Comments

From the solutions of individual works you should form the document, where are presented the question and after that the solution and answer, then next question and answer and so on.

Although the exercises are composed based on the datasets and analyses performed in practices with *R*, for individual works you can use some other statistical package (some analyses can be made even with *Excel* – in the first group of exercises 100% – but you can also use *Statistica* or *SPSS* or …).

Individual work 1

Dataset: http://ph.emu.ee/~ktanel/DK_0007/ studentsR_eng.csv,
or the same as *Excel* fail: http://ph.eau.ee/~ktanel/DK_0007/studentsR_eng.xls

Exercises

1.1. Calculate the mean and the median of the trait 'headcirc' depending on the math grade (trait 'math').

1.2. Is porridge eating and non-eating students' weights' variability different?

1.3. How are the students' math grades distributed – construct the frequency table and the bar plot.

1.4. Are the relative frequencies of math grades different in different specialities (LKI=animal science, LAT=veterinary science)?

1.5. Construct the pie plot of sex and write close to the sectors 'Men' and 'Women'.

1.6. Construct the weights' histograms and densities for students eating and not eating the porridge.

1.7. Construct the frequency table of students' heights divided into 6 intervals.

1.8. Make box-plots on head circuits for students with different math grade.

1.9. Find the 95% CI of women's average weight. Is the average weight different from 60 kg (test the hypothesis)?

1.10. Are the average body mass indexes of men and women statistically significantly different?

1.11. Is the body mass index distribution different from the normal distribution?

Individual work 2

The exercises base on the dataset about Estonian fishes ['fish' = 'kala' in Estonian], which is a part of dataset composed by Mariann Nõlvak in 2004-2006 and was used in her master thesis in 2007.

Dataset: http://ph.emu.ee/~ktanel/DK_0007/kala.xls

The traits (with names in Estonian in the first row) in the dataset are following:
- fish number ('kala_nr');
- species ('liik') with 6 different values (in Estonian: haug [= 'pike' in English], särg [roach], latikas [bream], luts [burbot], ahven [perch] and koha[pikeperch]);
- type ('rühm') has two values: roovkala [fish of prey?] and lepiskala [?];
- 5 fishing places (trait 'pyygikoht' with values Võrtsjärv, Kärevere, Kastre, Praaga and Peipsi järv);
- fishing season (kevad-suvi [spring-summer] and sügis-talv [autumn-winter]);
- weight ('kaal') and length ('pikkus') of fishes;
- sex ('sugu');
- infestation with the larvae of broad tapeworm *Diphyllobothrium latum* ('diphyl' = 0 or 1);
- number of *Diphyllobothrium latum* found (trait 'diph_arv').

Exercises

Study the pikes ['haug' in Estonian].

2.1. Are the fishing season and infestation with *Diphyllobothrium latum* related?

a) Construct the 2-dimentional frequency table with absolute frequencies and also with relative frequencies. Comments?

b) Test the hypothesis about the statistical significance of the relationship using χ^2-test and Fisher exact test.

c) Calculate the odds ratio and its 95% confidence interval to compare the infestation in autumn-winter season ('sügis-talv') with the infestation in spring-summer season ('kevad-suvi'). Conclusions?

2.2. Construct the box-plots to illustrate the differences (or similarities) of pikes' weights in different places. When the *R* treats the values as outliers and marks in diagram with separate points (try to find from *R Help*)?

2.3.

a) Study the relationships between pikes' weight and length with Pearson correlation and between pikes' weight and number of *Diphyllobothrium latum* with Spearman and Kendall correlation analyses. Are these correlations statistically significant?

b) Find the linear regression equation between pikes' weight and length. Is this equation statistically significant? Illustrate the regression analysis with the graph showing the regression line and its confidence interval. Predict the weight of pike with length 60 cm.

Individual work 3

Exercises

Dataset: http://ph.emu.ee/~ktanel/DK_0007/kala.xls

Study the pikes [‘haug’ in Estonian].

3.1. Study the effects of fishing place, season and sex on the length of pikes. Are the effects of mentioned factors statistically significant?

3.2. Are the average lengths of pikes caught from Lake Peipsi ja Lake Võrtsjärve different? Is this difference statistically significant?

3.3. Find the model based average length (with 95%-confidence intervals) of male and female pikes in autumn-winter season (‘sygis-talv’) in Lake Võrtsjärv?

Dataset: http://ph.emu.ee/~ktanel/DK_0007/lehm.xls

The dataset contains milk yields and fat and protein percentages of the first lactation Estonian Holstein cows. The factor variables are birth year, sire and owner.

3.4. Considering the birth year as fixed factor and sire and owner as random factors estimate the relative importances of sire and owner effects of yield traits. For which yield trait is the owner effect (effect of feeding and keeping conditions) the biggest and for which trait is the sire effect (genetic effect) the biggest.

Individual work 4

Dataset (*R* dataset format): http://ph.emu.ee/~ktanel/DK_0007/puud.rda
or packed *Excel*-fail: http://ph.emu.ee/~ktanel/DK_0007/puud.rar

In this dataset one row corresponds to the one tree. The traits are the following:
- column named'A' – trees' age (years);
- column 'D' – trees' diamater (cm);
- column 'H' – trees' height (m);
- column 'ARENGUKL' – growth class with values A (woodless), N (underwood), L (next from the underwood, 'latimets' in Estonian), K (middle-aged), V (maturing), Y (mature), S (unclear), – (missing value);
- column 'PE' – tree species with values HB (aspen), KS (birch), KU (fir), LH (larch), LM (alder), LV (grey alder), MA (pine), RE (willow), SA (ash), TA (oak);
- column 'KKT' – habitat type.

Exercises

4.1. Make the new binary trait 'Mature' with values 1 (tree is mature, 'ARENGUKL'=Y) and 0 (tree is not mature, 'ARENGUKL'≠Y).

a) The townsman-forest owner doesn't know the age of his forest. But he knows that mainly his forest contains firs and also he knows the trees' diameters.
Find the logistic regression model to predict the probability of the firs to be ready for felling (mature) based of diameter. Illustrate results.

b) Find the optimal estimated probability value to distinguish the mature and non-mature trees. How big is in this point the sensitivity and specificity of the decision criterion.

4.2. Perform the cluster analysis to cluster the tree species based on their average diameter at ages 20 and 50 years, growth between ages 60 and 80 estimated from the 4^{th} order polynomial (if the maximum age of tree species is less than 80 years equate the growth with zero) and maximal height (in the first stage the new dataset of tree species with mentioned traits must be constructed and at the second stage the cluster analysis must be performed).