The Power of Numbers:
Introduction to Quantitative Social Science

Instructor: Sara Lopus
Email: saral@demog.berkeley.edu
Days, Time, Room: TuTh 3:30-5:00pm [203 Wheeler]
Office Hours: Th 5:15-6:30pm [484 Barrows] and by appointment

Overview:

Social statistics have become more widespread and more visible in recent years. From Nate Silver's quantitatively-driven prediction of the 2008 and 2012 elections, to the book *Moneyball*, to the sometimes euphoric claims made for "big data," numbers are now everywhere in the public sphere. But despite their ubiquity, these numbers are not always well understood. Some statistics seem so transparent that we do not think much about them. Others seem so opaque that we give up. Many of the numbers that circulate as common knowledge are not even right. Yet, they have the appearance of precision, and a certain social power, and so they stay on. Wrong numbers can have important social consequences. As citizens, professionals, social activists, and civic leaders, we need to recognize bad statistics and produce better ones.

There are many facets of society that can be effectively understood quantitatively. There are even most facets that can *only* be understood quantitatively. ***This course will provide students with a set of skills to understand, evaluate, use, and produce quantitative data about the social world.*** It is intended specifically for social science majors, and focuses on social science questions. ***You do NOT need a strong mathematical, statistical, or computing background to succeed in this course***. What you do need is a basic curiosity about how society is organized and a willingness to try something new. Our aim is to show you that quantitative social science is useful, can be fun, and is something that you can do.

By the end of this course, students will be able to:

- Understand, evaluate, and produce basic graphs
- Find good-quality and relevant data on the Web
- Manipulate data in a spreadsheet, including producing pivot tables
- Understand and calculate basic statistical measures of central tendency, variation, and correlation
- Understand and apply basic concepts of sampling and selection
- Recognize an impossible statistic
- And, based on all of these more concrete skills: **Begin to think about social science questions in a quantitative way**

Books and resources:
You must have a laptop that can access the web to take this course. If you want to take this course and do not have a laptop, see the instructor immediately. There are two textbooks for this course:

> Silver, Nate. (2012) <u>The Signal and the Noise: Why so many predictions fail—but some don't.</u> Penguin Books.
> Wheelan, Charles. (2013) <u>Naked Statistics: Stripping the dread from the data.</u> Norton.

All other readings will be posted on bCourses. We will focus on working through real applications, and we will rely heavily on web-based resources and tools. Links to these tools are provided in the calendar portion of this syllabus.

Expectations and evaluation:
Classroom time will be a combination of lecture, discussion, and practical work. Sometimes you will be required to read in advance of class; sometimes you will be required to watch a video lecture or try out a web-based application in advance of class. All of this prep work must be done as listed in the syllabus. If you are not prepared, the class will not be useful for you, and this will be reflected in your participation grade.

Group work: Over the course of the semester, you will work in groups on a comprehensive, collaborative project using a real dataset. As you learn new topics in class, you will use your dataset to apply the course material. Your group project will build as the sum of in-class activities and some home assignment. At the end of the semester, each group will do a class presentation in which you identify some interesting relationships you have discovered, presented in graphical forms. This project will be an opportunity to familiarize yourself with data, move it around, and work with it in useful ways.

Individual homework: There will be 10 individual homework assignments designed to introduce new topics or reinforce material covered in recent lectures. Grading of these assignments will focus more upon effort than correct answers, since the purpose is often to prepare you for activities performed in the next class. Late homework will not be accepted.

Optional extra credit: One or more optional extra credit assignments will be offered during the semester. If you choose to submit an extra credit assignment, it will be graded as a homework. Your homework grade will then be calculated based upon the ten highest scoring submissions. An optional extra credit assignment can therefore make up for one missed homework assignment or for a homework assignment on which you received a low grade.

Exams: There will be two exams: a mid-term and a final exam. These exams will be much more conceptual than computational, and they will focus on your understanding of the core concepts of the course.

Grading (1000 points possible):
Group work (35% of grade)

Written group work (2 assignments, 100 points each)	200 points
Group presentation	150 points

Individual homework (25% of grade)

10 assignments, 25 points each	250 points

Exams (30% of grade)

Midterm Exam	150 points
Final Exam	150 points
Participation (10% of grade)	100 points

Honor Code:
The student community at UC Berkeley has adopted the following Honor Code:
"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." We hope and expect that you will adhere to this code. This is Berkeley. We are too good to cheat.

There will be a good bit of collaborative work in this course. While this kind of collaboration is an important tool in learning and realistic preparation for the contemporary world of work, it also can make it hard to know what is acceptable collaboration and what is taking unfair advantage of others. Throughout the course, we will indicate whether a given assignment is to be completed alone or in cooperation with others. You will avow on each assignment that you complied with those instructions. If at any point you have any questions about how the honor code applies, or how best to fulfill your obligations as a member of the UC Berkeley community, please ask the instructor. Almost certainly, someone else has the same question.

Schedule of activities

Class Session	Date	Do before class	Do in class	Learning objectives
1A	01/20		*Lecture* How to read graphs. *Class Activity* Interactive scatterplot on [Gapminder website](#).	*Introduce* Reading graphs

1B	01/22	*Read*
		• Selection from Cook 2013
		• Wheelan Chapter 1 (*"What's the Point?"*)
			Lecture
			Units of analysis:
			   • Gapminder website (country-level data),
			   • [Eurostat](#) (county-level)
			   • [CalEnviroScreen](#) (zipcode-level), and
			   • [Demographic and Health Surveys](#) (individual-level).
			Video
			"200 Countries, 200 Years, 4 Minutes"
				Introduce
				• Units of analysis (scale)
				• Maps as graphs
				Reinforce
				Reading graphs

2A	01/27	*Read/Watch:* [Stark Chapter 5 (Multivariate Data, Scatterplots, Exercises 5-1 to 5-4)](#) **Homework #1** Due 01/26 before 10pm	*Lecture* Spreadsheet basics *Group activity* Join other students who chose the same scale for group project; practice spreadsheet basics.	*Introduce* Columns and rows, basic operations *Reinforce* Units of analysis (scale)
2B	01/29	*Read* [Stark Chapter 3 (Variables; Exercise 3-1)](#)	*Discussion* Are data qualitative? Quantitative? Categorical? *Group activity* What type of data is each variable?	*Introduce* Types of data
3A	02/03	*Read/Watch* • [Stark Chapter 4 (Measures of Location; Videos of Exercises; Exercises 4-1 through 4-5)](#) • Wheelan Chapter 2 (first half) *"Descriptive Statistics"* pp15-23	*Lecture* Means/modes *Group activity* Calculate means/modes for five variables from your dataset.	*Introduce* • Central tendency (mean, mode…) • Doing descriptive statistics *Reinforce* Columns and rows, basic operations
3B	02/05	*Read/Watch* Best Chapter 1, section "The public as innumerate audience"	*Discussion* Familiarizing ourselves with the scales of things about ourselves, our	*Introduce* • What are the scales of things? • Cocktail napkin

			community, and the world. *Partner Activity* Cocktail napkin estimates	estimates *Reinforce* • Central tendency (mean, mode…) • Doing descriptive statistics
		Homework #2 Due 02/04 by 10pm		
4A	02/10	*Read/Watch* • [Stark Chapter 4 (Spread or Variability; The Range, IQR, and SD; Videos of Exercises; Exercises 4-6 to 4-8)] • Wheelan Chapter 2 (second half) *"Descriptive Statistics"* pp23-35	*Discussion* What can does a standard deviation tell us? *Activity* 1. Plot the range and SD of everyone's Chihuahua estimates from the last class. 2. Estimate what you expect the range and mean of various things to be. *Group activity* With your dataset, determine some ranges, variances, and standard deviations.	*Introduce* Range, variance, standard deviation *Reinforce* • Central tendency (mean, mode…) • Doing descriptive statistics
4B	02/12	*Read* • Huff "How to lie with statistics" pp60-73 • Wheelan Chapter 3: *"Deceptive Description"* ***Homework #3*** Due 02/11 by 10pm	*Group activity* Share the graphs you brought.	*Introduce* Truths and lies in graphs
5A	02/17	*Read* • Tufte "The Visual Display of Quantitative Information" (*Pages to be determined*) • Wheelan Chapter 10 "Polling: How we know that 64 percent of Americans support the death penalty (with a sampling error +/- 3 percent)" pp180-183 • Best Chapter 1, section "Thinking about statistics as social products"	*Discussion* Data we trust. *Group activity* With your dataset, make one graph depicting a single variable. Make another graph that displays a relationship between two variables.	*Introduce* • Making Good Graphs • Dealing with sources we know *Reinforce* • Truths and lies in graphs • Doing descriptive statistics
5B	02/19	*Read/Watch* • [Stark Chapter 24 (Simple Random	*Group activity* How would we identify a random sample of rock	*Introduce* Setting limits and randomizing

		Samples, Systematic Random Samples, Exercise 24-4) • Wheelan Chapter 7 "The Importance of Data: 'Garbage in, garbage out'" pp110-113 ***Homework #4*** Due 02/18 by 10pm	musicians, intravenous drug users, polo players, private universities, or abortion clinics? Define the boundary conditions and sampling strategy. What would be the effect of each of these decisions?	
6A	02/24	*Read/Watch* • [Stark Chapter 13 (Theories of Probability, Random Events, Equally Likely Outcomes, Frequency Theory, Exercises 13-1 to 13-5)](#) • Silver Chapter 2 ("Are You Smarter than a Television Pundit?") • Silver Chapter 4 pp134-138 ("How to know if your forecasts are all wet") • Wheelan Chapter 5: *Basic Probability* • Wheelan Chapter 5 ½: *The Monty Hall Problem*	*Lecture* Examples 13-1 through 13-4 from Stark's Chapter 13 *Group Activity* Use urns, dice, playing cards, and [stattrek.com](#) to reinforce the fundamentals of probability theory. *Class Activity* As a class, create a decision tree Poll the class: Would you take the risk?	*Introduce* Understand basic probability (urns, expected value…)
6B	02/26	*Read/Watch/Listen* • [Stark Chapter 24 (Sampling, Parameters and Statistics, Why Sample?, Sampling from Hypothetical Populations, Exercises 24-1 and 24-2)](#) • Wheelan Chapter 6: *Problems with Probability* • [Radio Lab ("A very lucky wind" from Stochasticity)](#)	*Discussion* Sampling and the hypothetical counterfactual *Class activity* Sampling	*Introduce* Samples and universes *Reinforce* Understand basic probability (urns, expected value…)
7A	03/03	*Read/Watch* • Berk "An introduction to sample selection bias	*Class activity* Draw from an urn in which selection bias will influence the results. As a	*Introduce* Selection biases *Reinforce* • Making good graphs

			class, observe the resulting shape of the associated scatterplot.	• Columns and rows, basic operations
		in sociological data" • Stark Chapter 24 (Bias in Surveys, Exercise 24-3) ***Homework #5*** Due 03/02 by 10pm		
7B	03/05	***Group Assignment #1*** Due at the start of class today ***Optional extra credit*** Due 03/04 by 10pm	Midterm Review	--
8A	03/10	--	Midterm	--
8B	03/12	*Read* • Stark Chapter 23 (The Normal Curve) • Wheelan Chapter 8 ("The central limit theorem") • Wheelan Chapter 9 ("Inference") ***Homework #6*** Due at the start of class today	*Activity* Perform in-class exercises using interactive Normal Curve in Stark's Figure 23-1	*Introduce* Normal distribution, p-value
9A	03/17	*Read/Watch/Listen* • Stark Chapter 27 (Hypothesis Testing: Does Chance Explain the Results?, Examples of Hypothesis Testing Problems, Significance Level and Power, Test Statistics and P-Values, Exercises 27-1 to 27-3) • Wheelan Chapter 11 ("Regression Analysis") • Wheelan Chapter 12 ("Common regression mistakes") • The Data skeptic podcast #4 [p-values]	*Discussion* Discuss Examples 27-1 to 27-3 in Stark's Chapter 27 *Lecture* Interpreting a regression analysis output table	*Reinforce* Normal distribution, p-value
9B	03/19	*Read/Listen* • Silver Chapter 8 ("Less and less and less wrong") • The Data Skeptic	*Lecture* Review categorical vs. continuous data *Class activity* Types of error and	*Introduce* Hypothesis tests with continuous data (Student's t) *Reinforce*

		Podcast: #2 [Type I/Type II errors] (start with minute 2:30) • The Data Skeptic Podcast: #24 [The T-Test]	Bayesian probabilities *Group activity* Choosing an alpha to avoid Type I or Type II error	• Types of data • Normal distribution, p-value
10A-10B	03/23 – 03/27	SPRING BREAK	SPRING BREAK	SPRING BREAK
11A	03/31	*Read/Watch* Stark Chapter 31 (The chi-square statistic, The chi-square test for goodness of fit, Exercises 31-2, 31-4) **Homework #7** Due in class today	*Group activity* Compare your results for the Student's t-tests you performed at home. *Group activity* Write a hypothesis that explains a phenomenon relating to categorical data in your dataset. Perform a chi-square test.	*Introduce* Hypothesis tests with categorical data (Pearson's chi-square) *Reinforce* Hypothesis tests with continuous data (Student's t)
11B	04/02	*Read* Wheelan Chapter 4: *"Correlation"* **Homework #8** Due 04/01 by 10pm (bCourses)	*Class activity* Take an online survey on a few topics. Explore correlations between preferences. *Discussion* "Correlation does not imply causation." But why not?	*Introduce* Correlation *Reinforce* • Making good graphs • Doing descriptive statistics
12A	04/07	*Read* • Best Chapter 1, section "Thinking about statistics as social products" • Huff Chapter 10 ("How to Talk Back to a Statistic") • Silver Chapter 12 ("A Climate of Healthy Skepticism") • The Data Skeptic Podcast: #36 [Data Provenance]	*Lecture* Methods of evaluating data skeptically. Ranges, outliers, and other fishy characteristics. *Activity* Take an anonymous online survey	*Introduce* Dealing with sources we don't know… or trust yet!
12B	04/09	*Read/Watch* • Steele Chapter 2 ("Identity and Intellectual Performance") • Excel Pivot Table Tutorial	*Partner Activity* In pairs, draw a graph like that described in the reading (SATs, GPAs, race) *Class Activity* Use pivot tables to manipulate the anonymized data from last class period's survey.	*Introduce* Pivot tables
13A	04/14	**Homework #9**	*Discussion*	*Introduce*

		Due 04/13 by 10pm on bCourses	What makes advanced data visualization great? At what point does additional information confuse rather than inform? *Group Activity* Use a pivot table to experiment with data visualization in tables.	Advanced data visualization *Reinforce* Pivot tables
13B	04/16	*Read/Listen* • Wheelan Chapter 10 "Polling: How we know that 64 percent of Americans support the death penalty (with a sampling error +/- 3 percent)" pp169-180 • The Data Skeptic Podcast: #21 [Selection Bias]	*Activity* Take classroom polls on subjects, and then see how the sample compares to the overall classroom opinions. How does this vary based upon our selection of the sample? What if we sample only those who sit in the front? Only those who got to class late? Only those who are wearing hats?	*Introduce* Sampling as social process
14A	04/21	*Read* • Lewis Chapter 4 ("Field of ignorance") **Group Assignment #2** Due at the start of class today	*Group activity* Consider how the data in your dataset were collected. Would there be a role for big data in complementing the information in your dataset? What kinds of variables would be interesting, and where might the data already exist?	*Introduce* Big data, data science.
14B	04/23	*Read* • Freeman "Statistical Models and Shoe Leather" • Silver Chapter 5 ("Desperately seeking signal") • Wheelan Chapter 7 "The Importance of Data: 'Garbage in, garbage out'" pp113-126 **Homework #10** Due on 04/22 by 10pm on bCourses	*Discussion* The limitations of big data; the importance of research design in a world with (too much?) data available	*Reinforce* Big data, data science.
15A	04/28	--	Group presentations	--
15B	04/30	--	Group presentations	--

RRR	05/05	--	TBA: Final exam review	--
RRR	05/07	--	TBA: Final exam review	--
Final	05/15 7-10pm	--	Final exam	--

Readings include:

Berk, R. (1983). "An introduction to sample selection bias in sociological data." <u>American Sociological Review</u> **48**(3): 386-398.

Best, Joel. (2012) <u>Damned Lies and Statistics</u>. University of California Press.

Cook, Gareth (ed.) and David Byrne (introduction). (2013) <u>The Best American Infographics</u>. Mariner Books.

Foreman, John. (2014) <u>Data Smart: Using data science to transform information into insight.</u> Wiley and Sons.

Freedman, D. A. (1991). "Statistical Models and Shoe Leather." <u>Sociological Methodology</u> **21**: 291-313.

Huff, Darrell (1954) <u>How to Lie with Statistics</u>. New York: Norton.

Lewis, Michael. (2003) <u>Moneyball: The art of winning an unfair game.</u> W.W. Norton Company,Inc. New York, NY.

Silver, Nate. (2012) <u>The Signal and the Noise: Why so many predictions fail—but some don't.</u> Penguin Books.

Stark, Philip B. 1997-2013. <u>SticiGui</u>
http://www.stat.berkeley.edu/~stark/SticiGui/index.htm

Steele, Claude. (2010) <u>Whistling Vivaldi: How stereotypes affect us and what we can do.</u> New York: Norton.

Tufte, Edward. (2001) <u>The Visual Display of Quantitative Information</u>. Cheshire, Conn: Graphics Press.

Wheelan, Charles. (2013) <u>Naked Statistics: Stripping the dread from the data</u>. Norton.