Essentials of Applied Data Analysis

IPSA USP Summer School 2018

January 08-12, 2018

Leonardo Sangali Barone
Post-Doc Fellow at the Department Political Science, USP
leonardo.barone@usp.br

Contents

1 Course Overview

This course is designed for students who are interested in reviewing their training in data analysis and statistics. It prepares students for courses offered in the IPSA-USP Summer School that require the application of basic concepts of probability, random variables and statistical inference. The course will take place in the week preceding the commencement of the regular courses in the Summer School.

In the course, we will review the following topics: descriptive statistics and basic data analysis; probability theory and applications to social science problems; random variables and distributions; confidence intervals and hypothesis testing; and (a very brief introduction to) regression analysis.

By the end of the intensive one-week course, students should be able to: 1- understand and provide solutions to basic probability and inference problems; and 2- apply the fundamental concepts in probability theory and statistics to social science research questions.

This course departs from the premise that the most effective way to learn statistics is by actively engaging in doing the statistical analysis. For each topic, we will have lectures that will be followed by sessions in which students will use data to answer questions that are important to political scientists. Similar to other 1st week IPSA-USP Summer School courses, this course takes a "hands on" approach. Students will apply the concepts taught in lectures to analyse problems in quantitative social science research using software packages including Stata and R.

2 Prerequisites

The course presumes students have some training in high school mathematics.

3 Outline of the Course

1. Math review: notations, algebra, functions and set theory
2. Introduction to data analysis in the social sciences
3. Descriptive analysis
4. Probability: sets and counting

5. Probability: definitions and axioms
6. Probability: conditional probability and independence
7. Probability: Bayes' theorem and the bayesian approach
8. Probability: random variables
9. Probability: probability distributions
10. Probability: expectation and variance
11. Probability: large sample theorems
12. Inference: standard error
13. Inference: point and interval estimation
14. Inference: hypothesis testing
15. Inference and prediction: linear regression

4 Readings

We are going to use two main books in the course:

- (*Imai*) Kosuke Imai. *Quantitative Social Science: An Introduction.* Princeton University Press, 2017
- (*M&S*) Will H. Moore and David A. Siegel. *A Mathematics Course for Political and Social Research.* Princeton University Press, July 2013

I highly recommend reading entirely both *M&S* and *Imai* during the first year of your graduate program. These two books cover, respectively, the content of a full and extensive math bootcamp and the most important topics of an introductory quantitative analysis course for a political science or sociology program. That's more math, statistics and programming than you will probably need in the first years of your career.

This year we will experiment with the recently published *Imai*'s' book. We will work *M&S* in the first day of class.

The first 3 chapters of *M&S* covers the basic mathemtics we need to know prior to learning statistics. Besides these introductory chapters, we will also use 3 more chapters as support material, though these are not required reading. The Mathematics for Social Sciences course regularly offered by IPSA-USP Summer School usually covers the remaining chapters of this book. As an alternative to *M&S*, you can read Gill's book, indicated below.

In most of the lectures we will work with *Imai*'s. The book has a higlhy innovative approach to introductory quantiative social science and we will (almost) try to stick to it. We will go through chapters 3, 6, 7 and 4 , in this order.

In the last two years the main reading has been Sheldon Ross's book, indicated below. It is an excelent introductory statistics book and it covers the entire program. Although Ross doesn't write books for social science students, the book's language is very accessible and the examples are sufficiently good. You can use it as an alternative reading to both *Imai* and *M&S.* I indicate for each class the correspondent reading for this books.

By the way, there is plenty of alternatives. I indicate some below (Sirkin; Agresti and Finlay are also good choices that I have used in the past). Statistics books are a matter of taste: pick the book that you fell comfortable with and that you can understand. If you have used a book in the past that you like, locate the topics of the syllabus in the book and go with it.

Please, be aware that we are covering in just a few hours a ton of material. Organize yourself in advance and come prepared. I recommend that you read all of the pre-course readings (of course) and the readings for the first day before we start.

Alternative readings

- (*M&S*) Sheldon M. Ross. *Introductory Statistics.* Academic Press, January 2010
- Jeff Gill. *Essential Mathematics for Political and Social Research.* Cambridge University Press, April 2006
- Alan Agresti and Barbara Finlay. *Statistical Methods for the Social Sciences.* Pearson Prentice Hall, 2009 (*)
- Seymour Lipschutz. *Probability.* McGraw-Hill Book Company, 1968 (*)
- Paul Kellstedt and Guy Whitten. *The Fundamentals of Political Science Research.* Cambridge University Press, May 2013 (*)
- R. Mark Sirkin. *Statistics for the Social Sciences.* SAGE, 2006

- Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, October 2015 (*)

Fun readings

- Leonard Mlodinow. *The Drunkard's Walk: How Randomness Rules Our Lives*. Penguin UK, April 2009 (*)
- Alex Bellos. *Alex's Adventures in Numberland*. A&C Black, April 2011 (*)
- David Salsburg. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt and Company, May 2002 (*)
- Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016 (*)

(*) editions in Portuguese available.

5 Course Schedule

Day 1 – Data analysis, causality, measurement and applied descriptive analysis in the social sciences

Topics: Notation; Algebra review; Functions and sets; Introduction to data analysis in the social sciences and examples; descritive analysis.

We will start our class by reviewing the mathematial notation, introductory algebra, functions and set theory.

Conventionally, introductory statistics courses start with descriptive analysis. Following *Imai*'s approach, we will learn a little bit of data analysis by going though some popular examples from published research before we go into how to describe data.

Descriptive analysis will be the topic of our first laboratory. It's better to learn it by doing. Learning how to manage data, generate summary statistics, tables, graphs and other types of data visualization with statistical software is our first application goal.

Readings: *M&S* - chapter 1 (p. 3-27); chapter 2 (p. 28-43); chapter 3 (p. 44-74 *Imai* - chapter 3 (3.1 to 3.6).

Alternative readings: *Ross* - chapter 2 (p. 17-70); chapter 3 (p. 71-138).

Day 2 – Probability fundamentals

Topics: Axioms and definitions of probability. Conditional Probability and Bayes' Rule. Independence.

In the second day we will learn the basics of probability. This is, by far and large, the most important class of the course. We will focus on the understanding of the basic rules and, specially, on the ideas of independence and conditional probability. Besides some computational applications, we will work a lot with paper and pencil.

Readings: *Imai* - chapter 6 (6.1 and 6.2).

Alternative readings: *M&S* - chapter 9 (p. 175-197); *Ross* - chapter 4 (p.139-208).

Day 3 – Random variables, probability distributions and large sample theorems

Topics: Random Variables. Discrete and continuous distributions. Bernoulli, Binomial and Normal distributions. Expectation and variance. Large Sample Theorems.

Random variables – and the correspondent distributions – are the building blocks of data analysis. We will work on this topic on the third day and prepare the basis for sampling theory and statistical inference. We will also work a little bit with simulations and software applications. Finally, we will learn the basics of sampling theory and the fundamental theorems, in particular the Central Limit Theorem, that are in the basis of hypothesis testing. One more time, we will spend some time with paper and pencil applications plus some computational examples.

Readings: *Imai* - chapter 6 (6.3 and 6.4).

Alternative readings: *M&S* - chapter 10 (p. 198-241); chapter 11 (p. 242-272); *Ross* - chapter 5 (p. 209-259); chapter 6 (p. 261-296).

Day 4 – Uncertainty and statistical inference: estimation and hypothesis testing

Topics: Standard error. Point and interval estimation. Hypothesis Testing.

This is the core class of our course in terms of real world applicability (but not the most important, in my point of view). This fourth class is going to be a very brief introduction to estimation and hypothesis testing. We will extensively work with laboratory examples in the fourth day.

Readings: *Imai* - chapter 7 (7.1 and 7.2).

Alternative readings: *Ross* - chapter 7 (p. 297-330); chapter 8 (p. 331-386); chapter 9 (p. 387-442); chapter 10 (p. 443-502).

Day 5 – Prediction, linear regression and uncertainty

Topics: Linear regression; LR and prediction; LR and hypothesis testing.

Finally, we will go back to Chapter 4 of *Imai*'s book to learn a little bit of linear regressions. We will quicly go through the mechanics of linear regression, but our focus will be how to build and interpret models using regression analysis for both prediction (chapter 4) and hypothesis testing (chapter 7).

Readings: *Imai* - chapter 4 (4.1 and 4.2); chapter 7 (7.3).

Alternative readings: *Ross* - chapter 12 (p. 537-573); chapter 9 (p. 387-442)

6 Course Time Schedule

Class starts at 9h00. At 12h00 we have a lunch break. We usually go back to class at 13h30. On Tuesday and Thurday there will be an information sessions from 13h30 to 14h30. Monday is the only exception, when everything happens half an hour later due to registration (8h30 to 9h00). Class finishs at 18h00, except on Friday, when it will finish at 17h00 (but I will probably hold you an extra hour to rush and finish the last topic). The first information session is "Research Funding Support through Fapesp" with Professor Eduardo Cesar Marques (University of São Paulo, CEM and Fapesp). The second is "Publishing Academic Research to Maximize Impact in Shaping Public Policy" with Bruno Cautrès (Sciences Po) and Lorena Barberia (University of São Paulo).

	Morning	**Info. Session**	**Afternoon 1**	**Afternoon 2**
Mon, Jan. 08th	Data Analysis in the Social Sciences		Intro to statistical packages	Applied descriptive analysis
Tue, Jan. 09th	Introduction to Probability	Research Funding Support through Fapesp	Probability problems	Probability problems
Wed, Jan. 10th	Random Var., Distributions and Large Sample Theo.		Applied Basic Statistics	Random variable problems
Thu, Jan. 11th	Estimation and Hypothesis Testing	Publishing Academic Research to Max. Impact in Shaping Public Policy	Applied Hypothesis Testing	Applied Hypothesis Testing
Fri, Jan. 12th	Linear Regression basics		Applied Regression Analysis	Applied Regression Analysis

References

[1] Alan Agresti and Barbara Finlay. *Statistical Methods for the Social Sciences.* Pearson Prentice Hall, 2009.

[2] Alex Bellos. *Alex's Adventures in Numberland.* A&C Black, April 2011.

[3] Jeff Gill. *Essential Mathematics for Political and Social Research.* Cambridge University Press, April 2006.

[4] Kosuke Imai. *Quantitative Social Science: An Introduction.* Princeton University Press, 2017.

[5] Paul Kellstedt and Guy Whitten. *The Fundamentals of Political Science Research.* Cambridge University Press, May 2013.

[6] Seymour Lipschutz. *Probability.* McGraw-Hill Book Company, 1968.

[7] Leonard Mlodinow. *The Drunkard's Walk: How Randomness Rules Our Lives.* Penguin UK, April 2009.

[8] Will H. Moore and David A. Siegel. *A Mathematics Course for Political and Social Research.* Princeton University Press, July 2013.

[9] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown, 2016.

[10] Sheldon M. Ross. *Introductory Statistics.* Academic Press, January 2010.

[11] David Salsburg. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* Henry Holt and Company, May 2002.

[12] R. Mark Sirkin. *Statistics for the Social Sciences.* SAGE, 2006.

[13] Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach.* Cengage Learning, October 2015.