Syllabus
Data Analysis
(4 ECTS)

Alisa Melikyan (amelikyan@hse.ru, https://www.hse.ru/org/persons/205653)
School of Software Engineering
Meeting Minute # ___dated _____ 2019

1. Course description

Pre-requisites
Basic knowledge of statistics is required.

Abstract

The course is taught to students of a master degree of Computer science faculty in NRU HSE in the third and fourth modules of the first year of training. The number of credits is 4. Training in an audience takes 64 hours, including 24 hours of lectures and 40 hours of seminars. The control includes in-class tasks, a homework, a control work, and an examination work. The main purpose of the course is to teach students how to use different data analysis methods to analyze real data.

2. Learning Objectives

The objectives of the course are to:

- give students an introduction to the most widely used data analysis methods;
- explain the data analysis methods using real data and concentrating on complications that may occur during the analysis in real-life research;
- teach students how to organize their own research project using the knowledge obtained during the course;
- explain how to use data analysis tools in the most effective way to perform the research tasks.

3. Learning Outcomes

As a result of study, the student should know how to:

- select appropriate methods of data analysis depending on the research question and types of empirical data;
- select an appropriate data analysis tool to conduct analysis of different types of data;
- prepare empirical data for their further analysis;
- formulate research hypotheses and construct models;
- interpret the results of data analysis;
- report research results to the audience.

4. Course Plan

№ п/п	Themes	Total hours	Hours of training in an audience		Self-study Work
			Lecture hours	Practical hours	Home work
1	Introduction to data analysis	2	2		
2	Descriptive data analysis	25	3	6	16
3	Investigating relationship between variables	31	5	6	20
4	Regression analysis	40	6	12	22
5	Factor analysis	27	4	8	15
6	Cluster analysis	27	4	8	15
	Total:	152	24	40	88

Detailed course content

Theme 1. Introduction to data analysis

- Statistical packages and programming languages for data analysis;
- Data sources;
- Working with data (exploring data, entering new data, coding variables, preparing data for analysis, export/import of the data, modifying data).

Theme 2. Descriptive data analysis

- Frequency analysis;
- Graphical analysis;
- Statistical characteristics: central tendency estimations, dispersion, standard deviation, standard error of mean, confidence interval, percentile values, measuring symmetry and pointiness of distribution;
- Normal distribution, Z-standardization, Kolmogorov-Smirnov test of normality;
- Working with multiple response questions.

Theme 3. Investigating relationships between variables

- Cross tabulation analysis;
- Formulation and testing hypothesis;
- Level of significance and first type error;
- Chi-square test;
- Correlation coefficients: bivariate, part and partial;
- T-tests;
- ANOVA;
- Non-parametric tests.

Theme 4. Regression analysis

- Objectives of regression analysis;
- Graphical representation of regression line;
- Simple and multiple linear regression;
- Logistic regression;

- Interpreting results of regression analysis;
- Multicollinearity:
- Heteroscedasticity;
- Dummy variables;
- Regression model limitations and diagnostics.

Theme 5. Factor analysis

- Factor analysis steps;
- Evaluating applicability of data for factor analysis;
- Methods of factor analysis;
- Factor loading, rotation;
- Saving factors as new variables;
- Interpreting factors.

Theme 6. Cluster analysis

- Cluster analysis steps;
- Evaluating applicability of data for cluster analysis;
- Methods of cluster analysis: hierarchical and k-means;
- Saving cluster membership information as new variable;
- Characterizing clusters.

5. Reading List

a. Required

Dougherty, C. Introduction to econometrics / C. Dougherty. – 5th ed. – Oxford; New York: Oxford University Press, 2016. pp. ISBN 978-0-19-967682-8. (http://opac.hse.ru/absopac/index.php?url=/notices/index/302956/default)

Mirkin, B. Core concepts in data analysis: summarization, correlation and visualization / B. Mirkin. – London [etc.]: Springer, 2011. pp. 390. ISBN 978-0-85729-286-5. (http://opac.hse.ru/absopac/index.php?url=/notices/index/218662/default)

b. Optional

Idris, Ivan. *Python Data Analysis*, Packt Publishing, Limited, 2014. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1826990.

Mckinney, W. Python for data analysis: data wrangling with pandas, numPy, and IPhython / W. Mckinney. – 2nd ed. – Sebastopol: O'Reilly, 2017. pp. 524 c. ISBN 9781491957660. (http://opac.hse.ru/absopac/index.php?url=/notices/index/321747/default)

Pevalin, David, and Karen Robson. *Stata Survival Manual*, McGraw-Hill Education, 2009. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=480615.

Treiman, Donald J. *Quantitative Data Analysis : Doing Social Research to Test Ideas*, John Wiley & Sons, Incorporated, 2009. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=706553.

Verbeek, M. A guide to modern econometrics / M. Verbeek. – 4th ed. – Chichester: John Wiley & Sons, 2012. pp. 497. ISBN 978-1-11-995167-4. (http://opac.hse.ru/absopac/index.php?url=/notices/index/225411/default)

6. Grading System

A 10-point grading scale is used to evaluate the results of students' work.

10-point scale	Russian grading framework
10	Excellent
9	Excellent
8	Excellent
7	Good
6	Good
5	Satisfactory
4	Satisfactory
3	Fail
2	Fail
1	Fail

Academic progress is evaluated by means of a cumulative system, where the final grade is made up of the ongoing assessment results and the examination grade.

Academic control forms

- Tasks in class (TC): tasks which are performed in class and are aimed at developing students' skills in data analysis.
- Homework (HW): a research project which is performed at home and presented in class. The grade for the homework is based on the evaluations of the presentation and the written report on the results of data analysis.
- Control Work (CW): a written work which is performed in class at the end of the 3rd Module and covers the topics which were studied in the 3rd Module.
- Examination Work (EW): a written work which is performed in class at the end of the 4th Module and covers all the topics of the course.

Formula for final grade's calculation

Final grade = 0,3*grade(EW) + 0,2*grade(TC) + 0,3*grade(CW) + 0,2*grade(HW)

There are no blocking grades.

Examples of control tasks

1. Create a frequency table. Calculate the following statistical characteristics: mode, median, mean, range, standard deviation, S. E. mean, interquartile range, quartile deviation, decile ratio.

2. Evaluate the symmetry of distribution of the variable. Indicate whether the distribution is positively or negatively skewed and what does it mean in terms of the shape of the distribution. Indicate is the distribution significantly different from the symmetrical distribution and the reason of your conclusion?

3. Evaluate the pointyness of distribution of the variable. Indicate whether the distribution is leptokurtic or platykurtic and what does it mean in terms of the shape of the distribution and is the distribution significantly different for the "normal" distribution and the reason of your conclusions?

4. Do the Kolmogorov-Smirnov test to conclude whether the distribution of the variable is significantly different from the normal. Formulate hypothesis. Make conclusions.

5. Create a contingency table between two variables and interpret the results. Select two pairs of categorical variables to run Chi-square statistical test. Formulate hypotheses. Interpret the results of analysis. Make conclusions.

6. Do the bivariate correlation analysis. Calculate Pearson's, Kendall's and Spearman's correlation coefficients. Evaluate the significance of the coefficients. Indicate the coefficient of determination for every coefficient. Interpret the results.

7. Do the partial correlation analysis.

8. Do the multiple regression analysis using at least two predictors. Select the appropriate variables. Write down the regression equation. Assess the goodness-of-fit of the model. Are all the gradients and intercepts of the model statistically significant? Do the diagnostics of the model. Are the residuals normally distributed? Test the multicollinearity.

9. Do the factor analysis. Interpret the factors and save them as new variables. Use saved factors for cluster analysis. Define the number of clusters. Describe the clusters' characteristics.

7. Examination type

In the course of the written examination work the students should demonstrate the ability to:

- prepare the data for the further analysis, modify and transform the data in necessary;
- do the descriptive analysis of the data;
- select the appropriate method of data analysis in accordance with the type and characteristics of the data and the research question;
- formulate hypothesis;
- perform the data analysis;
- interpret and present the results of analysis.

For the home assignment the students should prepare the research report and present it in class.

The report should be prepared on the basis of the results of the quantitative data analysis. It should contain the research questions and hypotheses as well as the main results of the data analysis.

The data should be analyzed using descriptive statistics (frequency analysis, statistical characteristics, graphs), analysis of relationship between variables (crosstabs, correlations), regression analysis, ANOVA, nonparametric tests, factor and cluster analysis.

If needed preliminary data transformation should be done using recoding, calculation of new variables on the basis of existing variables, filters, aggregation, ranking, etc.

There is no need to use as many methods of analysis as the students can. It's better to select 1-2 methods and do an accurate in-depth analysis of the data.

The analysis should be done on the basis of the data collected or downloaded from open sources. The dataset should contain not less than 10 variables with different scales of measurement and not less than 100 cases.

The report should be prepared in the group of 2-3 students. All the group members should report the results in class. The results of analysis should be presented on the slides (not less than 10 slides).

8. Methods of instruction

The following methods of instruction are used:

- Lecturer method: the lecturer method is used during the lectures to present the methods of analysis and demonstrate their practical usage.

- Practical tasks: students are doing practical tasks in class and at home.

- Collaborating: students are working on the homework in groups.

- Classroom discussion: each student is given equal opportunity to interact and put forth their views. A discussion taking place in a classroom can be either facilitated by a teacher or by a student. A discussion could also follow a presentation or a demonstration.