

# Statistics II | seminar 1

Work with data frames, descriptive statistics,  
linear regression model, t-test

**Markéta Zoubková, Ondřej Pokora**

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

9. 9. 2022

### Example (1)

1. Load data from file `people.csv` into a table. Notice column separator and encoding (UTF-8). Work with this table for the rest of the seminar.
2. Explore data types of variables in the table.
3. Check the data types and change them if needed. Data type `factor` is used for categorical variables and `numeric` or `integer` for numerical variables. Correct data types or/and values of variables *Weight*, *Sex* and *BMI*.
4. Recall exploratory data analysis (EDA). Numerical variables: compute sample means, standard deviations, variances and medians, draw boxplots and histograms. Categorical variables: compute frequency tables, draw frequency bar charts.
5. Draw scatter plots of Height (x), Weight (y) and Height (x), BMI (y) with coloring based on the variable *Sex*.
6. Variable *BMI* shows some missing (NA) or mistaken observations. Find and return these rows of the table. Using the BMI formula and variables *Weight* and *Height* add new variable *BMI.new* with correctly computed values.

### Example (2)

7. Use linear regression model with regression line to analyze the dependence of *Weight* on *Height* and *Sex*. Which regression coefficients are statistically significant? (Use significance level  $\alpha = 0,05$ .) Interpret the meaning of each coefficient. Data and model represent graphically, including 95% prediction intervals.
8. Recall (two-sample) t-test. Test a hypothesis about differences in mean BMI of men and women. Formulate null and alternative hypothesis and decide the result of the test.
9. Test previous hypothesis for another binary variables. E.g., test a hypothesis about differences in mean BMI when grouped by *iPhoneOwner* variable and explain the result.

# Statistics II | seminar 2

## One-way analysis of variance (ANOVA)

**Markéta Zoubková, Ondřej Pokora**

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

19. 9. 2022

### Example (1)

Total weight of potatoes grown from one plant was measured for four potato varieties (denoted 1–4).

Variety	Weight (in 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		

Data file [potatoes.csv](#): At a significance level of 5 %, test that the expected value of weight is not distinct across the varieties. Count effects and arithmetic means of the potato varieties. Recall ANOVA table and the meaning of the variables in it. Determine which pairs of varieties differ in case of rejecting the null hypothesis. Use both Tukey and Scheffe test.

**Overall:** Remember to use descriptive statistics and visualisation to get to know your data. Make sure that all of the conditions related to the chosen testing method are satisfied.

### Example (2)

Data file [salesman.csv](#) contains monthly sales (in thousands Kč) for three salesmen in half-year period.

At a significance level of 5 %, test null hypothesis that the expected values of the monthly sales for all salesmen are equal. Determine pairs of salesmen exhibiting different sales in case of rejecting the null hypothesis.

### Example (3)

Data file [coagulation.csv](#): Clinical study observed relationship between a diet and time needed for blood coagulation. The research was performed on 24 animals fed with four different diets (A–D).

At a significance level of 0.05, test null hypothesis that the expected values of the coagulation time for all diets are equal. In case of rejecting the null hypothesis, determine pairs of diets that have significantly different coagulation time.

### — Example (4, manual calculation, R only as calculator and quantile table)

It is given  $a = 5$  independent random samples of sizes  $n_i = 5; 7; 6; 8; 5$ , where  $i$ -th sample is of a distribution  $N(\mu_i, \sigma^2)$ ,  $i = 1, \dots, 5$ .

The total sum of squares  $S_T = 15$  and the sum of squares of the residual error  $S_e = 3$  are known.

Compute the treatment sum of squares  $S_A$  and at a significance level of 5 %, test hypothesis that expected values are equal in all five considered samples.

### — Example (5) —

Three species of mouse were tested on aggression based on their behavior in a maze. At the start each mouse was placed to the center of the square maze with sides equal to 1 m and divided to 49 identical squares. While the mouse was trying to escape the maze for 5 minutes, the number of crossed squares was observed.

Data file [mice.csv](#) contains the number of crossed squares for each species of mouse noted in separated columns. There are some missing values (NA) for the third species.

At a significance level of 0,05, test hypothesis that the expected values of the number of crossed squares for all three species of mouse are equal. Determine which pairs of species significantly differ in case of rejecting the null hypothesis.

### Example (6)

Consider linear regression models

$$M_1 : y = \beta_0 + \beta_1 x,$$

$$M_2 : y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

$$M_3 : y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

for values from file [data01.csv](#). Using the analysis of variance compare these models and select the most suitable one for the data.

### Example (7)

Annual depth data (ft) for lake Huron between the years 1875 and 1972 are stored in variable `LakeHuron` in *R* environment. Using the linear regression model fit the polynomial function of degree 8 to the data. Then using the analysis of variance look into the possibility of reducing the degree of the polynomial.



### Example (8)

Sugar beet yield (in hundredweight per hectare) was monitored for 126 factories in Czechia in relation to the used amount of fertilizer  $K_2O$  (in kilograms per hectare).

Interval data are in file [sugarbeet.csv](#) in 4 columns:

1. lower limit of an interval  $K_2O$  (kg/ha),
2. upper limit of an interval  $K_2O$  (kg/ha),
3. number of factories that use  $K_2O$  in given interval,
4. average yield for sugar beet (q/ha).

Consider linear regression models

$$M_1 : y = \beta_0 + \beta_1 x,$$

$$M_2 : y = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Choose the midpoint of the interval for the values of  $x$ .

Using the analysis of variance select the most suitable model for the data.

## Results

1. ANOVA rejects  $H_0$ . Distinct pairs: Tukey: 2-1, 3-1. Scheffe: 3-1.
2. ANOVA rejects  $H_0$ . Distinct pairs: 3-1, 3-2.
3. ANOVA rejects  $H_0$ . Distinct pairs: A-B, A-C, B-D, C-D.
4. ANOVA rejects  $H_0$ .  $S_A = 12$ ,  $F_A = 26$ .
5. ANOVA rejects  $H_0$ . Distinct pairs: all.
6.  $M_2$
7. Polynomial function of degree 7.
8.  $M_2$

# Statistics II | seminar 3

## Two-way analysis of variance (ANOVA)

**Markéta Zoubková, Ondřej Pokora**

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

26 September 2022

### Example (1)

Data file [hay.csv](#):

Analyze the hay yields (in tons per hectare) in relation to the soil type (neutral / acidic) and fertilizer (none / dung / calcium fertilizer). Each combination was realized four times and independently of each other. Perform two-way analysis of variance with and without interactions and one-way analysis of variance. Test null hypothesis at a significance level of 0.05:

- ▶ soil type and fertilizer are independent, i.e. they do not interact.
- ▶ soil type has no effect on yields;
- ▶ fertilizer has no effect on yields;

Determine distinct groups in case of rejecting the null hypothesis. Which factor causes the difference in expected value of the yields?

### Example (2)

Work with the tabel in variable `ToothGrowth` (in *R* environment) that contains the results of the study on the effect of the vitamin C on guinea-pig's teeth growth:

`len` = length of odontoblasts (cells of the outer surface of the tooth);

`supp` = administration of vitamin C (`OJ` = in orange juice, `VC` = in ascorbic acid);

`dose` = dose of vitamin C (3 groups: 0.5; 1.0 a 2.0 mg/day).

Perform two-way analysis of variance with interactions in relation to the administration and dose of vitamin C. Test corresponding null hypotheses at a significance level of 0.05. Determine distinct groups in case of rejecting the null hypothesis that the expected values of odontoblasts length are equal. Which factors and their interaction cause the difference in expected value of odontoblasts length?

### Example (3)

Data file `goods.csv`:

Certain goods sales were observed over the same period of time in relation to two factors. The goods were sold packed in a bag or box and with the support of advertising campaign only in the press or in the press and also on television or without any advertising at all. File contains profit data in millions Kč from the goods sales under the stated conditions. Perform two-way analysis of variance with and without the interactions. Test null hypotheses at a significance level of 0.05:

- ▶ there is no interaction between the type of the packaging and the type of the advertisement;
- ▶ expected values of the profit do not depend on the type of the packaging;
- ▶ expected values of the profit do not depend on the type of the advertisement.

Determine distinct groups in case of rejecting the null hypothesis that the expected values are equal. Which factor causes the difference in expected value of the profit?

### Example (4)

Data file [corn.csv](#):

When determining the yield of corn, measurements were made on three different types of seeds and five different methods of fertilization. Two measurements were made for each combination. Perform two-way analysis of variance with and without interactions.. Test null hypotheses at a significance level of 0.05:

- ▶ there is no interaction between the type of seed and fertilization method;
- ▶ expected values of the yields are equal for all types of seeds;
- ▶ expected values of the yields are equal for all fertilization methods.

Determine distinct groups in case of rejecting the null hypothesis that the expected values are equal. Which factor causes the difference in expected value of the yields?

## Results

- Two-way ANOVA does not reject  $H_0$  for Soil, but it does for Fertilizer, therefore Fertilizer has effect on Yield. Distinct pairs: none-calcium, none-dung.  
Two-way ANOVA with interactions moreover rejects  $H_0$  for interactions. Distinct pairs for Fertilizer: all. Distinct pairs for interactions: 8 (Tukey), 7 (Scheffe).
- Two-way ANOVA rejects  $H_0$  for both factors. Distinct pairs: all.  
Two-way ANOVA with interactions moreover rejects  $H_0$  for interactions. Distinct pairs for interactions: 11.
- Two-way ANOVA rejects  $H_0$  for both factors. Distinct pairs for Advert: PressTV-None, PressTV-Press.  
Two-way ANOVA with interactions cannot be performed due to the low number of observations.
- Two-way ANOVA rejects  $H_0$  for both factors. Distinct pairs for Variety: C-A, C-B. Distinct pairs for Fertilizer: 2-1, 4-2, 5-2.  
Two-way ANOVA with interactions does not reject  $H_0$  for interactions, therefore two-way ANOVA without interactions is sufficient.



# Statistics II | seminar 4

Rank-based tests and methods

**Markéta Zoubková, Ondřej Pokora**

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

3 October 2022

### Example (1)

Data file `minute.csv`. Ten research participants had to guess independently of each other and without prior training when a minute has passed after the sound signal. Test hypothesis that half of the participants had a period of one minute underestimated and the second half had it overestimated. Check functions `sort` and `rank`. Perform the sign test and one-sample Wilcoxon signed-rank test.

### Example (2)

Data file `field.csv`. Two fertilization methods were tested on 13 fields of the same soil quality. The first method  $A$  was tested on 8 fields and the second method  $B$  was tested on the remaining fields. Data file contains wheat yields (in tons per hectare) of tested fields. Determine if the fertilization method has effect on the wheat yields. Perform two-sample Wilcoxon rank-sum test (Mann-Whitney test).

### Example (3)

Use data file `potatoes.csv` from the previous seminar. Test hypothesis using the Kruskal-Wallis and the median test that the expected value of weight of the potatoes grown from one plant does not depend on the variety. Furthermore, determine significantly different pairs of varieties.

#### Example (4)

Data file [octane.csv](#). Octane rating was observed for 10 samples of gasoline. Test  $H_0 : \tilde{x} = 98$  using the sign test and the Wilcoxon signed-rank test.

#### Example (5)

Data file [bloodpressure.csv](#). Systolic blood pressure was measured for 8 randomly chosen people before and after a medical procedure. Test hypothesis that the medians of the systolic blood pressure before and after the medical procedure are equal.

#### Example (6)

Data file [machines.csv](#). There are 6 machines from 3 different manufacturers in the factory. Perform the Kruskal-Wallis test for the hypothesis that the expected value of machine efficiency does not depend on the manufacturer.

#### Example (7)

Data file [activesubstance.csv](#). The amount of active substance was observed for products of two suppliers. Test hypothesis that the medians of the amount of active substance are equal for both suppliers.

### Example (8)

Data file [minute2.csv](#). Thirty research participants had to guess independently of each other and without prior training when a minute has passed after the sound signal. Data file contains measured times in seconds. Test hypothesis that half of the participants had a period of one minute underestimated and the second half had it overestimated. Use the sign test and one-sample Wilcoxon signed-rank test with one-sided and two-sided  $H_1$ .

### Example (9)

Data file [nickel.csv](#). Four laborants measured the containment of nickel in steel. Perform non-parametric one-way analysis using the Kruskal-Wallis test and the median test. Determine the pairs of laborants whose results significantly differ.

### Example (10)

Data file [potatoes2.csv](#). Potato yields of 4 potato varieties were observed. The yields in tons per hectare are recorded in the file. Perform the Kruskal-Wallis test and the median test. Determine the pairs of varieties whose expected yields significantly differ.

### — Example (11) —

Data file [paper.csv](#). Four labs measured the smoothness of the paper. Perform non-parametric one-way analysis using the Kruskal-Wallis test and the median test. Determine the pairs of labs whose results significantly differ.

### — Example (12) —

Data file [IQvitaminB.csv](#). The effect of vitamin B on IQ was observed in 24 pairs of children. One child out of a pair was administered the vitamin B and the other one was administered the placebo. Compute the difference between the IQ in pairs. Test hypothesis that the medians of the differences in IQ between both groups of children are equal. (Try also one-sided alternatives.)

### — Example (13) —

Data file [rats.csv](#). One of the methods of wound sealing was chosen after surgery for twenty rats – the stitches or a compression bandage. Data on tension in the wound edges were collected. Test hypothesis that the medians of tension for both methods of wound sealing are equal.

### — Example (14) —

Data file [sales.csv](#): numbers of sold pieces of 5 different products in 7 different stores. Use Friedman test to test the hypothesis that the product type has no effect on the sales in particular stores (blocks).

## Results

1. The sign test does not reject  $H_0$  and the Wilcoxon signed-rank test does.
2. Wilcoxon rank-sum test rejects  $H_0$ .
3. Kruskal-Wallis test rejects  $H_0$  and median test does not. Distinct pairs of varieties: 3-4, 3-1, 2-1, 4-1.
4. Both test do not reject  $H_0$ .
5. Paired Wilcoxon signed-rank test does not reject  $H_0$ .
6.  $H_0$  that medians are equal is not rejected.
7. Wilcoxon rank-sum test does not reject  $H_0$ .
8. Both tests reject  $H_0$ . After repeating the test for one-sided  $H_1$  we show that median is  $< 60$ .
9. Both methods reject  $H_0$ . Distinct pairs: D-A, D-C, B-C, (B-A).
10. Both methods reject  $H_0$ . Distinct pairs: B-A, B-D, C-D, A-D, (B-C).
11. Both methods reject  $H_0$ . Distinct pairs: all pairs except B-C, (A-B).
12. Paired Wilcoxon signed-rank test does not reject  $H_0$ .
13. Wilcoxon test does not reject  $H_0$ .
14. Friedman test does not reject  $H_0$ .

# Statistics II | seminar 5

## Goodness-of-fit tests

**Markéta Zoubková, Ondřej Pokora**

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

10 October 2022

### Example (1)

Data file [families.csv](#). It was randomly selected 84 families from a set of families of 5 children and the number of boys was detected for each family. 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)

Data file [line.csv](#). 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). At a significance level of 0.05, test hypothesis that waiting time has exponential distribution using Pearson's chi-squared test and test for exponential distribution.

### Example (3)

Data file [trains.csv](#). The number of arriving trains during 1 hour was observed at the station for 5 days (i.e. 360 hours). At a significance level of 0.05, test hypothesis that the number of arriving trains during 1 hour follows Poisson distribution using Pearson's chi-squared test and test for Poisson distribution.



#### Example (4)

Data file [Brno.csv](#) contains a number of residents of Brno (data from the year 2001) sorted by birth month. At a significance level of 0.05, test hypothesis that the number of births in each month corresponds to the number of days in the month (consider a non-leap year).

#### Example (5)

Data file [dice.csv](#). The absolute frequencies of the numbers thrown for 60 dice rolls were observed. At a significance level of 0.05, test hypothesis that the dice is homogeneous (fair).

#### Example (6)

Data file [emergency.csv](#) contains a number of patients in the emergency room during 8 hour shift in a total of 75 days. Test hypothesis that the number of patients has Poisson distribution using Pearson's chi-squared test and test for Poisson distribution.

### Example (7)

Data file [dice2.csv](#). The number of sixes getting in 12 dice was observed for 4096 rolls with 12 dice. At a significance level of 0.05, test hypothesis that the number of sixes in one roll has binomial distribution  $\text{Bi}(12; \frac{1}{6})$ .

### Example (8)

Data file [pond.csv](#). Five traps was set into a pond glowing with white, yellow, blue, green and red light. The number of fish caught in each trap is contained in the data file. At a significance level of 0.05, test hypothesis that the color of light has no effect on the number of fish caught.

### Example (9)

Data file [supermarkets.csv](#). 300 customers of a supermarket chain were questioned which day of the week they shop the most. Data file contains these numbers. At a significance level of 0.05, test hypothesis that the customers shop evenly every day of the week.

### Example (10)

Data file [football.csv](#). Total number of goals scored in a total match time was observed for 84 football matches. Data file contains these numbers. At a significance level of 0.05, test hypothesis that the number of goals scored has Poisson distribution using Pearson's chi-squared test.

### Example (11)

Variables JANT and JULY in data file [pollution.csv](#) contain January and July temperatures in °F. Convert them to °C and use the Lilliefors test (also the Kolmogorov-Smirnov test) and Pearson's chi-squared test at a significance level of 0.05 to test that these variables have normal distribution. Plot an empirical and a theoretical distribution functions.

### Example (12)

Variable Rainfall in data file [Hawaii.csv](#) contains amount of precipitation in Hawaii. Use the Kolmogorov-Smirnov test and Pearson's chi-squared test at a significance level of 0.05 to test that the amount of precipitation has chi-squared distribution with 18 degrees of freedom. Plot an empirical and a theoretical distribution functions.

## Results

1.  $K = 3.1$ ,  $\chi^2_{0.95}(5) = 11.1$ . Assumptions  $\Rightarrow$  merge outer pairs of categories  $\Rightarrow K = 2.4$ ,  $\chi^2_{0.95}(3) = 7.8$ . We do not reject the null hypothesis.
2.  $K = 4.8$ ,  $\chi^2_{0.95}(6) = 12.6 \Rightarrow$  Pearson's chi-squared test does not reject the null hypothesis.  
 $Q = 35.7 \rightarrow$  Simple test rejects the null hypothesis.
3.  $K = 9.7$ ,  $\chi^2_{0.95}(6) = 12.6$ ;  $Q = 331.1$ . We do not reject the null hypothesis.
4.  $K = 1506.2$ ,  $\chi^2_{0.95}(11) = 19.7$ . We reject the null hypothesis, month of birth distribution is not uniform distribution.
5.  $K = 2.8$ ,  $\chi^2_{0.95}(5) = 11.1$ . We do not reject the null hypothesis.
6. Assumptions  $\Rightarrow$  merge the last 2 categories  $\Rightarrow K = 8.7$ ,  $\chi^2_{0.95}(8) = 15.5$ ;  $Q = 1158.6 \rightarrow$  Both tests do not reject the null hypothesis.
7.  $K = 5.5$ ,  $\chi^2_{0.95}(7) = 14.1$ . We do not reject the null hypothesis.
8.  $K = 14.1$ ,  $\chi^2_{0.95}(4) = 9.5$ . We reject the null hypothesis, color has effect on the number of fish caught.
9.  $K = 78$ ,  $\chi^2_{0.95}(6) = 12.6$ . We reject the null hypothesis, preferences of the shopping days are not uniform.
10.  $K = 2.1$ ,  $\chi^2_{0.95}(3) = 7.8$ . We do not reject the null hypothesis.

# Statistics II | seminar 6

## Correlation analysis

**Markéta Zoubková, Ondřej Pokora**

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

17. 10. 2022

Perform correlation analysis using the Pearson correlation coefficients, partial correlation coefficients, their significance tests and visualizations (scatterplots and correlograms). Interpret the results. Calculate also Spearman's and Kendall rank correlation coefficients.

### — Example (1) —

Data file [households.csv](#). The following variables were observed in 7 households: number of members, income and food and drink expenses (in thousands Kč per 3 months). Calculate sample multiple correlation between expense and remaining variables.

### — Example (2) —

20 children in different ages underwent pedagogical-psychological research within which among others they wrote a dictation. Data file [children.csv](#) contains children's weight, age and points earned in dictation. Calculate sample multiple correlation between Points and remaining variables.

### — Example (3) —

Data file [children2.csv](#). Memory capacity, IQ and speed reading ability were observed for 20 children of different ages. Calculate sample multiple correlation between IQ and remaining variables.

#### Example (4)

Data file [enrollment.csv](#). The number of filed university applications (ROLL) was observed in one of the states of USA for 29 years (YEAR) in relation to the average unemployment rate (UNEM, in %), the average wage (INC, in USD) and the number of students (HGRAD) who successfully graduated from high school in a given year. Focus on the dependence of the number of university applications (the coefficient of multiple correlation) on the rest of the variables. Calculate sample multiple correlation between HGRAD and remaining variables.

#### Example (5)

Data file [heptathlon.csv](#) contains heptathletes' results from a competition. Perform correlation analysis of given variables of results of 7 disciplines and column score. Examine their associations and interpret the results. Calculate sample multiple correlation between score and disciplines' results.

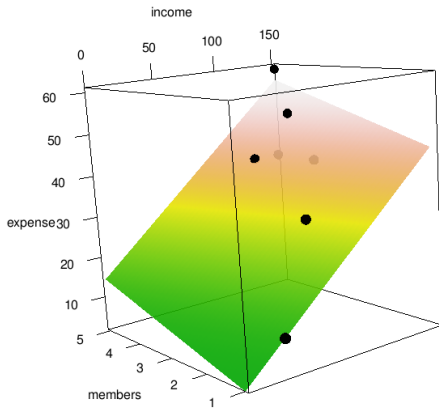
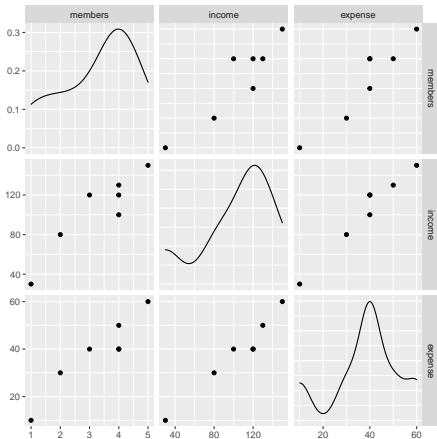
#### Example (6)

Data file [decathlon.csv](#) contains decathletes' results from competitions. Perform correlation analysis of given variables of results of 10 disciplines and column Points. Examine their associations and interpret the results. Calculate sample multiple correlation between Points and disciplines' results.

## Results

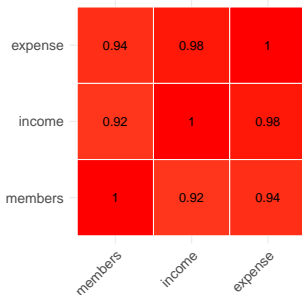
Sample multiple correlations:

**1.** 0.983; **2.** 0.990; **3.** 0.731; **4.** 0.959; **5.** 0.997; **6.** 1.





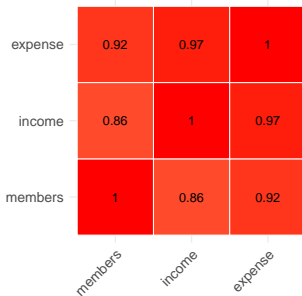
Pearson's correlations



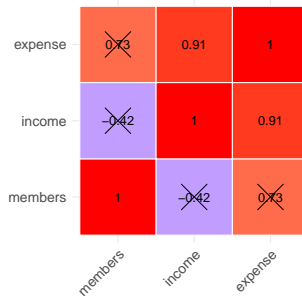
Partial Pearson's correlations

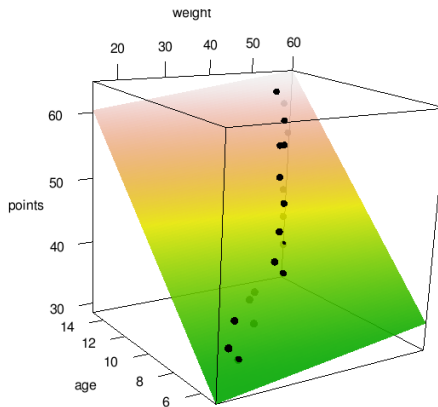
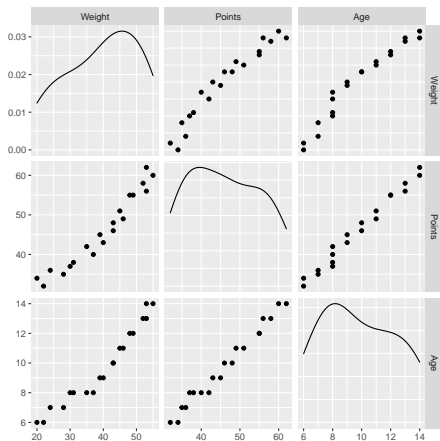


Spearman's correlations

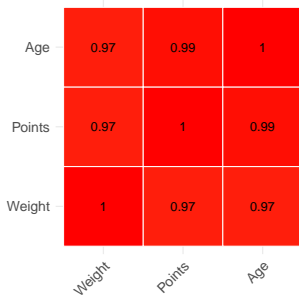


Partial Spearman's correlations

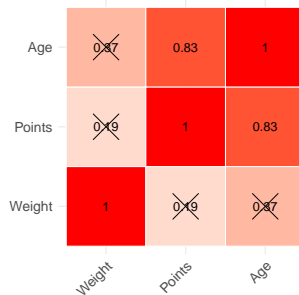




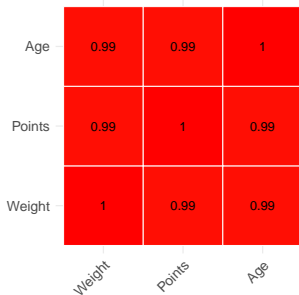
Pearson's correlations



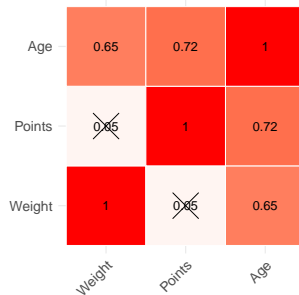
Partial Pearson's correlations

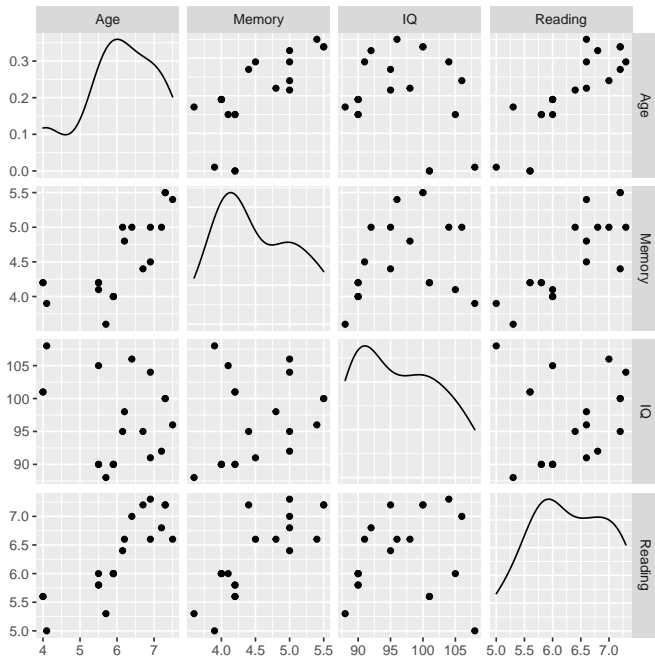


Spearman's correlations

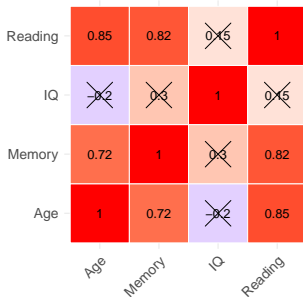


Partial Spearman's correlations

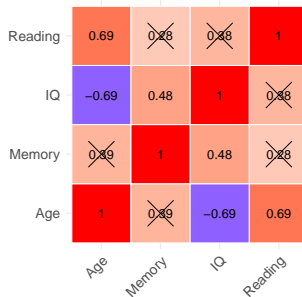




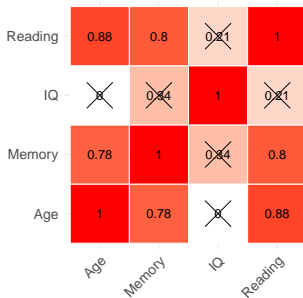
Pearson's correlations



Partial Pearson's correlations

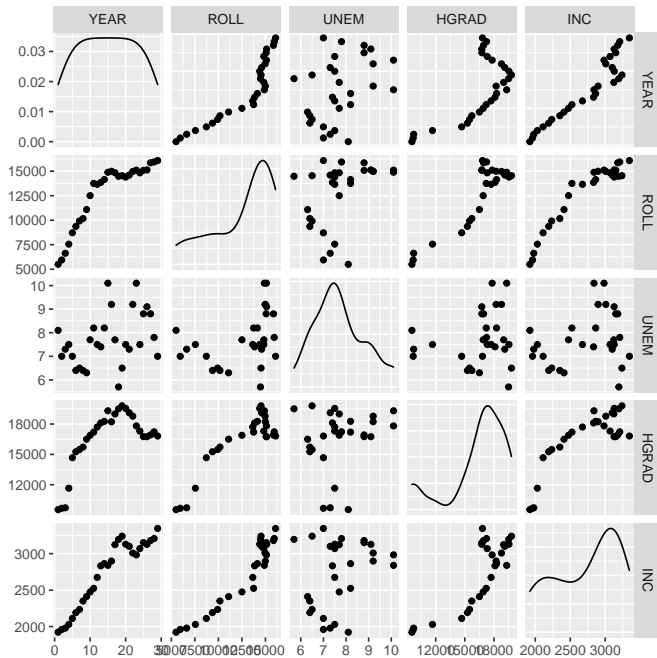


Spearman's correlations



Partial Spearman's correlations





Pearson's correlations

INC	0.94	0.95	<del>0.28</del>	0.82	1
HGRAD	0.67	0.89	<del>0.18</del>	1	0.82
UNEM	0.38	0.39	1	<del>0.18</del>	<del>0.28</del>
ROLL	0.9	1	0.39	0.89	0.95
YEAR	1	0.9	0.38	0.67	0.94
	YEAR	ROLL	UNEM	HGRAD	INC

Partial Pearson's correlations

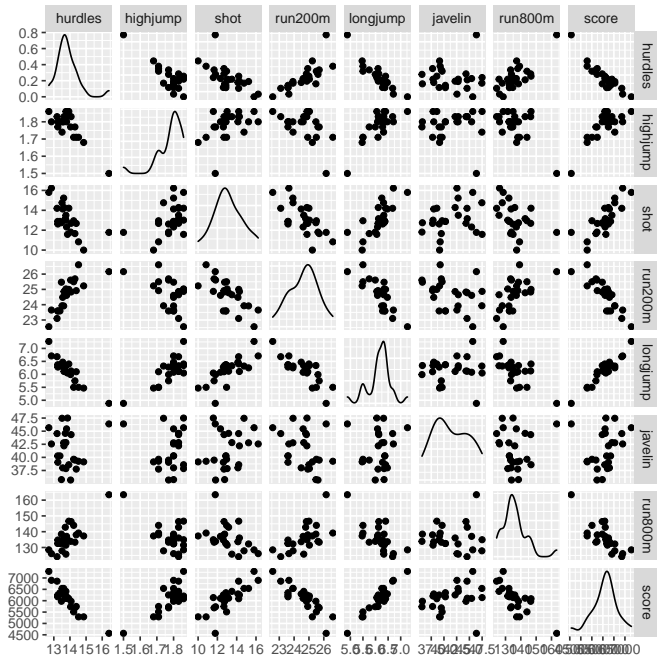
INC	0.69	<del>0.2</del>	<del>-0.31</del>	<del>0.29</del>	1
HGRAD	<del>-0.69</del>	0.8	<del>-0.36</del>	1	<del>0.29</del>
UNEM	<del>0.1</del>	0.52	1	<del>-0.36</del>	<del>-0.31</del>
ROLL	0.5	1	0.52	0.8	<del>0.2</del>
YEAR	1	0.5	<del>0.1</del>	<del>-0.69</del>	0.69
	YEAR	ROLL	UNEM	HGRAD	INC

Spearman's correlations

INC	0.93	0.88	<del>0.2</del>	0.69	1
HGRAD	0.57	0.56	<del>0.24</del>	1	0.69
UNEM	0.39	0.52	1	<del>0.24</del>	<del>0.2</del>
ROLL	0.96	1	0.52	0.56	0.88
YEAR	1	0.96	0.39	0.57	0.93
	YEAR	ROLL	UNEM	HGRAD	INC

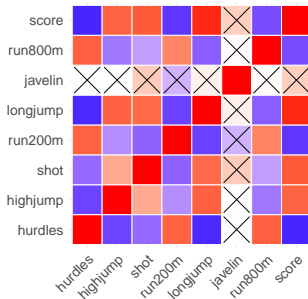
Partial Spearman's correlations

INC	0.61	<del>0.15</del>	<del>-0.58</del>	0.62	1
HGRAD	<del>-0.32</del>	<del>-0.04</del>	<del>0.34</del>	1	0.62
UNEM	<del>0.38</del>	0.53	1	<del>0.34</del>	<del>-0.58</del>
ROLL	0.64	1	0.53	<del>-0.04</del>	<del>0.15</del>
YEAR	1	0.64	<del>0.38</del>	<del>-0.32</del>	0.61
	YEAR	ROLL	UNEM	HGRAD	INC

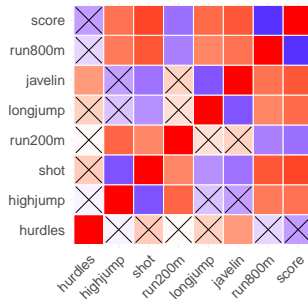




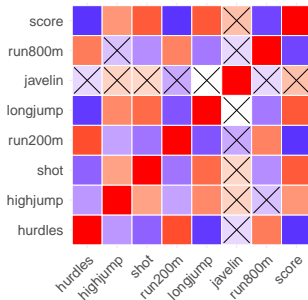
Pearson's correlations



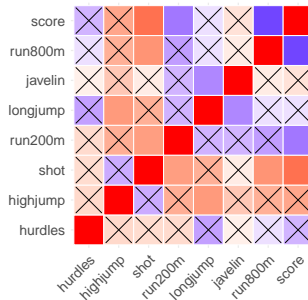
Partial Pearson's correlations

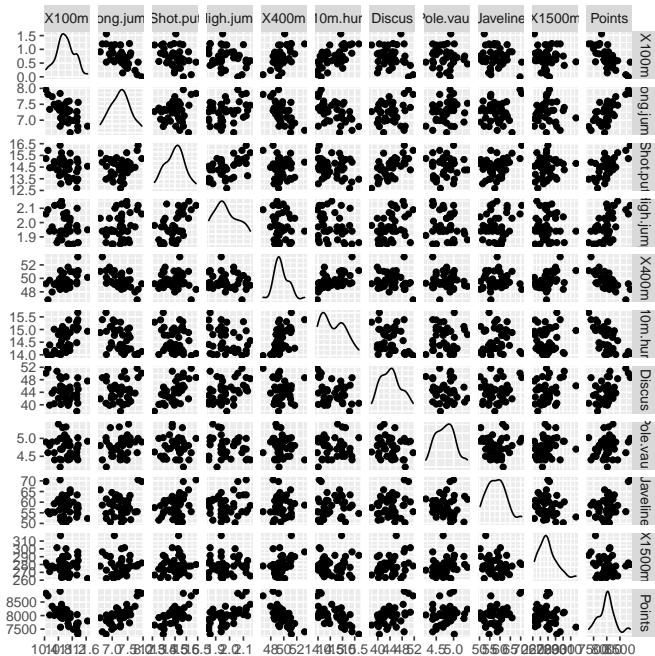


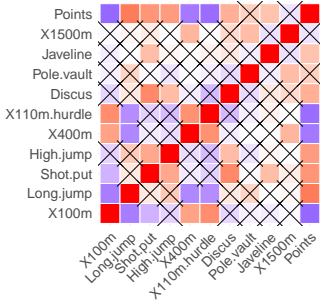
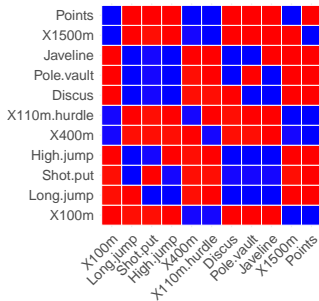
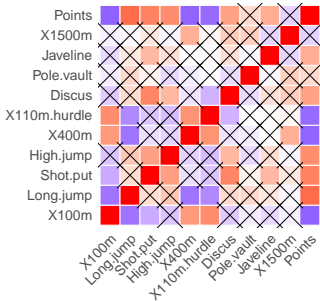
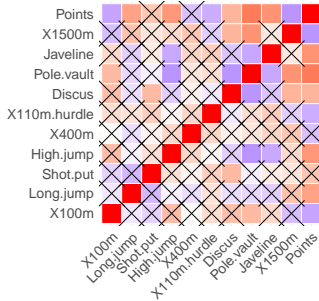
Spearman's correlations



Partial Spearman's correlations





[illegible]

# Statistics II | seminar 7

Autocorrelation, multicollinearity, stepwise regression

**Markéta Zoubková, Ondřej Pokora**

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

24 October 2022

### Example (1)

Data file [Hawaii.csv](#) contains a certain bird's species numbers on Hawaii between years 1956 and 2003. Using the linear regression model with AR(1), model the dependence of the square root of the number of birds on time (in years). Check for autoregression of order 1 in the data and in case of finding it, remove it.

### Example (2)

Data file [expenses.csv](#): data collected from 20 randomly selected households. Columns in the table contains the following variables: food and drink expenses ( $Y$ ), number of members ( $X_1$ ), number of children ( $X_2$ ), average age of workers ( $X_3$ ) a household income ( $X_4$ ). Select model with the most well-conditioned regressors using stepwise regression.

### Example (3)

Data file [water.csv](#): The water losses in distribution networks were determined in the years 1953–1983. Model dependence of the water loss ( $Y$ ) on the produced amount of water ( $x$ ). Check for autoregression of order 1 in the data and in case of finding it, remove it.

#### Example (4)

Data stored in the variable `mtcars` in *R* are for modeling the dependence of passenger car fuel consumption (`mpg`, number of miles per gallons) on the engine characteristics:

---

<code>cyl</code>	number of cylinders
<code>disp</code>	cylinders' volume (cubic inches)
<code>hp</code>	engine power (horsepower)
<code>drat</code>	differential ratio
<code>wt</code>	vehicle weight (in hundreds lbs)
<code>qsec</code>	acceleration (seconds from 0 to 1/4 mile)
<code>vs</code>	cylinder arrangement (1 = 'V'-type, 0 = inline)
<code>am</code>	gearbox (0 = automatic, 1 = manual)
<code>gear</code>	number of gears
<code>carb</code>	number of carburetors

---

Test for multicollinearity in the model. Select a suitable model using stepwise regression. Check for normality of residuals.

### Example (5)

Data file `cement.csv`: chemical composition of Portland cement:

---

<code>y</code>	heat of hydration in calories per 1 gram of cement
<code>x1</code>	<i>tricalcium aluminate</i> $3\text{CaO}.\text{Al}_2\text{O}_3$ in %
<code>x2</code>	<i>tricalcium silicate</i> $3\text{CaO}.\text{SiO}_2$ in %
<code>x3</code>	<i>tetracalcium alumino ferrite</i> $4\text{CaO}.\text{Al}_2\text{O}_3.\text{Fe}_2\text{O}_3$ in %
<code>x4</code>	<i>dicalcium silicate</i> $2\text{CaO}.\text{SiO}_2$ in %

---

Test for multicollinearity in a linear model on dependence of the heat of hydration on content of 4 given cement components. Select a suitable model for `y` using stepwise regression. After that, check for normality of residuals.

### Example (6)

Macroeconomic data of the USA from 1947–1962 are stored in variable `longley` (built-in in *v R*). Using a linear regression model, examine the dependence of employment on GDP and the population older than 14 years: `Employed ~ GNP + Population`. Consider AR(1) in the data.

### Example (Continuation of Examples 4–6 from Seminar 6)

Using stepwise procedure, build suitable linear models in Examples from **Seminar 6**. Model the following target variables using the other variables as predictors:

4. ROLL
5. score
6. Points

### Results

2. Multicollinearity is not rejected, suitable model:  $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3$ , normality of residuals is not rejected.
4. Multicollinearity is not rejected. Suitable models are e.g.  
 $\text{mpg} = \beta_0 + \beta_1 \text{wt} + \beta_2 \text{qsec} + \beta_3 \text{am}$ ,  $\text{mpg} = \beta_0 + \beta_1 \text{wt} + \beta_2 \text{cyl} + \beta_3 \text{hp}$ ,  
 normality of residuals is not rejected.
5. Multicollinearity is not rejected, suitable model:  
 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$ , normality of residuals is not rejected.



# Statistics II | seminar 8

## Principal component analysis

**Markéta Zoubková, Ondřej Pokora**

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

31. 10. 2022

Perform correlation analysis and principal component analysis in the following examples. Draw a correlogram, screeplot, proportion of variance explained and biplot. Identify principal components (first 2).

How many principal components are necessary to consider to explain at least 80 % of variance? What percentage of the data variance is explained by the first 2 principal components?

### — Example (1) —

Data file [indicators.csv](#): 16 economic and financial indicators for a set of 29 European countries from the year 2007 that was published by EUROSTAT in the year 2009.

### — Example (2) —

Data file [pollution.csv](#): USA climate and demographic data from the year 1960. Use the first 15 variables (do not consider variable [MORT](#)).

### — Example (3) —

Data file [pain.csv](#): 121 patients from the year 1991 complaining about a headache, responses are on a scale of 1 to 9. Do not consider variables [isub](#) and [dayslost](#).

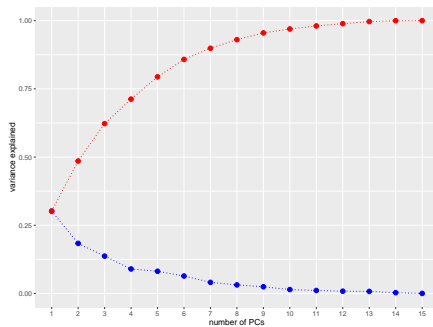
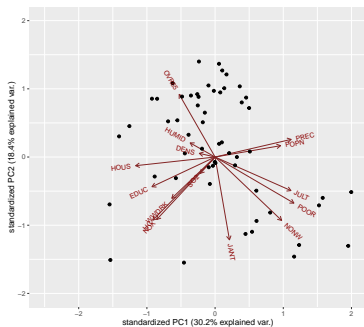
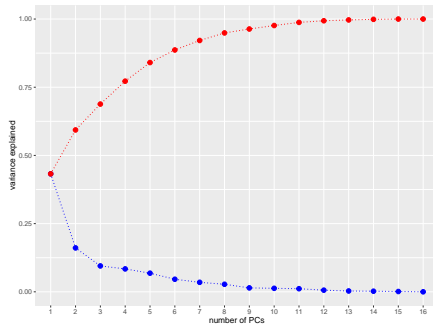
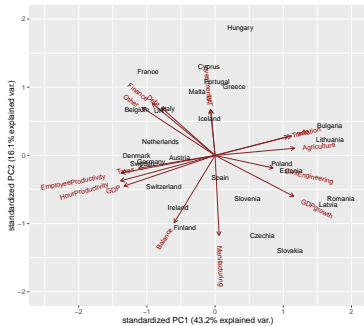
#### Example (4)

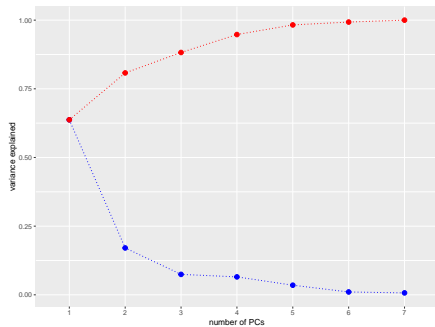
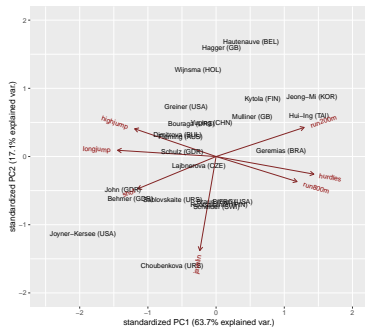
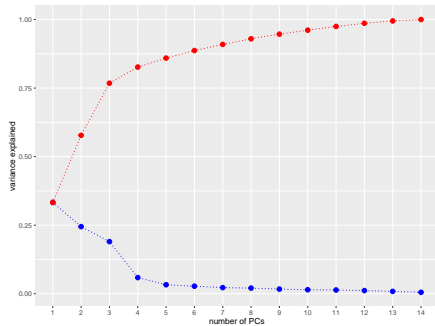
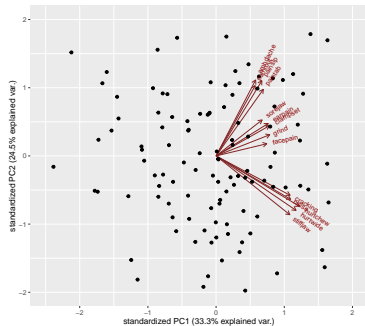
Data file [heptathlon.csv](#): 25 heptathletes' results from 7 disciplines and total score in one athletic competition. Do not consider variable [score](#). Find coefficients for the heptathlete Lajbnerová in the first two principal components.

#### Example (5)

Data file [decathlon.csv](#): results from individual disciplines, ranking and total score for 41 decathletes in one athletic competition. Use only ten variables of the results from individual disciplines. Find coefficients for the decathlete Šebrle in the first two principal components.

# Results







# Statistics II | seminar 9

Logistic regression and other generalized linear models (GLM)

**Markéta Zoubková, Ondřej Pokora**

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

6. 11. 2022

### Example (1)

Data file `beetle.csv`: mortality of the *confused flour beetle* (*Tribolium confusum*) due to exposure to gaseous carbon disulfide  $\text{CS}_2$ . Model the dependence of mortality on the amount of  $\text{CS}_2$  using the (i) logistic regression, (ii) probit link function and (iii) CLogLog link function. Examine statistical significance of variables, plot final regression functions, compare models.

dose	the dose of $\text{CS}_2$ (in mg/l)
population	number of beetles tested
killed	number of beetles killed



### Example (2)

Data file [car\\_income.csv](#): information about a new car purchase during last 12 months in relation to the household income and the age of a previous car.

(i) Plot the dependence of purchase on other variables and find a suitable logistic model. Is every variable statistically significant? (ii) Create a suitable factor from variable age, use it in the model and look at its statistical significance.

---

purchase	indicator of a new car purchase (1 = yes, 0 = no)
income	annual household income (in thousands dollars)
age	age of a previous car (years)

### Example (3)

Data file [heart.csv](#): presence of myocardial infarction in relation to the age of a patient. Model the dependence of chd on age using the (i) logistic regression, (ii) probit link function and (iii) CLogLog link function. A suitable model can be obtained by aggregating the data by (age). Examine statistical significance of variables and plot final regression function.

---

age	age of patient (years)
chd	indicator of myocardial infarction (1 = yes, 0 = no)

### Example (4)

Data file [hospital.csv](#): patient recovery in relation to the infection severity and the hospital. Find a suitable logistic model for variable Treatment\_Outcome.

Infection_Severity	severity of infection
Treatment_Outcome	indicator of recovery (1 = recovered, 0 = death)
Hospital	code of the hospital (1, 2, 3)

### Example (5)

Data file [cancer.csv](#): the number of skin cancer cases in women in relation to the age and the city in the USA where the patients lived. Find a suitable logistic regression model for the variable Cases. Plot the results. Calculate probability of the disease occurrence for 60 years old woman living in Minneapolis and the woman of the same age living in Dallas.

Cases	number of cases
Town	city (0 = Minneapolis, Minnesota, 1 = Dallas, Texas)
Age	age category
Population	total number of women of a given age category in the particular city

[Minneapolis: 0.00117, Dallas: 0.00276.]

### Example (6)

Data file [sharks.csv](#): the number of shark attacks in Florida between years 1946 and 1999. There are the following variables:

Year	year
Population	population size
Attacks	number of shark attacks
Fatalities	number of deaths by shark attack

Start with scatterplot of number of shark attacks per 1 million people depending on the time. Model number of shark attacks – use binomial and poissonian model with canonical link function. Consider a cubic polynomial for the design matrix for variable Year.

Draw predictions for both models with 95% confidence bands for regression function. Examine the problem with too high/low variance and redefine the model if necessary. Plot the results. Estimate the number of shark attacks (per 1 million people) in 2013 using the final model including 95% confidence interval.

### Example (7)

Data file `bees.csv`. One of the important characteristics when investigating bee activity is the number of bees that leave the hive to work in the outside environment. The study dealt with the measurement of this variable during several sunny days depending on the time of day. Data set contains the following variables:

<code>number</code>	number of bees that left the hive
<code>time</code>	time when the data was collected

Model the dependence of the number of bees that leave the hive on the time of day using the poissonian model with canonical link function. Consider a quadratic polynomial for the design matrix for variable `time`. Look for the *overdispersion* and alter the model if necessary.

### Example (8)

Data file [species.csv](#): 90 pastures of size  $25 \times 25$  m were observed differing in biomass, soil pH and species richness (the number of plant species). It is known that species richness decreases with increasing biomass. The question remained whether the rate of decline is related to the soil pH. Therefore, each pasture was classified according to the pH value of soil to the three levels (low, mid and high) and 30 pastures for each level were selected for the experiment. Variable Biomass is the long-term average of June measurements of biomass values. Data set contains the following variables:

pH	pH value of soil (low, mid, high)
Biomass	amount of biomass
Species	number of plant species

Plot the dependence of variable species on other variables. Find suitable poissonian model with logarithmic, identity and square root link function. Which variables are statistically significant? Use non-significance to simplify the model. Plot final models. Estimate the number of plant species for Biomass= 9 and pH=mid using all three models.

### Example (9)

Data file [henharrier.csv](#): predator *hen harrier* (*Circus cyaneus*) hunting for a capercaillie depending on a capercaillie population. Let  $Y_i$  denote the proportion of capercaillie consumed and  $x_i$  the population of capercaillie in the given area. Theories dealing with the behavior of these predators suggest using this formula for modeling

$$E(Y_i) = \mu_i = \frac{\alpha x_i^3}{\delta + x_i^3},$$

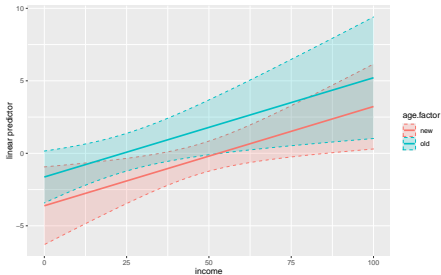
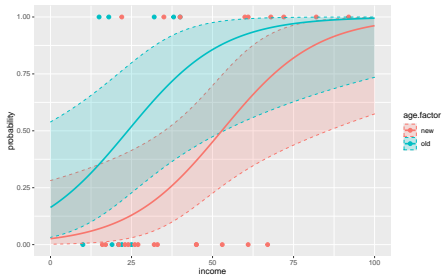
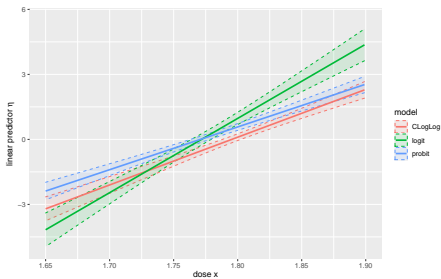
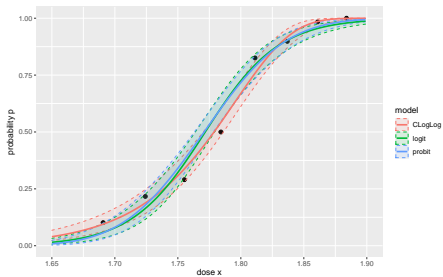
where  $Y_i$  is Gamma-distributed. It is necessary to estimate unknown parameters  $\alpha$  and  $\delta$ . Using the link function inverse, we get

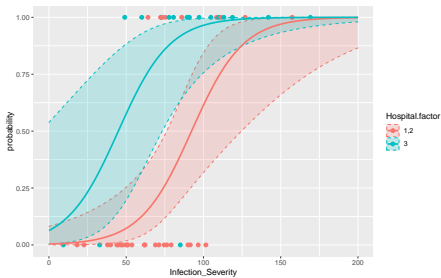
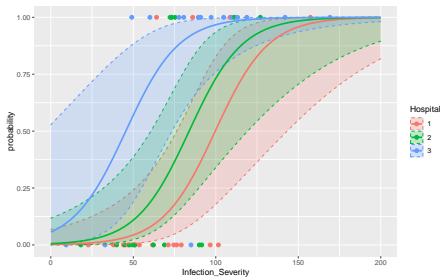
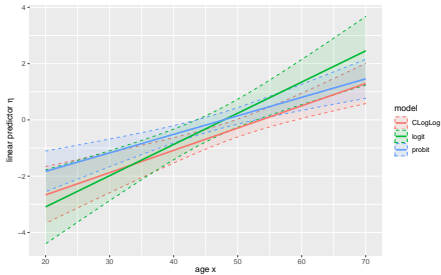
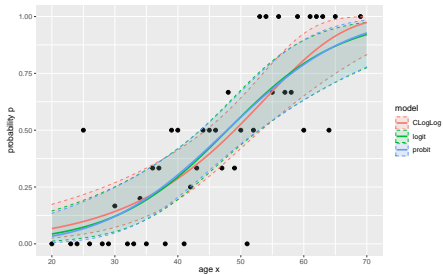
$$\frac{1}{\mu_i} = \frac{1}{\alpha} + \frac{\delta}{\alpha x_i^3}.$$

By defining the parameters  $\beta_0 = 1/\alpha$  and  $\beta_1 = \delta/\alpha$ , we get a linear relationship

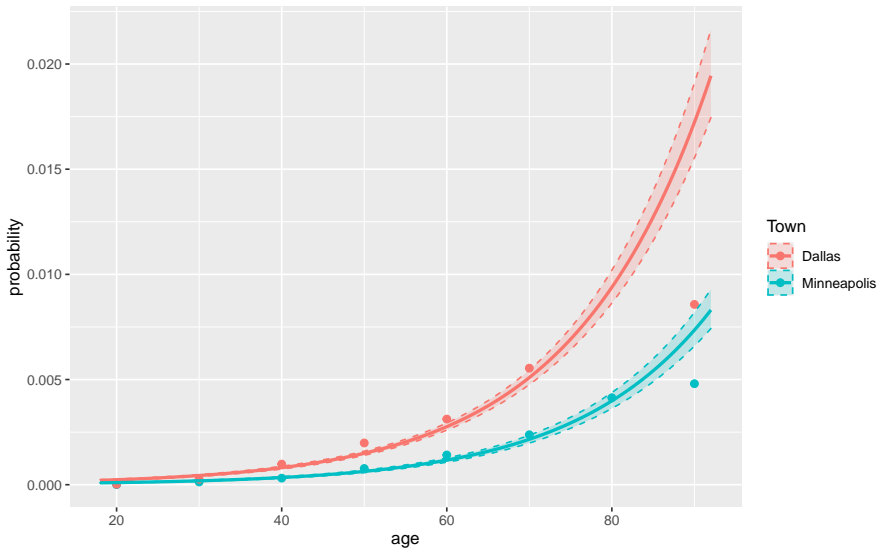
$$\frac{1}{\mu_i} = \beta_0 + \beta_1 \frac{1}{x_i^3}.$$

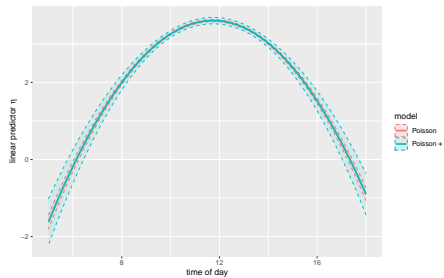
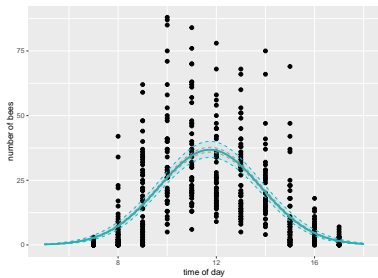
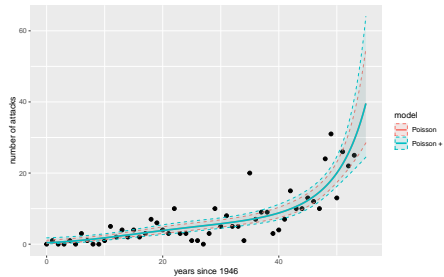
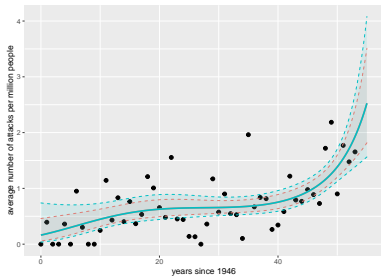
# Results

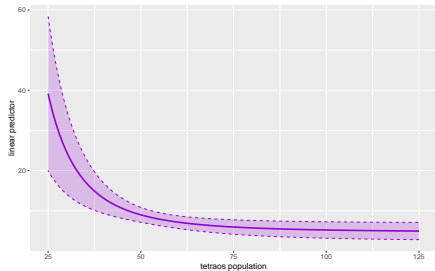
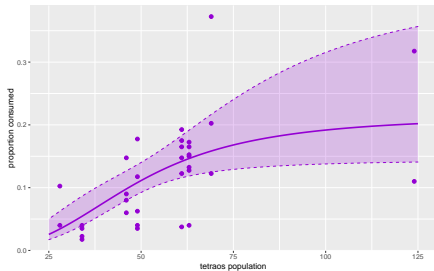
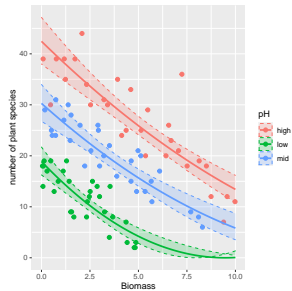
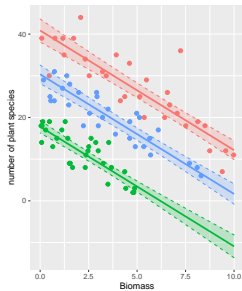
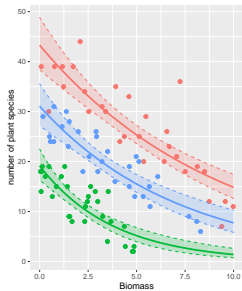












2. *Hint:*  $\text{age} \leq 3$  and  $\text{age} > 3$
6. *Overdispersion.* Estimation: 33.96 attacks per 1 million people, confidence interval:  $[3.207; 359.55]$ .
7. Residual deviance 4879.3 is disproportionately higher than the number of degrees of freedom 501  $\Rightarrow$  *overdispersion*.
8. The number of species estimates for logarithmic link: 8.895; identical link: 4.513; square root link: 7.414.

# Statistika II | seminar 11

Contingency tables, tests of independence

**Markéta Zoubková, Ondřej Pokora**

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

21. 11. 2022

### Example (1)

A random sample of 360 study applicants was collected in sociological survey. They examined, among other things, social groups the applicants comes from and types of universities they apply to. The results are summarized in the contingency table:

Type of university	Social group				$n_{i.}$
	I	II	III	IV	
academic	50	30	10	50	140
technical	30	50	20	10	110
economic	10	20	30	50	110
$n_{.j}$	90	100	60	110	360

At a significance level of 0.05, test the hypothesis that type of university and social group are independent. Compute Cramér's V.

### Example (2)

Let  $(X_1, Y_1), \dots, (X_{16}, Y_{16})$  be random sample from a two-dimensional normal distribution. The Pearson correlation coefficient  $r_{XY}$  is  $-0.87$ . If we make the transformation  $U_i = 1 + 3X_i$ ,  $V_i = -3 - Y_i$ ,  $i = 1, \dots, 16$ , what will the value of  $r_{UV}$  be?

### Example (3)

admitted	impression		$n_{i.}$
	good	bad	
yes	17	11	28
no	39	58	97
$n_{.j}$	56	69	125

125 study applicants were judged by the impression they made on selection board in the oral entrance exam. At a significance level of 0.05, test the hypothesis that admission to the studied university is independent of the impression made in the entrance exam.

### Example (4)

Two doctors assessed the condition of seven patients after the same surgery. They proceeded in such a way that the highest number was given to the worst case.

Patient number	1	2	3	4	5	6	7
Doctor 1	4	1	6	5	3	2	7
Doctor 2	4	2	5	6	1	3	7

Compute the sample Spearman correlation coefficient  $r_s$  and at a significance level of 0.05, test hypothesis that the assessed conditions by doctors are independent.

### Example (5)

We have test results from two subjects obtained from eight randomly selected students of a certain field.

Student number	1	2	3	4	5	6	7	8
Test score 1	80	50	36	58	42	60	56	68
Test score 2	65	60	35	39	48	44	48	61

At a significance level of 0.05, test hypothesis that the test results are not positively correlated.

### Example (6)

Iron content in 600 ore samples was determined by two analytical methods with sample correlation coefficient 0.85. The literature reports that the correlation coefficient of these two methods should be 0.9. At a significance level of 0.05, test null hypothesis  $H_0: \rho = 0,9$  against two-sided alternative  $H_1: \rho \neq 0,9$ .



### Example (7)

The concentration of substances A and B in the urine of patients suffering from a certain kidney disease was observed as part of medical research. The sample correlation coefficient between concentration of both substances was 0.65 for 100 healthy individuals and 0.37 for 142 patients suffering from mentioned kidney disease. At a significance level of 0.05, test null hypothesis that correlation coefficients are the same for both groups.

### Example (8)

An employee of the human resources department of a certain company investigates whether there is a relationship between the number of absent days per year ( $Y$ ) and the age of the worker ( $X$ ). Therefore he randomly selected data on 10 workers.

ID	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Assuming that the given data are realizations of a random sample of size 10 from a two-dimensional normal distribution, compute the sample correlation coefficient  $r$  and test hypothesis that  $X$  and  $Y$  are independent random variables at a significance level of 0.05. Compute 95% asymptotic confidence interval for the true correlation coefficient  $\rho$ .

### Example (9)

Perform a test of independence between academic rank and sex at a significance level of 0.05 and compute Cramér's V if the following data on the number of cases are available:

sex	academic ranks		
	assistant professor	associate professor	professor
male	32	15	8
female	34	8	3

### Example (10)

12 different software companies offer accounting programs. The programs were assessed by an expert committee and a committee composed of professional accountants.

experts	6	7	1	8	4	2,5	9	12	10	2,5	5	11
accountants	4	5	2	10	6	1	7	11	8	3	12	9

Compute the sample Spearman rank correlation coefficient and test hypothesis that ranks assessed by both committees are independent at a significance level of 0.05.

**Example (11)**

There are 15 workers in workshop. The number of shifts worked per month ( $X$ ) and the number of manufactured products ( $Y$ ) were determined.

$X$	20	21	18	17	20	18	19	21	20	14	16	19	21	15	15
$Y$	92	93	83	80	91	85	82	98	90	60	73	86	96	64	81

Roughly verify the 2-dimensional normality in the data. Compute the sample correlation coefficient between  $X$  and  $Y$ , construct 99% asymptotic confidence interval for it and test hypothesis that  $X$  and  $Y$  are independent at a significance level of 0.05.

**Example (12)**

400 randomly selected employees of a food company were asked about the causes of dissatisfaction at the workplace.

worker category	the main cause of dissatisfaction				
	working environment	workplace relationships	organization of work	wage	other
laborers	80	50	75	40	55
THP	10	10	25	30	25

At a significance level of 0.05, test hypothesis that the cause of dissatisfaction is independent of worker category. Compute Cramér's  $V$ .

## Results

1.  $K = 76.84$ ,  $V = 0.3267$ . We reject the null hypothesis that the type of university and the social group are independent, because  $K > 12.6 = \chi^2_{0.95}(6)$ .
2.  $r_{U,V} = 0.87$ .
3.  $OR = 2.298$ ,  $\ln OR = 0.832$ ,  $D = 1.895$ . We do not reject the null hypothesis that the admission and the impression are independent, because  $D < 1.96 = u_{0.975}$ .
4.  $r_S = 0.857$ , critical value:  $r_{0.975}(7) = 0.745$ . We reject the null hypothesis that the ranks are independent, because  $0.857 \geq 0.745$ .
5.  $r_{X,Y} = 0.6668$ ,  $T = 2.1917$ . We reject the hypothesis that there's not positive correlation between test results, because  $T > 1.9432 = t_{0.95}(6)$ .
6.  $Z = 1.2562$ ,  $U = -5.2976$ . We reject  $H_0$ , because  $|U| > 1.96 = u_{0.975}$ .
7.  $Z_A = 0.7753$ ,  $Z_B = 0.3884$ ,  $U = 2.9242$ . We reject  $H_0: \rho_A = \rho_B$ , because  $|U| > 1.96 = u_{0.975}$ .
8.  $r_{X,Y} = -0.9325$ ,  $T = -7.3053$ . We reject  $H_0$ , because  $|T| > 2.306 = t_{0.975}(8)$ . 95% CI for  $\rho$ :  $[-0.9842, -0.7336]$ .
9. We do not reject the null hypothesis that the sex and the academic rank are independent.  $V = 0.187$ .
10.  $r_S = 0.715$ . We reject the null hypothesis.
11.  $r_{X,Y} = 0.927$ . We reject the null hypothesis that the variables are independent. 95% CI for  $\rho$ :  $[0.7131, 0.983]$ .