Data Analysis for Public Policy Research

Dr. Keith Smith
keith.smith@gess.ethz.ch
IFW C45.2

Camille Fournier de Laurière
camille.fournierdelauriere@ir.gess.ethz.ch
IFW C41.1

Autumn 2025

Wednesdays, 14.15-16.00
September 17 - December 17, 2025
IFW B42
Course Moodle
Zoom (if needed)

1 Course Description

This course introduces students to the necessary fundamentals of mathematics and statistics, and their applications, to conduct quantitative policy evaluations. The course will provide a survey of theoretical foundational concepts and techniques. The applied part of the course will focus on implementing these techniques in R, as well as developing the practical skills in the language required to be able to independently conduct basic data based research projects.

1.1 Learning Objectives

Throughout the course, you will develop the following skills and statistical methods:

- Summarise data using graphs and descriptive statistics
- Conduct hypothesis tests to test for differences in means and proportions
- Interpret p-values and conduct significance tests
- Identify relationships between categorical and continuous variables

- Perform linear regression, including multiple regression with categorical variables, interaction terms and transformations
- Interpret substantive effects of regression models
- Apply statistical methods in R, including visualisation of results
- Present summary of analysis applying statistical methods

1.2 Semester Hourly Breakdown

This course is 4 ECTS, which equates to 120 total hours (inclusive of course lectures, preparations, exams, and studying). As a guideline, we suggest the following 'time budget' for this course:

- Course lectures (12 x 2h) = 24h
- Preparing for lectures - readings + problem sets (12 x 3h) = 36h
- Mid-term exam preparation (18h) = 18h
- Mid-term exam (1 x 2h) = 2h
- Project presentation preparation = 18h
- Project presentation week (1 x 2h) = 2h
- Final exam preparation (18h) = 18h
- Final exam (1 x 2h) = 2h
- **Total = 120 hours**

If you find that the preparations for any of these activities are taking much longer than this budget, please contact us *earlier rather than later* so we can help support you and provide guidance.

2 Expectations

This is a core course for students in the Masters in Science, Technology and Policy program. Students are not required to have a background in statistics or mathematics for this course.

Attendance in all course lectures is required (please inform us before the class if you have an emergency or excused conflict which requires your absence).

As a graduate-level course, we will quickly move forward from more introductory statistical methods into increasingly complex concepts. The statistical theories and methods covered in this course are sequential, meaning that if you fall behind, it is important to catch up quickly. Each week you will be assigned readings ($\sim$ 30-50 pages), as well as problem sets (4-6) to help you review this material. It is the responsibility of the student to keep current with each week's literature and assignments, and to seek assistance when required.

3 Learning Approaches

3.1 Course Lecture Structure

Each course lecture will begin with an ungraded 'quiz' ($\sim$ 10 minutes). This quiz will be similar to the problem sets you were asked to complete as part of your preparations for the week. Next, we will have lecture over the week's course content. This first lecture will focus on the statistical techniques of the week (largely overviewing and clarifying the textbook readings).

Then, after break, we will complete, as a class, problem set(s) which illustrate the week's content. During this time, we will complete the problem sets for the lecture's 'quiz', providing the correct solution. Lastly, we will then have the second lecture session, which will further demonstrate how the week's concepts can be analysed within the applied setting using R.

Example course lecture structure:

- 14:15-14:30 - Welcome and weekly quiz
- 14:30-15:00 - Lecture
- 15:00-15:15 - Break
- 15:15-15:25 - Course content example (weekly quiz solution)
- 15:25-16:00 - Applying techniques in R

3.2 Student Individual Preparation

The course lectures are reviews of the required weekly course material. Pragmatically, we cannot cover all the material and content required during these 2 hour sessions. As such, we expect that you read the **required readings and complete the problem sets** before each class lecture and come prepared with any questions or clarifications you may have.

We will not grade the weekly problem sets, these are for your review and to ensure that you understand the required concepts. The solutions to the problem sets are available on Moodle. But, you can ask us to review your problem sets at any time, or come to ask us for help if you have any questions.

3.3 Group Projects

In groups of 4-5, you will develop a research question, and apply statistical techniques in R to create a 10-15 minute presentation of your design and results. This project will help further develop your skills in group collaboration and assisting each other in problem solving.

4 Statistical Concepts and Methods

4.1 Required Readings

Each week of the course will utilise readings from textbooks and scientific papers. The required readings for each week are listed in the syllabus and are available on the course

Moodle. You are expected to read all the required readings before the course lecture, and come prepared with any questions or areas of uncertainty that arise from these readings.

The course material we will cover largely comes from three different statistics textbooks:

Agresti, Alan. Statistical Methods for the Social Sciences. Fifth/Sixth edition. Global edition. Harlow, England: Pearson, 2024. ETH Library

Gill, Jeff. Essential Mathematics for Political and Social Research. Cambridge: Cambridge University Press, 2006. ETH Library

Wooldridge, Jeffrey M. Introductory Econometrics: a Modern Approach. Fifth edition. Melbourne: South Western Cengage Learning, 2013. ETH Library

Firebaugh, G. (2018) Seven Rules for Social Research. Princeton, NJ: Princeton University Press.ETH Library

We would encourage you to further engage with these textbooks as well, to support and extend your knowledge. All of these textbooks are either available in hard copy, online, or both at the ETH Library (see links above).

4.2 Reading Extensions and Further Support

We have also included several readings which extend the course content and applications, as well as provide further support and background (listed as suggested in the syllabus). While we encourage you to read these as well, these materials are not required.

5 Applied Statistical Analysis in R

5.1 Software

All students will need to install R on their laptops and bring their laptops to class. We require that you install R as well as the more user-friendly R Studio environment. To make installation easier, you should install R first, and RStudio should detect your latest rversion automatically. *Laptops will be needed from the first session on-wards.* For a fairly detailed discussion of R, please see An Introduction to R.

5.2 Reading support materials

Further, we have included readings to support application of the statistical techniques in R. For many weeks, these readings are also required (where noted). You are expected to practice applying these techniques during your class preparation time outside of class.

Many of the materials for how to apply these techniques in R come from the following sources:

Crawley, Michael J. The R Book. Second edition. Hoboken, N.J: John Wiley and Sons Inc., 2013. ETH Library

Grolemund, Garrett. Hands-on Programming with R. 1st edition. Sebastopol, CA: O'Reilly Media, 2014. ETH Library

Lennert, Felix. An(other) introduction to R. 2022. Online.

Wickham, Hadley, and Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. First edition. Beijing, China: O'Reilly, 2017. ETH Library, second edition available online here.

If you would like further context, or are interested in extending your knowledge, please refer to these full texts, as they are excellent resources.

6 Course Assessment

6.1 Midterm and Final Examinations (35% of final grade each - total of 70% of final grade)

Two written, in-person exams will be held on **November 5, 2025 and January 28, 2026** from 14.00-16.00 in IFW B42. For the exam, please bring a pencil and calculator. We will provide the relevant formulas and statistical tables necessary to complete the exam.

6.2 Group Project Presentation (30% final grade)

Within groups of 4-5 students, you will develop an empirical research project which will assess your understanding of applying statistical techniques in "real-world" social science research settings. You will need to develop the project cumulatively. First developing a research question, then identifying data, developing an estimation strategy, and lastly, presenting the final project and results.

The project will be assessed based upon a 12-15 minute final group presentation (30% of your final grade), which will be presented during the final course lectures, December 10 or December 17, 2025.

To help guide your development of this project, you are required to submit 3 shorter written assignments for each portion of the project throughout the term (e.g., developing research questions, operationalizing concepts into data, proposed estimation strategy). Each of the written assignments will be roughly 1 page each (see group project assignment briefs in course Moodle). We will review these assignments, and provide verbal feedback to each group in individualised meetings. The written assignments are ungraded and are intended to guide your group with constructive feedback. You are still required to complete each of the assignments. Any written assignment that is missed by a group will result in a **5% deduction** in your final grade.

7 Use of AI-tools and Large Language Models

Large language models, LLMs such as ChatGPT, Gemini, Claude, are increasingly being used in statistical applications, for example, to help generate code or to suggest approaches to data analysis. In this course, we emphasise learning how to use LLMs effectively as a *tool to support your understanding, rather than as a substitute for it.* When used thoughtfully, LLMs can do provide many benefits, such as: helping you debug code or suggest alternative techniques; clarify statistical concepts in plain language or from a

different angle than in the book and lecture; develop substantive example of statistical techniques; provide immediate feedback on problem sets.

However, like any tool, LLMs must be used with your own expertise, knowledge, judgment and supervision. LLMs are not substitutes for understanding and critical thinking, and the outputs often can contain errors, misleading suggestions, and when used improperly, can actually make completing a task longer and more difficult.

In short, LLMs are best used to support your learning and analysis in this course, and not to replace it. For further guidance, here is a non-exhaustive list of tasks for which we allow or encourage the use of LLMs, and those that are not permitted.

7.1 Allowed usage of LLMs

- Help with coding in R (e.g., debugging, understanding syntax, exploring alternative approaches)
- As an informal tutor, to ask for help in understanding statistical concepts (e.g., regression assumptions, confidence intervals), and to explain or clarify difficult ideas in plain language
- To informally translate difficult to understand text into your native language (if non-native English speaker)
- Generate practice questions (more problem sets), help develop study guides or tools.
- Provide informal feedback on your completed problem sets, to identify where (and why) your answer may differ from the solution listed in the answer.
- Copy editing, spell-checking, checking grammar.

7.2 Not permitted uses of LLMs

- Submitting text or software code that is primarily generated by an LLM. This could mean using text or code from an LLM that is either direct copy-paste (verbatim) or 'lightly' edited.
- Using an LLM to replace your independent design, coding, analysis or interpretation. Do not ask an LLM to, for example, "perform a linear regression of y on x1, x2, x3, and write a summary of the results".
- When in doubt, the key rule is that **the original work must be generated by a human**. You can use an LLM to support this work, but not to generate it on its own.

If you have any questions about whether using an LLM for a certain task is appropriate, please ask us. This is also new for us, and we are also trying to figure out what is useful and appropriate, so we appreciate the continued dialogue.

8 Course Agenda

[Schedule is pending and can change throughout the term]

Lecture Week 1 - 17 September

Introduction to Course, First steps towards using R

- Required
  - Read and review course syllabus
  - Download R (first) and R Studio (second)
- Applied Statistics in R
  - Introduction to R and RStudio, getting familiar with the User Interface, first commands, create and load dataframes

Lecture Week 2 - 24 September

Descriptive Statistics : Measures of central tendency, distributions, variance

- Required Reading
  - Agresti - Chapter 1 : Introduction to Statistics, p.13-20
  - Agresti - Chapter 2 : Types of Variables, p.23-26
  - Agresti - Chapter 3 : Descriptive Statistics, p. 41-67
- Suggested Reading + Extensions
  - Agresti - Chapter 2 : Sampling, p.26-36
  - Wooldridge : Basic Mathematical Tools, Appendix A
- Applied Statistics in R
  - Use mean, standard deviation, median functions, plot descriptive statistics

Lecture Week 3 - 1 October

Probabilities and Sampling Distributions

- Required
  - Agresti - Chapter 4.1-4.2 : Probabilities, p.79-83
  - Agresti - Chapter 4.4-4.6 : Sampling Distribution, p.91-105
- Suggested Reading + Extensions
  - Gill - Chapter 7

Lecture Week 4 - 8 October

Normal Probability Distribution and Z-Scores

- Required
  - Agresti - Chapter 4.3 : Normal Probability Distributions, p.84-91
  - Wooldridge - Appendix B5
  - Firebaugh - Chapter 1 : There should be the possibility of surprise in social research
  - Data Visualization: R for Data Science (up to Section 1.2.5, included) link
- Applied Statistics in R
  - Simulate and plot a distribution, calculating Z-scores, more plotting

Lecture Week 5 - 15 October

Point and interval estimation, CIs for a mean, Significance testing, 1-sample mean tests
Group Presentation : Research Question Proposal due

- Required
  - Agresti - Chapter 5.1 : Point and interval estimation, p.115-117
  - Agresti - Chapter 5.3 : Confidence intervals for a mean, p.125-132
  - Agresti - Chapter 6.1 - 6.2 : The five parts of a significance test, test for a mean, p. 151-164
  - Agresti - Chapter 6.4-6.5 : Decisions and types of errors in tests, limitations of significance tests, p.167-174
- Suggested Reading + Extensions
  - Agresti - Chapter 6.6 : Finding P(Type II Error), p.175-177
  - Wooldridge, Appendix C5-C7

Lecture Week 6 - 22 October

Comparing groups with 2-sample mean t-tests

- Required
  - Agresti - Chapter 7.1-7.5 : Comparing means of groups, p. 191-210
- Suggested Reading + Extensions
  - Agresti - Chapter 7.7 : Nonparametric statistics for comparing groups, p. 213-216
- Applied Statistics in R
  - Perform statistical tests: mean and median tests

Lecture Week 7 - 29 October

Proportion tests

- Required
  - Agresti - Chapter 5.2 : Confidence interval for a proportion, p. 118-124
  - Agresti - Chapter 6.3 : Significance test for a proportion, p. 164-167
  - Agresti - Chapter 7.6 : Methods for comparing proportions, p. 210-213
- Suggested Reading + Extensions
  - Agresti - Chapter 6.7 : Small-sample test for a proportion, p.177-181
- Applied Statistics in R
  - Other statistical tests: proportions

Midterm Exam - 5 November

- ***Written, in-person exam, IFW B42, 14.00-16.00, 30% of final grade***
- Sampling, Distributions, probabilities, z-scores
- Inferences : Hypothesis testing, z- and t-tests

Lecture Week 8 - 12 November

Linear Relationships, bivariate regression, correlations

- Required
  - Agresti - Chapter 9 : Linear Regression and Correlation, p.271-302
- Suggested Reading + Extensions
  - Wooldridge - Chapter 2.1-2.5
- Applied Statistics in R
  - How to quantify and plot bivariate associations

Lecture Week 9 - 19 November

Group Presentation : Data and Operationalisation due Multivariate Regression, R^2, Comparing models, Standardized coefficients

- Required
  - Agresti - Chapter 11.1-11.3 : Multiple regression, p.335-353
  - Agresti - Chapter 11.6-11.8 : Partial correlation and standardized coefficients, p.359-366
- Suggested Reading + Extensions

  - Wooldbridge Chapter 3 - Multiple Regression Analysis : Estimation
  - Wooldbridge Chapter 4 - Multiple Regression Analysis : Inference
  - Wooldbridge Chapter 5 - Multiple Regression Analysis : OLS Asymptotics
- Applied Statistics in R
  - Multivariate regression

Lecture Week 10 - 26 November

Interaction terms and predicted values

- Required
  - Agresti - Chapter 11.4 : Modelling Interaction Effects, p.353-357
  - Agresti - Chapter 13.4 : Adjusted means, 429-433
  - Agresti - Chapter 11.5 : Comparing Regression Models, p.357-359
- Suggested Reading + Extensions
  - Wooldbridge Chapter 6 - Multiple Regression Analysis : Further Issues
- Applied Statistics in R
  - Making predictions and interaction terms in multivariate regressions

Lecture Week 11 - 3 December

Group Presentation : Analytical strategy due Association between categorical variables

- Required
  - Agresti - Chapter 8 : Analyzing association between categorical variables, p.229-251
- Applied Statistics in R
  - Categorical variables, factors, Chi2, Kendall's Tau

Lecture Week 12 - 10 December

Multiple regression with categorical variables and Group Presentations (first 2 groups)

- Required
  - Agresti - Chapter 12.1 : Regression Modeling with Dummy Variables for Categories, p.379-383
  - Wooldbridge Chapter 7 - Multiple Regression Analysis with Qualitative Information: Binary (or Dummy) Variables
- Applied Statistics in R
  - Categorical variables in regressions, dummies

Lecture Week 13 - 17 December

Group Presentations (Remaining 4 groups)

Final Exam - 28 January 2026

- ***Written, in-person exam, IFW B42, 14.00-16.00, 40% of final grade***
- Cumulative exam, including the concepts from mid-term exam plus:
- Categorical associations
- Correlations and linear regression
- Multivariate analyses