WARSZAWSKA SZKOŁA ANALIZY DANYCH

Warsaw School of Data Analysis Training Cycle Description

1. Introduction

The significance of data analysis-related skills in a curriculum of a contemporary employee has dramatically increased in recent years. Its importance has been most visible in the sector of information and communication technologies. In fact, the rapid development of information technologies made data analysis (or "data science") an important employee asset in almost all economy sectors. This is due to the fact that various types of data are currently collected both by companies and by public institutions. As a consequence, there is a high and increasing demand for highly skilled data scientists. Education need to closely follow, if not anticipate, these developments. The importance of data-analytic skills has been highlighted in numerous official documents (see e.g. conclusions from European Council Summit from October 2013: EUCO 169/13) as well as various reports (e.g. McKinsey Global Institute "The next frontier for innovation, competition, and productivity", 2011 lub Nadolny, M. "Data Economy: Gospodarka oparta na danych"). Apart from commercial importance on the market data analysis is becoming a crucial part of modern governance, or rather "evidence-based governance" – a term signifying the role of data in public institutions and governments. Policy making calls for data-based planning as well as data-driven evaluation practices. These tendencies, again, call for a workforce skilled in processing, analyzing, and communicating knowledge based on, or at least confronted with data. Evolution of modern science including, but not limited to, mathematisation of social and economic sciences made data analysis a crucial part of a curriculum of contemporary scholar. Thus data science is crucial in the Science itself.

The demand for data analytic skills is very broad. This obviously poses a challenge for educators. The WSAD program is to provide an answer to that challenge focused on the field of social and economic sciences as well as employees of research & development (R&D) institutions and companies dealing with social or economic data.

KAPITAŁ LUDZKI
CZŁOWIEK – NAJLEPSZA INWESTYCJA!
UNIWERSYTET WARSZAWSKI
ICS RUG / UU / RU
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY

Projekt współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

The educational offer of courses in statistics and data analysis is very fragmented. Different university departments have their own programs, often very diverging. This holds both when comparing different departments of the same university (e.g. Psychology, Sociology, and Economics), but also similar departments (say, Psychology), but at different universities. Lack of contact between different disciplines create a spurious impression of field-specific specialization and barriers for the flow of ideas between the disciplines. At the same time, statistics as a field of mathematics focused on, in general, dealing with data-related decision-making, is often thought in a way which is not focused on practical skills, including the use of modern data analytic software tools.

Above observations lead to the conclusion that WSAD general educational program:

1. Should have an interdisciplinary character. The choice of data analytic method or model should be a consequence of an analytical structure of the research question at hand, be the question scientific or not. It often turns out that research questions posed in different sciences in fact have the same analytical structure. Education in multidisciplinary groups facilitates learning and highlighting such commonalities.

2. Should focus on practical skills rather than theoretical knowledge. Theoretical knowledge is crucial and should be provided as a foundation for particular solutions. Practical skills should involve learning state-of-the-art software tools for data analysis.

3. Include some forms of remote participation.

4. Attractive in form and content.

2. Target groups

The target groups of WSAD educational activities are:

- PhD candidates at the University of Warsaw (UW).
- PhD candidates at other Polish universities.
- Persons outside of academic community employed in R&D sector.

We explicate two main reasons for the above choice of target groups:

1. Supporting educational potential of UW through the development of analytical skills of young academic employees.

2. Supporting the education of R&D personnel on the market. Acknowledging the human capital demands of the data-based economy as well as opening the access to knowledge as a public good.

Enrollment in WSAD educational program will require basic statistical knowledge. Recruitment will be conducted according to principles of gender equality.

Demands and expectations of target groups were identified based on:

- Survey conducted by Pracownia Ewaluacji Jakości Kształcenia UW. PhD candidates were asked to evaluate the quality of their courses. Mean score when evaluating "course attractiveness" was 3.1. For "course usefulness for own research" mean score was 2.9 (7-point scale 1="definitely bad", 7="definitely good") (PEJK, 2011; sample size 576, 2/3 women).
- Own research. UWs educational offer for PhD candidates lacks analytical component.

The creation of WSAD educational program, highlighting practical skills, is a response to an expressed demand for courses that are relevant for candidates' own scientific research.

Additionally, before the start of the summer school, a survey will be conducted to reveal detailed individual preferences of the participants in terms of the topics of individual workshops. The results will allow to, among other things, optimally assign participants to summer school courses.

3. General educational goals

The general goal of PhD programs, in particular at University of Warsaw and at the ICS, is to turn out high quality researchers and university faculty. WSAD educational program in quantitative methods should supplement this goal by providing ample opportunities for students to develop skills in the following areas:

- Theoretical attitudes and insights
- Research skills
- Academic skills
- General work orientation
- Miscellaneous skills

Theoretical attitudes and insights

A broadening of the scientific horizon via the research projects of other participants

Enlarge theoretical and substantial insights trough the research projects of other participants

Discovering new interdisciplinary, methodological and statistical approaches to research questions in social science.

Gaining insight in the theory of advanced statistical techniques such as R, Social Network Analysis, and Structural Equation Models

Gaining insight in the theory of text analysis for the social sciences

Research Skills

Ability to (re)construct theoretical models and to generate testable hypotheses from theoretical models

Ability to understand and apply advanced techniques of measurement and analysis;

Ability in modeling socio-economic phenomena en problems

Ability in the application of R

Ability in the application of social network analysis

Ability in the application of text analysis for the social sciences

Abiltiy in the application of structural equation models

Ability to critically assess and discuss statistical analyses designs and strategies in social science research projects, both one's own project and the projects of others.

Academic Skills

Ability in presenting scientific outcomes using new visualization techniques

Enhancing clear Academic writing for various audiences

Presenting research outcomes for various audiences

General work orientation

Ability to contribute to team work and to the production of collective goods

Adequate time management to reach the goals set in limited space and time

Learning by doing

Learning from colleagues

Experience in an international contexts

Participation in the international research network in the domain of one's own project.

Added Skills

Application of new skills to their own research questions as well as to new questions

Meet other PhD students with varying research topics, designs and strategies

Take home new skills and insights

Develop an international network through collaboration with PhD students from other countries

Becoming an international academic in a cooperative and constructive context

4. EDUCATIONAL CONTENT

The project consists of three phases:

1. Introductory Data Analysis course (online)
2. Summer School
3. Research Workshop Extension (online)
4. Massive Open Online Course (MOOC)

Introductory Course

The aim of the "Introduction to Data Analysis" course is to prepare the student for using quantitative methods of data analysis in analysing social survey data. Students will gain both a general overview of probability and statistics, and detailed knowledge regarding methods of acquiring quantitative information in management and social sciences. A practical aim is for students to become capable of both performing statistical analyses, and using results of data analyses.

The core knowledge is presented in 6 lessons **defined by the basic skills important in the data analysis**, designed in such a way, that they do not require prior mathematical training.

- MODULE 1: Describe an interesting phenomenon with variables
- MODULE 2: Think in a probabilistic way
- MODULE 3: Use samples to infer about populations
- MODULE 4: Determine relationship of two quantitative variables
- MODULE 5: Look for interactional impact of two independent variables
- MODULE 6: Analyze the third variable effect

Summer School of Data Analysis - Introduction

Following the introductory online course of data analysis, a 12-day intense training programme will take place in a contracted training facility. During the course, each of 120 participants will take part in at least 96 hours of workshops and lectures. The workshops will cover a range of topics in the area of quantitative and qualitative research, statistics and methodology.

WARSZAWSKA SZKOŁA ANALIZY DANYCH

The aim of the Summer School of Data Analysis is to prepare students for practical use of statistical data analysis methods in analyzing social survey data. Students will gain both a general overview and detailed knowledge regarding methods of quantitative data analysis methods, data visualization methods and tools, as well as interpreting the results of data analyses.

The educational programme is divided into a number of lectures and fifteen different workshops, each focusing on a different aspect of social data analysis, and each conducted by a renowned expert in the field. The workshops will prepare students for the use of the general linear model (GLM), as well as hierarchical models and structural equation modeling to analyze data from social research. Students will gain detailed knowledge of path analysis, confirmatory factor analysis and structural modeling. In addition, students will learn tools for structural modeling, and how to construct composite indicators and test their quality by using, e.g. Cronbach's Alpha.

Separate track of workshops focuses on gaining proficiency in the use of statistical analysis software: R and SPSS. For both R and SPSS, introductory and advanced courses will take place during the Summer School.

List of Summer School workshops:

1. Introduction to R for beginners
2. Going Deeper into R: Data Processing and Basic Statistical Modeling
3. Spatial Modeling of Socio-economic Phenomena
4. Text analysis for the social sciences
5. Introduction to Social Network Analysis
6. Introduction to the Statistical Analysis of Social Networks
7. Introduction to the Statistical Analysis of Dynamic Social Networks
8. Data visualization techniques. Theory and Practice
9. Introduction to data processing with SPSS
10. Advanced data processing with SPSS
11. Discovering Patterns in Your Data – Generalized linear mixed models and other techniques with R
12. Wiggles and curves: The art of data exploration
13. How to write clearly
14. Structural Equation Models
15. Advanced Data Analysis 3/Research Track

On top of the above workshops, a range of lectures will be available for the students.

60 Students taking part in the "Academic Track"/"Research Workshop" will have an opportunity to analyse real-life social survey data. The aim of the "Advanced Data Analysis 3" course is to prepare a ready-to-publish research report in English, based on empirical data from social surveys: ISSP (International Social Survey Programme), ESS (European Social Survey) and PGSS (Polish General Social Survey).

Students will conduct analyses of real data from social surveys: ISS , ESS and PGSS . Students will be working in groups of five, analysing real data and preparing articles for publication in English. The topic of each group's work will be established with an advice of the person who conducts the workshop .

Research Workshop – Online Extension

Following the Summer School training, 60 participants of the "Academic Track", will have an opportunity to work over the course of 6 months, with 15 teachers, on finalizing the research papers developed during the course of the Summer School, and further expanding the knowledge and skills of quantitative data analysis.

A dedicated e-learning platform will be available for the 60 participants of the Research Workshop, in order to make it possible for them to finalize their articles and make them publication-ready.

The "Research Workshop Extension" online course will also provide participants with an opportunity to prepare for a final exam, required to receive a certificate of completion of the Warsaw School of Data Analysis (WSAD).

Massive Open Online Course

Activities of the summer school as well as research workshop, including its online extension. Have a focused character: take place in a specific place and time (summer school) or are focused on development academia-specific skill sets (research workshop). To extend the reach of WSAD educational activities two open internet courses (MOOC) are planned. They will of a more popularizing character, accessible by non-specialists. The content will focus on general skills of reasoning and communicating with data and data-related products.

5. Educational methods

Educational methods employed at WSAD will include practical courses in data analysis, and will involve using a computer both by the instructor and by the participants.

Course forms

For optimal use of knowledge and skills of course instructors and to ensure the maximal educational effect has been obtained, course forms and other didactic methods used should be appropriately

chosen and differentiated. The differentiation should be based on the specifics of the topics covered as well as the specifics of the particular group of participants. Below we present several possibilities:

Large group workshop

Teaching large groups of participants enforces more lecture-like method. The interaction with participants is limited as divergent character of potential questions will make addressing them very time-consuming.

Small group workshop

In small groups more intense interaction with participants is possible. Different didactic methods can be chosen, including data and exercises. It is also possible to conduct larger number of exercises by the participants on their own. This format of a course is more appropriate for more advanced participants.

Didactic methods

Below we propose several options of didactic methods.

Single block for learning a single topic

If possible, the majority of topics covered during a course should be presented according to the following block structure:

- Presentation
- Examples
- Live examples of analysis or an exercise performed by the instructor, or an exercise performed by the students
- Summary

It is important to split the material into small chunks and to maintain a logically coherent structure of individual parts.

Such a educational block should take from 45 to 90 minutes and should be self-contained. This allows to communicate knowledge in small, comprehensive parts. Proposed structure ensures that the topic will be understood thanks to the presentation and examples, and because of the possibility for the student to conduct the analysis by himself, or at least to see how such an analysis should be performed. The summary ensures that the most important conclusions of the educational block are explicitly stated and not overlooked.

Coherent structure of the block makes comprehension and understanding much easier.

Live presentation by the instructor or individual work based on a prepared script is more beneficial for the **beginners**. Such an approach facilitates making initial learning steps such that the students do not have to contemplate each individual step for too long.

In case of more **advanced** students, it is more beneficial to perform the tasks individually under the supervision of the instructor.

Presentation of a topic

We suggest the following format for preparing a presentation:

- Short presentation about the key issues.
- Presentation of how an analysis should be performed with the software tool used (e.g. necessary R expressions).
- During the presentation the students should by motivated by the instructor to share their own experiences with the group.

Exercise and its solution

We suggest the following workflow when presenting exercises:

1. Instructor's presentation of the problem to be solved.
2. An attempt at solving the problem by the students on their own computers (optional).
3. Instructor's supporting the students when they work on their own (optional).
4. Instructor's presentation of a pre-prepared solution.
5. Comments on alternative interesting approaches to solving the posed problem by the students (optional).

Solving an example problem using "toy data"

Using small and appropriately selected data sets (real or simulated, so-called "toy data") often makes the problem more accessible. Complexity of typical large datasets often causes the key aspects of the statistical method to be blurred by data-specific problems (number of variables, coding, missing data, etc.).

For beginners, analysis of complex real data often hinders an effort to present a statistical model discussed. Problems omnipresent in large real datasets are often so extensive, that understanding the model itself can be very difficult.

Live problem presentations

We suggest preparing live exercises that are solved by the instructor in front of the group. Such a form makes understanding the problem and possible approaches to solving much easier than a lecture-like presentation. The students will be able to experience first-hand that some of the operations are not that complex to perform. This has also a positive motivational effect.

Self-performed exercises

In small workshop groups the best form are self-performed exercises solved under the supervision of the instructor. Each student receives an individual feedback on his work, or helpful assistance if necessary. In larger groups the students can be divided in smaller groups and perform an exercise collectively.

Working in subgroups

Completing tasks that require conceptual effort or that require divergent skills are ideally performed in subgroups of 3-5 students. Such a format is very good in case of exercises in data analysis. In each group the students can select members responsible for particular phases/subtasks of the analysis (e.g. specifying the goals, programming, presentation of results, etc.).

Questions and Answers

- The instructor should encourage the students to ask questions during the course at all times, not only at the end of each block/section.
- Possibility of asking questions facilitates the interaction.
- As a consequence, we are creating a much better flow of knowledge from the instructor to the students.
- Obviously, it is necessary to appropriately manage time devoted to asking questions and providing answers.

Educational materials

- The students should receive educational materials in the form of printed hand-outs of the presentations as well as in electronic form (e.g. slides and R scripts).
- They should also receive data files used throughout the course.
    - Data used in exercises should be selected appropriately to the topic.
    - To explain features of R language use small real or simulated data.
    - For presentations of statistical methods small and well-behaved real data are optimal.
- Thanks to above features the students will be able to repeat the exercises on their own after the course and use the data in their own work.

6. Organization of education

Main objectives of preparatory e-course:

- To deliver to the students the necessary minimum of statistical knowledge that enables to effectively follow more advanced courses that will take place during the summer school.
- WSAD participants will come from different universities and companies, and with different disciplinary backgrounds and expertise. Therefore it is crucial to "equalize" the level of statistical knowledge before the summer school.

Main objectives of the summer school:

- Education in modern statistical methods of data analysis.
- Participation of instructors from UW, ICS, and other internationally recognized universities.
- Integration of scientific and business communities around the issues of data analysis.
- Networking among the participants
- Courses taught in English.

According to the PEJK survey mentioned earlier, 55% of PhD candidates have some employment in parallel to doctoral studies. This may constitute a barrier in participating in the summer school because of the necessity of commuting to WSAD summer school courses. This barrier will be mitigated by the remote courses.

The summer school instructors will come from:

- scientific community, in particular: University of Warsaw, ICS, and other internationally recognized universities.
- business community.

7. Verification of educational effects

Summer school education will be verified and evaluated through:

- Oral exam, positive effect of which will result in obtaining WSAD certificate.
- Passing individual courses will allow to obtain ECTS points.
- Research reports prepared in small groups of the academic track.

8. Training cycle evaluation

Each of the training cycle phases will be evaluated. The evaluation forms should contain questions regarding:

- Self-assessment of the level of prior statistical knowledge,
- Assessment of the training programmes and their implementation,
- Assessment of the increase of one's knowledge and skills in the field of data analysis,
- Assessment of one's expectations regarding WSAD training cycle phase being met,
- Indication of the most interesting and the least interesting elements of the training cycle,
- Assessment of the organizational aspects of training cycle,
- Recommendations for the future.

It is suggested that the evaluation should be conducted with the use of paper-and-pencil questionnaire in order to assure the highest response rate possible.

9. Outline program of preparatory course „Introduction to Data Analysis"

Course summary

The aim of the course is to prepare the student for using methods of statistical data analysis in analysing social survey data. Students will gain both a general overview of probability and statistics, and detailed knowledge regarding methods of acquiring quantitative information in management and social sciences. A practical aim is for students to become capable of both performing statistical analyses, and using results of data analyses.

General Information

A necessary minimum of the methodological knowledge has been extracted for the course, the understanding of which is essential for statistical inference and data analysis.

The core knowledge is presented in 6 modules (lessons) **defined by the basic skills important in the data analysis**. They do not require mathematical training, and even those, who have an aversion to mathematics should experience no difficulty in using our original method of teaching based on the acute selection. With such an approach, **statistics is treated instrumentally - as a tool helping to answer interesting research questions.** We've tried to reduce the number of mathematical formulas to a minimum.

To make the learning process easier:

- we have assumed an equal sample size in all the formulas, because analyses of large data sets are performed by statistical packages;
- we often omit the indexes in the summation formulas, replacing them with a commentary,
- the examples relate to ridiculously small samples in order to simplify the calculations. Instead of a high-tech materials, which can be found in thick and widely available study books, **we introduce the important statistical concepts in a form of simple – interesting (not only in our opinion) examples.**

Similar to the process of learning a foreign language, **basic concepts should be memorized** even if at that moment students might not understand why they are important.

One cannot learn statistics in one week - human brain needs time to make the absorbed information implanted in our mind. Statistics cannot be learned randomly. Unlike in the humanities, systematic study is the key to success in study of statistics. This does not mean that the student cannot proceed to the second level (module 2), when he understands less than 100 % of the material in the first level – the full understanding may come much, much later. In such a simplified statistics course, students will have to take a lot of things for granted. Before he can get to the next level, student will need to memorize symbols, definitions and examples – regardless of whether he understands them fully or partly. Students who feel they cannot make it to the next level, until they understand everything perfectly, encounter a lot of problems, because our didactic approach, in its definition, assumes many simplifications. For example we don't discuss in length assumption for all analyses we introduce. We neither prove any theorem used nor discuss all the options. The course contains only the bare minimum the student needs to start an adventure with the data interpretation, but it also covers more advanced, useful topics, like different type of "the third variable" effect.

Even the best texts or lectures cannot replace ones' own learning effort. Statistics is like playing an instrument or a ball – you cannot master it only by watching how the others do it – you have to practice! Although our goal is to analyse the data using statistical packages, we will still ask our students to perform calculations of simple examples by hand, to make them understand the essence of statistical reasoning. The examples are designed to make the calculations as simple as possible (For example: it's rather unlikely to happen in real data, that the standard deviation is an integer value, as one can see in some tasks prepared for the students).

To analyse large data sets, a statistical package is required, as well as the ability to use a computer and the chosen program. The printouts we are going to use come from the analyses performed by the Statistical Package for Social Sciences (SPSS). It is a very powerful tool for data analysis which, therefore, requires from beginners (but not exclusively) the ability to ignore a lot of information. The printouts contain a lot of statistics, that are unneeded for a beginner. A novice user wanting to understand everything that SPSS prints out, is prone to be overwhelmed. We do not assume that a

graduate of this course will be able to analyse data on their own. Rome wasn't built in a day. The goal will be achieved if he/she is able to carry out the selected analyses presented throughout the course.

The concept of the course is largely based on G. Wieczorkowska and J. Wierzbiński's handbook "Statistics: From Theory to Practice", Warsaw: Scholar 2011, but we've used also many other source materials, both printed and available on the Internet. We tried to make the definitions used in the course coherent with the on-line study book Online Statistics Education: An Interactive Multimedia Course of Study (http://onlinestatbook.com), but it was not always possible. In case of discrepancy between this study book and the course materials, the definitions given in the materials of the course are the base for the answers to quiz's questions.

Practice track

On top of the theoretical knowledge presented as the core of the "Introductory Data Analysis" course, additional PRACTICE TRACK has been prepared for more advanced students, especially for those, who would like to take part in the "Academic Research Track" of the Summer School. Within this track, students are presented with four datasets, prepared as subsets of social research surveys:

- **Social Diagnosis survey**
  The Social Diagnosis is a project to support our diagnostic work with detailed data derived from institutional indicators concerning households and the attitudes, mind-sets and behaviours of their members. We investigate households and their occupants aged 16 and above using two separate questionnaires.

  The project takes into account all the significant aspects of the life of individual households and their members, both the economic (income, material wealth, savings and financing), and the not strictly economic (education, medical care, problem-solving, stress, psychological well-being, lifestyle, pathologies, engagement in the arts and cultural events, use of new communication technologies as well as and many others). In this sense the project is interdisciplinary, drawing on the work of the main authors of the Social Monitoring Council (*Rada Monitoringu Społecznego*) and a team appointed by a Council of experts made up of economists, a demographer, a psychologist, sociologists, an insurance specialist, a health economics expert and statisticians.

- **International Social Survey Programme**
  The ISSP data comes from the *International Social Survey Programme* (ISSP) project – "Religion" subset. ISSP contains subject-specific modules, present only in selected waves of the survey. The "Religion" module was first conducted in 1991, and then it was repeated in 1998 and 2008. The data you will be using comes from many countries and from all three editions of the survey.

- **PolPan Survey**
  The Polish Panel Survey POLPAN is a unique program conducted by Polish Academy of Sciences. It consists of panel surveys carried out since 1988 in 5-year intervals, and focuses

on describing social structure and its change during the last 25 years in Poland. To date, there is no other research worldwide, in which life histories of individuals from a nationally-representative sample of adults would be collected for such a long time span, reaching 25 years, while also opening the possibility of panel research on the renewal samples of the young.

The POLPAN study is unique also with respect to the scope of the collected data. Socio-demographic information of respondents and their families is supplemented by items on socio-political attitudes, some of them present in cross-national studies. At the same time, POLPAN includes the nonverbal Raven test, which captures intellectual flexibility (an essential IQ component), and the Nottingham Health Profile, which measures certain aspects of physical and mental health.

The range of problems covered in POPLAN is very broad. Questionnaires deal with interdisciplinary problems that can be labelled as follows: the “old” and “new” elements in the social structure; changes in the class structure; social mobility; differences in the standard of living; the process of adaptation to a market economy; the impact of the location in social structure on political attitudes and behaviour; perception of social conflicts; winners, losers, and the European integration; health issues; emigration.

- **European Social Survey**

  The European Social Survey (ESS) is an academically driven cross-national survey that has been conducted every two years across Europe since 2001.

  Following an application to the European Commission which was submitted by the UK on behalf of 14 other countries, the ESS was awarded ERIC status on 30th November 2013.

  The Director of the ESS ERIC is Rory Fitzgerald and the ESS ERIC headquarters are at City University London.

  The survey measures the attitudes, beliefs and behaviour patterns of diverse populations in more than thirty nations.

Within the practice track students will be presented with a set of data analysis tasks to be solved using the given data sets. To assist students with these tasks, example solutions have been prepared by experts working with these datasets on a daily basis.

Being fully aware, that some of the students might have a practical knowledge of SPSS not good enough to tackle the tasks, two on-site workshop meetings with students are scheduled, aimed to improve students skills with the SPSS software, to discuss the example solutions to practical data analysis task, as well as to provide the students with an opportunity to present their data analysis concepts, based on one of the above data sets.

Finally, many topics covered in the core six modules of the course, are directly applicable to the data analysis problems in the "Practice Track", and students who undertake to tackle these problems are referred to these materials to assist them with achieving better solutions.

All "Practice Track" problems, materials, and on-site meetings are open to students following the "Core Track" of the course.

Estimated Workload

The expected workload for the IDA workshop conducted in a foreign language consists of:

Core Track:

- 10h – preparation,
- 30h – Online classes
- 20h - Conducting data analyses

Practice Track:

- 12h – Classroom workshops
- 20h – Conducting data analyses
- 8h – Preparation and presentation of a data analysis project

TOTAL: 100h (60h core track + 40h practice track)

Descriptions of learning modules

The core track of the course consists of six modules. The modules cover the following topics:

MODULE 1: Describe an interesting phenomenon with variables

- Variable types: dependent, independent
- Measurement scales
- Experimental & correlational research

Estimated time: 10h

MODULE 2: Think in a probabilistic way

- Variable distributions (frequencies); histograms
- Empirical and theoretical probability distribution
- Interpreting probabilities using various types of distributions
- Central tendency and dispersion measures
- Normal distribution

Estimated time: 10h

MODULE 3: Use samples to infer about populations

- Distribution of a statistic, Central Limit Theorem
- Testing differences between two means (F test)

Estimated time: 10h

MODULE 4: Determine relationship of two quantitative variables

- Multiple and simple regression analysis
- Interpretation of residuals, coefficients of correlation and determination

Estimated time: 10h

MODULE 5: Look for interactional impact of two independent variables

- Multi-factor analysis of variance
- Interactions in regression analysis

Estimated time: 10h

MODULE 6: Analyse the third variable effect

- Mediation and moderation effects
- Confounding and suppression effects
- Adjusted R^2

Estimated time: 10h

LEARNING OUTCOMES:

Students will learn to:

- understand and interpret results of statistical analyses (correlational and experimental)

- understand the ideas behind sample-based statistical reasoning
- read probabilities from empirical and theoretical distributions of variables and to infer on them
- interpret multivariate data analyses (regression analysis, analysis of variance, interaction effects, time series)
- match an appropriate method of analysis with a research problem and measurement type.

10. OUTLINE PROGRAM OF THE SUMMER SCHOOL

Course title	Description, goals and requirements	Estimated number of course hours
Spatial Modelling of Socio-economic phenomena	Main goal of this course is to introduce students with the methods and applications of spatial modelling in social sciences. Using spatial statistics and econometrics together with mapping allows for: 1. visualization of the modelled data 2. observing and measuring spatial clusters, hot-spots etc. 3. observing spatial regimes (heterogeneity over space; clubs) 4. improving quality of estimators (unbiased and consistent) in regression 5. controlling for spatial diffusion process, spatial interactions and spatial correlation Course will focus on: 1. short introduction to spatial packages in R software, on-line collections of maps and data 2. visualization on map of socio-economic phenomena to catch spatial patterns (spatial autocorrelation, spatial heterogeneity) 3. spatial statistics to detect spatial clustering, spatial regimes and autocorrelation 4. spatial weights matrix – building neighborhood relations 5. spatial dependence models – basic econometric modelling, including spatial relations Strongly required: basic knowledge of R software (data import, operations on data, simple plotting)	25

Introduction to R for Beginners	Statistical system R (http://www.r-project.org/) is a powerful and free tool. It is applied virtually in all fields of science and business. You will get very practical skills. After the workshop you will be able to analyse single-handedly your own data using R: reading them, processing, and writing back the results. We will work hands-on with computers. Basically no lectures. Specifically, you will learn: • Working with R: how to deal with it, help system, further information • Using RStudio • Basics of R language • Reading data in different formats • Summaries and descriptive statistics • Basics of plotting • Writing data and results	30

Going Deeper into R: Data Processing and Basic Statistical Modeling	Prerequisites: Completed "Introduction to R for Beginners" or having the equivalent knowledge, basic understanding of linear regresssion.	30

During this workshop you will get important practical skills: processing data for further analysis and facilitating data analysis with help of R scripts. You will understand how to work effectively with data in R and how to build first linear regression models. These altogether will build up a strong basis for using R in your work and learning more of things you need in future. We will work hands-on with computers. Basically no lectures.

Specifically, you will learn:

- Basics of data wrangling: merging, subsetting, transformation, and aggregation
- Types and properties of objects in R
- Vectorization and indexing
- Working with results of statistical procedures
- Linear regression
- Basics of writing R scripts

Data processing with SPSS	The main objective of the course is learning to work effectively with data sets with SPSS. The course covers the basic issues related to the operations on the data files, data processing, diagnostic and description. The idea is that each WSAD participant will be able to concentrate on the undertaken problem without thinking about “how it was done”. The basic criterion for effective working with SPSS is using syntax. Saving steps of analysis (syntax commands) aligns the structure of our work. Syntax helps the description and interpretation of results. Editing text commands allows to quickly define analyzes of a large number of variables and define fast and reproducible analyzes of high complexity. Topics. 1. Using Help F1 2. File operation and working directory. Data files. Import external data. Selections and filtering. Merging. Data-set description. Output files. Export to html, txt, doc, xls. Syntax files. Pasting. Running. Auto saving commands. 3. Data description. Viewing of meta-data. Missing values. Display command (variable label list). Displaying labels or/and variable names or values in outputs. Descriptive statistics. Summaries. Examines. 4. Tables. Distributions. Cross-tabulations. Special tables. 5. Recode and compute commands. Transformation functions. Quantiles. 6. Basic analysis. Chi square. Conditional means. Analysis of variance. Regression. 7. Charts. SPSS native. Export of output values and using external graphics.

Text analysis for the Social Sciences

30

Text analysis is a data collection technique just like many others. Usually there are no direct responses to a (research) question like agree, agree a little bit till disagree with an argument. The answers are found in a piece of text, where the investigator should try to find them. In general these are not answers to direct questions (like open-ended questions in a survey), but the texts contain opinions or facts on a specific issue that are relevant for an investigator. This results especially in data that allow describing certain developments in time, as development of democracy based on information in editorials in newspapers. It is also possible to describe differences between groups, view of women versus men on certain issues.

The workshop is meant as an introduction into quantitative text analysis. In there several qualitative elements will receive attention.

Generally social scientists use for their data collection some kind of survey research. Here the data are collected from individual respondents. However, there are several other methods to collect data. Text analysis is one of these. In such studies data might be sampled from individual respondents (via diaries, letters, etc.), but more often the data come from articles in newspapers, documents from a government, minutes of meetings, discussion groups, and so on. These articles usually are assumed to represent the opinions that are hold by a broad public. At the end of the workshop the participants should have an idea of how to perform a simple text analysis study and must be able to evaluate more complex studies.

In the workshop a short overview is given of the actual state of the art of text analysis mainly within research in sociology, political science, and communication studies. This overview is based on the literature. After being confronted with the traditional content analysis or instrumental thematic text analysis where concepts or themes are considered from the perspective of the investigator (Holsti, Krippendorff, and Weber) the student will learn about the representational form of analysis, where the concepts are considered from the point of view of the sender of the message. The instrumental and representational approach are not only applied to the thematic analysis (with focus on frequency of occurrence and co-occurrences of concepts

Discovering Patterns in Your Data – Generalized linear mixed models and other techniques with R	Within the course I will present the general linear model with particular emphasis on mixed experimental design analyses. I present material in a way that will give a new intuition about what ANOVA and regression do and how these techniques can be used in research settings - students who claim solid knowledge of ANOVA and/or regression still report learning quite a bit from the course. We will challenge the assumptions about data independency. The topics covered during the course include repeated measures, hierarchical models, FlexMix analysis and dyad analysis. Data analysis is an exploratory process where the end product is a description of what happened in the study. In this course we will learn simple procedures to uncover what the "data are telling us they focus on building skills and highlight subtle distinctions in the concepts we cover. Prerequisites: An introductory statistics course is a must. You should review the material on hypothesis testing, measures of location, measures of spread, and exploratory data analysis. All examples and homework will be prepared in R, so the willingness to learn R is necessary in order to complete the problem sets.	40

Data visualization techniques. Theory and Practice	Prerequisites: Basic knowledge about R (it is enough to know how to read data and plot anything). Basic knowledge about statistics (what is mean/median/quartile, fractions, linear regression) The goal is to improve participants' data visualization skills. During lectures following topics will be discussed: major points in history of statistical graphic, how we perceive elements of the chart, how different data characteristics can be presented graphically and why some ways are better than others, why bad charts are bad, how to create a chart in R with ggplot2 package. During practice sessions participants will work in groups. Each group will work on a data visualization of some relations between variables from an international survey (most likely from PIAAC study). Groups will start their projects with paper and pencil in order to create a prototype / brainstorm ideas. Then they will work with R to create a data based version of proposed prototypes. And finally they will tune the visualization in Inkscape.

Structural Equation Modelling with Mplus	Structural equation model (SEM) represents a general statistical approach to the evaluation of theoretical models fit to empirical data. SEM with latent variables embodies simultaneous equations w multiple exogenous and endogenous variables (path models), along with measurement error models (confirmatory factor analysis). Thus SEM represents a synthesis of methods developed in econometrics and psychometrics. This course introduces methods and applications of SEM with the M*plus* software. The course will provide exposure to the fundamentals and extensions of SEM, with an emphasis on applications in applied research settings. The course covers such topics as data requirements, M*plus* syntax, specification and identification of models with observed and latent variables, multigroup models, models for longitudinal and clustered data, and model estimation, testing and reporting. The course will be of interest to those who plan to utilize the general structural equation model with latent variables or its specializations. Familiarity with elementary matrix algebra will be useful, though not essential, for understanding M*plus* syntax. Coursework Prerequisites: A course on intermediate statistics and/or practical experience with regression and factor analysis.	30

Introduction to Social Network Analysis	Prerequisites: Taking advantage of all the features of the course will require knowledge of R on the level of "Going deeper into R...". R-unrelated topics should be accessible to everybody. Social Network Analysis (SNA) is an approach to study groups of actors, be they individuals or organizations, through the analysis of relations between them that combine into complex networks. Relations linking actors can come in various forms, such as kinship, friendship, collaboration, information seeking, authority, but also co-membership in associations or having something else "in common". General goal of SNA is to understand the functioning of a group of actors by studying the pattern in which they are connected to one another. The workshop will introduce SNA and demonstrate various basic techniques using R. In particular, the following topics will be covered: 1. Types of research problems addressable with SNA 2. Social network data collection techniques 3. Network data representations 4. Overview of various descriptive measures characterizing (a) actor's position in a network, (b) groups of actors, and (c) network as a whole. Items (3) and (4) will be illustrated using R.	30

Introduction to the Statistical Analysis of Social Networks	Social network studies can be designed in many different ways and so the number of social network analysis methods is immense. In this course, participants will get an overview of state-of-the-art methods for the statistical analysis of social networks and acquire practical skills in two well-established methods for the analysis of complete social network data (as opposed to ego-centered networks). First, exponential random graph models (ERGMs) and their estimation with the software package Statnet will be introduced. ERGMs can be used to analyze cross-sectional network data that was collected at one point in time. Second, stochastic actor-oriented models (SAOMs) will be introduced and how they can be estimated with the software package RSiena. SAOMs can be used to model longitudinal network data that was collected repeatedly over time. A number of social network research questions will be formulated and tested empirically on data sets from social network school studies. Exemplary research questions are: Do friends of friends tend to be friends (transitivity)? Do school children rather have friendship relations with other children of the same gender (homophily)? Do those with many friends attract even more friends over time (preferential attachment)? Day 1: – Recap: Social Network Analysis in R + Practical – Introduction to statistical methods for social networks Day 2: – Introduction to Exponential Random Graph Models (ERGMs) – Practical with Statnet Day 3: – Introduction to dynamic network analysis with SIENA, part 1 – Practical with RSiena Requirements: Participation in the course “Introduction to Social Network Analysis” in week 1 or basic knowledge in social network analysis and R. A computer with the latest version of R and the software packages igraph, Statnet and RSiena	30

Introduction to the Statistical Analysis of Dynamic Social Networks	In this course, participants will deepen their understanding of stochastic actor-oriented models (SAOMs) and the RSiena software. The course builds upon the course “Introduction to the Statistical Analysis of Social Networks”. SAOMs can be used to model longitudinal network data that was collected repeatedly over time (longitudinal social network data). RSiena is a program for the statistical analysis of network data, with the focus on social networks. Networks here are understood as entire (complete) networks, not as personal (ego-centered) networks: it is assumed that a set of nodes (social actors) is given, and all ties (links) between these nodes are known - except perhaps for a moderate amount of missing data. In particular, the course focuses on the evolution of social networks, the co-evolution of social networks and individual behaviour variables, the co-evolution of multiple one-mode networks and the co-evolution of one-mode and two-mode networks. All these topics will be illustrated in practicals that make use of empirical social network data. Examples of these applied topics are ethnic homophily in schools, the co-evolution of smoking and friendship and the co-evolution of gossip and friendship. Day 1: – Introduction to dynamic network analysis with SIENA, part 2 – Practical with RSiena Day 2: – Analysis of co-evolution of networks and behavior – Practical with RSiena Day 3: – Analysis of multiple one-mode networks – Analysis of one-mode and two-mode networks – Practical with RSiena Requirements: Participation in the course “Introduction to Social Network Analysis in week 1” or basic knowledge in social network analysis and R. Participation in the course “Introduction to the Statistical Analysis of Social Networks” A computer with the latest version of R and the software package igraph and RSiena	30

Wiggles and curves: The art of data exploration	Science is art using numbers. Alas, much of the art has been lost in the social sciences to the mechanics of traditional statistical practices. By focusing on statistically significant differences among central tendencies of aggregates, these practices can obscure more subtle features of social science data that might offer useful insights about the people who generated the data. The purpose of this workshop is to learn some simple statistical techniques that explore data subtleties — techniques that deviate from traditional statistical practices in three ways. First, the techniques do not estimate how well samples of data might generalize to populations. Instead, the techniques indicate how well predictions generalize to samples of data. Second, the techniques do not require data to be aggregated across people before they are analysed. Instead, they allow data to be analysed individually before they are aggregated. Third, the indicants generated by these techniques are remarkably easy to calculate and to interpret. People wishing to take this workshop should know enough about traditional statistical practices to be skeptical of them; a course or two in omni-mega-multivariate statistical analyses using commercial software such as SPSS® will do. Participants will practice the alternative techniques using data from the European Social Survey and the International Social Survey.	30

How to Write Clearly

Most social science literature is boring and obscure. It need not be. Most social science is fascinating. Sadly, however, most social scientists never learn to write about it well. The purpose of this workshop is to teach participants some of the techniques for writing clearly about social science. We will discuss the psychology of writing, and what distinguishes good writing from bad. We will examine published examples of good and bad writing, and show how to improve some of the bad stuff. Participants will also write short paragraphs in class, some of which will be shown anonymously to participants for class critique. By the end of the workshop, each participant will write a brief research report. People wishing to take this workshop should know enough English to write it badly or better. They should not be shy about making grammatical, stylistic or rhetorical errors. A good sense of humour would also be helpful.

Advanced Data Analysis 3	**Prerequisites**:	36

1. Completion of “Introduction to Data Analysis” course. Completion of “Social Data Analysis” 1 and 2 courses.
2. Basic knowledge of mathematics; a PC or a MAC with Excel or Open Office Calc; optionally SPSS or PSPP. Internet connection.

The aim of the course is to prepare a ready-to-publish research report in English, based on empirical data from social surveys: ISSP (International Social Survey Programme), ESS (European Social Survey) and PGSS (Polish General Social Survey).

Students will conduct analyses of real data from social surveys: ISS , ESS and PGSS . Students will be working in groups of five, analyzing real data and preparing articles for publication in English. The topic of each group’s work will be estabilished with an advice of the person who conducts the workshop .

Objectives of the workshop:

- OBJECTIVE 1 : Select an interesting research question, leading to hypotheses that can be tested on available data
- OBJECTIVE 2 : Develop a bibliography adequate for the analysis of the research problem.
- OBJECTIVE 3: Formulate a theoretical model, construct indicators, test hypotheses arising from the model
- OBJECTIVE 4: Understand and use the formal structure of a research article to describe the research problem.
- OBJECTIVE 5: Make a methodologically correct description of the results of conducted analyses.
- OBJECTIVE 6: Interpret of the results in the context of the bibliography.

II OUTLINE PROGRAM OF REMOTE WORKSHOPS

Research Workshop – Online Extension

Following the Summer School training, 60 participants of the "Academic Track", will have an opportunity to work over the course of 6 months, with 15 teachers, including three experts from abroad, on expanding their knowledge of quantitative data analysis and finalising the research papers developed during the course of the Summer School.

The "Research Workshop Extension" online course will also provide participants with an opportunity to prepare for a final exam, required to receive a certificate of completion of the Academic Track of the Warsaw School of Data Analysis (WSAD).

Training for tutors will be conducted with the use of Moodle online platform. The training will take place on campus in the form of a remote, online class.

60 hours of online workshops, for each group, will be conducted to further develop and finalise the articles and scientific papers. The task will be to realize remotely by 15 people, (refund will be made of expenditure), one expert from the ISR of the United States (in the framework of the work).

The educational platform for virtual classes will be maintained for the duration of 6 months.