Introduction to Statistics

A brief introduction to applied statistics for quantitative social science research

Laboratorio de Investigación para el Desarrollo del Ecuador

Instructor: Daniel Sánchez, MA

Module length: 15 hours

Course level: Intermediate

Prerequisite knowledge: Basic arithmetic, algebra, cartesian graphing, Stata scripting, Git.

Corequisites: *Introduction to R* module.

GitHub repository: https://github.com/laboratoriolide/intro-to-stats

1 Overview

This course introduces statistics at a basic level, focusing on its application to quantitative social science research. The course assumes no prior knowledge of statistics, but students will reach an intermediate level of undergraduate statistics by the end of the course. The material will not be proof-heavy, however, it will be rigorous enough to prepare you for the more advanced modules in the program (econometrics, causal inference, etc.)

A good grasp of statistics cannot be achieved with theory alone. I will provide practical implementations of what we cover during class mainly using R and, to a significantly lesser extent, Stata. The companion module, *Intro to R*, will focus on technical aspects of the language, while this course will focus on theory, intuition and applications specific to statistics. To best understand the material, students should have a basic understanding of arithmetic, algebra and cartesian graphs (i.e. x-y plots). This was provided in the introductory mathematics course at the beginning of the program (Math for Social Sciences)[1].

[1] Materials are available in the Google Drive and LIDE GitHub org for review if needed.

2 Module contents

The following is a planned outline of the course. This may change depending on the pace of the class. Each lecture will have one or two assigned readings, which will all be academic articles submitted to the course's GitHub repository.

- **Lecture 1**: Intro
  - Introduction to statistics: what is it and why do we need it in social science research?
  - Observations, variables, data types, data formats
  - Populations vs. samples
  - Descriptive vs. inferential statistics
  - Univariate descriptive statistics: measures of central tendency, dispersion, position and distributional shape
  - **Assigned reading**: Course syllabus, Brown (1998).
- **Lecture 2**: More on descriptive statistics and statistical data visualization
  - Categorical data: frequency tables, bar charts, pie charts, contingency tables
  - Bivariate descriptive statistics: cross-tabulations, covariance, correlation, etc.
  - Data visualization: histograms, box plots, scatter plots, cumulative distributions
  - **Assigned reading**: Lupu and Zechmeister (2020), prepare to replicate charts and apply descriptive statistics to the dataset (file provided on GitHub).
- **Lecture 3**: A brief introduction to probability
  - Why probability?
  - Basic probability: experiments, counting, set theory, conditional probability
  - Random variables and probability distributions (discrete and continuous)
  - Probability distribution functions (PDFs) and cumulative distribution functions (CDFs)
  - The normal distribution and its properties, the empirical rule
  - The standard normal distribution and z-scores
  - **Assigned reading**: Redacción Primicias (2020), will be used for a class exercise.
  - Time-permitting: bootstrapping.
- **Lecture 4**: Statistical inference
  - What is statistical inference?
  - Sampling distributions
  - The central limit theorem and the law of large numbers
  - Expected value, variance and standard errors, use of operators.
  - Using simulation to understand sampling distributions
  - **Assigned reading**: None, see DataCamp platform for assignments.
- **Lecture 5**: Point and interval estimation

  - What is estimation?
  - Bias, efficiency and margin of error
  - Significance levels (alpha)
  - Margin of error with sampling distributions
  - Confidence intervals and their interpretation, hypothesis testing with CIs
  - Avoiding dynamite plots
  - **Assigned reading**: Barbayannis et al. (2022), with attention to the research question, charts and discussion, less on methods being used.

- **Lecture 6**: Hypothesis testing
  - What is hypothesis testing?
  - Type I and Type II errors
  - p-values and critical values
  - Single-sample hypothesis tests: one and two-tailed tests
  - One sample z-test for mean inference
  - Time-permitting: Statistical power, sample size, effect size and simulation exercises
  - **Assigned reading**: Ahmed, Andersson, and Hammarstedt (2013), with attention to Table 1.

- **Lecture 7**: More on hypothesis testing
  - The t-distribution, degrees of freedom and the one sample t-test
  - Two sample z and t-tests for mean inference
  - Paired t-tests (dependent samples)
  - Time permitting: statistical tables, proportion tests and chi-squared tests
  - Choosing the right test for your research question
  - **Assigned reading**: Broockman and Skovron (2018), with attention to Table 2.

- **Lecture 8**: Experimental design and ANOVA
  - What is experimental design?
  - Why experiments? Causality in social science
  - One-way and two-way ANOVA
  - **Assigned reading**: Stanley, Neck, and Neck (2023),
  - Time-permitting: factorial ANOVA

- **Lecture 9**: Basic Ordinary Least Squares (OLS) regression
  - What is regression? Is it always OLS?
  - Simple linear regression
  - Coefficient interpretation in simple regression
  - OLS features and properties: residuals, fitted values, unit transformations
  - Linearity in parameters
  - Time permitting: multiple regression
  - **Assigned reading**: Holcombe and Boudreaux (2015), with attention to regression (1) on Table 1.

- **Lecture 10**: Case study presentations
  - See below for details on the case study. No assigned reading for the last day of lecture.

3 Evaluation

The module will be graded following the same weighting as the one proposed by the program's regulation handbook. The final grade will be calculated as follows:

Component	Percentage
Attendance & Participation	15%
Assignments	50%
Case Study	35%

3.1 Attendance and participation

I do not have any special requirements for attendance nor participations other than the requirements set by the program. Consult the program's regulation handbook for more information. I encourage you to participate in class and ask questions, as this will help you understand the material better. Statistics typically inspires frustation, so it is important to ask questions when you are confused - as any other quantitative course, the material builds up on itself, so better to understand things sooner rather than later.

3.2 Assignments

There will be weekly DataCamp assignments that reinforce the material taught in this module as well as that of the sister *Intro to R* module. The assignments will focus on R implementations of statistical methods, statistical theory or other relevant topics. These will be either DataCamp courses, projects or exercises, and are graded on a pass/fail basis. Please check the DataCamp platform for exact due dates. Further, as mentioned above, Stata exercises will be provided, but not graded.

3.2.1 DataCamp assignments

The following is a tentative list of DataCamp assignments. This may change depending on the pace of the class, so please check the DataCamp platform for the most up-to-date assignments.

- Introduction to Statistics in R **(will be optional, for 1pt extra credit due September 22)**
- Foundations of Probability in R
- Sampling in R: optional, due October 13th.
- Hypothesis Testing in R
- Foundations of Inference in R, optional, due October 20th.
- Experimental Design in R, optional, due October 27th.

3.3 Case study

The case study is due the last day of class. The case study involves using the skills gained in this module as well as the *Intro to R* module to analyze the provided case study dataset in groups of 5-6 students. Each group will present their findings in 10-minute academic-style presentation. followed by a 5-minute questions and answers (Q&A) session by the instructor and the rest of the class. A handout will be provided with the case study details and rubric.

4 Course materials

All course materials will be provided in the course's GitHub repository. This includes lecture slides, readings, datasets, assignments and any other relevant material. The repository will be updated regularly, so please check it often for new material. I recommend using a Git client, such as GitHub Desktop or GitKraken, to keep your local repository up-to-date.

4.1 Readings

A reading will be assigned for each lecture. These can be academic or general articles that will be used as to illustrate the concepts covered in class. Students are expected to read the assigned readings before class and be prepared to discuss them. Files for all readings will be provided in the course's GitHub repository.

When readings are academic articles, students should not expect to fully understand the methods used in the paper. This module is an introduction to statistics, so students will progressively learn the methods used in the readings as the course progresses. Further, students should not be taking too much time to read papers as it will quickly become overwhelming considering the module's course load, instead, developing the ability to skim and understand the main points of a paper is more important. Students should focus on the specific part the reading was assigned for, as well as the paper's objective, a general overview of the methods and any applicable public policy implications. For example, for Ahmed, Andersson, and Hammarstedt (2013), the focus is on understanding Table 1.

4.2 Textbook

There is no required textbook for this course, as I will provide slides for all lectures However, I recommend the following books for those who want to delve deeper into the material. These were used as references for the course.

- *Statistics for Business and Economics*, Anderson, Williams, and Cochran (2020).
- The statistics review in *Introductory econometrics: A modern approach*, Wooldridge (2020).
- *Using R for Introductory Econometrics*, Heiss (2020)[2]
- *Discovering statistics using R*, Field et al. (2012).
- *Statistics for economists: a beginning*, Floyd (2010).
- The Effect, Huntington-Klein (2022)[3].
- The Library of Statistical Techniques LOST[4].

4.3 Software

The course will mainly use R (R Core Team 2024) for practical implementations of the statistical concepts covered in class. R is a free and open-source software which has growingly become the standard for advanced statistical analysis. The sister module to this course, *Intro to R*, introduces the language in depth, focusing on technical aspects, while here we mostly use it as a tool to understand statistical concepts.

Further, the course will also feature demonstrations of statistical concepts using Stata, a widespread software package across academic environments. Further, I will provide practice exercises to be solved with this language. These will not be graded, but feedback will provided to those that want it. I will not introduce the use of Stata, as it is assumed all students have already taken and succesfully passed the introductory course on Stata from our last module, *Intro to Stata*. However, I will answer questions and provide personalized help to those who need it, conditional on available time. It is recommended that students review Huntington-Klein (2022) for comparisons on R-Stata syntax in statistical analysis.

Finally, I will use Git and GitHub for version control and to distribute course materials. I will not introduce or evaluate the use of Git, as it is assumed all students have already taken and succesfully passed the introductory course on Git.

[2]This book contains R implementations of Wooldridge's textbook, with code uploaded here.

[3]This textbook, while focused on causality, describes much of the modern R development environment for statistics and econometrics. It also contains Stata and Python code.

[4]An open-source website with a collection of statistical techniques and their implementations in R, Python and Stata. It is a great resource for those who want to learn more about specific statistical methods.

5 Keyboard layout

We will routinely need to type symbols like "/", "<-", "%>%", and others. Make sure you are comfortable with your keyboard layout and that you can type these symbols easily. This may seem trivial, but it is important for the course, as we absolutely cannot afford to lose time finding symbols on the keyboard. You may need to change your keyboard layout to the correct language so that the computer follows the physical layout of your keyboard. For Windows users, this is easily done by pressing `Win + Space` and selecting the correct layout (see here)[5].

6 Communication

All communications to the instructor should be made through the course's Slack channel. I hope to respond to questions within 72 hours, but please be patient if we take longer. I do not monitor email regularly, so please use Slack for all communications if you need a timely response.

References

Ahmed, Ali M., Lina Andersson, and Mats Hammarstedt. 2013. "Are Gay Men and Lesbians Discriminated Against in the Hiring Process?" *Southern Economic Journal* 79 (3): 565–85. https://doi.org/10.4284/0038-4038-2011.317.

Anderson, David R, Thomas A Williams, and James J Cochran. 2020. *Statistics for Business & Economics*. Cengage Learning.

Barbayannis, Georgia, Mahindra Bandari, Xiang Zheng, Humberto Baquerizo, Keith W. Pecor, and Xue Ming. 2022. "Academic Stress and Mental Well-Being in College Students: Correlations, Affected Groups, and COVID-19." *Frontiers in Psychology* 13 (May). https://doi.org/10.3389/fpsyg.2022.886344.

Broockman, David E., and Christopher Skovron. 2018. "Bias in Perceptions of Public Opinion Among Political Elites." *American Political Science Review* 112 (3): 542–63. https://doi.org/10.1017/S0003055418000011.

Brown, Cara L. 1998. "Sexual Orientation and Labor Economics." *Feminist Economics* 4 (2): 89–95. https://doi.org/10.1080/135457098338482.

Field, Andy, Jeremy Miles, Zoe Field, and Zoë Field. 2012. *Discovering Statistics Using R*. 1st edition. London ; Thousand Oaks, Calif: Sage Publications.

Floyd, John E. 2010. "Statistics for Economists: A Beginning." https://chrisbillakings.wordpress.com/wp-content/uploads/2015/06/tmp_ecstats1727528223.pdf.

[5] Mac and Linux Users, sorry, you're on your own here.

Heiss, Florian. 2020. *Using R for Introductory Econometrics*. Düsseldorf: Independently published. https://www.urfie.net/.

Holcombe, Randall G, and Christopher J Boudreaux. 2015. "Regulation and Corruption." *Public Choice* 164: 75–85.

Huntington-Klein, Nick. 2022. *The Effect: An Introduction to Research Design and Causality*. 1st edition. Boca Raton: Chapman and Hall/CRC. https://theeffectbook.net/.

Lupu, Noam, and Elizabeth J. Zechmeister. 2020. "Chapter 3. Social Media and Political Attitudes in the Latin American and Caribbean Region." In *The Political Culture of Democracy in Ecuador and in the Americas, 2018/19: Taking the Pulse of Democracy*, edited by Juan Carlos Donoso, Paolo Moncagatta, Arturo Moscoso, Simón Pachano, J. Daniel Montalvo, and Elizabeth J. Zechmeister, 1st ed., 1:45–72. AmericasBarometer Country Studies 2018/19. Quito: Latin American Public Opinion Project. https://www.vanderbilt.edu/lapop/ecuador/AB2018-19-Ecuador-Country-Report-Eng-V2-W-200903.pdf.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Redacción Primicias. 2020. "Mapa electoral: cuánto pesa el voto de las 24 provincias." Primicias. October 12, 2020. https://www.primicias.ec/noticias/politica/mapa-electoral-pesa-voto-provincias/.

Stanley, Matthew L, Christopher B Neck, and Christopher P Neck. 2023. "Loyal Workers Are Selectively and Ironically Targeted for Exploitation." *Journal of Experimental Social Psychology* 106: 104442.

Wooldridge, Jeffrey M. 2020. *Introductory Econometrics: A Modern Approach / Jeffrey M. Wooldridge*. Seventh edition. Cengage.