

# Data Analysis in Sociology

NEW COURSE, yay!



NATIONAL RESEARCH  
UNIVERSITY  
SAINT PETERSBURG

Olesya Volchenko

[ovolchenko@hse.ru](mailto:ovolchenko@hse.ru)

Dmitry Arkatov

[darkatov@hse.ru](mailto:darkatov@hse.ru)

January 2023

# Welcome!

- This course can give you tools to get insights about data (and sell them for money/biscuits).
- Sometimes it is going to be tough and challenging; sometimes much fun and group work.
- This course is part of the Data Analysis in Sociology series and part of the core of this BA.

# Why I Am Teaching This

- The [LCSR](#) lab is a research centre for advanced methods of comparative social research. At LCSR, we endorsed R back in 2012
- I have taught courses in data analysis since 2015
- You are going to be the 8<sup>th</sup> generation of graduates with R! We keep contact with graduates working in industry and academia
- A shorter version of the course is part of the [Radboud Summer School](#) in the Netherlands

# In This Presentation

- I will give an overview of the whole course, its structure, grading, and organization.
- Questions are good. (Remember the story about the rabbit and the lion?)

# Data Analysis - Is It Worth Your Time?

Yes, because it helps you make decisions +

Draw conclusions about the information you have +

Estimate the risks of possible action +

Describe what you observe and compare it +

Analyze relationships between this and that +

Make your thinking more logical and consequential +

# Two frequent quests after graduation

*1. You get into marketing research.*

“The company has been collecting some data from its customers. Last year, the net-promoter score fell by 2 points, and your boss wants you to figure out why, and how to keep it up. You are to analyse the data, produce a report, and suggest solutions that will bring profit to the company.

*2. You enroll at a Master’s or Doctoral programme.*

“You are to complete your own independent piece of research and produce a thesis where you contribute to understanding a certain social mechanism.”

In both cases, you should know how to approach the data. By succeeding in this class, you are making yourself more valuable in any job you may want.

# Types of Questions Across Disciplines

*Sociology:* Do men really get a higher pay than women in this company, controlling for age, training, and experience?

*Marketing:* Does this new design of X sale considerably better than the previous one? What explains sales growth better — ads, distribution, or shelf share?

*Medicine:* Do patients who take X medicine will get fewer side effects if they take Y instead?

# Data Analysis - Why Care About It?

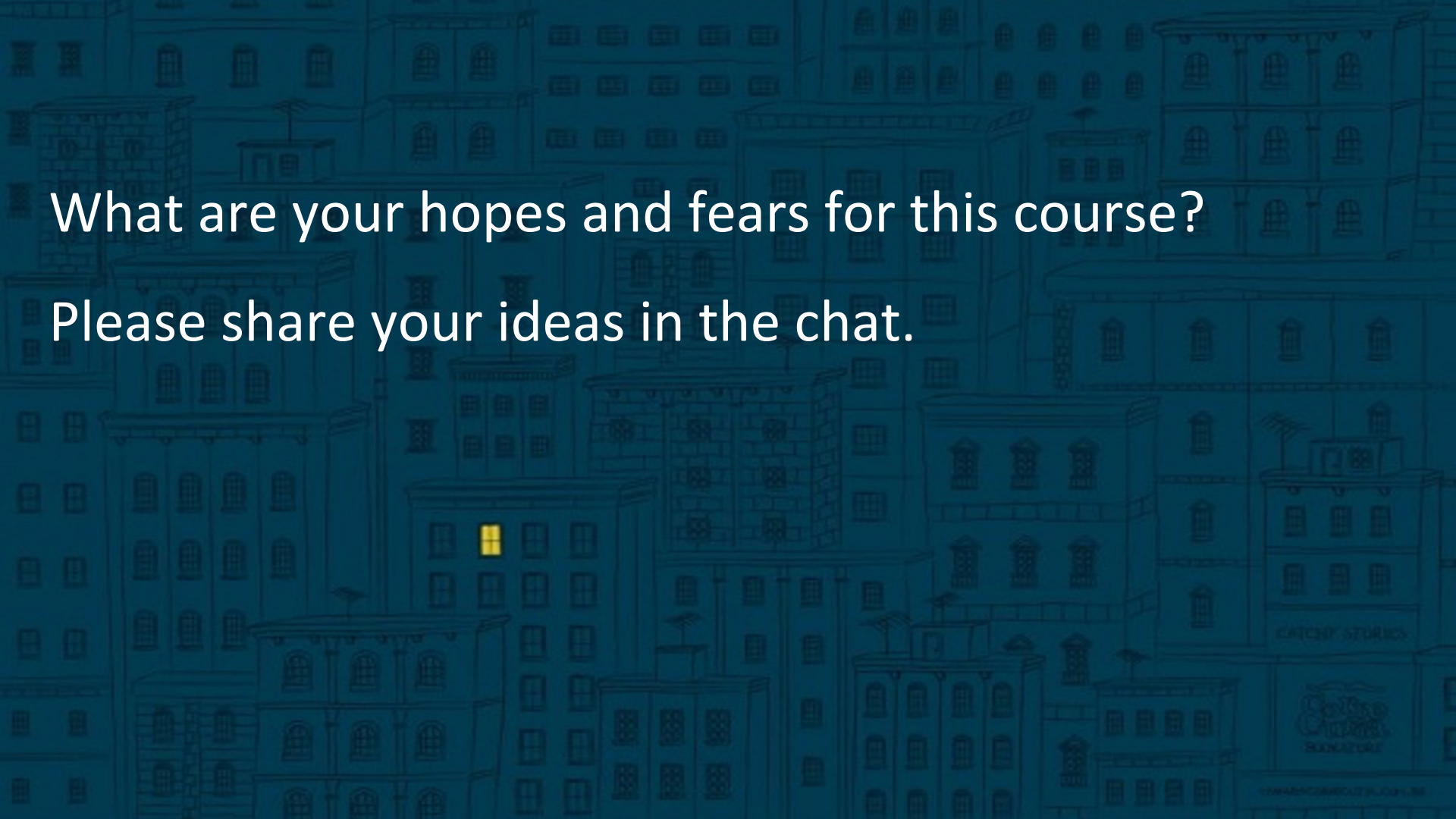
- deals with real-world data
- hugely various
- powerful
- shares basic concepts across the methods
- uses various software solutions (once you master one type of it, it gets easier later)



# Data Analysis - Why Care About It?

## **The goals of data analysis:**

- to help to understand *patterns* in data and
- to make decisions – by predicting risks, profits, etc.



What are your hopes and fears for this course?  
Please share your ideas in the chat.

# Course Contents

- Research hypothesis testing
- Describing data
- Measures of association
- Comparing mean scores across groups
- Correlational analysis
- Nonparametric tests for smaller samples
- Linear regression

# Logic of the Course\*

Describe and visualise the data, find associations

Compare means across groups

Identify relations between variables

Make predictions with linear regression

\*All this all is real classics for an introductory course.  
This is a cornerstone of working with data nowadays

# By the end of this course you should be able to:

1. understand basic statistical tests (and look clever)
2. perform these tests in R on your own (independently)
3. read scientific papers using the (1) t-test, (2) chi-square, (3) linear regressions, and (4) one-way ANOVA
4. understand what they mean and what they do not mean, and, importantly,
5. be cool about it.

# ‘Data Analysis in Sociology’ as a course of this Bachelor programme

- It is one of the key points in your study plan
- Those who attend and engage have better results
- This course continues to advanced levels in the two senior years
- It connects to the ‘Methods of Sociological Research’ and links you to further elective courses that can bring you fame, money, or whatever you will make of it.

# This year is like building a foundation to something interesting and reliable





If you skip the  
classes, then you  
know what





In the following years you will learn how to:

- make predictions with qualitative outcomes
- research something not directly observable
- create indices
- reduce larger groups of questions to a few
- classify observations by their characteristics
- create two-dimensional data maps
- and more

This is a slide when questions are asked about how to succeed in this course

# When & Where: Modules 3-4 this year

## **Module 3**

- Research vs. statistical hypotheses
- Types of variables
- Central tendency measures
- Chi-square
- Two-means comparison
- Correlations

## **Module 4**

- One-way ANOVA
- Post hoc comparisons
- Linear regression
- Main and interactive regression effects in linear regression models

# There is a Course Calendar

- i.e., a weekly planner of assignments, tests, and deadlines
- It is available [here](#)

# There are 5 types of activities

- **Lectures** (Q&A, quizzes, new material, summaries)
- **Seminars** (tests, tutorials on problem solving in R)
- **Practice Sessions** (hands-on practice in R, solve and check together)
- **Online Practice** (additional assignments on DataCamp after the classes, see the invite link in class materials)
- **Group Projects** (4 of them, in set teams, peer-reviewed)

# Grading Components

$$\begin{aligned} \textit{Grade} = & 30\% * \text{Group projects} + \\ & 20\% * \text{Mid-Term test} + \\ & 20\% * \text{In-class activity} + \\ & 30\% * \text{Final Exam} \end{aligned}$$

If you plagiarize, you will fail.

Don't steal from others even if they "don't mind"

# Mid-Term (beginning of module 4)

- It is a large test covering all the previous topics, both theoretical and practical. Any question discussed at the classes may get into the test – to be updated this year!
- We will do our best to give short tests at seminars and quizzes at the lectures.

# Final Exam

The final exam consists of 4 problems of data analysis. You will have about 60 minutes and a personal computer to solve them (it will be a very busy process, so our goal is to get you prepared):

- Similar problems will be provided to you during practice sessions.
- The exam is done individually, as if you were solving these problems in a real-life setting in a rush.



# In-class activity

Each seminar/practical session you'll get a grade from 0 to 2

- 0 - if you have not attended the seminar/you have not contributed to the discussion
- 1 - you've made a minor contribution to the discussion
- 2 - you've made a major contribution to the discussion

Make sure that your camera is ON during the seminars

# Group Work

How do you feel about working in a team?



# Teams

- 3-4 students each
- No more than 1 student from the “Data Science” minor per team
- individual responsibility zone in each project ([group contracts?](#))
- one country from the latest [European Social Survey](#)
- conduct analyses on the chosen topic (e.g., inequalities in health, welfare attitudes, migration, ageism, etc.)
- have a unique team name
- meet 1-3 times before each project to complete a report

# What are group projects in this course?

## Here is an algorithm:

1. You have survey data about a country (the same country throughout the course).
2. Learn about a new type of analysis in class.
3. Get a data analysis task (e.g. compare two group means) for which you develop a plausible research hypothesis, test it and deliver results.
4. Create a report about this and submit it to a discussion thread in canvas by the deadline.
5. Review another group's project for both good practice and points to improve.
6. Get the evaluation and feedback for the project. You can use the comments to update the project. At the end of the course you combine all the reports into a portfolio.

# What are group projects in this course?

<http://rpubs.com/werbitsky/499205>

(LMS view on the right, 2019)

Please, never copy-paste  
your projects,  
do your original work.

We have a very good memory  
and zero tolerance for cheating.  
Those who could support our  
words were expelled.

Vice Versa Team  
Topic: migration in Germany

- Alexandra Shanina
- Milena Oleshko
- Zakharova Victoria



Вложения: Vice\_Versa\_Final.Rmd Vice\_Versa\_Final.html

Re: Project 5. Linear Regression: the Ultimate Genealogy  
Отправил Бахарева А. П. (apbakhareva\_1@edu.hse.ru) в 25 май 2019, 15:44:52

2BK: Bakhareva, Borisenko, Kireeva, Kuzmicheva

Topic: Politics in Ireland

Rpubs: <http://rpubs.com/werbitsky/499205>

Вложения: 2BK-final.zip

Отправил Кукарцев Ю. Е. (yuekukartsev@edu.hse.ru) в 25 май 2019, 15:53:13

Team DAS\_is\_fantastisch: Mariya Pronyuk, Yuriy Kukartsev, Vilkhovenko Alexander, Belorukova Linara

<http://rpubs.com/pcgeniusjesus/finalPR>

Вложения: DASIsFANSTASTICSH\_Final\_project.rmd ESS8FR.sav

# Project Topics

1. Describe Your Data
2. Basic Tests for Data Comparison
3. From Correlation to Linear Regression
4. The Methods Portfolio

# What will group projects be about?

- Real data are fun. Sometimes bad fun, too.
- Each team picks a country from the European Social Survey, a large comparative survey in Europe and around.
- For each test we cover in class, you will find suitable variables in the data for your country, and perform analysis.
- At the end of the course, you present your results to the class.

# What will group projects be about?

Round 10 of the ESS was fielded in 2020

<http://www.europeansocialsurvey.org/>

There are many cool [findings](#)!



# You decide on the country and research questions

	R1 2002	R2 2004	R3 2006	R4 2008	R5 2010	R6 2012	R7 2014	R8 2016	R9 2018
Media and social trust	•	•	•	•	•	•	•	•	•
Politics	•	•	•	•	•	•	•	•	•
Subjective well-being...	•	•	•	•	•	•	•	•	•
Gender, Household	•	•	•	•	•	•	•	•	•
Socio demographics	•	•	•	•	•	•	•	•	•
Human values	•	•	•	•	•	•	•	•	•
Immigration	•						•		
Citizen involvement	•								
Health and care		•							
Economic morality		•							
Family work and well-being		•			•				
Timing of life			•						•
Justice and Fairness in Europe									•

Research questions can be combined from different areas and batteries, both standard and the so called “modules” focusing on particular topics.

# Main Dataset for Analysis

- European Social Survey  
<http://www.europeansocialsurvey.org/>
- Got the the Descartes Prize for rigorous methodological standards (=very cool)
- You can [register](#) on the site to support the project
- Download the data for your country and survey year
- Among the options available, download the SPSS file

# Recommended Reading (at HSE elibrary)

## **Data Analysis:**

*Tabachnick, B.G., & Fidell, L.S. (2014). Using Multivariate Statistics: Pearson New International Edition (Vol. 6th ed). Harlow, Essex: Pearson.*

## **R:**

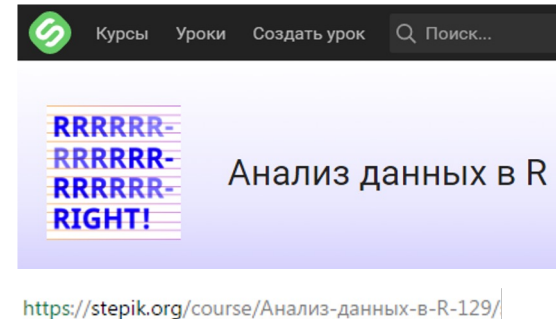
*Stowell, S. (2014). Using R for Statistics. Berkeley, CA: Apress.*

# Internet-Based Resources

- Most of what you need to learn can be found online. But the books provide solid guidelines.
- Online class: [datacamp.com](https://datacamp.com) (access pending)
- We recommend all the books on R you can find at the library – probably we had them bought for this course.
- Reference guides and a large community:
  - <http://statistics.ats.ucla.edu/stat/>
  - <http://stackoverflow.com>

# Internet Courses Can Enhance Your Experience

- There are plenty of free MOOCs on statistics. Some of them are really great! Get a liking for them, and listen to them to get a better understanding.
- youtube - Statistics One (Prof. Andrew Conway, Princeton U)
- Stepik.org (A.Karpov, Mail.ru group)



# Feedback I expect during the course

- Learning is communicating: Questions are good.
- Good questions are exciting and give me food for thought.
- There are many elements to a good course, and we are constantly working to improve it
- If you feel that something is wrong (or that you are going to miss some of the classes) let us know as soon as possible

# Software

**Day-to-day:** R in RStudio IDE –

works for every operating system and is free

– you can learn on your granny's computer

– you can learn at a boring party

– you can use it later, still for free.

**→ at home, before Seminar 1: First, download and install R**

<https://www.r-project.org/>

**→ Second, download and install RStudio**

<https://www.rstudio.com/products/rstudio/download/>

# Two facts about data analysis you will have to live with:

1. IBM SPSS and SAS are the most popular software in the social sciences worldwide. Python is the industry workhorse for computer scientists.

They are the respected (standard) default; IBM SPSS and SAS have proprietary, expensive licenses (\$1000+/year); traditional software has a given set of functional tools, while a programming language needs you to program + know more than just this language.



# Two facts about data analysis you will have to live with:

2. R is on a steep rise for statistical analysis. Everywhere in the world. Especially since 2020.

Which means: IRL you will need to be flexible and probably deal with many software tools in your work.

But it also means: you will need to learn how to work with at least one of them. Starting with R gives many advantages.

# Two tasks you need to do to prepare for the first seminar: Task 1

- create teams of 3-4 people (and remember to get a cool memorable team name)
- pick a country from the ESS 2020 and assign yourselves to a specific topic (please do not pick the name of the questionnaire section as a topic, be more creative)
- Fill in the [form](#)

Two tasks you need to do to prepare for the first seminar: Task 2

- Enrol with DataCamp

Register via invite link (it will be posted in the announcements soon)

# Canvas as LMS for this course

<https://canvas.instructure.com/>

- You'll get an invitation to join the class today or tomorrow
- You'll see your assignments and announcements here
- You'll be able to submit them here as well

Questions?

Now, let's move to  
the substantive part  
of the lecture!

# Part I of this course is Descriptive Statistics:

## **1. Research hypotheses vs. statistical hypotheses**

- The cycle of research. Posing and testing hypotheses.
- Variable types and their descriptive stats.

## **2. Central tendency measures. Means as a model**

- Mean, median, mode. Standard normal distribution and its use. Z-scores.
- Moments of distributions. Interpretation of z-scores. Mean as a model.

# Research questions and hypotheses

*„Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time“*

B. Efron

*„All models are wrong, but some are useful“*

G.E.P. Box

Statistics deals with data collection, analysis, interpretation, and presentation. Statistics is also said to ‘quantify uncertainty’.



# Research questions and hypotheses

- The research process starts with **research questions**, e.g. "Do computer games affect personal relations?"
- A research-based answer to that question assumes the following:
  - A research hypothesis
  - Data collection
  - Data analysis
  - Conclusion as to the research hypothesis
  - Reporting the results to the audience

# Research questions and hypotheses

**A research hypothesis** is a statement about the expected or predicted relation between variables.

A **variable** is a quality of an item, event, individual, etc. that can take on different values.

*Example: The variable „a Russian alphabet letter“ can take on 33 values.*

# Research questions and hypotheses: examples

Research Question	Research Hypothesis
Is students' performance on tests more influenced by their motivation or their learning strategies?	Students who are taught effective learning skills will perform better on tests than students offered incentives to do well.
Do college students and faculty differ in their beliefs about the prevalence of student academic misconduct (plagiarism)?	Faculty members' beliefs about the frequency of student academic misconduct will be lower than students' beliefs.
Is there a link between adolescents' exposure to violence in their family and their academic achievement?	The more adolescents are exposed to violence in their family, the lower their levels of academic achievement.

# Research questions and hypotheses: examples

Research Question	Research Hypothesis
Is one method of disciplining children more effective than another?	Children will rate a disciplining strategy that emphasizes logic and reason as more effective than one based on rewards and punishment.
Does playing online games affect one's interpersonal relationships?	Heavy users of online games have less fulfilling interpersonal relationships than users spending little or no time playing online games.
Does providing substance abuse treatment to drug users have an effect on safety in the workplace?	Drug users are less likely to have work-related accidents after undergoing substance abuse treatments than before the treatment.

# Types of Variable Measurement

„Disciplining strategy“, „academic performance“, „good interpersonal relationship“ – these variables that need to be measured.

**Measurement** is assigning categories or numbers to items or events according to certain rules.

*Example:* to measure „height“, one can use „number of *cm* from feet to the tip of the head, without shoes“; to measure „academic performance“ – average grade/exam/rating, etc.

# Types of Variable Measurement

As a rule, authors identify four types of variable scales:

- **Nominal** (names of colours, marital status)
- **Ordinal** (race winners, order of appearance)
- **Interval** (psychological scales)
- **Ratio** (wavelength, income, IMDb rating)

# Types of Variable Measurement

As a rule, authors identify four types of variable scales:

- **Nominal variables** differentiate between categories or types, “names”
- Nominal variables are further divided into binary (2 categories) and multinomial (many categories).
- Example: gender (male, female, etc.)
- Example: education strategy (effective learning, incentives, etc.)

# Types of Variable Measurement

As a rule, authors identify four types of variable scales:

- **Ordinal variables** take on values that could be ordered relative to some other value; these are ranks (gold-silver-bronze), volume (small-medium-large at McDonalds), and the like.
- Ordinal scales can show that one value is larger than another. However, it cannot show how large the difference is.



# Types of Variable Measurement

As a rule, authors identify four types of variable scales:

- **Interval variables** have values separated by equal distances in a numeric continuum. The difference between 1-3 and 7-9 will be the same on such scales.
- „Zero“ is just one of the values in a row.
- Examples: Celsius scale, aggressiveness scale, etc.

# Types of Variable Measurement

As a rule, authors identify four types of variable scales:

- **Ratio** scales also have values separated by equal distances (1-3 is the same as 7-9).
- They contain the *absolute zero*, which mean the total lack of the measured quality (8 is four times larger than 2 on this scale).
- Examples: height, weight, distance, time, number of correct answers to a test.

# Types of Variable Measurement: Why?

- Classical methods of data analysis often work best with continuous, ratio and interval variables.
- The „higher“ a variable scale (nominal<ordinal<interval<ratio), the *more methods* are normally available.
- *Scales of higher order can be downgraded*, but not the other way round. Example: age in years (ratio) > equal age groups (interval) > old/mid-age/young (ordinal) > old/young (binary)
- If a variable could be measured with various scales, *use the higher scale* as it contains more information.

# Types of Variable Measurement: Exercise

Name the variable scale for the following:

- Type of the operating system (iOS/Android/...)
- Probability to complete a bachelor's degree
- A blogger's popularity (by the number of subscribers)
- Your rank by GPA
- Your group number