Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

TEACHING PROCESS OF AN APPLIED STATISTICS COURSE FOR STUDENTS OF THE ANTROPOLOGY STUDY PROGRAMME AT MASARYK UNIVERSITY

BENDOVÁ Veronika (CZ), BUDÍKOVÁ Marie (CZ)

Abstract. The point of these proceedings is to describe the teaching process of an applied statistics course that is offered by the Institute of Mathematics and Statistics on the Masaryk University in Brno to students of the Anthropology study programme. Teaching statistics to students of non-mathematical study programmes is in general a difficult matter. Therefore, we aim to make it as easy for the students as possible. To fulfill this goal, we employ a collection of solved statistics exercises, worksheets, along with friendly and open mindset. We succeed in motivating students to use statistical methods in practice despite raised difficulty of the course.

Keywords: course of Applied Statistics, antropology, R software

Mathematics subject classification: Primary 97K80; Secondary 97Dxx

1 Introduction

An important task of the Institute of Mathematics and Statistics in Brno is to develop statistical skill set of students of non-mathematical study programmes, among which are students of the anthropology study programme. Graduates of said programme have to be able to understand the contents of scientific studies that are based on quantitative data obtained by measurement of real objects. It is also expected that they will be able to present results of these studies in a way that is both statistically correct and anthropologically comprehensible. Because of these requirements, two mandatory courses have been integrated into the curricula of anthropology study programmes: Applied statistics I has been integrated into the Bachelor's study programme, and Applied statistics II has been integrated into the follow-up Master's programme.

Lectures of courses Applied statistics I and II are open to all student of non-mathematical study programmes. Illustrative examples, discussed during the lectures, are based on topics from various fields of study, e.g., economics, psychology, and medicine. The STATISTICA software is used for computations of the examples. This software is also used during practical seminars designed for students of all study programmes with the exception of anthropology students.

A special seminar that caters to specific requirements of anthropology students has been created for Applied statistics I and II, since their dominant computation tool is the R statistical software [1]. During these seminars, students analyze one-dimensional and multidimensional data using the R software. Practice real world data, used in the seminars, have been meticulously selected in collaboration with the Department of Anthropology. Students learn on model examples how to get a basic overview of the dataset using methods of descriptive statistics, how to correctly formulate goals and hypotheses, how to choose the right statistical test, and how to interpret obtained results.

We consider knowledge of basic terminology and methods of mathematical statistics, as well as a degree of proficiency in practical skill in working with the R software, to be a key part of absolvent's profile of the modern anthropology study programme for the 21^{st} century.

2 Syllabus of the course

The special seminar for students of the anthropology study programme, that encompasses usage of the R software, is currently taking place only for the Applied statistics I course. A similarly designed practical seminar for the follow-up course Applied statistics II is in preparation, and will be offered during the spring term of 2018. Therefore, we will only consider the former course in this text. Lectures of the Applied statistics I course, as well as the practical seminars, take place each week for 2 hours. As the compulsory literature is used the text [2] or [3]. Students should be able to comprehend advanced statistical literature [4] after completing this course.

The syllabus of Applied statistics I is as follows: 1. Exploratory data analysis, diagnostic plots. 2. Probability space, classic and conditional probability, independent events. 3. Random variable and description of their probability distribution, numerical characteristics, selected discrete and continuous distribution. 4. Basic terms of mathematical statistics – random selection, statistics derived from random selection, point and interval estimations of parameters. 5. Normality tests. 6. Parametric tests with one, two, or more independent random selections. 7. Non-parametric tests with one, two, or more independent random selections. 8. Dependence analysis of two random samples – dependence in contingency tables, Spearman's and Pearson's correlation coefficient.

3 Teaching methods

All study materials, i.e. texts of the lectures, notes for the seminars, practical examples, R-scripts written during the seminars, additional practice examples and files containing associated datasets, are available to the students in the Information system of the Masaryk University.

Each lecture starts with motivation, followed by explanation of theoretical terms, description of statistical methods, and presentation of several examples. Practical seminars take place in computer seminar rooms. The most important theoretical matter is revised during each seminar. Then, its application is shown on practical examples. The examples are granted to the students in the form of worksheets, with which they work. Solution to the examples is created by students in cooperation with the seminar tutor. First, the assigned example is decomposed on the whiteboard. The tutor asks questions that lead to the correct decomposition of the example. Students provide answers, and thus actively participate on the decomposition. The R software is then used for statistical computations, which are carried out by the tutor and visualized on the projector. Students type the R code, projected by the tutor, on their computers, which leads to the students getting familiar with the R software. Formulation of conclusions of the analysis, as well as interpretation of results, is done collectively by students under the supervision of the seminar tutor.

4 Course completion

The course concludes with an exam, which consists of a theoretical and a practical part. Students are allowed to use all study materials from lectures and practical seminars during the exam. In the theoretical part, students solve 8 examples for no more than 60 minutes. The point of the practical part is to present results of a dataset analysis that is carried out independently by each student. The presentation is evaluated on the basis of choosing the right statistical methods, the quality of contents and graphics of the presentation, and the performance of the speaker themselves.

The dataset for the presentation is obtained by the students in different ways. Some use the dataset from their bachelor's thesis, while others download freely available data from the internet. The remaining students ask their lecturer or seminar tutor. Presentation topics tend to vary. For example, these were the topics of some of the presentations from the fall term of 2016: 1. Processing of length and width dimensions of skulls and their comparison. 2. Circumference of hand and height of newborns. 3. Comparison of birth weight of first born and second born boys. 4. Determination of correlation dependence between length of ear and length of nose. 5. Incidence of dermatoglyphic patterns and ridge-count values.

5 Midterm homework

Another requirement for successful completion of the course is to solve a set of examples, assigned to the students as homework during the second half of the term. The point of this homework is to motivate students to use the R software in a self-reliant way and to think about an assigned examples in conditions similar to those that resemble conditions of scientific work, i.e., calm environment with the availability of all materials, as well as with the potential help of one's colleagues.

The homework contains 6 to 8 examples (the count is based on the difficulty of considered examples) that encompass all of the matter that was taught up to the point of assignment of the homework. Students successfully pass the homework if they get a score of at least $75\,\%$. The assignment is scored on the basis of correctness of the solution, graphical quality of plots, readability of the R code, and interpretation of results. Examples of the homework are assigned along with correct numerical results, as well as a view on required plots. Students can work on the assignment together. The ultimate point is to make the students think about the examples together and discuss them. The only practice that is forbidden is copying of the R-code. The time limit for handing in the solution of the homework is 14 days.

6 Sample of used examples

Since teaching of statistics supported by the R software bears a large challenge for students of non-mathematical study programmes (and thus for student of anthropology), we have decided to create a collection of examples with guides for solving each one using the R software. Usage of statistical methods, which are shown during each lecture, is always illustrated on an solved example that strongly emphasizes correct formulation of hypotheses on the basis of research goals, as well as valid interpretation of results received in numeric, tabular, or graphic manner. Students have sufficient number of examples for each statistical method available along with results, but without a solution guide.

Below are sample of three examples used in the course: An example from the collection, an example from the worksheet used in the practical seminars, and an example from the homework.

6.1 A solved example from the collection

We have english set of documented skelets, which includes osteometric data of clavicles (Parsons, 1916). The expected value ($\mu = 151.74\,\text{mm}$) and the standard deviation ($\sigma = 11\,\text{mm}$) of the right-side clavicle length of males were estimated from the data set. Suppose the dataset have normal distribution, compute the probability, that the right-side clavicle length of males is (a) equal to $150\,\text{mm}$; (b) less than $140\,\text{mm}$; (c) greater than $160\,\text{mm}$.

a) Because the data have normal distribution, the probability $\Pr(X = x) = 0$, and also

$$\Pr(X = x) = \Pr(X = 150) = \mathbf{0}.$$

b) We have two ways to compute the probability that right-side clavicle length of males is less than $140\,\text{mm}$. The first way is to use cumulative distribution function for the normal distribution $N(151.74, 11^2)$, the second is to use cumulative distribution function for the standardized normal distribution $N(0, 1)$.

$$\begin{aligned}
\Pr(X < 140) = \Pr(X \leq 140) &= \mathsf{pnorm}(140, 151.74, 11) = \mathbf{0.1429} \\
\Pr(X < 140) = \Pr(X \leq 140) &= \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{140-\mu}{\sigma}\right) = \\
&= \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{140-151.74}{11}\right) = \\
&= \Pr(U \leq -1.067) = \Phi(-1.067) = \\
&= \mathsf{pnorm}(-1.067, 1, 0) = \mathbf{0.1429}.
\end{aligned}$$

c) We have two ways to compute the probability that right-side clavicle length of males is greater than $160\,\text{mm}$. The first is to use cumulative distribution function for the normal distribution $N(151.74, 11^2)$, second way is to use cumulative distribution function for the standardized normal distribution $N(0, 1)$.

$$\begin{aligned}
\Pr(X > 160) = 1 - \Pr(X \leq 160) &= 1 - \mathsf{pnorm}(160, 151.74, 11) = \mathbf{0.2264} \\
\Pr(X > 160) = 1 - \Pr(X \leq 160) &= 1 - \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{160-\mu}{\sigma}\right) = \\
&= \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{160-151.74}{11}\right) = \\
&= \Pr(U \leq 0.7509) = \Phi(0.7509) = \\
&= \mathsf{pnorm}(0.7509, 1, 0) = \mathbf{0.2264}.
\end{aligned}$$

Interpretation of the results:
1. Because the data have normal distribution, the probability, that right-side clavicle length of males is equal to $150\,\text{mm}$, is $0\,\%$.
2. The probability, that right-side clavicle length of males is less than $140\,\text{mm}$, is $14.29\,\%$.
3. The probability, that right-side clavicle length of males is greater than $160\,\text{mm}$, is $22.64\,\%$.

6.2 An example from the worksheet used in the practical seminars

We have origin dataset of the craniometric data of skull width (in mm) of 13 females and 7 males from ancient egyptian population (Schmidt 1888, archival materials, adjusted for this example). We can see the data in the table 1.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13
females	133	134	132	141	135	135	136	137	137	136	139	118	120
males	132	132	133	128	149	132	137						

Tab. 1: Skull width of ancient egyptian females and males.

On the significance level $\alpha = 0.05$ test the null hypothesis, that skull width of ancient egyptian females is the same as the skull width of ancient egyptian males. Draw the boxplot (see the figure 1).

Normality test

```
## [1] 0.004327
## [1] 0.024478
```

The p-value of normality test for skull width of females ($p = 0.0043$) is less than significance level $\alpha = 0.05$, so H_0 is rejected. The p-value of normality test for skull width of males ($p = 0.0245$) is also less than significance level $\alpha = 0.05$, so H_0 is rejected. Both the datasets have not normal distribution.

Because the datasets have not normal distribution, we have to use ~~parametric~~ / non-parametric method to test the null hypothesis, that skull width of females is equal to the skull width of males.

H_0 : $x_{0.5} - y_{0.5} = 0$

H_1 : $x_{0.5} - y_{0.5} \neq 0$

```
## [1] "U1=36; U2=55"
## [1] "W: (-infty ; 20>"
## [1] "CI: (-11.9999 ; 5)"
## [1] "p-value: 0.4487"
```

1. Rejected region approach:
The minimum of test statistics U_1, U_2 is *36*, the sample size of females is n_1= *13*, the sample size of males is n_2 = *7*, rejected region W is $(-\infty; 20\rangle$. Because the value U_1 $36 \notin W$, the null hypothesis of medians equality *is not rejected* on significance level α = *0.05*.

2. Confidence interval approach:
The confidence interval is $(-8; 5)$. Because $c = 0 \in CI$, the null hypothesis of medians equality *is not rejected* on significance level α = *0.05*.

3. The p-value approach:
Because the p-value = *0.4487* is *greater* than α = 0.05 the null hypothesis H_0 of medians equality *is not rejected* on the significance level α.

Boxplot

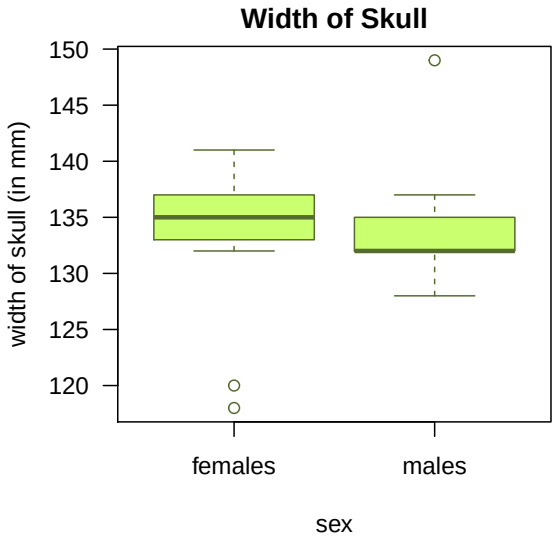

Fig. 1: Boxplot of skull width of ancient egyptian females and males.

Interpretation of the results:
There *is not* statistically significant difference between skull width of females and skull width of males. That means, the skull width of females *is the same* as skull width of males from ancient egyptian population.

6.3 An example from the homework

The dataset *102-nose-length-and-width.txt* includes data about nose length and nose width of the malaysian and peruvian population.

1. Compute the value of median, upper quartile and lower quartile of nose length of the peruvian population. (3 points)
2. Compute the value of arithmetic mean, standard deviation, skewness and kurtosis of nose length of the peruan population. (4 points)
3. The characteristics computed in part (1) and (2) insert to the table. (2 points)
4. Draw the histogram and boxplot of nose length of the peruvian population (see the Figure 2). (8 points)
5. Compute the appropriate type of correlation coefficient for association between nose length and nose width of the peruvian population. (2 points)
6. Draw a dotplot describing the relationship between nose length and nose width of the peruvian population. (4 points)

7. Be sure to sufficiently comment on all results and plots. (5 points)

```
##     median l.quartile h.quartile    mean st.dev skewness kurtosis
## 50%     51         48         53 50.4565 3.0267  -0.3134  -0.4115
```

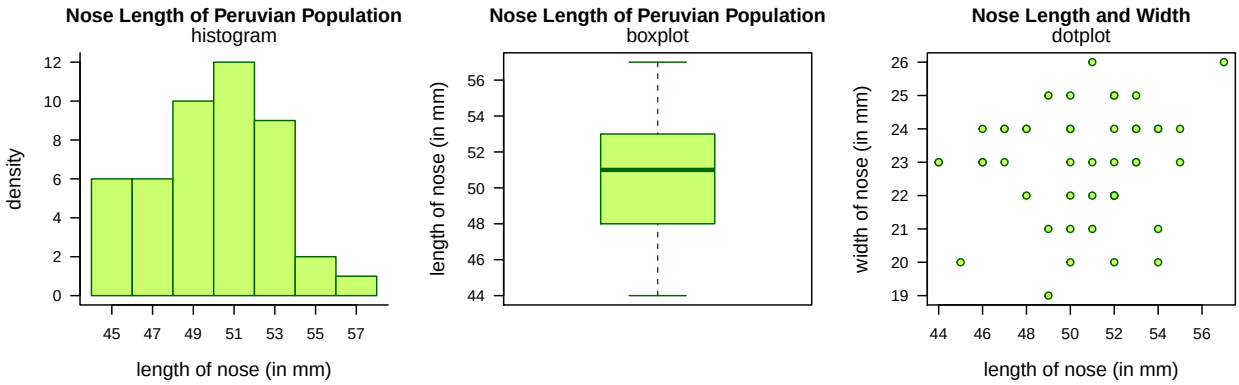

Fig. 2: Histogram and boxplot of nose length; dotplot of relationship between nose length and nose width of peruvian population.

```
## [1] "Pearson correlation coefficient: 0.1371"
```

7 Conclusion

Excellent anthropological research cannot do in this day and age without basic knowledge of statistics and data analysis. That is why the Applied statistics I course has been added to the list of mandatory courses of the anthropology Bachelor's study programme at the Masaryk University, while the anthropology Master's study programme was enriched by the mandatory Applied statistics II course. In these courses, students gradually get familiar with statistical methods suitable for analysis of anthropological data, along with the R statistical software. Learning statistics and the R software is, however, fraught with difficulty for students of non-mathematical study programmes. We therefore aim to ease the comprehension of statistical matter using a collection of solved problems and worksheets used in practical seminars. We also prefer friendly and open approach to every student. Our goal is to motivate students to use statistical methods and to accept the usefulness of statistics in their future careers. That is why students like to contact us after successfully completing the course to consult with us the statistical methods for analyzing datasets for their Bachelor's or Master's thesis.

References

[1] R CORE TEAM. R: *A Language and Environment for Statistical Computing. The R Project for Statistical Computing*, 2017 [cited 2017 Nov 28]. URL: https://www.R-project.org/

[2] BUDÍKOVÁ M., KRÁLOVÁ M., MAROŠ B.: *Průvodce základními statistickými metodami*, Praha, Grada, 2010, ISBN 978-80-210-7752-2, 272 s.

[3] BUDÍKOVÁ M., LERCH T., MIKOLÁŠ Š.: *Základní statistické metody*, Brno, Masarykova Univerzita, 2009, ISBN 978-80-210-3886-8, 170 s.

[4] KATINA S., KRÁLÍK M., HUPKOVÁ A.: *Aplikovaná štatistická inferencia I.*, Brno: MUNI Press, 2015, ISBN 978-80-210-7752-2, 306 s.

Current address

Bendová Veronika, Mgr.
Institute of Mathematics and Statistics
Faculty of Science
Masaryk University in Brno
Kotlářská 2, 611 37 Brno, Czech Republic
E-mail: bendova.veroonika@gmail.com

Budíková Marie, RNDr., Dr.
Institute of Mathematics and Statistics
Faculty of Science
Masaryk University in Brno
Kotlářská 2, 611 37 Brno, Czech Republic
E-mail: budikova@math.muni.cz