Санкт-Петербургский филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет Санкт-Петербургская школа социальных наук и востоковедения
Национального исследовательского университета «Высшая школа экономики»
Департамент социологии

Рабочая программа дисциплины
Анализ данных в социологии
(преподается на английском языке)

для образовательной программы «Социология и социальная информатика»
направления подготовки 39.03.01 «Социология»
уровень бакалавриат

Разработчики программы:
Широканова А.А., к.социол.н., ashirokanova@hse.ru
Тенишева К.А., к.социол.н., ktenisheva@hse.ru
Волченко О.В., маг.социол.н.,ovolchenko@hse.ru

Согласована методистом ОСУП
«30» августа 2018 г.

Т.Г. Ефимова _____

Утверждена Академическим советом образовательной программы

«30» августа 2018 г., № протокола_____1_____

Академический руководитель образовательной программы

Д.А. Александров _____

Санкт-Петербург, 2018

Аннотация

<table>
<tr><td>Название дисциплины</td><td colspan="3">Анализ данных в социологии (преподается на английском языке)</td></tr>
<tr><td>Образовательная программа</td><td colspan="3">«Социология и социальная информатика»</td></tr>
<tr><td>Тип дисциплины</td><td colspan="3">Обязательная</td></tr>
<tr><td>Требования к уровню знаний студентов, необходимых для освоения дисциплины (пререквизиты)</td><td colspan="3">Студенты должны иметь базовые знания по теории вероятности, методологии и методам социологического исследования, социологической теории.</td></tr>
<tr><td>Объем з.е.</td><td colspan="3">10</td></tr>
<tr><td rowspan="2">Объем в часах</td><td>Аудиторная работа</td><td>Самостоятельная работа</td><td>Всего</td></tr>
<tr><td>152</td><td>228</td><td>380</td></tr>
<tr><td>Краткое описание курса</td><td colspan="3">Дисциплина читается три года. Первый год ориентирован на начинающих и нацелен на формирование и развитие умений по решению типичных задач при анализе социальных данных в программной среде R. Изучаются вводные темы (типы переменных, проверка гипотез, описательные статистики), некоторые методы и статистики (хи-квадрат, t-тест, непараметрические методы; однофакторный дисперсионный анализ, линейная регрессия). Второй и третий год посвящен многомерному анализу данных в социологии для продолжающих обучение. Второй год строится вокруг двух главных тем: факторного анализа и статистического предсказания, включающего линейную регрессию и моделирование структурными уравнениями. Также обсуждаются вопросы создания индексов и определения каузальности. Третий год посвящен методам анализа категориальных данных и включает виды предсказательных моделей (бинарная логистическая регрессия), классификации и представления данных (многомерное шкалирование, анализ соответствий, кластерный анализ).
Целью курса является обучение студентов осознанному использованию возможностей статистических методов анализа данных. Данный курс также является отправной точкой для студентов, нацеленных на более углубленное изучение методов статистики или планирующих применение количественных методов в собственных исследованиях.</td></tr>
<tr><td>Образовательные результаты по дисциплине</td><td colspan="3">В результате успешного освоения данной дисциплины студенты смогут использовать в профессиональной деятельности методы статистического анализа, используемые для описания данных, установления связи между переменными, структуры связи переменных, предсказания, сокращения размерности пространства и классификации в программной среде R.</td></tr>
<tr><td>Краткое содержание дисциплины</td><td colspan="3">1. Описательная статистика
2. Сравнение средних
3. Введение в обобщенные линейные модели
4. Статистический вывод и основы регрессионного анализа</td></tr>
</table>

	5. Линейная регрессия 6. Эксплораторный факторный анализ 7. Конфирматорный факторный анализ 8. Моделирование структурными уравнениями. Путевой анализ 9. Логистическая регрессия 10. Анализ соответствий и многомерное шкалирование 11. Кластерный анализ
Образовательные технологии	1. Проблемное обучение: разбор случаев 2. Метод проектов 3. Работа в малых группах, peer-to-peer assessment 4. Онлайн-обучение в группе на DataCamp
Формы контроля	Тесты, проекты, письменный экзамен (решение задач, подготовка проектного портфолио)
Литература	Основная 1. Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227 . 2.Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663 3. Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064 4.Stowell, Sarah (2014). Using R for Statistics. Apress. https://link.springer.com/book/10.1007%2F978-1-4842-0139-8 Дополнительная 1. .Agresti, Alan (2013). Categorical Data Analysis, 2nd edition, John Wiley & Sons, Inc. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1168529 2. Agresti, Alan, and Finlay, Barbara. (2007). Statistical Methods for the Social Sciences, Fourth Edition, Pearson Prentice Hall. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418314 3.Beh, Eric J., and Rosaria Lombardo. (2014). Correspondence Analysis: Theory, Practice and New Strategies, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1780725 4. Brown, T.A. (2015). Confirmatory factor analysis for applied research, Guilford Publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1768752 5. Crawley, M. (2013). The R Book, Second Edition. John Wiley & Sons. https://library.books24x7.com/toc.aspx?bookid=51275 6.Little, Todd D. (ed.) (2013). The Oxford Handbook of Quantitative Methods. Volume 2: Statistical Analysis, Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/97801999348

	98.001.0001/oxfordhb-9780199934898
Преподаватель	Волченко О.В., маг.социол.н., ovolchenko@hse.ru Тенишева К.А., к.социол.н., ktenisheva@hse.ru Широканова А.А., к.социол.н., ashirokanova@hse.ru

Course Syllabus

Title of the course	Data Analysis in Sociology (offered in English)		
Title of the Academic Programme	BA 'Sociology and Social Informatics'		
Type of the course	Mandatory		
Prerequisites	Students are expected to have taken some sort of introductory course for probability theory, social research methods, and social theory.		
ECTS workload	10		
Total indicative study hours	Directed Study	Self-directed study	Total
	152	228	380
Course Overview	This course lasts for three years. The 1st year aims at beginners and serves to develop skills necessary to solve typical problems in analysing social data in R software environment. The course goes from introductory topics (variable types, hypothesis testing, descriptive statistics) to some statistics and methods (chi-square, t-test, nonparametric statistics, one-way ANOVA, and linear regression). The 2nd and 3rd years provide an intermediate-advanced statistical analysis for quantitative research in sociology. In the 2nd year, the course covers two main topics - factor analysis and statistical prediction, including linear regression and structural equation modelling. We also discuss key issues in statistical analysis, such as creating indices and identifying causality based on the results of the analysis. The 3rd year focuses on multivariate analysis of categorical data. It includes special types of prediction models (logistic regression), techniques of dimension reduction (correspondence analysis and multidimensional scaling) and classification (cluster analysis). The course covers the building blocks of quantitative data analysis with the goal of training students to be informed consumers and producers of quantitative research. This course is also the starting point for students interested in pursuing advanced methods training or planning to use quantitative methods in their own research.		
Intended Learning Outcomes (ILO)	After completing this course, students will be able to apply for professional purposes multivariate data analysis methods and techniques used to describe data, establish and test variable structures, to predict, to reduce data dimensions and to classify data in the R software. As a result of studying the discipline, the student will be able to: • Conduct statistical analyses in RStudio; • Choose appropriate methods and techniques for certain types of variables and certain aims of the analysis; • Give meaningful interpretation of statistical results: regression coefficients, tables, plots and diagrams (produced in R); • Perform data transformations;		

	• Represent graphically the results of the statistical analyses; • Create analytical reports describing all the stages of analysis and interpreting its results.
Teaching and Learning Methods	1. Problem-solving in case studies 2. Project portfolio 3. Work in small groups, peer-to-peer assessment 4. Online learning in a DataCamp virtual class

Content and Structure of the Course

№	Topic / Course Chapter	Total	Directed Study		Self-directed Study
			Lectures	Tutorials	
1	Research hypotheses vs. statistical hypotheses. Variable types	18	2	2	14
2	Central tendency measures	20	2	4	14
3	Chi-square	18	2	2	14
4	Two means comparison	20	2	4	14
5	One-way ANOVA	22	4	6	12
6	Linear regression	26	4	10	12
7	Linear regression with multiple predictors	28	4	12	12
8	Introduction to GLM	4	0	2	2
9	Linear regression: OLS. Diagnostics	14	0	4	10
10	Linear regression: Interaction effects	14	0	4	10
11	Exploratory factor analysis	18	4	4	10
12	Confirmatory factor analysis	20	4	6	10
13	Introduction in SEM	18	4	4	10
14	SEM: model specification	16	2	4	10
15	Path analysis	16	2	4	10
16	SEM with latent variables	16	2	4	10
17	Putting it all together	16	2	4	10
18	Overview of categorical data analysis	8	2	2	4
19	Binary logistic regression	18	4	4	10
20	Multidimensional scaling	16	2	4	10
21	Correspondence analysis	16	2	4	10
22	Cluster analysis	18	4	4	10
Total study hours		380	54	98	228

Indicative Assessment Methods and Strategy	Tests, projects, and written exams (problem solving, project portfolio). Second-year students are evaluated by in-class activity, group projects, and exam. Evaluation of the 3rd year students is based on four criteria: in-class activity, projects, and exam. Evaluation of the 4th year students

	relies on projects, tests, previous exams, and the final exam.
Readings / Indicative Learning Resources	Mandatory 1. Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227 . 2.Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663 3. Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064 4.Stowell, Sarah (2014). Using R for Statistics. Apress. https://link.springer.com/book/10.1007%2F978-1-4842-0139-8 Optional 1. .Agresti, Alan (2013). Categorical Data Analysis, 2nd edition, John Wiley & Sons, Inc. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1168529 2. Agresti, Alan, and Finlay, Barbara. (2007). Statistical Methods for the Social Sciences, Fourth Edition, Pearson Prentice Hall. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418314 3. Beh, Eric J., and Rosaria Lombardo. (2014). Correspondence Analysis: Theory, Practice and New Strategies, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1780725 4. Brown, T.A. (2015). Confirmatory factor analysis for applied research, Guilford Publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1768752 5. Crawley, M. (2013). The R Book, Second Edition. John Wiley & Sons. https://library.books24x7.com/toc.aspx?bookid=51275 6. Little, Todd D. (ed.) (2013). The Oxford Handbook of Quantitative Methods. Volume 2: Statistical Analysis, Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898

Indicative Self- Study Strategies	**Type**	**+/–**	**Hours**
	Reading for seminars / tutorials (lecture materials, mandatory and optional resources)	+	60
	Assignments for seminars / tutorials / labs	+	40
	E-learning / distance learning (MOOC / LMS)	+	36
	Fieldwork	-	
	Project work	+	70
	Other (please specify)	-	

	Preparation for the exam	+	22
Academic Support for the Course	Academic support for the course is provided via LMS, where students can find: guidelines and recommendations for doing the course; guidelines and recommendations for self-study; samples of assessment materials. There is also an open-accessed questions-and-answers board for students.		
Facilities, Equipment and Software	Lectures are supported by slide presentations demonstrated with a projector. Seminars and lab sessions are to be held in a fully-equipped computer class with personal computers available to every student in a group (in cases when there are more students than PCs they are welcome to bring their own computers). The necessary software is R (https://www.r-project.org/) and RStudio (https://www.rstudio.com) that are both open-source software available free of charge under the GNU Affero General Public License v3.		
Course Instructor	Dr. Anna Shirokanova, ashirokanova@hse.ru Dr. Ksenia Tenisheva, ktenisheva@hse.ru Olesya Volchenko, MA, ovolchenko@hse.ru		

Annex 1

Course Content

Topic 1. Research hypotheses vs. statistical hypotheses. Variable types
The cycle of research. Data analysis as part of the research process. Posing and testing hypotheses. Research hypotheses vs. statistical hypotheses testing. Directed and non-directed hypotheses. Dependent and independent variables. Variable scales: nominal, ordinal, continuous (interval and ratio). Descriptive statistics of a variable depending on its type.
Getting to know R and RStudio.
Core reading
Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227
Stowell, Sarah (2014). Using R for Statistics. Apress.
https://link.springer.com/book/10.1007%2F978-1-4842-0139-8
Additional materials
DataCamp Introduction to R, URL: datacamp.com
DataCamp R for the Intimidated, URL: datacamp.com
DataCamp Reading Data into R with readr, URL: datacamp.com

Topic 2. Central tendency measures
Mean, median, mode. Standard normal distribution and its use. Z-scores.
Moments of distributions. Distribution plots and reading them. Sources of bias in data.
Interpretation of z-scores.
Mean as a data model.
Creating objects, types of objects, basic functions in R. Descriptive statistics in R. Tidy data.
Core reading
Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227
Additional reading
Stowell, Sarah (2014). Using R for Statistics. Apress.
https://link.springer.com/book/10.1007%2F978-1-4842-0139-8

Topic 3. Chi-square
Observed and expected frequencies. Measures of association for categorical variables. Reading and interpreting chi-square tests. Assumptions of chi-square. Independence. Standardised residuals. Odds ratio.
Chi-square and other association measures in R.
Core reading
Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227
Additional materials
DataCamp Introduction to Probability and Data Labs, URL: datacamp.com
Stowell, Sarah (2014). Using R for Statistics. Apress.
https://link.springer.com/book/10.1007%2F978-1-4842-0139-8

Topic 4. Two means comparison
Independent and paired samples. Assumptions behind the t-test. One-sample t-test. Two-sample t-tests. Nonparametric tests for two samples and for multiple samples.
Reading and interpreting means comparison. Confidence intervals.
Means comparison in R
Core reading
Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227
Additional materials
DataCamp Introduction to Probability and Data Labs, URL: datacamp.com
Stowell, Sarah (2014). Using R for Statistics. Apress.
https://link.springer.com/book/10.1007%2F978-1-4842-0139-8

Topic 5. One-way analysis of variance (ANOVA)
Assumptions and usage of ANOVA. Between-group and within-group variance, their ratio. Planned and non-planned comparisons; corrections. Post hoc comparisons for equal and unequal variances. Reading and interpreting ANOVA. One-way ANOVA in R. Presenting the results of ANOVA. Getting to know RMarkdown: reports and slide shows.
Core reading
Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227
Additional reading
Stowell, Sarah (2014). Using R for Statistics. Apress.
https://link.springer.com/book/10.1007%2F978-1-4842-0139-8

Topic 6. Correlation and linear regression
Correlations. Research problems for correlational analysis. Correlation coefficients for different types of data. ANOVA, correlation, regression as linear models. Building a linear regression. Ordinary least squares. Fitting the regression line. Assumptions behind linear regression. Reading and interpreting regressions. Presenting and interpreting a linear regression. Categorical predictors in a linear regression. Dummy-coding.
Linear regression in R.
Plotting linear regressions in R (case studies).
Core reading
Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson.
http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064
Additional materials
DataCamp Inferential Statistics, URL: datacamp.com

Topic 7. Linear regression with multiple predictors

The concept of interaction effects for categorical by categorical, categorical by continuous, continuous by continuous variables. Effect coding. Centring. Multicollinearity.
Reading and interpreting interaction models in a linear regression.
Testing for interactions in R.
Reporting and interpreting a linear regression with interactions.
Core reading
Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson.
http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064
Additional reading
DataCamp Regression Models with swirl, URL: datacamp.com

Topics 8-9. Introduction to GLM. Linear regression: OLS. Diagnostics

Covariance and correlation. Basic concepts and logics of linear regression and GLM. OLS estimator of linear regression, interpretation and statistic test of OLS estimators, fitted values and residuals, R-squared, addressing nonlinearity in linear regression framework, standardized coefficients, drawing plots, practice in R.
Core reading
Agresti, Alan, and Finlay, Barbara. (2007). Statistical Methods for the Social Sciences, Fourth Edition, Pearson Prentice Hall.
http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418314
Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson.
http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064
Additional reading
Crawley, M. (2013). The R Book, Second Edition. John Wiley & Sons.
https://library.books24x7.com/toc.aspx?bookid=51275

Topic 10. Linear regression: Interaction effects

Main and multiplicative effects in regression models. Interaction effects, additive effects. Interpreting results. Choosing best model. Practice in R.
Core reading
Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227.
Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson.
http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064
Additional reading
Agresti, Alan, and Finlay, Barbara. (2007). Statistical Methods for the Social Sciences, Fourth Edition, Pearson Prentice Hall.
http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418314

Topic 11. Exploratory factor analysis

Dimensionality reduction. Manifest and latent variables. Factors, graphical representation of factors. Exploratory factor analysis. Factor scores, factor space, types of rotation. Optimal number of factors. Interpretation of the results. Creating indices based on factor analysis. Practice in R.
Core reading

Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227.
Brown, T.A. (2015). Confirmatory factor analysis for applied research, Guilford Publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1768752
Additional reading
Crawley, M. (2013). The R Book, Second Edition. John Wiley & Sons. https://library.books24x7.com/toc.aspx?bookid=51275

Topic 12. Confirmatory factor analysis

Difference between exploratory and confirmatory factor analyses. Factor structure. Testing your (or somebody else's) scales. Types of latent variables. Constructing factor model in lavaan package. Calculation of degrees of freedom, minimal number of cases. Non-correlated and correlated latent factors. Interpreting results. Model diagnostics. Cronbach's alpha. Practice in R.
Core reading
Brown, T.A. (2015). Confirmatory factor analysis for applied research, Guilford Publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1768752
Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663

Topic 13. Introduction to SEM

Structural equation modeling as extension of confirmatory factor analysis. Exogenous and endogenous variables. Testing causal assumptions. Partial correlation, heterogeneous correlations (polychoric, tetrachoric and polyserial correlations). Practice in R.
Core reading
Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663

Topic 14. SEM: model specification and identification

Formulating theory-based causal hypotheses. Causal inference. Specification concepts. Mediation and moderation effects. Measurement error: correlated and uncorrelated. Practice in R.
Core reading
Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663
Additional reading
Crawley, M. (2013). The R Book, Second Edition. John Wiley & Sons. https://library.books24x7.com/toc.aspx?bookid=51275

Topic 15. Path analysis

Concept of "path". Path analysis: only observed variables. Graphical representation. Identification of path model. Estimation of structural equation model. Model fit. Degrees of freedom, number of cases. Meaning of the indices. Corrected chi-square measures. Interpreting the results. Practice in R.
Core reading
Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663

Topic 16. SEM with latent variables

Introducing latent factors in the model. Identification of SEM. Estimation of structural equation model. Model fit. Meaning of the fit indices. Model modification. Interpreting the results. Practice in R.

Core reading

Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663

Topic 17. Putting it all together

Implementing all the methods to the real-life research. Combining factor analysis and regression analysis. Using SEM to test theoretical assumptions about causality. Advantages and disadvantages of the methods.

Topic 18. Overview of Categorical Data Analysis

Models for categorical outcome variables. Variety of goals of analysis with categorical data. Examples of empirical research for various methods, e.g. Poisson regression (count variable), binary logistic regression, ordinal regression, multinomial regression, correspondence analysis, conjoint analysis, multidimensional scaling, and cluster analysis. Typical goals of analysis and interpretation of results.

Core reading

Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064

Additional reading

Agresti, Alan (2013). Categorical Data Analysis, 2nd edition, John Wiley & Sons, Inc. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1168529

Topic 19. Binary Logistic Regression

Logistic regression. Objectives of logistic regression. Logistic curve. Assumptions of logistic regression. Transforming a probability into odds and logit values. Maximum likelihood estimation. Goodness-of-fit measures for logistic regression. Interpretation of results (linear and dichotomous predictors). Stepwise model building. Diagnostics. Procedures in R.

Core reading

Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064

Additional reading

Agresti, Alan (2013). Categorical Data Analysis, 2nd edition, John Wiley & Sons, Inc. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1168529

Topic 20. Multidimensional Scaling

Dimension reduction as an objective of data analysis. Idea of MDS. Perceptual map. MDS vs. factor and cluster analyses. Objectives of MDS. MDS algorithms. Decompositional and compositional approach. The number and selection of objects. Nonmetric vs. metric methods. Similarity data and preference data. Assumptions of MDS analysis. Selecting dimensionality. Ideal point. Goodness-of-fit measures. Interpreting MDS results. Procedures in R.

Core reading

Tabachnick, B. G., and Fidell, L. S. (2014). Using Multivariate Statistics: Pearson New

International Edition (Vol. 6th ed). Harlow, Essex: Pearson. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1418064

Additional reading

Little, Todd D. (ed.) (2013). The Oxford Handbook of Quantitative Methods. Volume 2: Statistical Analysis, Oxford University Press.

Topic 21. Correspondence Analysis

Objectives of correspondence analysis. Assumptions of correspondence analysis. Perceptual mapping. Principal components analysis. The row and column problems. Correspondence analysis displays. Correspondence analysis biplots. Multiple correspondence analysis. MCA maps. Measures of fit for MCA. Interpreting correspondence analysis results. Canonical correspondence analysis. Procedures in R.

Core reading

Beh, Eric J., and Rosaria Lombardo. (2014). Correspondence Analysis: Theory, Practice and New Strategies, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1780725

Additional reading

Little, Todd D. (ed.) (2013). The Oxford Handbook of Quantitative Methods. Volume 2: Statistical Analysis, Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898

Topic 22. Cluster Analysis

Objectives of cluster analysis (taxonomy description, data simplification, and relationship identification). Necessity of conceptual framework. Similarity measures. Proximity matrix. Decision-process in cluster analysis. Dendrograms. Cluster profiles. Distance measures for various types of variables. Assumptions of cluster analysis. Measures of overall fit. Between- and within-cluster variation. Hierarchical and non-hierarchical clustering algorithms. Determining the number of clusters. Interpretation of clusters. Cross-classification from several solutions. Procedures in R.

Core reading

Beh, Eric J., and Rosaria Lombardo. (2014). Correspondence Analysis: Theory, Practice and New Strategies, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1780725

Additional reading

Little, Todd D. (ed.) (2013). The Oxford Handbook of Quantitative Methods. Volume 2: Statistical Analysis, Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898

Assessment Methods and Criteria

Assessment Methods

1st year of the course

Types of Assessment	Forms of Assessment	Modules			
		1	2	3	4
Formative Assessment	Project			*	*
	Test				*
	In-class Participation			*	*
Summative Assessment	Exam				*

Project. Students create teams of 2-3 and work together on their project during the whole course, submitting and peer-reviewing them by each computer lab. Final projects are submitted in full and presented in the classroom. Each group selects one country from the European Social Survey, then picks the topic of interest within the scope of available survey questions (e.g. Health, Democracy, Religion, etc.) and performs all the tests covered in class on these data. One day before each computer lab, the due piece of work is to be submitted and blindly peer-reviewed by two other groups in LMS. The instructors would assign reviewers, while students might not know who would be their reviewers next time. Final projects are presented in two steps. At the first stage, the group submits the code with interpretations. After this, they present the findings and procedures in class. Students are expected to choose and perform correctly the ways to analyse and interpret the data, as well as to demonstrate their knowledge and skills in presenting these results to the audience. Individual contribution of each student is graded. Projects themselves should be submitted as scripts or RMarkdown objects; in-class presentations should be adapted for the slide shows (e.g. Prezi, LibreOffice Impress, etc.). Project details are available in LMS.

Test. All students fill in a comprehensive paper-and-pencil test covering all previous topics.

In-class activity during lectures and seminars. Students are expected to ask questions and participate in discussions, as well as help other students during practice sessions. Small regular tests held at seminars are also part of this grade. Regular active participation in the classes is graded as perfect (10); no participation at all is graded as 0.

Exam is aimed at checking the skills students should have obtained during the course. Its structure is close to the structure of projects but covers all the topics: standard problems including descriptive statistics, measures of association, comparing two or more means, and linear regression.

The grades are calculated by the following formula:

Cumulative score = 0.3*projects module 3 + 0.2*projects module 4 + 0.2 * class assignment + 0.3 * in-class activity

Final grade = 0.6*cumulative score + 0.4*exam

2nd year of the course

Types of Assessment	Forms of Assessment	Modules			
		1	2	3	4
	Project		*	*	
	In-class Participation		*	*	
Summative Assessment	Exam			*	

Project. There are three basic features assessed: correct calculations and correct code (syntax); correct interpretations – students must describe trends properly, assess significance of the results, and predict values of dependent variable correctly; and produce correct graphics, with proper types of plots and formatting applied. Homework is graded from 0 ("extremely poor"="fail") to 10 ("perfect"="pass") each. Proficiency in the English language does not affect the grade.

In-class participation during lectures and seminars. Students are expected to ask meaningful questions and participate in discussions, as well as help other students during practices. Regular active participation in the classes is graded as perfect (10); no participation is graded as 0.

Exam is aimed at checking the skills students should have obtained during the course. Its structure is close to the structure of home assignments, though it covers all the topics studied. Criteria for the assessment of the exam are the same as for home works: correct calculations, correct interpretation and correct graphics.

The grades are calculated by the following formula:

Cumulative score = 0.2*activity + 0.8*mean(projects)

Final grade = 0.8*cumulative score + 0.2*exam

3rd year of the course

Types of Assessment	Forms of Assessment	Modules			
		1	2	3	4
	Project			*	
	In-class Participation			*	
Summative Assessment	Exam			*	

Project. Two individual projects are due. A project applies one of the methods covered in the course (binary logistic regression or cluster analysis) and presents the results as a report. Two projects sum up to a student's portfolio. Specific project requirement are available in the LMS.

In-class participation. Every second seminar there is an in-class test on interpretation of binary logistic regression, multidimensional scaling, correspondence analysis, and cluster analysis.

Exam consists of two problems involving the methods covered in this course. Criteria for the assessment of the exam are the same as for home projects: correct specification, interpretation or results, and conclusions.

The grades are calculated by the following scheme:

Cumulative score = 0.2* exam grade in sophomore year + 0.2* exam grade in junior year + 0.2*project 1 + 0.2*project 2 + 0.2*activity

Final grade = 0.8*cumulative score + 0.2*final exam

Assessment Criteria

In-class Participation

Grades	Assessment Criteria
«Excellent» (8-10)	A critical analysis which demonstrates original thinking and shows strong evidence of preparatory research and broad background knowledge.
«Good» (6-7)	Shows certain evidence of preparatory research and background knowledge, a reasonable standard of expression.
«Satisfactory» (4-5)	Satisfactory overall, showing a fair knowledge of the topic. Significant hesitation in answering follow-up questions and/or incomplete or partly irrelevant answers.
«Fail» (0-3)	Very limited to insufficient evidence of relevant knowledge and skills in addressing the topic. Unable to offer relevant information or opinion in answer to follow-up questions.

Project Work

Grades	Assessment Criteria
«Excellent» (8-10)	A well-structured, analytical presentation of the project. Student shows strong evidence and broad background knowledge. In a group presentation, all members contribute equally and each contribution builds on the previous one clearly. Answers to follow-up questions reveal a good range and depth of knowledge beyond that covered in the presentation and show confidence in discussion.
«Good» (6-7)	Clearly organized analysis, showing evidence of good overall knowledge of the topic. The presenter of the project work highlights key points and responds to follow-up questions appropriately, making several minor mistakes. In group presentations, there is evidence that the group has met to discuss the topic and is presenting the results of that discussion, in an order previously agreed but lacks some knowledge to address the necessary points.
«Satisfactory» (4-5)	Takes a very basic approach to the topic, using broadly appropriate but suboptimal material, lacks focus. The presentation of project work is largely unstructured, and some points are irrelevant to the topic. Knowledge of the topic is limited and there may be evidence of basic misunderstanding. In a group presentation, most of the work is done by one student and the individual contributions do not add up.
«Fail» (0-3)	Fails to submit the project or to demonstrate enough knowledge to meet the project's requirements.

Written Assignments (Written Exam)

Grades	Assessment Criteria
«Excellent» (8-10)	All problems are solved correctly, all results are properly interpreted. The student makes a clear argument that responds effectively to all aspects of the problem. Few minor errors may occur.
«Good» (6-7)	The answer addresses most aspects of the topic with a clear argument which is not always correct. The response covers part of the task requirements.
«Satisfactory» (4-5)	The student addresses a minor part of the task which solves at least some part of the problems correctly. The answer demonstrates certain skills of correct problem-solving and interpretation of results.
«Fail» (0-3)	The student fails to demonstrate enough appropriate knowledge to be able to solve at least part of the problems correctly.

Samples of Assessment Material

In the first year of the course, students are to submit group projects developed within the same team per each technique covered in class. Each team works with the European Social Survey data for one country on one general topic during the whole course. Each project has its own requirements (see LMS for details). After submitting their own project, each team reviews the project of another team, on a rotating basis, and gives feedback. A typical project task would be to locate suitable variables within their country and topic, describe the team's hypothesis, the data, run, report, and interpret the result.

The exam after the 1st year of the course consists of 3-4 standard problems including descriptive statistics, measures of association, comparing two or more means, and linear regression. Sample exam problems are as follows:

Problem 1.
1.What are the mean weights of male and female respondents? Report the mean and the SD.
2.Show the boxplot of mass by sex.
3.Test whether males and females in the sample have the same smoking habits. Report the results.

Problem 2
4. Compare whether male and female respondents in the sample have different weight. Report the results.
5. Now compare whether different groups of smokers are of the same height or not. Report the results.
6. Show the pairwise comparison plot for question 5.

Problem 3.
7. What is the size of shared variance (R-squared) between the height and weight of the respondents?
8. Are the height, sex, and smoking habits good predictors of the weight of the respondents? Show it.
9. Is the additive or the multiplicative model based on height, sex, and smoking, better at explaining the weight? Report the results.
10. Draw the effect of the interaction (even if not significant) between height and sex on mass.

To assess their progress in the 2nd year of the course, students will be given home projects each two weeks. Students are expected to replicate analysis from a given scientific article using the methods discussed during the classes or to produce their own model following the lecture. Each project is based on the materials from preceding lectures and includes both theoretical questions and practical tasks that ought to be answered using R.

Example:

- *Replicate the regression table given in the article. Interpret the results.*
- *What is the null hypothesis? What is the alternative hypothesis in this test?*
- *Find the test statistic in the R output. Show how the resulting test statistic could be computed.*
- *What is the p-value and what does the p-value mean?*
- *Should the null hypothesis be rejected here? Can one conclude that the estimated relationship between household size and satisfaction would hold true in the population?*

Exam structure is similar to home tasks, though it covers all topics studied. Sample of exam questions (2nd year of the course, students are expected to answer at home):

- *Read the article carefully. Identify the measures used in the analysis, find them in the dataset.*
- *Were the variables transformed in any way? Prepare the data for analysis*
- *Replicate the regression table given in the article.*
- *Interpret the results. Do your results fit those presented in the article? What might be the reason of this difference?*

Sample of exam questions (3rd year of the course):

"You have 60 minutes to solve two problems. Hand in your solutions as an R script titled "YourLastName_Option.R". You may use your own or others' scripts. You may not use the help of other people by talking to them, texting them, or otherwise. Should you be noticed doing so, the exam is over and you get zero points for it.

1. Estimate and describe a logit model. File: exam1.txt
You have the data about religiosity (0=not religious, 1=religious), sex (1=male, 2=female), age (years), and marital status of the respondent (married, divorced, widowed, or single).
Build a model predicting whether the respondent is religious or not.
Report your final model and describe the results.

2. Identify and describe clusters. File: exam3.csv
You have the data about trust to political parties, to the European Parliament, and to the United Nations Organizations in 26 countries around Europe. Using cluster analysis, identify which countries are closer to each other in how their citizens trust these political institutions.
Report the final solution, how you reached it, and describe the clusters."

Recommendations to Course Instructors

The goal of this course above and beyond teaching specific methods is to enable students to use the methods covered in the course on a stand-alone basis whenever they need this in the future. Therefore, every reasonable effort should be made to make the material understandable and comprehensible, depending on the level of the student. Encouraging those students who have already understood new material to share their understanding with the others has demonstrated rewarding results. Regular Q&A sessions are needed at the beginning of each session, at lectures and labs alike. The general recommendation is to put emphasis on training the skills to perform the same types of analysis autonomously; therefore, the more time students get to practise their data analysis skills on different data sets, the more reliable the success of the course. Try using as many ways of approaching students as possible. Since the knowledge and readiness to learn in English is non-equal among students, be always prepared to stratify exercises as well as theory for different levels. Find and use the youtube.com tutorials. Small groups are always encouraged whenever possible and reasonable. For more ideas, consult the booklet: B. Rosenshine Principles of Instruction (2010). URL: http://www.ibe.unesco.org/sites/default/files/resources/edu-practices_21_eng.pdf

Recommendations to Students for Doing the Course

Completing this course is not like any other discipline you have studied. Here, the purpose of the course is to equip you with necessary methods when doing data analysis of some kind. You are offered an array of methods of data analysis which you are likely to use while staying in the social sciences and beyond. This course is meant for your better understanding of what happens when you call a function in R. Try using your own words when describing the methods or covering new material. Do not hesitate to pose your questions to the instructor but please restrict from blind copying from the book even when it seems a good idea. In doing your group projects, allocate the time to work together and talk to the instructor or teaching assistant at least once a week. Your results should be logical and rather easy to comprehend, whether it is you talking about them or another person learning about your project. When required to do peer review for other teams, try to provide helpful and timely feedback and suggest possible improvements to your groupmates' projects. While your comments would not be decisive in grading other projects, they will be crucial to fostering discussion and coordination between the teams. Additionally, try keeping a vocabulary on each topic covered, supplying the most important terms with examples. This will help you at the exam and in the future as your personal reference book. Explore relevant chapters on DataCamp and practise every day to master your programming skills.

Recommendations for Self-Study

Read and watch as much about the topic as possible. Having read on the same topic from several textbooks usually improves your understanding substantially. Watch and learn extra beyond the classes. Complete the excellent introductory methods courses on Coursera or DataCamp to recap on basic statistical concepts if you feel you could benefit from it. Try to use systematically the power of community at stackoverflow. Thick methods textbooks are now accompanied by useful websites with the data sets and additional learning materials provided (www.mvstats.com).

Inclusive teaching for the Organization of Learning Process for Students with Special Needs

The following types of comprehension of learning information (including e-learning and distance learning) can be offered to students with disabilities (by their written request) in accordance with their individual psychophysical characteristics:

1) *for students with visual impairment:* a printed text in enlarged font; an electronic document; audios (transferring of learning materials into the audio); an individual advising with an assistance of a sign language interpreter; individual assignments and advising.
2) *for students with hearing impairment: a* printed text; an electronic document; video materials with subtitles; an individual advising with an assistance of a sign language interpreter; individual assignments and advising.
3) *for students with physical impairment:* a printed text; an electronic document; audios; individual assignments and advising.