How to Teach Quantitative Methods to Social Science Students: The Princeton Experience

Kosuke Imai

Department of Politics
Center for Statistics and Machine Learning
Princeton University

Q-Step Symposium, University of Warwick

November 24, 2015

Why Teach Quantitative Methods to Social Science Students?

- Massive technological changes $\rightsquigarrow$ Internet and computing revolution
- **Past**: only statisticians and methodologists analyzed data
- **Today**: EVERYONE is analyzing data

  In God we trust. All others must bring data. — William Deming

- **Past**: government data, national survey data
- **Today**: more of old types of data and lots of new data
  - surveys
  - experiments
  - administrative records
  - social media data
  - GIS data
  - text, images, sounds, videos
- "Big (Social Science) Data" revolution inside and outside the academia
- We must teach students how to analyze data

How Well Are We Teaching? Let's Look at Some Data

- Non-politics introductory quantitative methods courses in social sciences:
  - 5 year average: 2008/09 – 2013/14
  - Economics, Psychology, Sociology, Public Policy

	Lectures	Assignments	Readings	Labs	Overall
Statistics	3.2	3.3	3.1	3.6	3.1
All courses	3.8	3.7	3.7	4.0	3.9

- Politics introductory quantitative methods courses:

	Lectures	Assignments	Readings	Labs	Overall
POL 245 (2014)	4.4	3.9	3.5	3.9	4.3
POL 245 (2015)	NA	4.0	3.4	NA	4.3
POL 345 (2011)	4.0	3.8	3.7	4.2	4.1

Why is Teaching Quantitative Methods Courses So Hard?

1. Students are NOT interested in statistics:

	Professor	Distribution Requirement	Departmental	Certificate Program	General Interest
Statistics	0%	20%	71%	3%	6%
All PU courses	6%	12%	32%	7%	42%

"Professor Imai tried hard to make statistics interesting. But, statistics is boring."

2. Students have weak mathematical and programming background

"as a person not naturally inclined towards statistics and probability, I don't feel at all qualified to pass judgement on how the course might have been improved."

New Teaching Strategies

1. Motivating students
   - Data analysis as a necessary tool for social science research
   - Data analysis as a useful skill for post-graduate career
2. Helping students learn efficiently
   - Short but frequent assignments
   - Hands-on instruction in computer labs
   - Outside-of-classroom assistance: online (Piazza and Perusall) or in-person

Traditional		**New**
paper-and-pencil statistics	$\rightsquigarrow$	data analysis
probability $\longrightarrow$ statistics $\longrightarrow$ data	$\rightsquigarrow$	data $\longrightarrow$ probability $\longrightarrow$ statistics
general $\longrightarrow$ application	$\rightsquigarrow$	application $\longrightarrow$ general $\longrightarrow$ application
toy examples	$\rightsquigarrow$	data from published research
lectures	$\rightsquigarrow$	computer labs
exams	$\rightsquigarrow$	projects

Freshman Scholars Institute as a Testing Case

- 6-week long summer school for 30 – 40 selected incoming freshmen
  - come from "disadvantaged" background
  - first generation college students, minority students
  - lack mathematical and computing background

- Goals:
  - transition them from high school to college
  - get them used to Princeton before the semester starts
  - offer head start by earning early Princeton course credits

- Similar programs at other schools: http://nyti.ms/1gjJOoU

- Hardest test of new teaching strategies at Princeton

Structure of the Course

- Module contents:
  1. Week 1: Introduction
  2. Week 2: Causality
  3. Week 3: Measurement
  4. Week 4: Prediction
  5. Weeks 5 and 6: Discovery

- Module format for each week:
  1. Two 50 minute lectures
  2. Two 80 minute computer lab sessions
  3. One 80 minute guest lecture from industry, small-group discussion: NYT, Facebook, Google, Political consulting firm, FiveThirtyEight, etc.
  4. Three optional tutoring sessions with additional exercises

- Assignments:
  1. 12 short non-graded pre-class assignments
  2. 4 problem sets with collaboration
  3. 1 take-home midterm without collaboration
  4. 1 final group project

The Textbook: A First Course in Quantitative Social Science

- Combines three essential components:
  1. social science research
  2. methodological concepts
  3. computer programming (using **R** and **RStudio**)
- Teaches data analysis before statistics:
  1. Introduction
  2. Causality
  3. Measurement
  4. Prediction
  5. Discovery
  6. Probability
  7. Uncertainty
  8. Next
- Contains about 50 data sets from published social science research
  1. Effects of raising minimum wage
  2. Hearts and minds in Afghanistan
  3. Forecasting election outcomes
  4. Who wrote the Federalist papers?
  5. Predicting race from surname
  6. Return to political office
- Additional exercises including **swirl** lessons available

Chapter 2: Causality

- Concepts:
  - causality and counterfactuals
  - randomized controlled trials and observational studies
  - confounding and selection bias
  - cross-sectional and before-and-after comparisons, difference-in-differences
  - mean, standard deviation, quantile
- Applications:
  - racial discrimination in labor market
  - social pressure and turnout
  - minimum wage increase and unemployment
  - efficacy of small classroom in early education
- Programming:
  - Introduction to **R** and **RStudio** (Chapter 1): loading and saving data, arithmetic operations, basic functions
  - logical and relational operators
  - conditional statements and subsetting

Chapter 3: Measurement

- Concepts:
  - random sampling
  - non-response bias
  - histogram and density
  - scatterplot and quantile-quantile plot
  - correlation
  - clustering and k-means algorithm
- Applications:
  - support for Taliban and the international forces in Afghanistan
  - academic integrity and changing minds on gay marriage
  - political ideology and political polarization
- Programming:
  - matrix and list
  - handling of missing data
  - basic graphics: barplot, histogram, scatterplot, quantile-quantile plot

Chapter 4: Prediction

- Concepts:
  - prediction error, classification error
  - linear regression and least squares
  - regression towards the mean
  - predicted values, residuals, and R^2
  - regression and causality
  - regression discontinuity design
- Applications:
  - pre-election polling and betting markets
  - facial appearance and election outcomes
  - women as policy makers in India
  - return to political office in Britain
- Programming:
  - loop
  - merging data sets

Chapter 5: Discovery

- Concepts:
  - text analysis: document-term matrix, tf-idf, topic discovery
  - network analysis: undirected/directed networks, weighted/unweighted networks, centrality measures, PageRank algorithm, community detection
  - spatial analysis: visualization through maps

- Applications:
  - Federalist papers and authorship prediction
  - Marriage network in Renaissance Florence
  - Twitter following among politicians
  - John Snow and Cholera
  - Expansion of Walmart

- Programming:
  - visualization through various **R** packages including animation

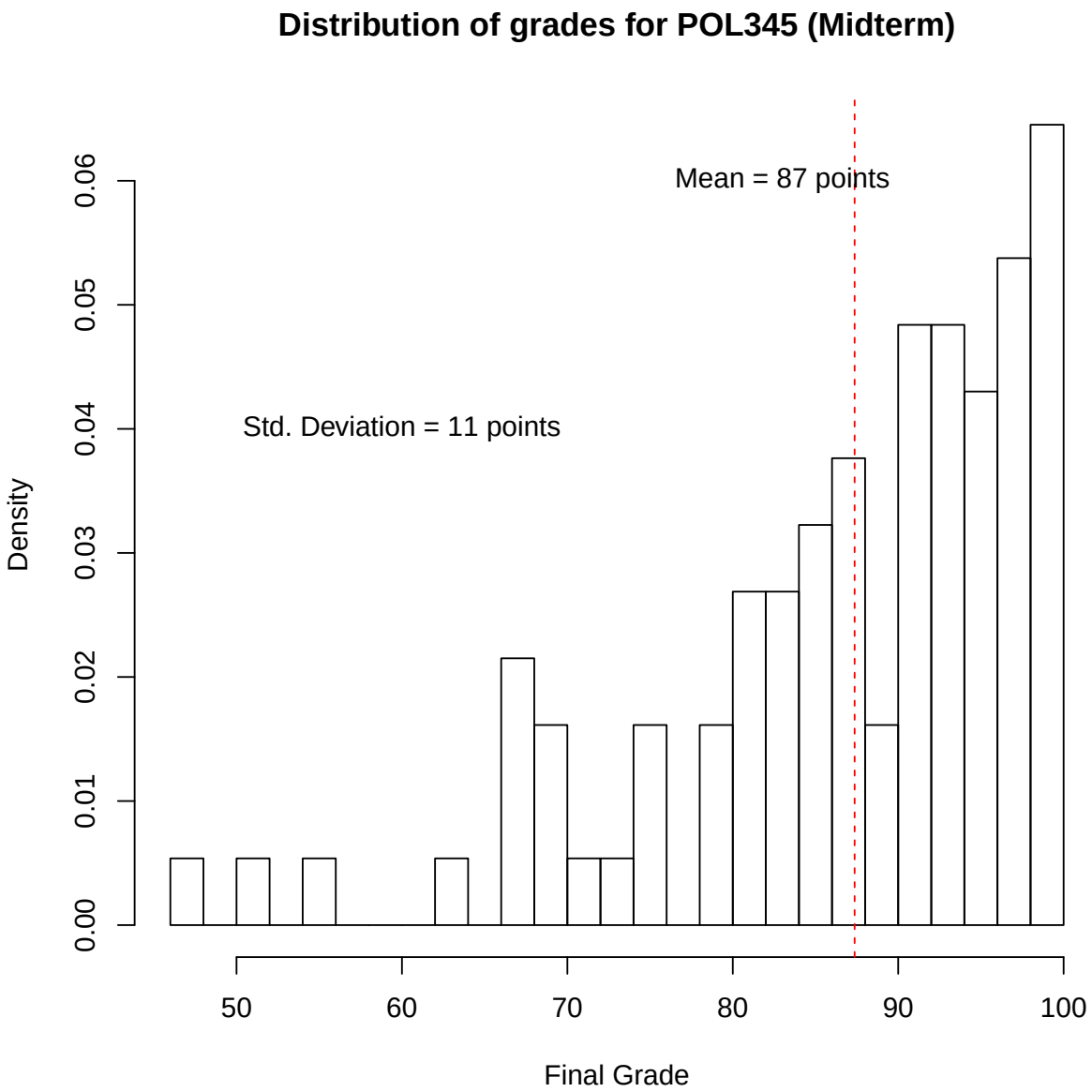
Chapter 6: Probability & Chapter 7: Uncertainty

- After lots of data analyses, students are motivated to learn difficult concepts
- Goal: quantify uncertainty of data analysis
- No new programming techniques, focus on methodological concepts
- A regular semester course (12 weeks):
  - first half $\rightsquigarrow$ data analysis
  - second half $\rightsquigarrow$ probability and statistics
- Chapter 6: Probability:
  - probability, conditional probability, and Bayes' rule
  - random variables, probability distributions
  - law of large numbers, central limit theorem
  - applications: predicting race using surname and residence, election fraud
- Chapter 7: Uncertainty
  - bias, standard error, confidence intervals
  - hypothesis testing
  - regression with uncertainty
  - applications: previous applications, who voted for Nazis, effects of 9/11

How Much Did Students Learn?

- Final project
  - group project (3 students)
  - start from data collection to data analysis
  - short write-up with 3 figures and 750 words
  - 5 minute presentation followed by Q&A
- Take-home exam
  - students must complete it within a week
  - open book, no collaboration, no assistance
- Electoral effects of Fox News (published in *Quarterly Journal of Economics*)
  1. examining balance of pre-treatment covariates
  2. examining balance using k-means algorithm
  3. recoding of a key variable, before-and-after comparison
  4. difference-in-differences
  5. placebo tests
- Emphasis on interpretation: semi open-ended questions

Impressive Performance



Other Measures of Success

- High numerical evaluation
- Students' feedback:

  "The course was a lot of fun and really interesting and I plan on taking the next level of the course."

  "I felt it gave me a very true sense of what to expect at Princeton."

- Diverse students in the next level of the course
- Increasing enrollment (over 3 years):
  - introductory course: 40 $\rightsquigarrow$ 100
  - advanced course: 5 $\rightsquigarrow$ 30
  - enrollment in graduate statistics courses
- Increasing use of quantitative methods in junior papers and senior theses
- Research assistantships, top PhD programs

Concluding Remarks

- Technological changes $\rightsquigarrow$ everyone must analyze data!
- paper-and-pencil statistics $\rightsquigarrow$ practical data analysis
- Goal: teach how cool quantitative social science research is
- Key: use of published research

- "A First Course in Quantitative Social Science"
  - brings together materials accumulated over years
  - early users: Columbia, American U., Dartmouth, NYU, Princeton, Stanford, Texas A&M, and UCSD
  - will be revised and published sometime next year
  - website with many more applications and contributions from instructors
  - comments and suggestions to kimai@princeton.edu