

Intro 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.

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

1 Course 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 sister module to this course, *Intro to R*, will focus on the practical implementations and technical aspects, while this course will focus on theory and intuition. 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.

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](#). See below for details on these readings.

- **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
 - **Assigned reading:** Course syllabus, Brown (1998)
- **Lecture 2:** Descriptive statistics and statistical data visualization
 - Measures of central tendency: mean, median, mode
 - Measures of dispersion: variance, standard deviation, range, IQR
 - Measures of distributional shape: skewness, kurtosis
 - 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
 - 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 (α)
- Margin of error with sampling distributions
- Confidence intervals and their interpretation, hypothesis testing with CIs
- **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:

3.1 Assignments (Prelim)

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. They are due by 11:59pm on the Sunday of the week they are assigned, however, please check the DataCamp platform for the exact due date. As mentioned above, Stata exercises will be provided, but not graded. ## Readings

There will be a reading assignment 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. 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.

3.2 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.

3.3 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 frustration, 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.

4 Readings and course materials

5 Software

As mentioned, 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 be provided to those that want it. I will not introduce the use of Stata, as it is assumed all students have already taken and successfully 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.

6 Module policies

6.1 Communication

All communications to the instructor or teaching assistant (TA) should be made through the course's Slack channel. We hope to respond to questions within 72 hours, but please be patient if we take longer.

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