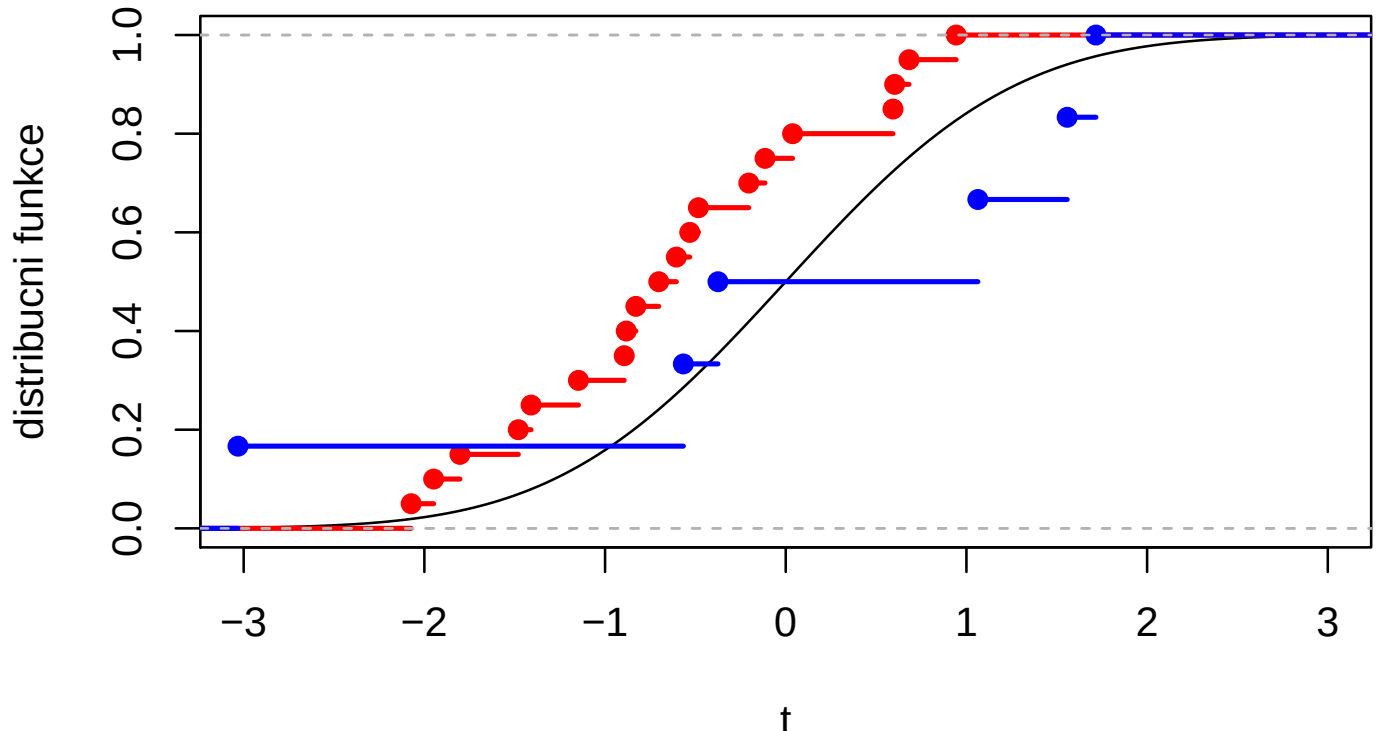
Empirical cumulative distribution function

Definition (ECDF *(empirická distribuční funkce)*)

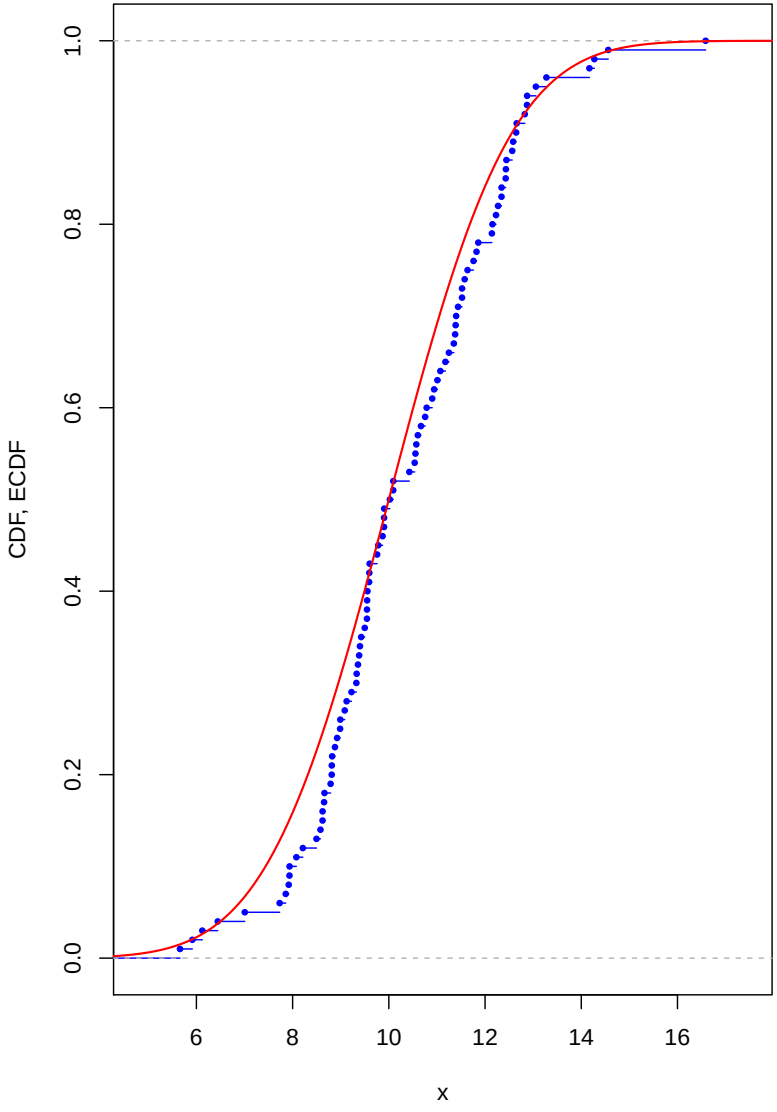
Let $(X_1, \ldots, X_n)$ be a random sample. Empirical cumulative distribution function / ECDF *(empirická distribuční funkce)* is defined

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}\{X_i \leq x\}, \qquad \text{where} \quad \mathbf{I}\{X_i \leq x\} = \begin{cases} 1, & X_i \leq x; \\ 0, & X_i > x. \end{cases}$$

ECDF is right-continuous step function. The steps correspond to the observed values $(X_1, \ldots, X_n)$.



Example: ECDF of simulated data




```
mu <- 10
sigma <- 2
X <- rnorm (100, mean=mu, sd=sigma)

F.emp <- ecdf (X)
plot (F.emp)

x <- seq (0, 20, by=0.1)
F <- pnorm (x, mean=mu, sd=sigma)
lines (t, F)
```

As the range n of the random sample increases, the empirical cumulative distribution function $\widehat{F}(x)$ approaches the true cumulative distribution function $F(x)$.

Kolmogorov-Smirnov test

Let $(X_1, \ldots, X_n)$ be a random sample from an **absolutely continuous probability distribution** with cumulative distribution function $F(x)$. Let $F_0(x)$ be the cumulative distribution function beeing tested,

$$H_0 : F(x) = F_0(x) \qquad H_1 : F(x) \neq F_0(x).$$

The test statistic is

$$D = \sup \left\{ \left| \widehat{F}(x) - F_0(x) \right| ; \quad -\infty < x < \infty \right\}.$$

Theorem (One-sample Kolmogorov-Smirnov test)

H_0 is rejected at the level of significance α, if $D \geq D_\alpha(n)$,
where $D_\alpha(n)$ is *critical value* of the one-sample Kolmogorov-Smirnov test.

For large n, ($n \geq 30$), we use $D_\alpha(n) \approx \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$.

What is the geometric meaning of the test statistic D?

Remark: The test statistic K has the same probability distribution as the supremum of *Brownian bridge*.

Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be two independent random samples from **absolutely continuous probability distributions** with cumulative distribution functions $F_X(x)$ and $F_Y(x)$. The equality of the cumulative distribution functions is tested,

$$H_0 : F_X(x) = F_Y(x) \qquad H_1 : F_X(x) \neq F_Y(x).$$

The test statistic is

$$D = \sup \left\{ \left| \widehat{F}_X(x) - \widehat{F}_Y(x) \right| ; \quad -\infty < x < \infty \right\}.$$

Theorem (Two-sample Kolmogorov-Smirnov test)

H_0 is rejected at the level of significance α, if $D \geq D_\alpha(n.m)$,
where $D_\alpha(n.m)$ is *critical value* of the two-sample Kolmogorov-Smirnov test.

For large m, n, $(m + 1 \geq 35)$, we use $D_\alpha(n, m) \approx \sqrt{\frac{n+m}{2\,m\,n}\,\ln\frac{2}{\alpha}}$.

What is the geometric meaning of the test statistic D?

Lilliefors test = K-S test of normality

H_0: random sample $(X_1, \ldots, X_n)$ comes from a gaussian (normal) probability distribution $\mathrm{N}(\mu, \sigma^2)$ with unknown parameters;

H_1: the random sample comes from non-gaussian (non-normal) probability distribution.

1. Parameters are estimated, $\widehat{\mu} = \overline{X}$, $\widehat{\sigma} = \sqrt{S^2}$,
2. and the K-S-like statistic is calculated,

$$L = \sup \left\{ \left| \widehat{F}(x) - \Phi\left(\frac{x - \widehat{\mu}}{\widehat{\sigma}} \right) \right| ; \; -\infty < x < \infty \right\},$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal $\mathrm{N}(0, 1)$ distribution.

Theorem (Lilliefors test)

H_0 is rejected at the level of significance α, if $L \geq L_\alpha(n)$,
where $L_\alpha(n)$ is *critical value* of the Lilliefors test.

Some other specific tests

Test for Poisson distribution

Let us remind: random variable X with Poisson distribution $X \sim \mathrm{Po}(\lambda)$ has equal expectation and variance,

$$\mathrm{E}(X) = \mathrm{Var}(X) = \lambda.$$

The test statistic uses this specific feature.

Theorem

Under the hypothesis of Poisson distribution of the observations,

$$\text{test statistic} \qquad Q = (n-1)\frac{S^2}{\overline{X}}$$

has asymptotically chi-squared distribution $\chi^2(n-1)$.

The hypothesis of the Poisson distribution of the sample is rejected at the asymptotic level of significance α, if

$$Q \le \chi^2_{\alpha/2}(n-1) \quad \text{or} \quad Q \ge \chi^2_{1-\alpha/2}(n-1).$$

Test for exponential distribution

Let us remind: the variance of random variable X with exponential distribution $X \sim \mathrm{Ex}(\lambda)$ is equal to the square of its expectation,

$$\mathrm{Var}(X) = [\mathrm{E}(X)]^2 = \frac{1}{\lambda^2}.$$

The test statistic uses this specific feature.

Theorem

Under the hypothesis of exponential distribution of the observations,

$$\text{test statistic} \qquad Q = (n-1)\frac{S^2}{\overline{X}^2}$$

has asymptotically chi-squared distribution $\chi^2(n-1)$.

The hypothesis of the exponential distribution of the sample is rejected at the asymptotic level of significance α, if

$$Q \le \chi^2_{\alpha/2}(n-1) \quad \text{or} \quad Q \ge \chi^2_{1-\alpha/2}(n-1).$$

Example 2


```
X <- rep (prumer.j, N.j)
Q <- (n-1) * var (X) / (mean (X))^2
[1] 35.72647

q1 = qchisq (0.025, n-1)
q2 = qchisq (0.975, n-1)
[1] 47.92416 93.85647

Q <= q1 | Q >= q2
[1] TRUE
```

At the asymptotic level of significance of 0.05, we reject the null hypothesis, that the waiting times follow the exponential distribution.

$$H_0 : F(x) = F_0(x) \qquad H_1 : F(x) \neq F_0(x).$$

Theorem (Anderson-Darling test)

H_0 is rejected at the level of significance α, if

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F_0(X_{(i)}) + \ln\left(1 - F_0(X_{(n-i+1)})\right)\right] \geq A_\alpha(n),$$

where $A_\alpha(n)$ is *critical value* of the one-sample Anderson-Darling test.

Theorem (Cramér-von Mieses test)

H_0 is rejected at the level of significance α, if

$$T = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{n} - F_0(X_{(i)})\right]^2 \geq T_\alpha(n),$$

where $T_\alpha(n)$ is *critical value* of the Cramér-von Mieses test.

Typically, both tests are used for **testing of normality** of random sample, i.e, F_0 is the distribution function of gaussian (normal) distribution.

Remark: Idea of test statistics is $\int_{-\infty}^{\infty}[\widehat{F}(x) - F_0(x)]^2\, w(x)\, f_0(x)\, \mathrm{d}x$.

Other normality tests

Shapiro-Wilk test

Shapiro-Wilk test is another very frequently used normality test, which uses order statistics $X_{(i)}$ and their specific properties in the normal probability distribution to calculate the test statistic.

Normality tests based on skewness and kurtosis

For random variable X, we define

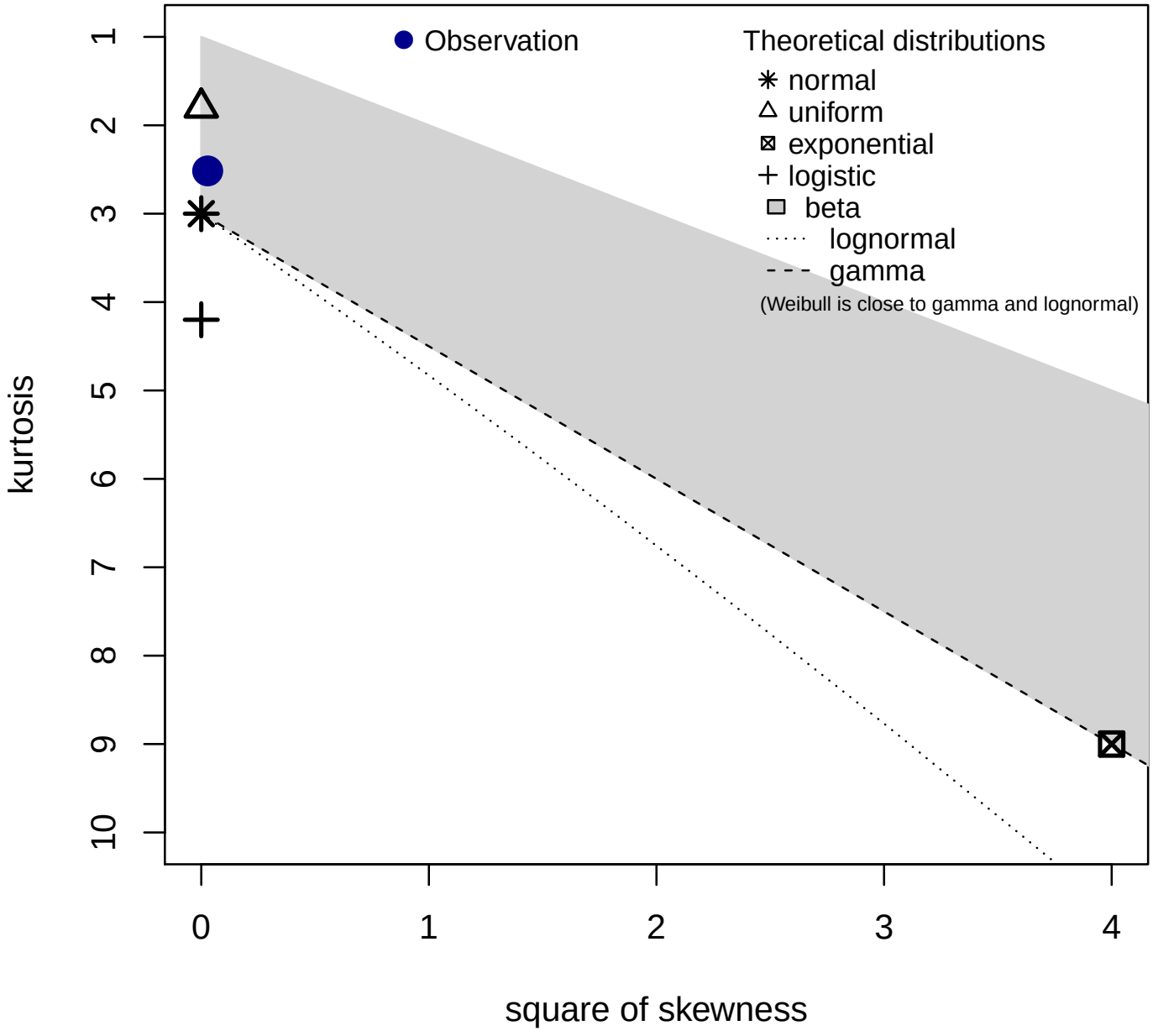
- skewness *(šikmost)*: $\alpha_3 = \dfrac{\mathrm{E}(X^3)}{[\mathrm{Var}(X)]^{3/2}}$,
- kurtosis *(špičatost)*: $\alpha_4 = \dfrac{\mathrm{E}(X^4)}{\mathrm{Var}(X)}$.

Specifically for gaussian $X \sim \mathrm{N}(\mu, \sigma^2)$, it holds $\alpha_3 = 0$ and $\alpha_4 = 3$.

Tests of normality using these features of the nornmal distribution:

- **D'Agostino's** K-squared **test**,
- **Jarque-Bera test**.

Cullen-Frey plot




```
library("fitdistrplus")
descdist(X)
```

Normality test in *R*

Pearson's chi-squared	`chisq.test(X, p=...)`	
Kolmogorov-Smirnov	`ks.test(X, "pnorm", mean=..., sd=...)`	
Lillieforse	`lillie.test(X)`	⋆
Pearson	`pearson.test(X)`	⋆
Anderson-Darling	`ad.test(X)`	⋆
Cramér-von Mieses	`cvm.test(X)`	⋆
Shapiro-Wilk	`shapirotest(X)`	
D'Agostino	`agostino.test(X)`	∗
Jarque-Bera	`jarque.test(X)`	∗
quantile-quantile plot	`qqnorm(X), qqline(X)`	

∗ `library("moments")`, ⋆ `library("nortest")`

Summary

- empirical and theoretical frequencies, calculations
- Pearson's chi-squared test, calculation of test statistics, adjustment of the degrees of freedom
- algorithms for discrete and absolutely continuos random variables
- empirical distribution function, Kolmogorov-Smirnov test, geometric interpration, Lilliefors test
- specific tests for Poisson exponential and normal distribution

Think about ...

1. Assume random sample $(X_1, \ldots, X_n)$ from an absolutely continuous probability distribution with cumulative distribution function $F_0(x)$.
2. Transform the observations, $Y_i = F_0(X_i)$.
3. What is the probability distribution of the transformed random sample $(Y_1, \ldots, Y_n)$?

MUNI FI

Statistics II | 5

Correlation coefficients, multiple linear regression

Ondřej Pokora

Department of Mathematics and Statistics, Faculty of Science, Masaryk University

10 October 2022 (updated 15 October 2022)

Pearson's correlation, variance-covariance matrix, correlation matrix

Probability theory:

- random variable *(náhodná veličina)* X having some probability distribution *(rozdělení pravděpodobnosti)* with non-random parameters (numbers)
- cumulative distribution function (c.d.f.) *(distribuční funkce)* $F(x) = P(X \leq x), P(a < X \leq b) = F(b) - F(a)$
- discrete X: probability mass function (p.m.f.) *(pravděpodobnostní funkce)* $p(x) = P(X = x)$
- absolutely continuous X: probability density function (p.d.f.) *(hustota pravděpodobnosti)* $f(x), P(a < X \leq b) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$

Statistics:

- random sample *(náhodný výběr)* $\boldsymbol{X} = (X_1, \ldots, X_n)$
- So-called sample statistics *(výběrové statistiky)* are functions of the random sample. They are random variables, i.e., have probability distributions.

Characteristics of numeric random variables and samples

- expected value / mean *(střední / očekávaná hodnota)*:

$$\mathrm{E}(X) = \sum_x x\, p(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x$$

- sample mean *(výběrový průměr)*: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

- variance *(rozptyl)*:

$$\mathrm{Var}(X) = \mathrm{E}[X - \mathrm{E}(X)]^2 = \sum_x [x - \mathrm{E}(X)]^2\, p(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} [x - \mathrm{E}(X)]^2\, f(x)\, \mathrm{d}x$$

- sample variance *(výběrový rozptyl)*: $S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

- standard deviation *(směrodatná odchylka)*: $\sigma_X = \sqrt{\mathrm{Var}(X)}$

- sample standard deviation *(výběrová směrodatná odchylka)*: $S_X = \sqrt{S_X^2}$

Pearson's correlation coefficient

- covariance *(kovariance)*: $\mathrm{C}(X, Y) = \mathrm{E}\left([X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]\right)$
- sample covariance *(výběrová kovariance)*:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

- Pearson's correlation coefficient *(korelační koeficient)*:

$$\rho(X, Y) = \frac{\mathrm{C}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \qquad \in [-1; 1]$$

- sample Pearson's correlation *(výběrový korelační koeficient)*:

$$r_{XY} = r(X, Y) = \frac{S_{XY}}{S_X \cdot S_Y} \qquad \in [-1; 1]$$

Remember: Sample statistics $\overline{X}$, S_X^2, S_X, S_{XY}, r_{XY}, etc. are random variables. They are estimates of the corresponding theoretical parameters.

Uncorrelation and stochastic independence

$H_0 : \rho_{X\,Y} = 0$, i.e., random variables are uncorrelated *(nekorelované)*;
$H_1 : \rho_{X\,Y} \neq 0$, i.e., random variables are correlated *(korelovanbé)*

Theorem

Under H_0, for $n \geq 3$, test statistic T has Student $\mathrm{t}(n-2)$ distribution,

$$T = r_{X\,Y}\sqrt{\frac{n-2}{1-r_{X\,Y}^2}} \;\sim\; \mathrm{t}(n-2).$$

H_0 is rejected at the level of significance α, if $|T| \geq t_{1-\alpha/2}(n-2)$.

- Pearson's correlation coefficient measures the strength of association between random variables X, Y and the direction of the relationship.
- **Stochastic independence of X, Y implies the uncorrelation.**
- H_0 not rejected, $r_{X\,Y} \approx 0 \Rightarrow X$, Y are uncorrelated
  But: **Uncorrelation does not imply stochastic independence of X, Y!**
- H_0 rejected, $r_{X\,Y} \approx \quad 1 \Rightarrow X$, Y correlated, positive relationship
- H_0 rejected, $r_{X\,Y} \approx -1 \Rightarrow X$, Y correlated, negative relationship

Test of (un)correlation using Z-transformation

$H_0 : \rho_{XY} = \rho_0$ for a given fixed value $\rho_0 \in [-1; 1]$,
$H_1 : \rho_{XY} \neq \rho_0$

Theorem (R. A. Fisher)

Under H_0, test statistic Z, so-called Z-transformation, has asymptotically normal distribution,

$$Z = \frac{1}{2} \ln \frac{1 + r_{XY}}{1 - r_{XY}} \overset{as.}{\sim} \mathsf{N}\left(\frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0}, \quad \frac{1}{n - 3} \right).$$

H_0 is rejected at the level of significance α, if

$$\frac{|Z - \mathsf{E}(Z)|}{\sqrt{\mathsf{Var}(Z)}} = \sqrt{n - 3} \left| Z - \frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0} \right| \geq u_{1-\alpha/2}.$$

Examples

Example (1)

Expenses (E) of 7 households (thousands CZK per 3 months) for food and beverages were observed depending on the number of household members (M) and the net income (I) of the household (thousands CZK per 3 months).

E	40	30	40	10	60	40	50
M	4	2	4	1	5	3	4
I	100	80	120	30	150	120	130

Analyze and quantify the association (correlation) of the variables.

Example (2)

20 children of different ages underwent pedagogical-psychological research, during which, among other things, they answered test questions and were weighed. Surprising was the value 0.968 of Pearson's correlation coefficient between the children's weight and the number of points achieved in the test. Does this mean that obesity has a positive effect on learning ability?

Example 1: correlation r_{ME}

Sample Pearson's correlation: $r_{ME} \doteq 0.942$

$$T = 0.942\sqrt{\frac{5}{1-0.942^2}} = 6.326 > t_{0.975}(5) = 2.571 \Rightarrow \rho_{ME} \text{ is significant}$$

$$Z\text{: } \sqrt{4}\left|0 - \frac{1}{2}\ln\frac{1+0.942}{1-0.942}\right| = 3.526 > u_{0.975} = 1.96 \Rightarrow \rho_{ME} \text{ is significant}$$

Expanses and number of members of household are correlated, with positive relationship.

```
cor(dt$expense, dt$members)
[1] 0.9762737
```

```
cor.test(dt$expense, dt$members)
  Pearson's product-moment correlation
data:  dt$expense and dt$members
t = 6.3263, df = 5, p-value = 0.001455
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6544368 0.9917452
sample estimates:
      cor
0.9428374
```

More variables

Let us assume l random variables $X_1, \ldots, X_l$ are examined.
We have an l-dimensional random sample of size n,

$$\boldsymbol{M} = \begin{pmatrix} X_{1\,1}, & \cdots, & X_{1\,l} \\ \vdots & & \vdots \\ X_{n\,1}, & \cdots, & X_{n\,l} \end{pmatrix},$$

where $X_{i\,j}$ denotes the i-th observation of X_j, $i = 1, \ldots, n$, $j = 1, \ldots, l$.

Definition

- Matrix $\boldsymbol{S}$ of sample covariances $S_{X_i X_j}$ of all pairs of random variables is called sample variance-covariance matrix *(výběrová kovarianční matice)*,

$$\boldsymbol{S} = \left\{ S_{X_i X_j} \right\}_{i,j=1}^{l}$$

- Matrix $\boldsymbol{R}$ of sample correlation coefficients $r_{X_i X_j}$ of all pairs of random variables is called sample correlation matrix *(výběrová korelační matice)*,

$$\boldsymbol{R} = \left\{ r_{X_i X_j} \right\}_{i,j=1}^{l}$$

Variance-covariance and correlation matrix

- Sample variance-covariance matrix $\boldsymbol{S}$ is squred $l \times l$, symmetrical positive definite matrix.
- The diagonal of $\boldsymbol{S}$ consists of sample variances $S^2_{X_1}, \ldots, S^2_{X_l}$.
- Sample correlation matrix $\boldsymbol{R}$ is squred $l \times l$, symmetrical matrix.
- The diagonal of $\boldsymbol{R}$ consists of l ones.

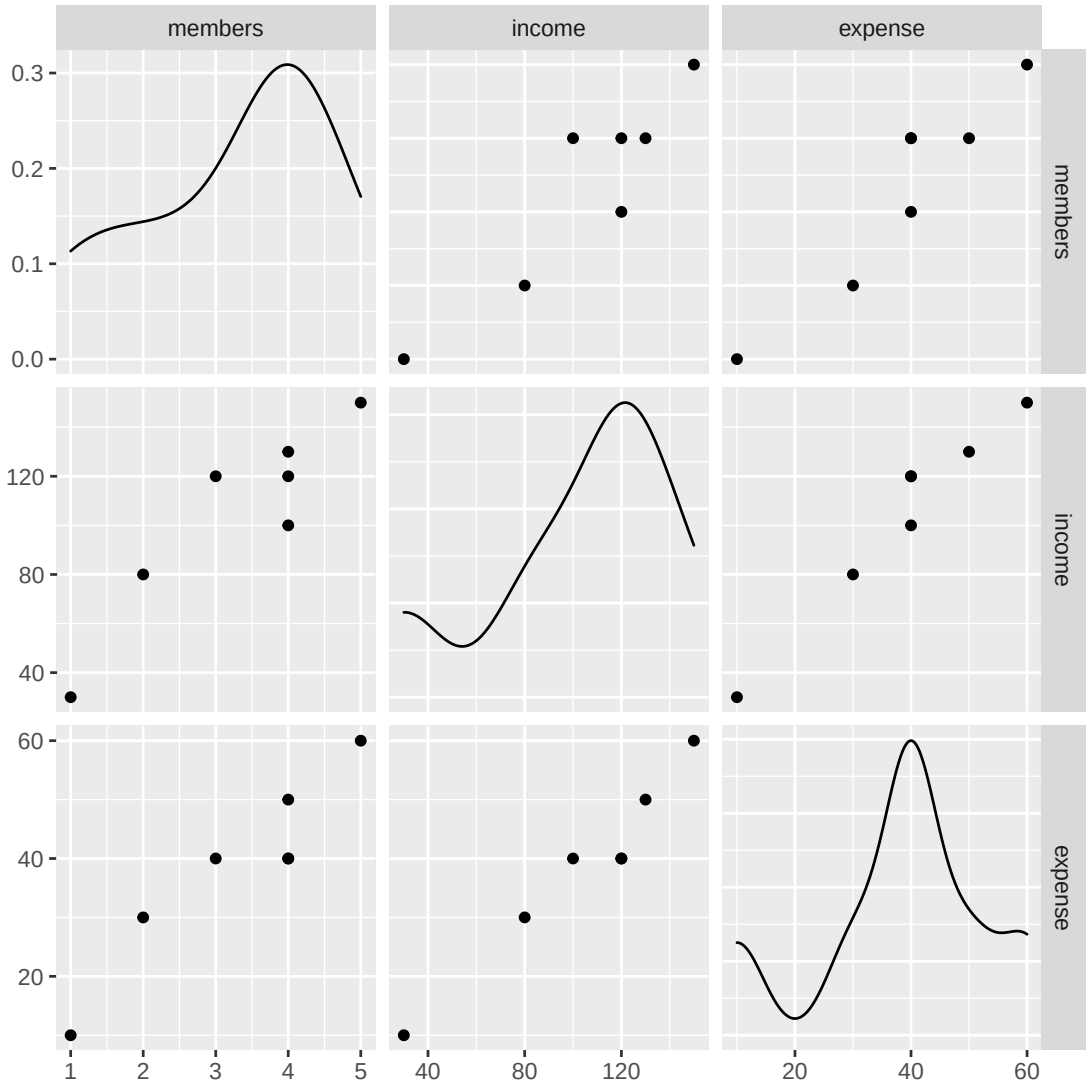
Graphical tools:

- Scatterplot is a matrix of two-dimensional point plots of the relationships of variable X_j on variable X_i, for each i, j.
- Correlogram is a visual representation of the sample correlation matrix $\boldsymbol{R}$. The color or size of a symbol indicates the value of the corresponding sample correlation coefficient. Correlogram can also display information about significance of the correlation coefficients.

Example 1: variance-covariance matrix, scatterplot


```
cov(M) # Variance-covariance matrix
           members     income   expense
members   1.904762    50.2381  20.47619
income   50.238095 1561.9048 607.14286
expense  20.476190  607.1429 247.61905
```

```
cor(M) # Correlation matrix
            members    income   expense
members  1.0000000 0.9210550 0.9428374
income   0.9210550 1.0000000 0.9762737
expense  0.9428374 0.9762737 1.0000000
```



Example 1: correlation matrix, correlogram


```
# Correlation matrix
           members    income   expense
members 1.0000000 0.9210550 0.9428374
income  0.9210550 1.0000000 0.9762737
expense 0.9428374 0.9762737 1.0000000
```

```
# p-values of tests of correlation
                members       income      expense
members 0.000000000 0.0032221604 0.0014548012
income  0.003222160 0.0000000000 0.0001644319
expense 0.001454801 0.0001644319 0.0000000000
```

Sample statistics in *R*

sample mean	$\overline{X}$	`mean(X)`
sample variance	S_X^2	`var(X)`
sample standard deviation	S_X	`sd(X)`
sample covariance	S_{XY}	`cov(X, Y)`
sample correlation	r_{XY}	`cor(X, Y)`, `cor.test(X, Y)`
sample variance-covariance matrix	$\boldsymbol{S}$	`cov(M)`
sample correlation matrix	$\boldsymbol{R}$	`cor(M)`, `Hmisc::rcorr(M)`
scatterplot		`GGally::ggpairs`
correlogram		`ggcorrplot::ggcorrplot`

Rank-based correlation coefficients

Spearman's rank correlation

Definition (Spearman's correlation)

Assume a pair of random samples, $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\boldsymbol{X} = (Y_1, \ldots, Y_n)$. Denote $\boldsymbol{R} = (R_1, \ldots, R_n)$, $\boldsymbol{S} = (S_1, \ldots, S_n)$ the ranks of particular samples.

Sample Spearman's rank correlation coefficient *(Spearmanův výběrový pořadový korelační koeficient)* of random variables X and Y is defined as Pearson's correlation of vectors of their ranks,

$$r_S(X, Y) = r(\boldsymbol{R}, \boldsymbol{S}).$$

If the ranks are not averaged, then $r_S = 1 - 6 \dfrac{\sum_{i=1}^{n} (R_i - S_i)^2}{n(n^2 - 1)}$.

- Spearman's correlation $r_S \in [-1; 1]$ quantifies rank correlation between random variables X and Y;
- is nonparametric analogy of Pearson's correlation;
- typically used for ordinal or non-gaussian data.

Test of rank correlation

Test of rank correlation *(Test pořadové korelovanosti)*

$H_0 : r_S = 0$,
$H_1 : r_S \neq 0$, i.e., random variables X and Y are rank-correlated *(pořadově korelované)*

Theorem

- Under H_0, test statistic $T_S = r_s \sqrt{\dfrac{n-2}{1-r_S^2}} \sim \mathrm{t}(n-2)$.

  H_0 is rejected at the level of significance α, if $|T_S| \geq t_{1-\alpha/2}(n-2)$.

- Under H_0, test statistic $Z_S = \sqrt{\dfrac{n-3}{1.06}} \cdot \dfrac{1}{2} \ln \dfrac{1+r_S}{1-r_S} \overset{as.}{\sim} \mathrm{N}(0;1)$.

  H_0 is rejected at the asymptotic level of significance α, if $|Z_S| \geq u_{1-\alpha/2}$.

`cor(..., method="spearman")`, `rcorr(..., type="spearman")`, `cor.test(..., method="spearman")`

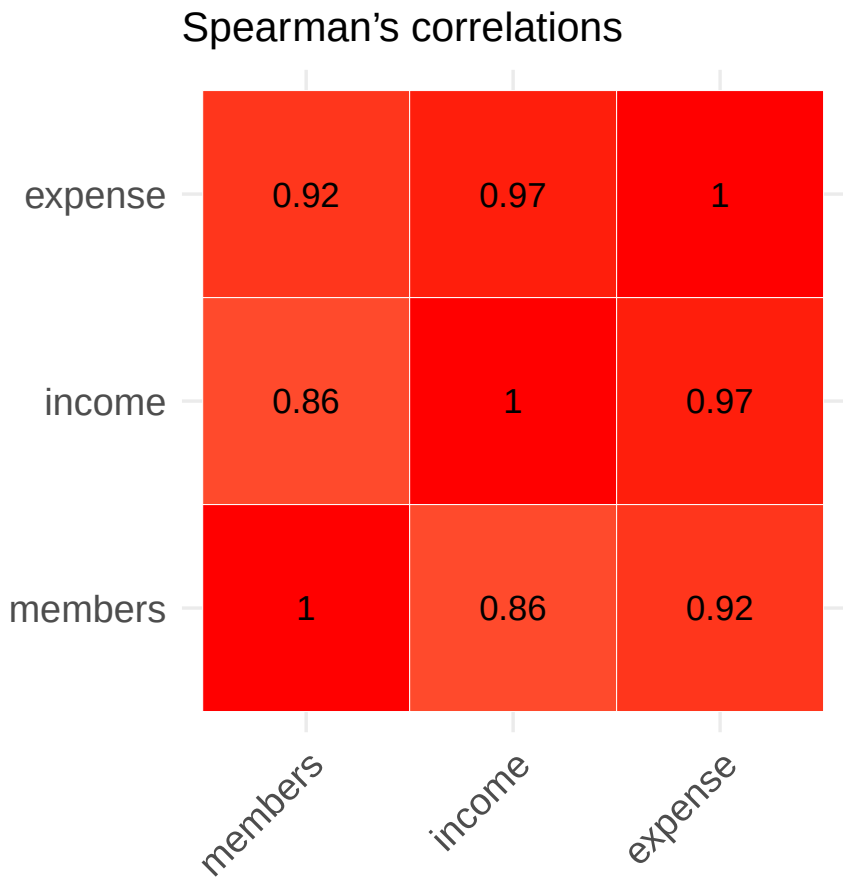
Example 1: Spearman's correlation

E_i	40	30	40	10	60	40	50
M_i	4	2	4	1	5	3	4
I_i	100	80	120	30	150	120	130
$R_i =$ rank of V_i	4	2	4	1	7	4	6
$S_i =$ rank of C_i	5	2	5	1	7	3	5
$T_i =$ rank of P_i	3	2	4.5	1	7	4.5	6

$$r_S(M, E) = r(\boldsymbol{R}, \boldsymbol{S}) = \frac{\sum_{i=1}^{n}(R_i S_i) - n\overline{R}\,\overline{S}}{\sqrt{\sum_{i=1}^{n} R_i^2 - n\overline{R}^2}\sqrt{\sum_{j=1}^{n} S_j^2 - n\overline{S}^2}} \doteq 0.923$$

```
[1] 0.5174525
# Spearman's correlation by definition
cor(rank(dt$expense), rank(dt$members))
[1] 0.9230769
cor(dt$expense, dt$members, method = "spearman")
[1] 0.9230769
cor.test(dt$expense, dt$members, method = "spearman") # Test of rank order
  Spearman's rank correlation rho
data:  dt$expense and dt$members
S = 4.3077, p-value = 0.003023
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
```

Example 1: Correlograms



Definition (Kendall's tau)

Sample Kendall's rank correlation coefficient / Kendall's tau *(Kendallův výběrový korelační koeficient)* of random variables X and Y is

$$\tau(X, Y) = \frac{n_+ - n_-}{\sqrt{n_0 - n_X}\sqrt{n_0 - n_Y}}, \qquad \text{where}$$

- $n_0 = \frac{1}{2}n(n-1) = \binom{n}{2}$ = number of all pairs,
- n_+ = number of concordant pairs,
- n_- = number of discordant pairs,
- $n_X = \frac{1}{2}\sum_i u_i(u_i - 1)$, $n_Y = \frac{1}{2}\sum_j v_j(v_j - 1)$, and u_i, v_j are numbers of particular ties in X and Y.

Pairs (X_i, Y_i) and (X_j, Y_j) are called:

- concordant *(konkordantní)*, if $X_i < X_j \;\&\; Y_i < Y_j$ or $X_i > X_j \;\&\; Y_i > Y_j$;
- discordant *(diskordantní)*, if $X_i < X_j \;\&\; Y_i > Y_j$ or $X_i > X_j \;\&\; Y_i < Y_j$.

Test of ordinal association

- Kendall's tau $\tau \in [-1; 1]$ quantifies ordinal association between random variables X and Y;
- in contrast to Spearman's correlation, it does not consider the distance between ranks;
- typically used for ordinal data, e.g., for two types of ranking.

Test of ordinal association *(Test ordinální asociace)*

$H_0 : \tau = 0$,
$H_1 : \tau \neq 0$, i.e., random variables X and Y are ordinally associated

Theorem

In the case of no ties in the data, H_0 is rejected at the asymptotic level of significance α, if $\sqrt{\frac{9n(n-1)}{2(2n+5)}}\,|\tau| \geq u_{1-\alpha/2}$.

For small n or in the presence of ties, corrected test statistics can be used.

```
cor(..., method="kendall"), rcorr(..., type="kendall"), cor.test(..., method="kendall")
```

Example 1: Kendall's tau

E_i	40	30	40	10	60	40	50
M_i	4	2	4	1	5	3	4
I_i	100	80	120	30	150	120	130

- $n_0 = \frac{1}{2}7 \cdot 6 = \binom{7}{2} = 21$ pairs in total
- $n_+ = 16$ concordant pairs (all except 1–3, 1–6, 1–7, 3–6, 3–7)
- $n_- = 0$ discordant pairs
- 3×4 in $M \Rightarrow n_M = \frac{1}{2}3 \cdot 2 = 3$; $\quad 3 \times 40$ in $E \Rightarrow n_E = \frac{1}{2}3 \cdot 2 = 3$
- $\tau(M, E) = \dfrac{n_+ - n_-}{\sqrt{n_0 - n_M}\sqrt{n_0 - n_E}} = \dfrac{16 - 0}{\sqrt{21 - 3}\sqrt{21 - 3}} \doteq 0.889$

```
cor(dt$expense, dt$members, method = "kendall")
[1] 0.8888889
cor.test(dt$expense, dt$members, method = "kendall") # Test of ordinal association
  Kendall's rank correlation tau
data:  dt$expense and dt$members
z = 2.6146, p-value = 0.008933
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8888889
```

Multiple linear regression model

Multiple linear regression model

Assume random variable Y dependens on l random variables $X_1, \ldots, X_l$,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_l X_l + \varepsilon = \beta_0 + \sum_{j=1}^{l} \beta_j X_j + \varepsilon$$

with random error ε.

n observations $(X_{i1}, \ldots, X_{il}, Y_i)$, $i = 1, \ldots, n$, are collected, where X_{ij} denotes the i-th observation of random variable X_j.

Multiple linear regression model:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \cdots + \beta_l X_{1l} + \varepsilon_1, \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \cdots + \beta_l X_{nl} + \varepsilon_n, \end{aligned}$$

written in matrix form as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & X_{11} & \cdots & X_{1l} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nl} \end{pmatrix}}_{\boldsymbol{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_l \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}, \quad \text{i. e., } \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_l)'$ = vector of $k = l + 1$ regression coefficients *(regresní koeficienty)*,
- $\boldsymbol{X}$ = regression / design matrix *(matice plánu)* of type $(n \times k)$ consists of a column of ones and of l columns of regressors *(regresory)*, $(X_{1\,1}, \ldots, X_{n\,1})', \ldots, (X_{1\,l}, \ldots, X_{n\,l})'$,
- $n > k$,
- $r(\boldsymbol{X}) = k = l + 1$, i. e., the design matrix has full rank *(plná hodnost)*, its columns are linearly independent.

Random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$:

- are nonsystematic: $\mathrm{E}(\varepsilon_i) = 0$, i.e., $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$,
- have homogeneous variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$,
- are mutually uncorrelated: $\mathrm{C}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$;
- variance-covariace matrix *(kovarianční matice)* of the vector of observations is $\mathrm{Var}(\boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$.
- Hence, observations are uncorrelated and have homogeneous variance.

Least squares method

Optimization: find such $\boldsymbol{\beta}$ which minimizes the sum of quadratic deviations,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i - \beta_0 - \sum_{j=1}^{l} \beta_j x_{ij} \right]^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \longrightarrow \min .$$

- Ordinary Least Squares (OLS) estimate *(odhad metodou nejmenších čtverců)*
$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_l) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$
- predicted / fitted values: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, i.e., $\widehat{Y}_i = \widehat{\beta}_0 + \sum_{j=1}^{l} \widehat{\beta}_j x_{ij}$,
- residuals *(rezidua)* $r_i = Y_i - \widehat{Y}_i$,
- residual sum of squares *(reziduální součet čtverců)*
$$S_e = S(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \sum_{i=1}^{n} \left[Y_i - \widehat{\beta}_0 - \sum_{j=1}^{l} \widehat{\beta}_j X_{ij} \right]^2 = \sum_{i=1}^{n} r_i^2,$$
- coefficient of determination *(index determinace)* R squared:
$$R^2 = \frac{S_{\widehat{Y}}}{S_T} = 1 - \frac{S_e}{S_T}, \text{ where } \quad S_{\widehat{Y}} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2, \; S_T = \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$
- adjusted - R bar squared: $\overline{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$.

Theorems

Theorem (Gauss-Markov)

OLS estimate $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is BLUE = Best Linear Unbiased Estimate *(nejlepší nestranný lineární odhad)* of vector $\boldsymbol{\beta}$ and its variance-covariance matrix *(kovariančí matice)* is $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.

Theorem

$\widehat{\sigma^2}_{\text{OLS}} = \dfrac{S_e}{n-(l+1)} = \dfrac{S_e}{n-k}$ is an unbiased estimate of the variance σ^2 of random errors.

Theorem

Additionaly, let us assume that the observations have n-dimensional gaussian (normal) disribution $\boldsymbol{Y} \sim \text{N}_n\left(\boldsymbol{X}\boldsymbol{\beta},\, \sigma^2 \boldsymbol{I}_n\right)$. Then:

- OLS estimate has gaussian distribution, $\widehat{\boldsymbol{\beta}}_{\text{OLS}} \sim \text{N}_k\left(\boldsymbol{\beta},\, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$,
- statistic $K = (n-k)\dfrac{\widehat{\sigma^2}_{\text{OLS}}}{\sigma^2} \sim \chi^2(n-k)$ has chi-square distribution,
- OLS estimate $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ and statistic K are independent.

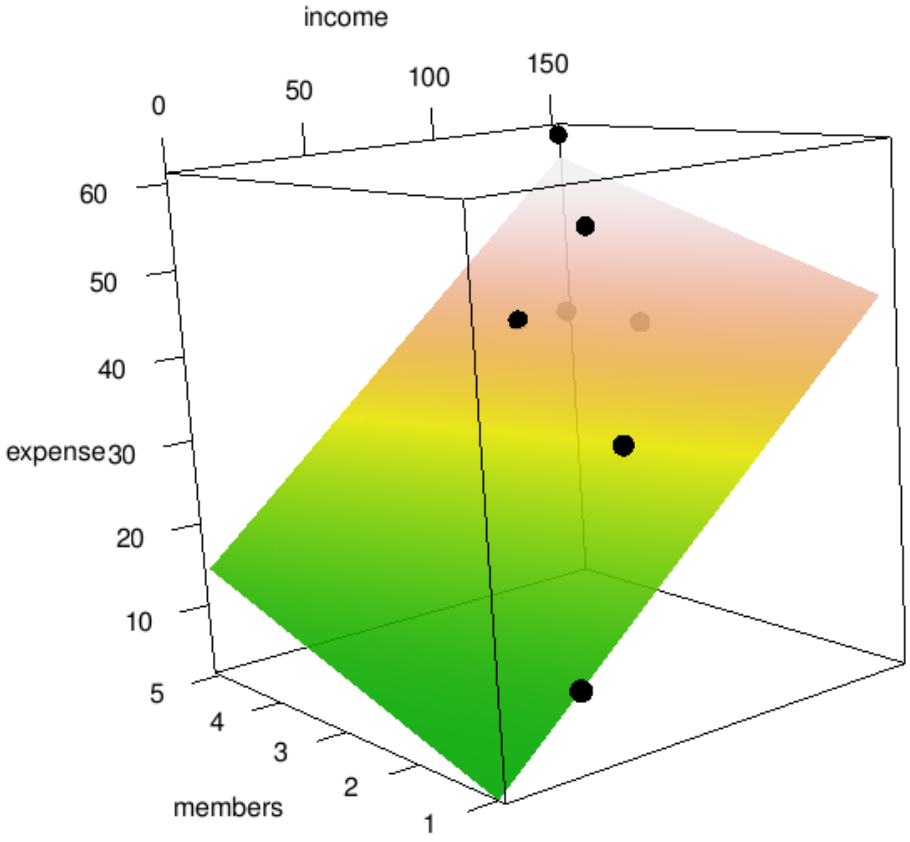
Best linear approximation

Definition

Random variable

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_p} X_l,$$

where $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_l$ are the OLS-estimates, is called best linear approximation *(nejlepší lineární aproximace)* of random variable Y using random variables $(X_1, \ldots, X_l)$.



The graph of the best linear approximation $\widehat{Y}$ in dependency on variables $(X_1, \ldots, X_l)$ is **l-dimensional hyperplane in a k-dimensional vector space**, $k = l + 1$.

Example 1: multiple linear regression

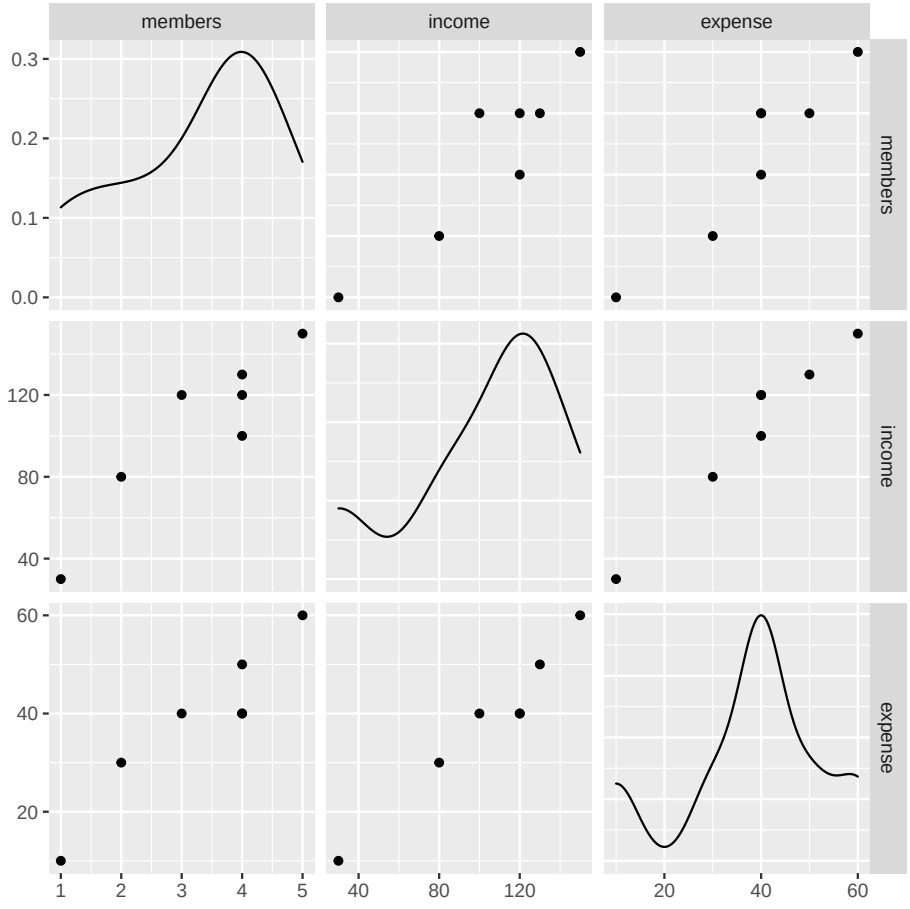
The household expenses are modeled in dependency on income and the number of household members.

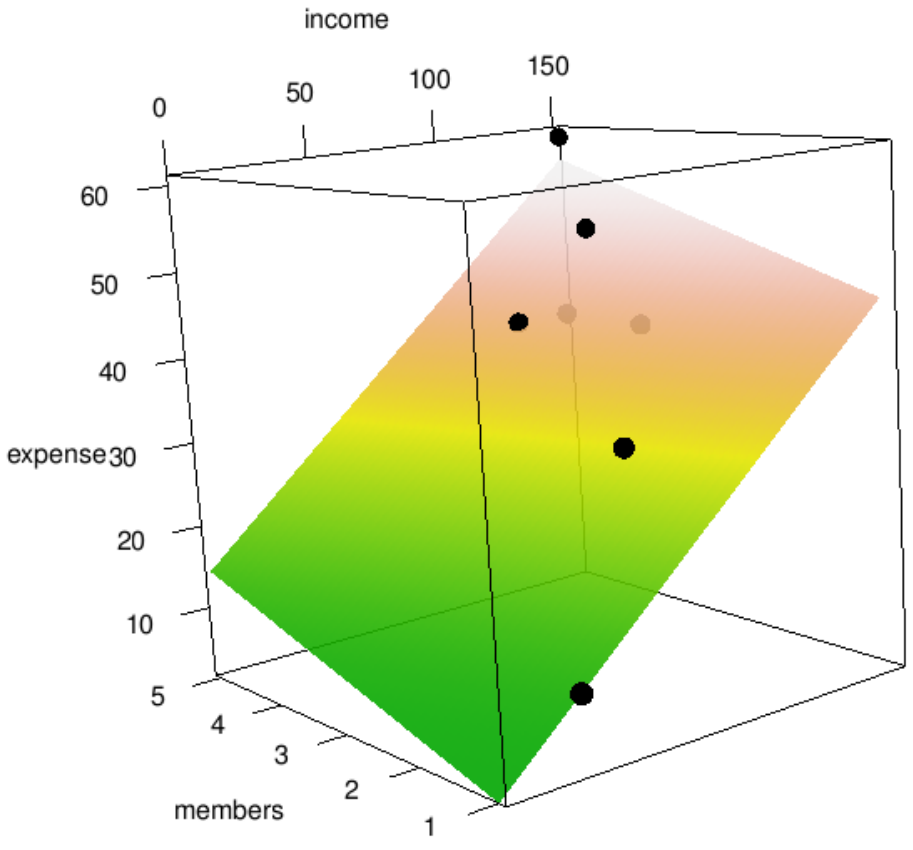
Multiple linear regression model: $E = \beta_0 + \beta_1 \cdot I + \beta_2 \cdot M + \varepsilon$

```
lm(formula = expense ~ income + members, data = dt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.74142    4.08119  -0.427   0.6916
income       0.28320    0.09473   2.990   0.0404 *
members      3.28056    2.71254   1.209   0.2931
---
Residual standard error: 3.571 on 4 degrees of freedom
Multiple R-squared:  0.9657,  Adjusted R-squared:  0.9485
F-statistic: 56.25 on 2 and 4 DF,  p-value: 0.001179
```

OLS-estimates: $\widehat{\beta_0} = -1.741,\ \widehat{\beta_1} = 0.283 > 0,\ \widehat{\beta_2} = 3.281;\ R^2 = 0.966$

Example 1: scatterplot and regression plane





Best linear approximation $\widehat{E}$ of E using I and M is **regression plane** given by

$$\widehat{E} = -1.741 + 0.283 \cdot I + 3.281 \cdot M$$

Multiple and partial correlation

Multiple correlation

Multiple correlation coefficient *(koeficient mnohonásobné korelace)* $\rho_{Y \cdot X_1 \dots X_l}$ is Pearson's correlation between variable Y and its best linear approximation $\widehat{Y}$ using variables $X_1, \dots, X_l$,

$$\rho_{Y \cdot X_1 \dots X_l} = \rho(Y, \widehat{Y}) \in [0; 1].$$

It quantifies the association between Y and the vector $(X_1, \dots, X_l)$. It is the **largest correlation** between Y and any linear combination of variables $X_1, \dots, X_l$.

Sample multiple correlation coefficient $r_{Y \cdot X_1 \dots X_l} = r(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) \in [0; 1]$ is sample Pearson's correlation between the observation vector $\boldsymbol{Y}$ and **fitted values** $\widehat{\boldsymbol{Y}}$ using predictors $X_1, \dots, X_l$ in multiple linear regression model

$$M : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Theorem (Test of significance of multiple correlation)

$H_0 : \rho_{Y \cdot X_1 \dots X_l} = 0$ against one-sided alternative is rejected at the level of significance α, if $F = \dfrac{n-l-1}{l} \cdot \dfrac{r^2_{Y \cdot X_1 \dots X_l}}{1 - r^2_{Y \cdot X_1 \dots X_l}} \geq F_{1-\alpha}(l, n-l-1)$.

Partial correlation

Partial correlation coefficient *(koeficient parciální korelace)* $\rho_{X\,Y\,\cdot\,Z_1\,\ldots\,Z_m}$ is Pearson's correlation between variables $(X - \widehat{X})$ and $(Y - \widehat{Y})$, where $\widehat{X}$, $\widehat{Y}$ are the best linear approximations of X, Y using variables $Z_1, \ldots, Z_m$,

$$\rho_{X\,Y\,\cdot\,Z_1\,\ldots\,Z_m} = \rho(X - \widehat{X},\ Y - \widehat{Y}) \in [-1; 1].$$

It quantifies the association between variables X and Y **excluding the effect of variables** $Z_1, \ldots, Z_m$.

Sample partial correlation coefficient

$$r_{X\,Y\,\cdot\,Z_1\,\ldots\,Z_m} = r(\ \underbrace{X - \widehat{X}}_{\text{residuals in } M_X}\ ,\ \underbrace{Y - \widehat{Y}}_{\text{residuals in } M_Y}\ ) \in [-1; 1],$$

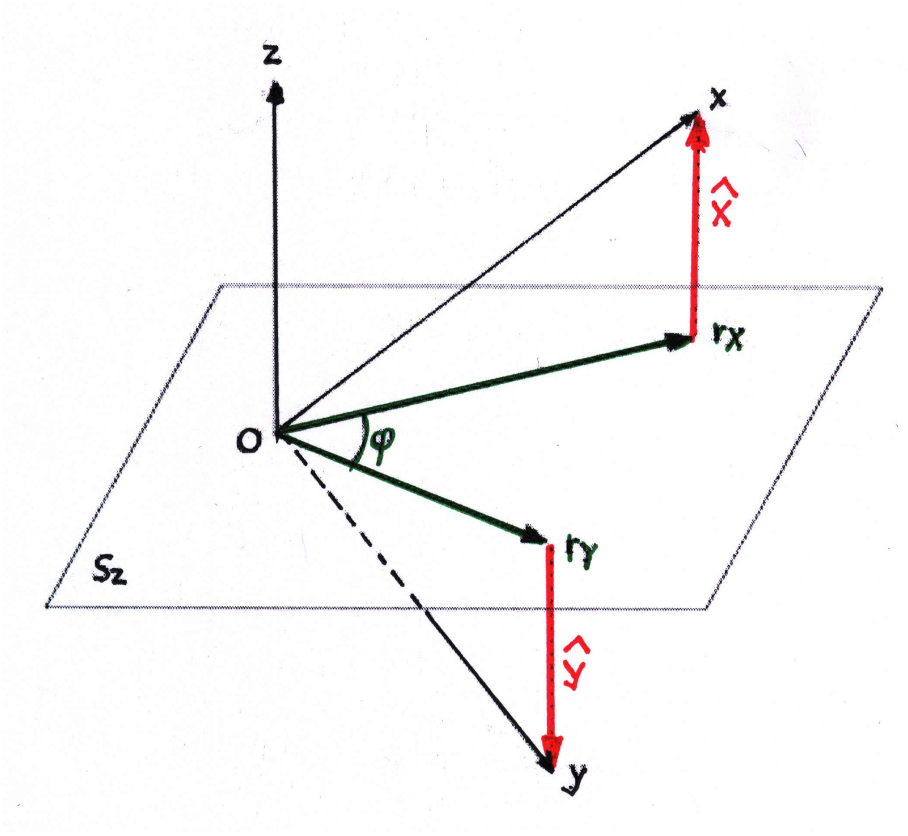
is sample Pearson's correlation between the vectors of **residuals** $(X - \widehat{X})$ and $(Y - \widehat{Y})$ in multiple linear regression models

$$M_X : X = \alpha_0 + \alpha_1 Z_1 + \cdots + \alpha_m Z_m; \qquad M_Y : Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_m Z_m.$$

Theorem (Test of significance of partial correlation)

$H_0 : \rho_{X\,Y\,\cdot\,Z_1\,\ldots\,Z_m} = 0$ against two-sided alternative is rejected at the level of significance α, if $\quad T = r_{X\,Y\,\cdot\,Z_1\,\ldots\,Z_m} \sqrt{\dfrac{n - m - 2}{1 - r^2_{X\,Y\,\cdot\,Z_1\,\ldots\,Z_m}}} \geq t_{1-\alpha/2}(n - m - 2).$

Geometric meaning of partial correlation



- x, y, z = three observations in 3D
- S_z = hyperplane perpendicular to z
- $\widehat{x}$ = best linear approximation of x using z
- $\widehat{y}$ = best linear approximation of y using z
- $r_x = x - \widehat{x}$ = residuals of x = projection of x into S_z
- $r_y = y - \widehat{y}$ = residuals of y = projection of y into S_z
- partial correlation = cosine of the angle of residuals in S_z,

$$r_{XY \cdot Z} = \cos\varphi = \cos|\sphericalangle\, r_X,\ r_Y|$$

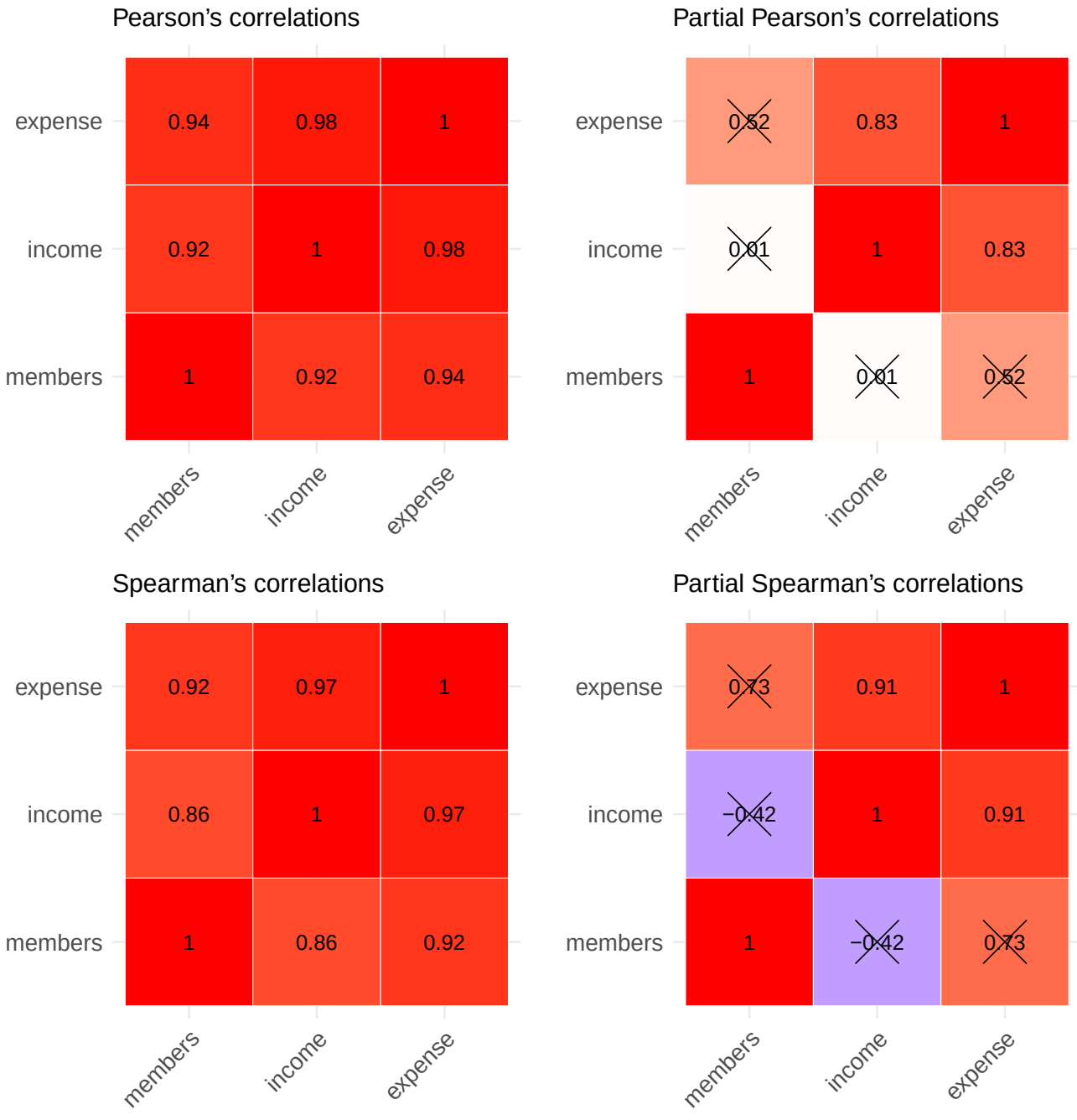
mutiple correlation $r_{Y \cdot X_1 \ldots X_l}$

```
model.M <- lm(Y ~ X1 + ... + Xl)  # model M
# multiple correlation between Y and X1, ..., Xl
# as correlation between Y and best linear approximation of Y using X1, ..., Xl in r
cor(Y, fitted.values(model.M))
```

partial correlation $r_{X\,Y \cdot Z_1 \ldots Z_m}$

```
model.MX <- lm(X ~ Z1 + ... + Zm) # model MX
model.MY <- lm(Y ~ Z1 + ... + Zm) # model MY
# partial correlation between X and Y excluding the effect of Z1, ..., Zm
cor(residuals(model.MX), residuals(model.MY))

library("ppcor")  # or using functions from "ppcor" library
# partial correlations of each pair of variables exluding all other variables
pcor(M)
# test of significance of partial correlation between X and Y excluding Z
pcor.test(X, Y, Z)
```

Example 1: Correlograms



Example 2: multiple linear regression

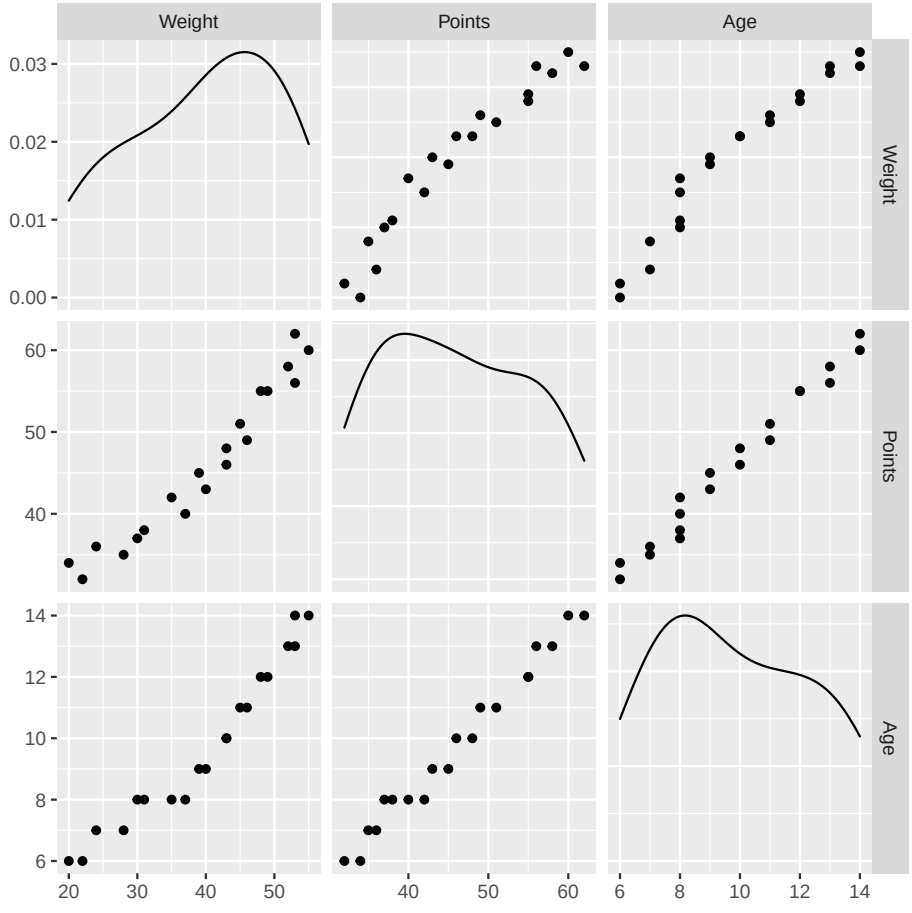
Number of point achieved in the test are modeled in dependency on weight and ages of the children.

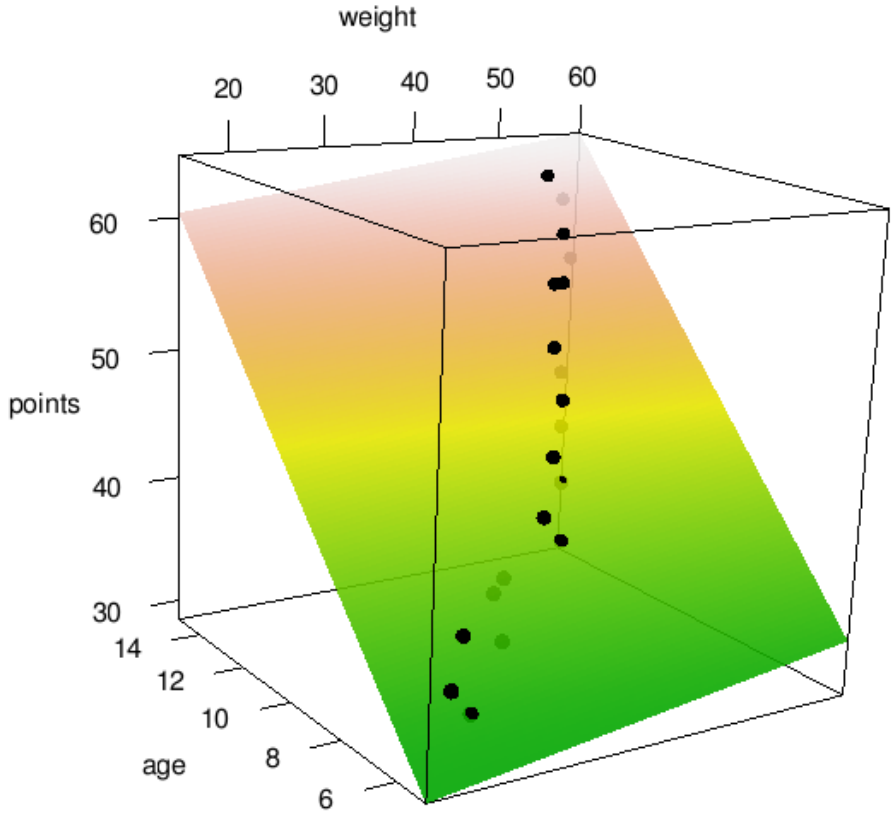
Multiple linear regression model: $Points = \beta_0 + \beta_1\, Weight + \beta_2\, Age + \varepsilon$

```
lm(formula = Points ~ Weight + Age, data = dt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.06490    1.23693   8.945 7.72e-08 ***
Weight       0.09466    0.12090   0.783    0.444
Age          3.19203    0.51058   6.252 8.77e-06 ***
---
Residual standard error: 1.377 on 17 degrees of freedom
Multiple R-squared:  0.9806,  Adjusted R-squared:  0.9784
F-statistic: 430.4 on 2 and 17 DF,  p-value: 2.753e-15
```

OLS-estimates: $\widehat{\beta_0} = 11.065 > 0,\ \widehat{\beta_1} = 0.095,\ \widehat{\beta_2} = 3.192 > 0;\ R^2 = 0.981$

Example 2: scatterplot and regression plane



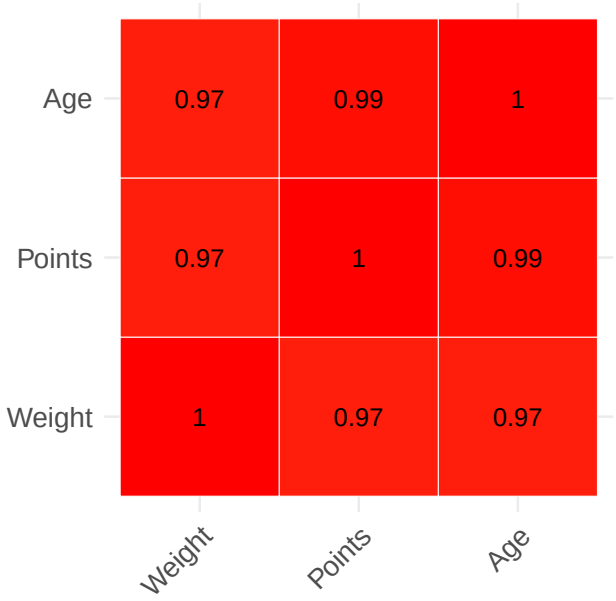


Best linear approximation $\widehat{Points}$ of *Points* using *Weight* and *Age* is **regression plane** given by

$$\widehat{Points} = 11.065 + 0.095\, Weight + 3.192\, Age$$

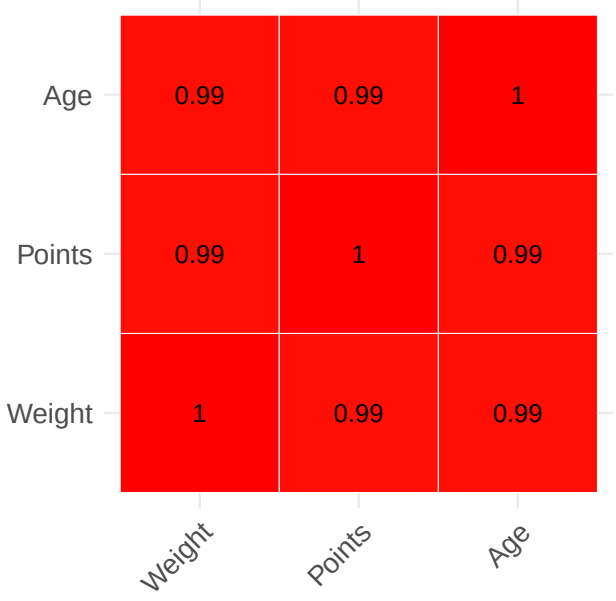
Eample 2: Correlograms

Pearson's correlations

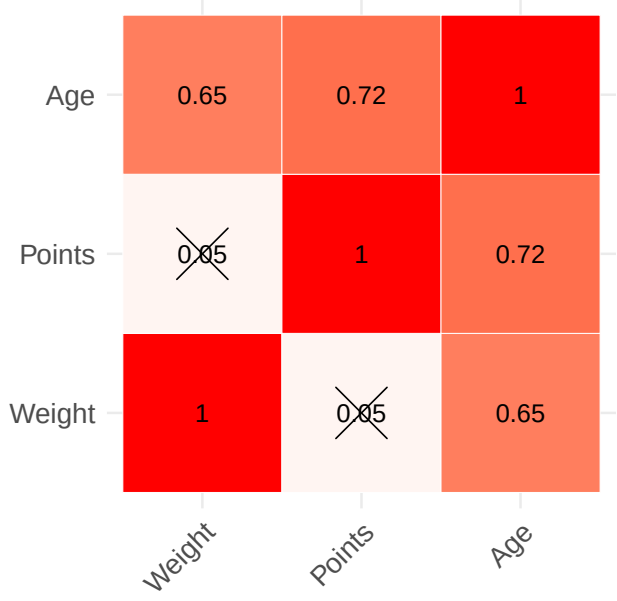


Partial Pearson's correlations

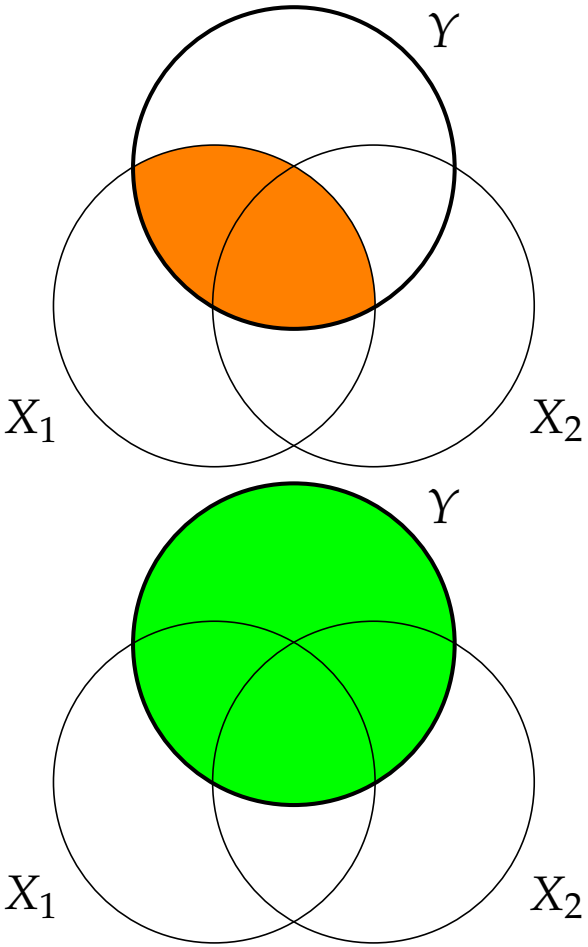
Spearman's correlations



Partial Spearman's correlations



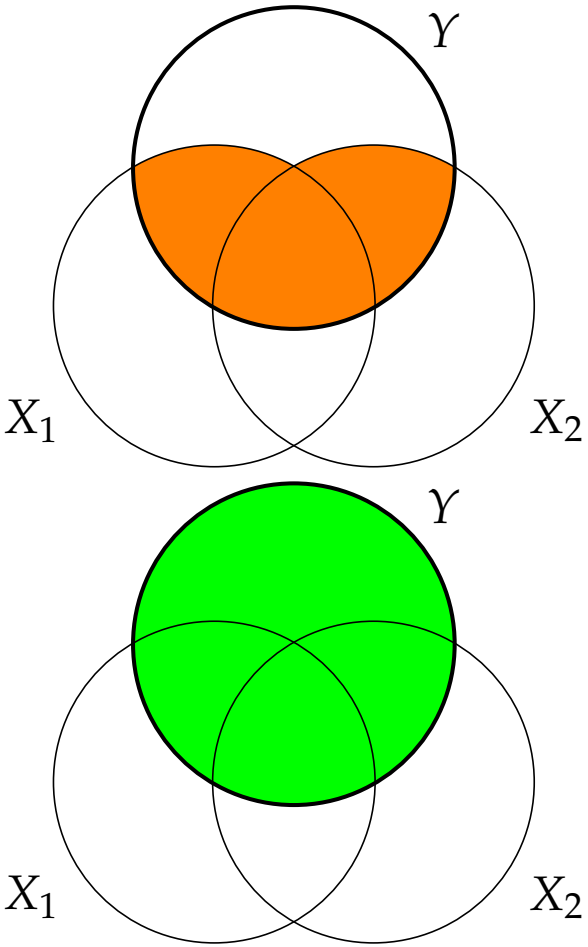
R squared and Pearson's correlation



- variable Y is modeled by variable X_1
- Pearson's correlation $r_{Y X_1}$ quantifies the association between Y and X_1
- coefficient of determination $R^2_{Y \cdot X_1}$ in model `Y ~ X1`
  quantifies the ratio of variability of Y which is explained by X_1
- the square $r^2_{Y X_1}$ of the Pearson's correlation is equal to the coefficient of determination,

$$r^2_{Y X_1} = \frac{R^2_{Y X_1}}{1} = R^2_{Y X_1}$$

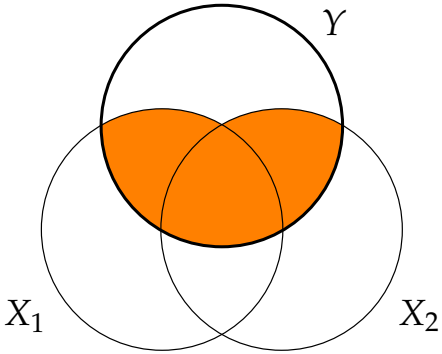
R squared and multiple correlation



- variable Y is modeled by variables X_1, X_2
- multiple correlation $r_{Y \cdot X_1 X_2}$ quantifies the association between Y and its best linear approximation $\widehat{Y}$ using X_1, X_2
- coefficient of determination $R^2_{Y \cdot X_1 X_2}$ in model `Y ~ X1 + X2` quantifies the ratio of variability of Y which is explained by X_1 and X_2
- the square $r^2_{Y \cdot X_1 X_2}$ of the multiple correlation is equal to the coefficient of determination,

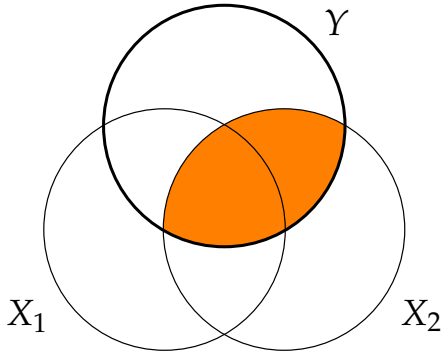
$$r^2_{Y \cdot X_1 X_2} = \frac{R^2_{Y \cdot X_1 X_2}}{1} = R^2_{Y \cdot X_1 X_2}$$

R squared and partial correlation



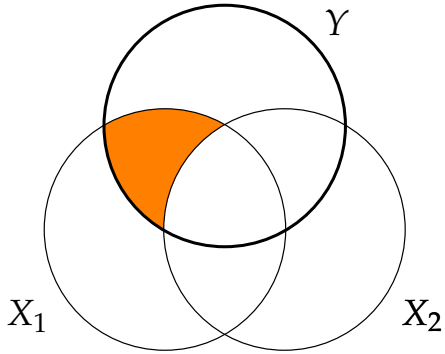
`Y ~ X1 + X2`

$R^2_{Y \cdot X_1 X_2}$

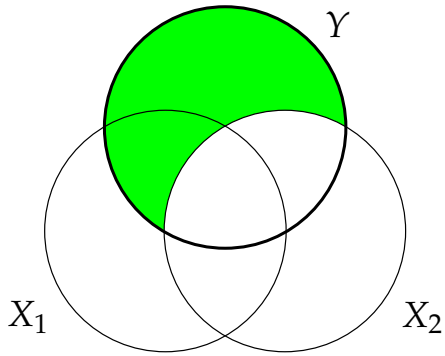


`Y ~ X2`

$R^2_{Y \cdot X_2}$



$R^2_{Y \cdot X_1 X_2} - R^2_{Y \cdot X_2}$



- variable Y is modeled by variable X_1 and excluding variable X_2
- partial correlation $r_{Y X_1 \cdot X_2}$ quantifies the association between Y and X_1 excluding the effect of X_2
- coefficient of determination $R^2_{Y \cdot X_2}$ in model `Y ~ X2` quantifies the ratio of variability of Y which is explained by X_2
- the square $r^2_{Y X_1 \cdot X_2}$ of the partial correlation is equal to the ratio ov variability of Y which is explained by X_1 independently on X_2,

$$r^2_{Y X_1 \cdot X_2} = \frac{R^2_{Y \cdot X_1 X_2} - R^2_{Y \cdot X_2}}{1 - R^2_{Y \cdot X_2}}$$

- Note that `Y ~ X2` is a submodel of `Y ~ X1 + X2`.

Example 1: multiple and partial correlation

Mutiple correlation $\rho_{E.MI}$

```
model.M <- lm(expense ~ income + members, data = dt)
cor(dt$expense, fitted.values(model.M)) # by definition, as correlation between vari
sqrt(summary(model.M)$r.squared)        # or, as square root of R squared
[1] 0.9826827
```

$$r_{E.IM} = 0.983,\ R^2_{E.IM} = r^2_{V.CP} = 0.966$$

Partial correlation $\rho_{EM.I}$

```
R.partial$estimate["expense", "members"] # from the partial-correlation matrix
model.M1 <- lm(expense ~ income, data = dt)
model.M2 <- lm(members ~ income, data = dt)
cor(residuals(model.M1), residuals(model.M2)) # by definition, as correlation betwee
sqrt(( summary(model.M)$r.squared - summary(model.M1)$r.squared ) /
  ( 1 - summary(model.M1)$r.squared )) # or, using relationship with R squared
[1] 0.5174525
```

$$\rho_{EM.I} = 0.517$$

Summary

- Pearson's correlation: definition, calculation, test of significance, interpretation
- Spearman's and Kendall's rank correlation: definition, calculation, test, interpretation, concordant and discordant pairs
- Correlation matrix, correlogram, scatterplot
- Multiple linear regression: model, best linear approximation, geometric interpretation
- Coefficients of multiple and partial correlation: definition, calculation, interpretation, relation to R-squared