First Year Training in Statistical Methods

RWJ Clinical Scholars Program at Yale School of Medicine
2015/2016

Primary Instructor: Douglas McKee, PhD Economics (douglas.mckee@yale.edu, 310-266-2438)
Lab Instructor: Laura Cramer, PhD Health Services Research, MS Biostatistics
(lauracramer@optonline.net, 203-241-2816)

Goals:

This year-long curriculum is designed for physicians who start with no training in quantitative methods. At the end of the year they should be able to:

- Understand the math and intuition behind the core quantitative methods used in medicine, public health, health care, and health policy research
- Understand a variety of advanced statistical methods and the problems they solve
- Choose appropriate statistical methods to answer substantive research questions
- Build sensible statistical models of causal processes
- Use Stata to manage data and apply both the basic and advanced methods we've covered
- Present results of statistical analysis in tables and figures
- Digest and critique modern quantitative research

No one year class can turn you into a seasoned researcher, but it can give you the tools to become one.

Class Structure:

The training will divided into three distinct parts.

Summer: 5 intensive weeks

We will cover the core methods and concepts of biostatistics. These will form the foundation of everything we do during the rest of the year. Lectures will be two hours long and meet twice a week. In addition there will be a two hour lab at the end of each week.

Fall: 10 weeks

Scholars will learn linear and logistic regression in depth and how to use the tools learned thus far to answer substantive questions. Lectures (2 hours) will be once a week and labs (2 hours) will be approximately every other week.

Spring: 12 weeks

Scholars learn a whole set of advanced statistical methods including multinomial models, count models, propensity score analysis, factor analysis and survival analysis. Scholars will also learn tools for dealing with missing data and complex survey weighting schemes. Lectures will again be once a week with labs every other week.

Lectures:

Lectures will include the basic math and intuition behind each method or concept we cover. Scholars will learn how to interpret estimation results and look at lots of examples. I will also demonstrate how to use Stata to manage data and apply the methods we cover in class. The style will be very interactive with lots of opportunities for scholars to ask questions and get answers.

Labs:

At the beginning of each lab session, scholars are given an assignment and a data set. Their job is to answer a substantive question by applying methods we've covered in class. There's no lecturing and no "skeleton program" that get filled in. The entire two hours is spent interacting with the computer with an expert (the lab instructor) nearby to answer questions. In this way, the experience is very different from working on a problem set where getting stuck on something small for hours at a time is common. Struggling with a problem is good for learning, but banging your head against a wall isn't an efficient use of time. In addition, scholars work in pairs and take turns driving. This keeps scholars focused and learning from each other.

The end product of most labs will be a report containing results (similar to what might be found in a published paper), textual interpretation and Stata code. Sometimes I'll provide an empty table that must be filled in and other times scholars will produce their own tables of results from scratch. It is expected that completing some lab reports may take time beyond the two hours spent in the lab. All lab reports should be emailed to Laura Cramer and Doug McKee by the end of the weekend following the lab. You can expect written feedback on the labs once or twice per term.

While it's possible to use any statistical analysis tool in a lab successfully, some packages are better than others. Stata allows easy browsing of data in a spreadsheet style interface. You can play with commands through the menus and when you choose one, it shows you the command-line equivalent. You can work interactively at the command-line or build programs (using those same commands) in an editor. The documentation is excellent and available online.

Progress Assessment:

Scholars need objective assessments of their progress through the training in part because students are notoriously bad at recognizing when they do or do not understand material presented in lecture. I have found exams to be a poor way to evaluate such progress since it is very difficult to design an exam that tests the "real world" skills being acquired here.

During the year, Laura and I will take a close look at a subset of the scholars' lab reports. We will provide feedback to scholars about whether they are on track or if they need to put more time into the class. Just as important, I will learn what topics need to be reviewed or re-taught. In addition to information, evaluation of these lab reports should provide motivation for scholars to review material and work outside the classroom.

Reading:

Every lecture will have accompanying written material that scholars can read to prepare for class, read afterward to get an alternative perspective, and use as reference material when doing their own research.

The core topics we cover in the summer are discussed in detail in two biostatisics textbooks. The Rosner book is a traditional math-based biostatistics textbook. The Motulksy book is just what it says: chock full of intuition with just a few equations. The class will follow the Rosner book more closely, but some of you may appreciate the alternative presentation in the Motulsky book.

- Rosner B. *Fundamentals of Biostatistics*, 7th Edition. Duxbury Press, 2010.
- Motulsky H. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking, 3rd Edition.* Oxford University Press, 2013.

Linear Regression is covered in some detail in both of the biostatistics books, but I think Allison does a better job. Pampel clearly explains how logistic regression works.

- Allison PD. *Multiple Regression: A Primer.* Thousand Oaks, CA: Pine Forge Press, 1999.
- Pampel. *Logistic Regression: A Primer* Sage Publications, 2000.

Angrist and Pischke cover a wide range of modern econometric techniques that are now being applied to health and health care. The book is mathematical, but is written for both academic and non-academic audiences:

- Angrist JD, Pischke JS. *Mostly Harmless Econometrics: An Empiricists Companion.* Princeton University Press, 2009.

Personally, I think you can learn Stata very efficiently through a combination of the excellent documentation and free online resources. But for those scholars who want a book, this one is up to date and pretty good:

- Acock, AC. *A Gentle Introduction to Stata, Fourth Edition.* Stata Press, 2014.

Some advanced topics are better learned with a text that walks you through actual code that analyzes data instead of only with intuition and math. For that reason, I use the following two books for most of the spring term:

- Long JS, Freese J. *Regression Models for Categorical Dependent Variables Using Stata, Third Edition*, Stata Press, 2014.
- Cleves M, Gutierrez RG, Gould W, Marchenko YV. *An Introduction to Survival Analysis Using Stata, Third Edition*, Stata Press, 2010.

We will use Chapter 13 (Principal Components and Factor Analysis) from Tabachnick and Fidell's book:

- Tabachnick BG, Fidell LS. *Using Multivariate Statistics, Sixth Edition*, Prentice Hall, 2012.

The lecture on Power Analysis will focus on basic concepts and actually doing power calculations in Stata. For that reason, you will only need to read the first chapter of Cohen's classic text (which I'll scan and distribute).

- Cohen J. *Statistical Power Analysis for the Behavioral Sciences, Second Edition*, LEA, 1988.

I will also recommend journal articles throughout the term.

Acknowledgements

This class owes a great debt to five very generous people:

1. Many of the early lectures on probability, statistics, and testing are derived from an econometrics class that Lanier Benkard taught at Yale in Fall 2010. While most of the example applications are new, the interactive and applied nature of his lectures provided a terrific starting point.
2. Vida Maralani (Yale Sociology) graciously shared her slides and notes for her graduate-level Quantitative Methods class. Much of my material on data visualization and Stata was derived from her slides and hand-outs.
3. Many of my lectures on linear and logistic regression come from a class I taught for two years in the Yale School of Public Health on Research Methods for Health Care Research. This was a class I inherited from Andy Epstein who handed over all of his slides, problem sets, and exams. I've modified and extended many topics, but the core owes a huge debt to Andy.
4. The labs over the summer are entirely built on data collected by Rani Hoff and she also suggested most of the themes in each lab.
5. The solutions for all the labs were carefully written and tested by Laura Cramer.

This class would be a pale shadow of itself if not for the work of all five of these people.
Thank you!

Schedule:

PART I: Core Biostatistics (Summer)

Week 1 Session 1: Introduction—Populations, Samples, Models, Variables and Statistics

Reading: Rosner, Chapter 1; Motulsky, Chapters 1, 2, 3, 8

Week 1 Session 2: Probability

Reading: Rosner, Chapter 3

Week 2 Session 1: Stata Basics

Reading: Acock, Chapters 1–5

Week 2 Session 2: Describing Data Visually and Numerically

Reading: Rosner, Chapter 2; Motulsky, Chapters 7, 9

Week 3 Session 1: Discrete and Continuous Random Variables

Reading: Rosner, Chapters 4, 5

Week 3 Session 2: Estimation and Confidence Intervals

Reading: Rosner, Chapter 6; Motulsky, Chapter 4

Week 4 Session 1: Hypothesis Testing—One Sample

Reading: Rosner, Chapter 7; Motulsky, Chapter 16

Week 4 Session 2: Hypothesis Testing—Two Sample

Reading: Rosner, Chapter 8; Acock, Chapters 6, 7.1–7.8

Week 5 Session 1: Hypothesis Testing—Nonparametric Methods and Categorical Data

Reading: Rosner, Chapters 9, 10; Acock, Chapter 7.11

Week 5 Session 2: Analysis of Variance (ANOVA)

Reading: Rosner, Chapter 12.1-12.5; Motulsky, Chapter 39; Acock, Chapter 9

PART II: Regression and Model Building (Fall)

Week 1: Review and Preview

Week 2: Bivariate Regression

Reading: Rosner, Chapter 11.1–11.3; Acock, Chapter 8

Week 3: Statistical Inference and Multiple Regression

Reading: Rosner, Chapter 11.4–11.8; Allison, Chapters 1, 2

Week 4: Data Management in Stata

Reading: Acock, Chapters 3, 4

Week 5: Joint Hypothesis Tests

Reading: Acock, Chapter 10

Week 6: Interactions and Transformations

Reading: Allison, Chapter 8

Week 7: Causality and Evaluation

Reading: Parker SW, Teruel GM. "Randomization and Social Program Evaluation: The Case of Progresa." *Annals of the American Academy of Political and Social Science,* 599:199-219, 2005.

Week 8: Model Building and Selection

Reading: Allison, Chapter 3

Week 9: Logistic Regression—Logic and Interpretation

Reading: Rosner, Chapter 13.8; Pampel Chapters 1,2; Acock Chapter 11

Week 10: Logistic Regression—Estimation and Model Fit

Reading: Pampel Chapter 3

PART III: Advanced Methods (Spring)

Week 1: Power Analysis

Reading: Cohen, Chapter 1

Week 2: Difference-in-Differences

Reading: Dowd B, Town R; "Does X Really Cause Y? Changes in Health Care Financing and Organization," brief, 2002. (pp. 1-10); Angrist and Pischke, Chapter 5.1–5.2

Week 3: Multinomial Models

Reading: Long and Freese, Chapter 6

Week 4: Ordered Models and Count Models (Poisson and Negative Binomial)

Reading: Long and Freese, Chapters 5, 8

Week 5: Missing Data and Sampling Weights

Reading: Allison P, *Missing Data*, 2001; Acock, Chapter 13; Stata User's Guide Chapter 20.22; Stata Survey Data (SVY) Manual, pp. 3–24

Week 6: Instrumental Variables

Reading: Angrist and Pischke, Chapter 4.1 and 4.4; Garabedian et al ("Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research," *Annals of Internal Medicine*, 2014)

Week 7: Regression Discontinuity

Reading: Angrist and Pischke, Chapter 6

Week 8: Propensity Scores

Reading: D'Agostino RB; "Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group," *Statistics in Medicine*, 1998.

Week 9: Factor Analysis

Reading: Tabachnick and Fidell: Chapter 13 (Principal Components and Factor Analysis)

Week 10: Basic Survival Analysis

Reading: Cleves M, et al Chapters 1–4, 8, 9

Week 11: Fixed Effects and Random Effects

Reading: Angrist and Pischke, Chapter 5

Week 12: Trend Analysis

Reading: Zeger, Irizarry, and Peng ("On Time Series Analysis of Public Health and Biomedical Data", *Annu. Rev. Public Health*, 2006)