Essentials of Applied Data Analysis

IPSA USP Summer School 2017
January 23-27, 2017

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 Excel and Stata.

2 Prerequisites

The course presumes students have some basic training in mathematics including arithmetic and algebra operations.

3 Outline of the Course

1. Descriptive Statistics in the Social Sciences
2. Toolkit for Data Analysis and Statistics
3. Probability: the basics
4. Probability: Conditional Probability and Independence

5. Probability: Bayes' theorem and the bayesian approach
6. Probability: Random Variables
7. Probability: Discrete and Continuous Probability Distributions
8. Statistical Inference: Sampling. Law of Large Numbers. Central Limit Theorem.
9. Statistical Inference: Point and Interval Estimation. Hypothesis Testing
10. Statistical Inference: Bivariate Regression Analysis

4 Readings

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

- (*M&S*) Will H. Moore and David A. Siegel. *A Mathematics Course for Political and Social Research*. Princeton University Press, July 2013
- (*Ross*) Sheldon M. Ross. *Introductory Statistics*. Academic Press, January 2010

I highly recommend reading the entire *M&S* during the first year of your graduate program. It covers the content of a full and extensive math bootcamp for a political science or sociology program. That's more math than you will probably need in the first years of your career (although not enough statistics!). As you will see below, the first 3 chapters of *M&S* book are required pre-course readings. During the course, we will also use 3 more chapters. The Mathematics for Social Sciences course regularly offered by IPSA-USP Summer School (however, not this year) usually cover the remaining chapters. As an alternative to *M&S*, you can read Gill's book, indicated below.

In most of the lectures I will work with *Ross*'s book. It is an introductory statistics book and it covers the entire program. Although Ross doesn't write books for social science students, the book language is very accessible and the examples are sufficiently good. There are a lot of alternatives for this book. I indicate some below (Sirkin; Agresti and Finlay are good alternatives that I have used in the past and I will take examples from these two books). Statistics book is 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 in the book and use 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 classes start.

Alternative readings

- 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
- Alan C. Acock. *A Gentle Introduction to Stata, Second Edition.* Stata Press, September 2008
- Adrian Colin Cameron and P. K. Trivedi. *Microeconometrics using Stata.* Stata Press, 2009
- R. Mark Sirkin. *Statistics for the Social Sciences.* SAGE, 2006
- Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach.* Cengage Learning, October 2015 (*)

Bayesian and some advanced readings

- Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Chapman and Hall/CRC, Boca Raton, December 2015
- Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis.* Texts in Statistical Science. Chapman & Hall/CRC, 3rd edition, 2013
- Simon Jackman. *Bayesian Analysis for the Social Sciences.* John Wiley & Sons, October 2009
- Sheldon M. Ross. *A First Course in Probability.* Pearson Education, 2014
- Sheldon M. Ross. *Introduction to Probability Models.* Elsevier, May 2014

- Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, October 2010

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 (*)

(*) editions in Portuguese available.

5 Course Schedule

Day 0 - Pre-Course math review

Topics: Notation; Algebra Review; Functions and Sets.

Please, review the basics of mathematical notation, algebra, functions and set theory prior to the beginning of the course. Moore and Siegel's book is an excellent pre-course reading. Gill is a good alternative. I will be available after January, 4th to answer questions by e-mail or by appointment in the week before Summer School starts.

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

If you want some interesting Holiday reading, I highly recommend both Leonardo Mlodinov and Alex Bellos' books (see Fun readings above).

Fun Readings: Mlodinov (the whole book); Bellos (chapters 9 and 10).

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

Topics: Introduction to data analysis in the social sciences and examples. Tool kit for the social science data analyst. Introduction to Stata.

We will start by reviewing the basics of data analysis in the social sciences and go through some examples of how statistics and mathematics can be useful in social research. This

first day will be a lot more light and we will lean a little bit of descriptive statistics and learn how to manage data, generate summary statistics, tables, graphs and other types of data visualization in Stata.

Please, understand that we will not review the mathematics in the first day. This is a pre-course task.

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

Day 2 – Probability and random variable fundamentals

Topics: Counting. Sets. Probability. Conditional Probability and Bayes' Rule. Independence.

In the second day we will learn the basics of probability. We will focus on the understanding of the basic rules and, specially, of the ideas of independence and conditional probability. Differently from the first day, we will not concentrate on working with Stata, but with paper and pencil.

Readings: *M&S* - chapter 9 (p. 175-197) OR *Ross* - chapter 4 (p.139-208).

Day 3 – Probability Distributions - discrete, continuous

Topics: Random Variables. Discrete Random Variables. Continuous Distribution Topics. From Binomial to Normal.

Random variables 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.

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

Day 4 – Statistical Inference - Sampling, Central Limit Theorem and Hypothesis Testing

Topics: Sampling. Sampling Distributions. Law of Large Numbers. Central Limit Theorem. Point and Interval Estimation. Hypothesis Testing.

This class is the core class of our course. 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. We will extensively work with laboratory examples in the fourth

day.

Readings: *Ross* - chapter 7 (p. 297-330); chapter 8 (p. 331-386); chapter 9 (p. 387-442)

Day 5 – Statistical Inference for two populations and Bivariate Regression Analysis

Topics: Hypothesis Testing. Population Regression Model. Sample Regression Model. Linear Regression with a Single Variable.

The last class is going to be a very brief introduction to hypothesis testing for two populations an to linear regression with a single explanatory variable. Again, we will extensively work with laboratory examples during the last day.

Readings: *Ross* - chapter 10 (p. 443-502); 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 go back to class at 13h30, except on Wednesday, when there will be an information session with Prof. Derek Beach ("Controversies and Innovations in Qualitative Research: Transparency and Replication") fron 13h30 to 14h30. 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).

	Morning	**Info. Session**	**Afternoon 1**	**Afternoon 2**
Monday, January 23th	Data Analysis in Social Sciences		Data Analyst Toolkit	Introduction to Stata
Tuesday, January 24th	Introduction to Probability		Probability problems	Probability problems
Wednesday, January 25th	Random Var. and Distributions	Controversies and Innovations in Quali. Research:Transparency and Replication	Applied Basic Stats (Stata)	Cond. Probability problems
Thursday, January 26th	Sampling, CLT and Ho Testing		Applied Ho Testing (Stata)	Inference problems
Friday, January 27th	Two Populations and Regression basics		Applied Regression with Stata	Data Analyst Toolkit II

References

[1] Alan C. Acock. *A Gentle Introduction to Stata, Second Edition*. Stata Press, September 2008.

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

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

[4] Adrian Colin Cameron and P. K. Trivedi. *Microeconometrics using Stata*. Stata Press, 2009.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, 3rd edition, 2013.

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

[7] Simon Jackman. *Bayesian Analysis for the Social Sciences*. John Wiley & Sons, October 2009.

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

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

[10] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Boca Raton, December 2015.

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

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

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

[14] Sheldon M. Ross. *A First Course in Probability*. Pearson Education, 2014.

[15] Sheldon M. Ross. *Introduction to Probability Models*. Elsevier, May 2014.

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

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

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

[19] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data.* MIT Press, October 2010.