Course on Design and Analysis of Sample Surveys

Andrew Gelman

19 Apr 2017

Abstract

Survey sampling is central to modern social science. In this course, we discuss how to design and analyze surveys, with a particular focus on my areas of expertise: public opinion polls in the United States and models for adjusting sample to population.

1. Goals for the students

By the end of the semester, you should be able to do the following things:

- Design a survey;
- Analyze data from a survey you have designed;
- Find and grab data from existing social surveys;
- Analyze data from existing social surveys.

2. Student responsibilities

- **Three or four times a week**, you will write an entry in your **survey sampling diary** (a special case of a *statistics diary*; see here: `http://andrewgelman.com/2015/01/07/2015-statistics-diary/`). Just set up a text or Word file and add to it every other day. The diary entries can be anything. They can be short slice-of-life observations ("Looking at faces on the subway this morning. Is it really true that people are less happy on Monday? How to measure this in a survey?"), quick questions ("Attitudes toward recreational drugs seem more permissive than in the past? Is this a real trend? If so, is it recent or has it been gradually happening for decades?"), research notes ("I'm comparing attitudes about military intervention in several European countries. Do I have to be concerned about question-wording effects in different languages?"), or things you're working on, difficult problems that you might be stuck on, or have an insight about. You can write as little or as much as you want each time. The only requirement is that you write something new in it, every other day. You're *not* allowed to go back a week later and fill in 3 entries at once. That would be cheating. Do it three or four times a week. Just type it in to the file.
- **Each week**, you will have two **homework assignments**. Each homework assignment needs to be uploaded to Courseworks and printed out and brought to class. Except when you are preparing slides, lay out the pages "portrait style" so they do not need to be rotated 90 degrees to be read. It's ok—encouraged, actually!—to include multiple graphs on a page and to mix graphics and text.
- **Before every class**, you will have **readings**. These include research articles, blog posts, various other online materials, and chapters from the two assigned books:
  - Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*, second edition. Wiley.

- Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* Wiley.

We'll also hand out chapters of these two books which are under preparation:

- Gelman, A., and Hill, J. (2017). *Regression and Other Stories.* Cambridge University Press, to appear.
- Gelman, A., and Hill, J. (2017). *Multilevel Regression.* Cambridge University Press, to appear.

- **Before every class**, you will have a **jitt** (just-in-time teaching assignment). Each of your jitts will be a set of three quick online items, separate from the main homework assignments, that are a mix of questions on the required reading, short exercises to get you ready for the upcoming class discussion, and feedback.

- **Each class** will involve your **active participation** in class discussion. Also, **bring your laptop computer** to class as we will be doing activities together in R.

- Attendance is required for regular meetings with the teaching assistant to review homework assignments and keep you up to speed on computing.

- **At the end of the semester**, you will have a final exam. An old exam is here: `http://www.stat.columbia.edu/~gelman/surveys.course/final2012.pdf` but this year's exam will be much different. It could be helpful to read these discussions of old exam questions: `http://andrewgelman.com/?s=my+final+exam+for+Design+and+Analysis+of+Sample+Surveys`

3. Structure of course

Introduction (week 1):
1a: Overview of the course
1b: Examples of surveys in the news
Statistics review (weeks 2–4)
2a: Basic statistics
2b: Linear regression basics
3a: More advanced linear regression
3b: Logistic regression
4a: Statistical graphics
4b: Causal inference overview
Classical design and analysis of surveys (weeks 5–7)
5a: Survey interviewing
5b: Survey measurement
6a: Simple and stratified random sampling
6b: Poststratification
7a: Survey weights
7b: Cluster sampling
Social and political science (weeks 8–10)
8a: Surveys in the United States
8b: Surveys in other countries
9a: Voting and political participation
9b: Public opinion

10a: Network sampling
10b: Survey experiments

Advanced analysis of survey data (weeks 11–14)

11a: Multilevel linear regression
11b: Multilevel logistic regression
12a: Item-response and ideal-point modeling
12b: Multilevel regression and poststratification
13a: Constructing survey weights
13b: Missing-data imputation
14a: Open problems in analysis of survey data
14b: Summary of the course

Class 1a: Overview of the course

Readings before class: None

Homework due at beginning of class: None

In class:

1. Lecture and discussion of several examples:
   (a) Tea party example (basic statistics, challenges in applying classical statistical principles)
   (b) Xbox example (American politics, survey nonresponse, Mister P)
   (c) Millenium Village (surveys in other countries, survey for causal inference)
   (d) Generations of presidential voting (elaborate analysis of survey data, open research questions)
2. Plan for the semester
   (a) Statistics review
   (b) Classical design and analysis of surveys
   (c) Social and political science
   (d) Advanced analysis of survey data
3. Special challenges with your own surveys:
   (a) Sampling frame
   (b) Finding potential respondents and getting them to respond
   (c) Measurement
   (d) Interviewing
   (e) Construction of weights, missing-data imputation, and poststratification
   (f) Ethics
4. Special challenges with surveys conducted by others:
   (a) Finding the data

   (b) Sampling frame and method of sampling
   (c) Clustering
   (d) Measurement
   (e) Weights and data adjustments

5. Structure of the course
   (a) Survey sampling diary
   (b) Homeworks
   (c) Statistical software
   (d) Readings
   (e) Jitts
   (f) Class participation

6. Discuss readings and next class

Class 1b: Examples of surveys in the news

Readings before class:

1. Andrew Gelman, notes on R: `http://www.stat.columbia.edu/~gelman/surveys.course/Rnotes.pdf`
2. Kumail Nanjiani, “Cheese heroin”: `http://www.youtube.com/watch?v=WVIC2gJTD9s`
3. Andrew Gelman, “Debunking the so-called Human Development Index of U.S. states”: `http://andrewgelman.com/2009/05/20/debunking_the_s/`
4. Andrew Gelman, “The General Social Survey is a great resource”: `http://andrewgelman.com/2011/10/14/the-general-social-survey-is-a-great-resource/`
5. Andrew Gelman, “Sports fans as potential Republicans?”: `http://andrewgelman.com/2009/01/27/sports_fans_as/`
6. Andrew Gelman, “Big corporations are more popular than you might realize”: `http://andrewgelman.com/2012/01/17/big-corporations-are-more-popular-than-you-might-realize/`
7. Andrew Gelman, “Social class and views of corporations”: `http://andrewgelman.com/2008/07/27/social-class-and-views-of-corporations/`
8. Andrew Gelman, “Where are the larger-than-life athletes?”: `http://andrewgelman.com/2012/01/12/where-are-the-larger-than-life-athletes/`
9. Andrew Gelman, “Controversy about average personality differences between men and women”: `http://andrewgelman.com/2012/01/12/controversy-about-average-personality-differences-between-men-and-women/`
10. Andrew Gelman, “Libertarians in space”: `http://andrewgelman.com/2012/01/03/libertarians-in-space/`

11. Andrew Gelman, "Surveys show Americans are populist class warriors, except when they aren't": http://andrewgelman.com/2011/12/23/surveys-show-americans-are-populist-class-warriors-except-when-they-arent/

12. Andrew Gelman, "This guy has a regular column at Reuters": http://andrewgelman.com/2011/12/20/this-guy-has-a-regular-column-at-reuters/

13. Andrew Gelman, "The most clueless political column ever—I think this Easterbrook dude has the journalistic equivalent of 'tenure'": http://andrewgelman.com/2011/10/14/the-most-clueless-political-column-ever-i-think-this-easterbrook-dude-has-the-journalistic-equivalent-of-tenure/

14. Andrew Gelman, "1.5 million people were told that extreme conservatives are happier than political moderates. Approximately .0001 million Americans learned that the opposite is true": http://andrewgelman.com/2012/08/1-5-million-people-were-told-that-extreme-conservatives-are-happier-than-political-moderates-approximately-0001-million-americans-learned-that-the-opposite-is-true/

15. Andrew Gelman, "Was it really necessary to do a voting experiment on 300,000 people? Maybe 299,999 would've been enough? Or 299,998? Or maybe 2000?": http://andrewgelman.com/2014/10/30/really-necessary-voting-experiment-300000-people-maybe-299999-wouldve-enough-299998-maybe-2000/

Homework due at beginning of class:

1. *Getting started in R.* Set up R and Rstudio on your laptop computer and do everything in the notes on R: http://www.stat.columbia.edu/~gelman/surveys.course/Rnotes.pdf

In class:

1. Discuss Jitts

2. Lecture and discussion of examples

3. Get started with R

4. Discuss readings and next class

Class 2a: Basic statistics

Readings before class:

1. Andrew Gelman, "What's the point of the margin of error?": http://andrewgelman.com/2015/01/23/whats-point-margin-error/

2. *Regression and Other Stories*, chapters 1–3

3. L. J. Zigerell, "R graph: plot"; http://www.ljzigerell.com/?p=1891

4. L. J. Zigerell, "R graph: confidence intervals": http://www.ljzigerell.com/?p=1916

5. Andrew Gelman and Hal Stern, "The difference between 'significant' and 'not significant' is not itself statistically significant": http://www.stat.columbia.edu/~gelman/surveys.course/GelmanStern2006.pdf

Homework due at beginning of class:

1. *Bayesian inference.* An election is coming. From a reputable forecast you get a prediction that candidate A will win 51% of the vote, with a forecast standard error of σ. You now conduct a survey of 500 people, of whom 53% support candidate A and 47% support candidate B. For simplicity, ignore nonsampling error—that is, assume the poll is a simple random sample of the population of voters, assume responses are accurate and that no voters will change their minds.

   (a) Suppose that, given the above information, your Bayesian forecast is that A will receive 53% of the vote. What must σ then be, and what is the standard error of your Bayesian forecast?

   (b) What is your Bayesian probability that candidate A will win the election?

2. *Simulation of regression with fake data.* Sample 100 random data points x from the normal distribution with mean 10 and standard deviation 5. Then simulate 100 data points y from the model, $y = 2 + 10x - x^2 + \text{error}$, where the errors are normally distributed with mean 0 and standard deviation 1.

   (a) Fit a linear regression to the data and fit a quadratic regression to the data. Load the `arm` package into R and display the fitted regressions using the `display()` function.

   (b) Use `plot()` to graph the data; then add the fitted linear and quadratic regression lines to the graph using `curve(a+b*x,add=TRUE)` and `curve(b0+b1*x+b2*x^2,add=TRUE)`. You should hand in a graph that includes the data, the straight line, and the quadratic curve.

In class:

1. Discuss Jitts

2. Lecture and discussion on basic statistics:

   (a) "How many people were in this survey?"

   (b) Estimates and standard errors

   (c) Weighted averages

   (d) Sample size calculations

   (e) The $(y + 2)/(n + 4)$ estimate

3. R on your laptop computer

   (a) Estimates, standard errors, and confidence intervals for proportions and comparisons

   (b) Data manipulations

   (c) Simulations

4. Discuss readings and next class

Class 2b: Linear regression basics

Readings before class:

1. *Regression and Other Stories*, chapters 4–5
2. Andrew Gelman and David Weakliem, "Of beauty, sex, and power: Statistical challenges in estimating small effects": `http://www.stat.columbia.edu/~gelman/research/published/power5r.pdf`
3. Andrew Gelman and John Carlin, "Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors": `http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf`
4. Andrew Gelman, "The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective": `http://www.stat.columbia.edu/~gelman/research/published/bayes_management.pdf`
5. Andrew Gelman, "Disagreements about the strength of evidence": `http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics12.pdf`
6. Andrew Gelman, "God, guns, and gaydar: The laws of probability push you to overestimate small groups": `http://andrewgelman.com/2010/07/god_guns_and_ga/`
7. David Hemenway, "The myth of millions of annual self-defense gun uses: a case study of survey overestimates of rare events": `http://www.stat.columbia.edu/~gelman/surveys.course/Hemenway1997.pdf`

Homework due at beginning of class:

1. *Getting started in Stan.* Set up Rstan on your laptop computer, following the instructions at `http://mc-stan.org/interfaces/rstan.html`. Make sure you can get the 8 schools example running as described there.

In class:

1. Discuss Jitts
2. Lecture and discussion on statistical inference and scientific claims
   (a) Equivalent sample size (beauty and sex ratio example)
   (b) Problems with p-values and statistical significance
   (c) Difficulties with estimation of small probabilities
   (d) Political science examples
3. R on your laptop computer
   (a) Simulate fake data
   (b) Fitting a simple regression
4. Discuss readings and next class

Class 3a: More advanced linear regression

Readings before class:

1. *Regression and Other Stories*, appendix A
2. Andrew Gelman, "What are the key assumptions of linear regression?": `http://andrewgelman.com/2013/08/04/19470/`
3. *Regression and Other Stories*, chapters 6–7

Homework due at beginning of class:

1. *Linear regression.* The file at `http://www.stat.columbia.edu/~gelman/surveys.course/pew_research_center_june_elect_wknd_data.dta` has data from Pew Research Center polls taken during the 2008 election campaign. You can read these data into R using the `read.dta()` function (after first loading the `foreign` package into R). For this homework problem, ignore the survey weights.

   Fit a linear regression (using the `lm()` function in R) to predict political ideology (on a 5-point scale: –2 = very liberal, –1 = liberal, 0 = moderate, 1 = conservative, 2 = very conservative, with nonresponses coded as 0's), given sex, age, and marital status. Use `display()` to display the result. In a short paragraph, describe the meaning of each coefficient in the fitted model.

In class:

1. Discuss Jitts
2. Lecture and discussion on linear regression
   (a) The assumptions of linear regression
   (b) Including and excluding predictors
   (c) Main effects and interactions
3. Working with survey data in R
   (a) Building regression models, interpreting models, graphing
4. Linear regression in Stan
5. Discuss readings and next class

Class 3b: Logistic regression

Readings before class:

1. *Regression and Other Stories*, chapter 8

Homework due at beginning of class:

1. *Logistic regression.* Using the Pew 2008 survey, fit a logistic regression (using the `glm()` function in R) to predict whether a person is liberal (that is, responds "liberal" or "very liberal" to the ideology question, excluding respondents who do not respond to this question), given sex, age, and marital status. Use the `display()` function to display the result. In a short paragraph, describe the meaning of each coefficient in the fitted model.

In class:

1. Discuss Jitts
2. Logistic regression
   (a) Building logistic regression models (arsenic well-switching example)
   (b) Divide-by-4 rule
   (c) Discrete choice model
3. Fitting logistic regressions in R, including fake-data simulation and graphing
4. Discuss readings and next class

Class 4a: Statistical graphics

Readings before class:

1. *Regression and Other Stories*, appendix B
2. L. J. Zigerell, "R graph: plot"; `http://www.ljzigerell.com/?p=1891`
3. L. J. Zigerell, "R graph: confidence intervals": `http://www.ljzigerell.com/?p=1916`
4. Andrew Gelman and Antony Unwin, "Infovis and statistical graphics: Different goals, different looks": `http://www.stat.columbia.edu/~gelman/research/published/vis14.pdf`
5. Andrew Gelman and Antony Unwin, "Tradeoffs in information graphics": `http://www.stat.columbia.edu/~gelman/research/published/visreply3.pdf`

Homework due at beginning of class:

1. *Plotting survey data in R.* Using the Pew 2008 survey, compute the percentage of respondents in each state (excluding Alaska and Hawaii) who are liberal. Then make the following three graphs, putting them on a single page:
   (a) A plot of estimated proportion liberal in each state vs. Obama's vote share in 2008 (data available at `http://www.stat.columbia.edu/~gelman/surveys.course/2008ElectionResult.csv`, readable in R using `read.csv()`), as a scatterplot using the two-letter state abbreviations (see `state.abb()` in R).
   (b) A plot of estimated proportion liberal in each state vs. sample size in each state (again as a scatterplot using the two-letter state abbreviations).
   (c) A map of estimated proportion liberal using colors in a U.S. map.

In class:

1. Discuss Jitts
2. Lecture and discussion on choices in statistical graphics
3. Making graphs in R
4. Discuss readings and next class

Class 4b: Causal inference overview

Readings before class:

1. Andrew Gelman and Adam Zelizer, "Evidence on the deleterious impact of sustained use of polynomial regression on causal inference": `http://www.stat.columbia.edu/~gelman/research/published/rd_china_5.pdf`

2. Andrew Gelman, "Income, education, and religion as 'background variables' or 'treatments'": `http://andrewgelman.com/2008/06/19/income_educatio/`

3. *Regression and Other Stories*, chapter 11

4. Andrew Gelman, "Pushing at an open door: When can personal stories change minds on gay rights?": `https://www.washingtonpost.com/blogs/monkey-cage/wp/2014/12/19/pushing-at-an-open-door-when-can-personal-stories-change-minds-on-gay-rights/`

5. Andrew Gelman, "LaCour and Green 1, This American Life 0": `http://andrewgelman.com/2015/12/16/lacour-and-green-1-this-american-life-0/`

6. Andrew Gelman, "College football, voting, and the law of large numbers": `http://andrewgelman.com/2012/10/25/college-football-voting-and-the-law-of-large-numbers/`

7. Andrew Gelman, "Are you ready for some smashmouth FOOTBALL?": `http://andrewgelman.com/2015/10/30/are-you-ready-for-some-smashmouth-football/`

Homework due at beginning of class:

1. Exercise 3 from chapter 11 of *Regression and Other Stories*.

In class:

1. Discuss Jitts

2. Lecture and discussion on causal inference

3. Classroom activity doing causal modeling in R

4. Discuss readings and next class

Class 5a: Survey interviewing

Readings before class:

1. Groves et al., chapters 7–9

2. Jay Livingston, "Poverty, perceptions, and politics": `http://montclairsoci.blogspot.com/2015/01/poverty-perceptions-and-politics.html`

3. Andrew Gelman, "President of American Association of Buggy-Whip Manufacturers takes a strong stand against internal combustion engine, argues that the so-called 'automobile' has 'little grounding in theory' and that 'results can vary widely based on the particular fuel that is used'": http://andrewgelman.com/2014/08/06/president-american-association-buggy-whip-manufacturers-takes-strong-stand-internal-combustion-engine-argues-called-automobile-little-grounding-theory/

4. Andrew Gelman, "Buggy-whip update": http://andrewgelman.com/2014/12/09/buggy-whip-update/

5. Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman, "Forecasting elections with non-representative polls": http://www.stat.columbia.edu/~gelman/research/published/forecasting-with-nonrepresentative-polls.pdf

Homework due at beginning of class:

1. *Survey interviewing.* Design a survey form and try it out on five friends. Write up what you learned.

In class:

1. Discuss Jitts

2. In pairs, discuss your experiences with your survey forms

3. Review confusing points in class so far

4. Discuss readings and next class

Class 5b: Survey measurement

Readings before class:

1. Andrew Gelman, "Don't trust the Turk": http://andrewgelman.com/2013/07/10/dont-trust-the-turk/

2. Groves et al., chapter 2

3. Andrew Gelman, "Counting churchgoers": http://andrewgelman.com/2006/07/counting_church/

4. Hadaway, C. K., Marler, P. L., and Chaves, M., "What the polls don't show: A closer look at U.S. church attendance": http://www.stat.columbia.edu/~gelman/surveys.course/HadawayMarlerChaves1993.pdf

5. Amos Tversky and Daniel Kahneman, "The framing of decisions and the psychology of choice": http://www.stat.columbia.edu/~gelman/surveys.course/TverskyKahneman1981.pdf

6. Andrew Gelman, "Age and happiness: The pattern isn't as clear as you might think": http://andrewgelman.com/2010/12/age_and_happine/

7. Frijters, P., and Beaton, T., “The mystery of the U-shaped relationship between happiness and age”: http://www.stat.columbia.edu/~gelman/surveys.course/FrijtersBeaton2008.pdf

8. David Blanchflower and Andrew Oswald, “Is well-being U-shaped over the life cycle?”: http://www.stat.columbia.edu/~gelman/surveys.course/BlanchflowerOswald2008.pdf

9. Stone, A. A., Schwartz, J. E., Broderick, J. E., and Deaton, A., “A snapshot of the age distribution of psychological well-being in the United States”: http://www.stat.columbia.edu/~gelman/surveys.course/StoneSchwartzBroderickDeaton2010.pdf

Homework due at beginning of class:

1. *Survey measurement.* Find a measurement effect in an existing survey.

2. *Logistic regression.* Using the Pew 2008 survey, fit a logistic regression using the `glm()` function in R to predict whether a person is liberal (that is, responds “liberal” or “very liberal” to the ideology question, excluding respondents who do not respond to this question), given five predictors: a constant term, sex (coded as 1 for male and 0 for female), age (coded as a continuous variable), marital status (coded as a continuous variable, 0 = unmarried, 1 = married, and 0.5 if there are any intermediate states such as living together but not married), and the interaction between sex and age. Use the `display()` function to display the result. In a short paragraph, describe the meaning of each coefficient in the fitted model.

In class:

1. Discuss Jitts

2. Discuss examples of survey measurement

3. Review logistic regression

4. Discuss readings and next class

Class 6a: Simple and stratified random sampling

Readings before class:

1. Groves et al., chapter 3

Homework due at beginning of class:

1. *Simulation and analysis of stratified sample.* Write an R function to take a random subsample of the 2010 General Social Survey using regions of the country as strata.

   (a) Perform a sample of size 100 with each stratum sampled in proportion to its population size (in this case, the “population” is just the full 2010 GSS). Use this subsample to estimate the proportion of people who favor a law which would require a person to obtain a police permit before he or she could buy a gun. Also compute the standard error for this estimate, first directly using the formula for the standard error of a cluster sample, then using the `survey` package in R. (These two standard errors should be identical.)

(b) Put step (a) above in a loop and do it 100 times. Check that your estimate is unbiased and that its standard deviation is approximately equal to the average standard error computed in the 100 simulations.

In class:

1. Discuss Jitts
2. Simple and stratified sampling in R:
   (a) Doing the sampling
   (b) Formulas for the estimate and standard error
   (c) Analyzing data
3. Systematic sampling as an example of how to use these ideas in practice
4. Discuss readings and next class

Class 6b: Poststratification

Readings before class:

1. Andrew Gelman and Thomas Little, "Improving upon probability weighting for household size": `http://www.stat.columbia.edu/~gelman/research/published/household.pdf`

Homework due at beginning of class:

1. *Analysis of a stratified sample.* A survey is taken of 100 undergraduates, 100 graduate students, and 100 continuing education students at a university. Assume a simple random sample within each group. Each student is asked to rate his or her satisfaction with his or her experiences, on a 1–10 scale. Write the estimate and standard error of the average satisfaction of all the students at the university. Introduce notation as necessary for all the information needed to solve the problem.

In class:

1. Discuss Jitts
2. Poststratification in R:
   (a) Unit weights and stratum weights
   (b) Connection to regression models
3. Discuss readings and next class

Class 7a: Survey weights

Readings before class:

1. Groves et al., chapter 10

Homework due at beginning of class:

1. *Regression analysis including survey weights.* Using the Pew 2008 data:
   (a) Compute the weighted average proportion liberal in each state and plot vs. the raw average; this should be a square plot (in R, `par(pty="s")`) with identical scales on x and y axes, and each state indicated by its two-letter abbreviation.
   (b) Using the `survey` package in R, fit a weighted regression (using the `svyglm()` function in R) to predict political ideology, given sex, age, and marital status. Compare to the results from an unweighted regression.

In class:

1. Discuss Jitts
2. Weighting in R:
   (a) Constructing weights
   (b) Analyzing data with weighting and poststratification
3. Discuss readings and next class

Class 7b: Cluster sampling

Readings before class:

1. Groves et al., chapter 4
2. Afrobarometer, "Sampling principles": `http://www.afrobarometer.org/survey-and-methods/sampling-principles`
3. Afrobarometer, "Malawi round 4 survey technical information": `http://www.afrobarometer.org/files/documents/survey_technical_information/mlw_r4_tif.pdf`
4. John Carlin, Mark Stevenson, Ian Roberts, Catherine Bennett, Andrew Gelman, and Terry Nolan, "Walking to school and traffic exposure in Australian children": `http://www.stat.columbia.edu/~gelman/research/published/CarlinStevensonParkerRobertsBennettGelmanNolan1997.pdf`

Homework due at beginning of class:

1. *Cluster sampling.* Suppose you have a library of 100 books and you want to estimate the frequency of the different words in this library. So you decide to take a random sample of 1000 words. Come up with a sampling scheme in which all words are equally likely to be selected (in proportion to their total number of appearances in the library).
2. *Simulation and analysis of cluster sample.* Write an R function to take a random subsample of the 2010 General Social survey using occupations as clusters.

(a) Take a cluster sample in the following way: first sample 20 occupations at random, then sample 50% of the respondents from each sampled occupation. From this sample, estimate the proportion of people in the population who favor a law which would require a person to obtain a police permit before he or she could buy a gun. Compute the standard error of this estimate.

(b) Repeat (a), but this time taking the sample as follows: first sample 20 occupations at random, then sample 5 people from each sampled occupation (or, if there are fewer then 5 people with that occupation category, sample all of them). Again get an estimate and standard error for the gun control question.

(c) Repeat (a), but this time first sample 20 occupations with probability proportional to size, then sample 5 from each sampled occupation (or, if there are fewer then 5 people with that occupation category, sample all of them). Again get an estimate and standard error for the gun control question.

In class:

1. Discuss Jitts
2. Cluster sampling in R:
   (a) Doing the sampling
   (b) Computing sampling probabilities
3. Discuss readings and next class

Class 8a: Surveys in the United States

Readings before class:

1. Mark Blumenthal, "Polling: Crisis or not, we're in a new era": http://www.huffingtonpost.com/mark-blumenthal/polling-crisis-or-not-wer_b_10328648.html
2. Groves et al., chapter 1
3. Felipe Osorio, Andrew Gelman, and Lucas Leeman, "Working with the General Social Survey in R": http://www.youtube.com/watch?v=mu2sEf12Eu4
4. Andrew Gelman, "Fixing the race, ethnicity, and national origin questions on the U.S. Census": http://andrewgelman.com/2013/08/12/fixing-the-race-ethnicity-and-national-origin-questions-on-the-u-s-census/
5. U.S. Census Bureau, "Census Bureau releases estimates of undercount and overcount in the 2010 Census": http://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html
6. Tom Smith, "The hidden 25 percent: An analysis of nonresponse on the 1980 General Social Survey": http://www.stat.columbia.edu/~gelman/surveys.course/Smith1983.pdf
7. Wei Wang and Andrew Gelman, "Xbox, Big Data, and the return of non-representative polling": http://data-informed.com/xbox-big-data-return-non-representative-polling

8. Andrew Gelman, “Political attitudes of the super-rich”: `http://andrewgelman.com/2008/11/02/political-attitudes-of-the-super-rich/`

9. Benjamin Page, Larry Bartels, and Jason Seawright, “Democracy and the policy preferences of wealthy Americans”: `http://stat.columbia.edu/~gelman/surveys.course/PageBartelsSeawright2013.pdf`

10. Andrew Gelman, “Hack pollster Doug Schoen illustrates a general point: The #1 way to lie with statistics is . . . to just lie!”: `http://andrewgelman.com/2011/10/27/the-1-way-to-lie-with-statistics-is-to-just-lie/`

11. Andrew Gelman, “Can you trust international surveys?”: `http://andrewgelman.com/2016/02/28/can-you-trust-international-surveys/`

Homework due at beginning of class:

1. The “Can you trust international surveys?” reference above discusses a project in which researchers looked for duplicate or near-duplicate records as an indication that data in many international surveys might be faked.

   Take one of these surveys:

   - World Values Survey 6: `http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`
   - Pew 2013 global survey: `http://www.pewglobal.org/category/datasets/?download=31111`
   - Zogby surveys on the Middle East: `http://sadat.umd.edu/new%20surveys/surveys.htm`

   Pick one of the above 3 surveys and pick two of the countries (they can be two poor countries or one poor country and one rich country). Analyze the data from the two countries you picked (do them separately, don’t try to analyze both countries at once), checking for duplicates as discussed in the above post. Then write up what you found.

In class:

1. Discuss Jitts

2. Discuss U.S. surveys

3. Design class project survey

4. Discuss readings and next class

Class 8b: Surveys in other countries

Readings before class:

1. Egor Lazarev, Anton Sobolev, Irina Soboleva, and Boris Sokolov, “Trial by fire: A natural disaster’s impact on support for the authorities in rural Russia”: `http://www.stat.columbia.edu/~gelman/surveys.course/LazarevSobolevSobolevaSokolov2014.pdf`

2. Michael Spagat, "The reliability of cluster surveys of conflict mortality: Violent deaths and non-violent deaths": http://stat.columbia.edu/~gelman/surveys.course/Spagat2009.pdf

3. Andrew Gelman, "Ethical and data-integrity problems in a study of mortality in Iraq": http://andrewgelman.com/2010/04/ethical_and_dat_1/

4. Michael Spagat, "Ethical and data-integrity problems in the second Lancet survey of mortality in Iraq": http://stat.columbia.edu/~gelman/surveys.course/Spagat2010.pdf

5. Andrew Gelman, "Peeking behind the curtain, or, What's (not) the matter with Portugal?": http://andrewgelman.com/2008/03/peeking_behind/

Homework due at beginning of class:

1. We will ask you to analyze some survey data that we will give you.

In class:

1. Discuss Jitts

2. Discuss surveys in other countries

3. Discuss readings and next class

Class 9a: Voting and political participation

Readings before class:

1. Andrew Gelman, "What difference would it make if everybody voted?": http://andrewgelman.com/2008/07/10/what-difference-would-it-make-if-everybody-voted-leighley-and-nagler-disagree-with-wolfinger/

2. Andrew Gelman, "Minor-league stats predict major-league performance, Sarah Palin, and some differences between baseball and politics": http://andrewgelman.com/2011/04/07/minor-league_st_1/

3. Andrew Gelman and Gary King, "Why are American Presidential election campaign polls so variable when votes are so predictable?": http://www.stat.columbia.edu/~gelman/surveys.course/GelmanKing1993.pdf

4. David Rothschild, Sharad Goel, Andrew Gelman, and Douglas Rivers, "The mythical swing voter": http://www.stat.columbia.edu/~gelman/research/unpublished/swing_voters.pdf

5. David Rothschild and Justin Wolfers, "Forecasting elections: Voter intentions versus expectations": http://assets.wharton.upenn.edu/~rothscdm/RothschildExpectations.pdf

Homework due at beginning of class:

1. We will give you a statistical modeling assignment in R and Stan.

In class:

1. Discuss Jitts
2. Discuss the survey experiment we will do
3. Play with multilevel modeling in Stan
4. Discuss readings and next class

Class 9b: Public opinion

Readings before class:

1. Benjamin Page and Robert Shapiro, "Changes in Americans' policy preferences, 1935-1979": `http://www.stat.columbia.edu/~gelman/surveys.course/PageShapiro1982.pdf`
2. Benjamin Page, Robert Shapiro, and Glenn Dempsey, "What moves public opinion?": `http://www.stat.columbia.edu/~gelman/surveys.course/PageShapiroDempsey1987.pdf`
3. Robert Shapiro and Benjamin Page, "Foreign policy and the rational public": `http://www.stat.columbia.edu/~gelman/surveys.course/ShapiroPage1988.pdf`
4. Delia Baldassarri and Andrew Gelman, "Partisans without constraint: Political polarization and trends in American public opinion": `http://www.stat.columbia.edu/~gelman/surveys.course/BaldassarriGelman2008.pdf`
5. Andrew Gelman, Daniel Lee, and Yair Ghitza, "Public opinion on health care reform": `http://www.stat.columbia.edu/~gelman/surveys.course/GelmanLeeGhitza2010.pdf`
6. Robert Shapiro and Sara Arrow, "Support for health care reform: Is public opinion more favorable for Obama than it was for Clinton in 1994?" `http://www.stat.columbia.edu/~gelman/surveys.course/ShapiroArrow2009.pdf`

Homework due at beginning of class:

1. *Regression and poststratification.* Set up a simulation in which you estimate a Yes/No survey response given sex, age (18–29, 30–44, 45–64, 65+), and ethnicity (non-hispanic white, black, hispanic, other): thus you have $2 \times 4 \times 4$ categories.
   (a) Get the poststratification table from the U.S. Census. If you can't figure out how to do this, make up reasonable numbers. But really you should be able to get the numbers: it's part of the assignment.
   (b) Simulate these 3 demographic variables from a sample of 1000 survey respondents. Specify a nonresponse pattern in which women, older people, and whites are more likely to respond than men, younger people, and minorities. Display n_j/n and N_j/N for each of the $J = 2 \times 4 \times 4$ cells.
   (c) Assume the true probability of Yes response follows a logistic regression with indicators for the levels of each demographic factor, and no interactions. Pick a particular question that might be asked and make up reasonable values for the logistic regression coefficients.
   (d) From the assumed logistic regression, simulate fake data for your 1000 respondents.

(e) Fit a logistic regression to these data. Estimate the coefficients, and estimate the proportion of Yes responses in each of the $2 \times 4 \times 4$ cells.

(f) Poststratify and estimate the proportion of Yes responses among U.S. adults.

(g) By fitting the logistic regression using `stan_glm()` and propagating uncertainty, give a standard error for your poststratified estimate.

In class:

1. Discuss Jitts
2. Discuss sampling and public opinion
3. Discuss the survey experiment we will do
4. Play with multilevel modeling in Stan
5. Discuss readings and next class

Class 10a: Network sampling

Readings before class:

1. Yotam Margalit and Andrew Gelman, "Social penumbras and political attitudes" [draft paper]
2. Sharad Goel, Winter Mason, and Duncan Watts, "Real and perceived attitude agreement in social networks": `http://www.stat.columbia.edu/~gelman/surveys.course/GoelMasonWatts2010.pdf`
3. Keith Hampton, Lauren Goulet, Lee Rainie, and Kristen Purcell, "Social networking sites and our lives": `http://www.stat.columbia.edu/~gelman/surveys.course/HamptonGouletRainiePurcell2011.pdf`
4. Tyler McCormick, Matthew Salganik, and Tian Zheng, "How many people do you know?: Efficiently estimating personal network size": `http://www.stat.columbia.edu/~gelman/surveys.course/MccormickSalganikZheng2010.pdf`
5. David Heckathorn, "Respondent-driven sampling: A new approach to the study of hidden populations": `http://www.stat.columbia.edu/~gelman/surveys.course/Heckathorn1997.pdf`
6. Sharad Goel and Matthew Salganik, "Assessing respondent-driven sampling": `http://www.stat.columbia.edu/~gelman/surveys.course/GoelSalganik2010.pdf`
7. Andrew Gelman, "'How many zombies do you know?': Using indirect survey methods to measure alien attacks and outbreaks of the undead": `http://www.stat.columbia.edu/~gelman/research/published/zombies.pdf`

Homework due at beginning of class:

1. Design a survey experiment and write the questionnaire. We will be using an internet survey design tool. Your survey should be about any topic of your interest, but should include:

   - Different types of questions (open-ended, multiple-choice, slider, etc.)
   - Experiment 1: Priming manipulation (think about something good and about something bad) with records for pre-treatment and post-treatment indicators (see Sample Experiment).
   - Experiment 2: Manipulation of question wording (for example, "welfare" vs. "aid to the poor")
   - Experiment 3: Randomization of the order of some questions

   Include a link to your survey and the full code.

In class:

1. Discuss Jitts
2. Discussion of the penumbra problem
3. Discussion of the survey questions
4. Discuss readings and next class

Class 10b: Survey experiments

Readings before class:

1. Andrew Gelman, "Thinking of doing a list experiment? Here's a list of reasons why you should think again": `http://andrewgelman.com/2014/04/23/thinking-list-experiment-heres-list-reasons-think/`
2. Adam Glynn, "What can we learn with statistical truth serum? Design and analysis of the list experiment": `http://www.stat.columbia.edu/~gelman/surveys.course/Glynn2013.pdf`

Homework due at beginning of class:

1. *Sample size calculation.* In a survey of n people, half are asked if they support "the health care law recently passed by Congress" and half are asked if they support "the law known as Obamacare." The goal is to estimate the effect of the wording on the proportion of Yes responses. How large must n be for the effect to be estimated within a standard error of 5 percentage points?

In class:

1. Discuss Jitts
2. Simulate and analyze survey experiments in R
3. Discuss readings and next class

Class 11a: Multilevel linear regression

Readings before class:

1. Andrew Gelman, "Regression: What's it all about?": http://www.stat.columbia.edu/~gelman/research/published/wakefield_regression.pdf
2. *Multilevel Regression*, chapters 1–2
3. Andrew Gelman, "Is it meaningful to talk about a probability of '65.7%' that Obama will win the election?": http://andrewgelman.com/2012/10/is-it-meaningful-to-talk-about-a-probability-of-65-7-that-obama-will-win-the-election/
4. Kari Lock and Andrew Gelman, "Bayesian combination of state polls and election forecasts": http://www.stat.columbia.edu/~gelman/surveys.course/LockGelman2010.pdf

Homework due at beginning of class:

1. *Multilevel data structures.* You will simulate fake data from the following hypothetical world: An educational experiment is being performed in 20 classrooms, of which 10 will get the treatment and 10 will get the control. There are 25 students in each class, and each student gets a pre-test and post-test.

   Assume the pre-test score for student i in class $j[i]$ can be written as $\alpha_{j[i]} + \eta_i$, where the α_j's are normally distributed with mean 50 and standard deviation 10, and the η_i's are normally distributed with mean 0 and standard deviation 15.

   Further assume that the relation between pre-test and post-test in the control group is post-test $= 10 + 0.7 *$ pre-test $+$ error, where the error is normally distributed with mean 0 and standard deviation 10.

   Finally, assume the treatment effect varies by class, with an average treatment effect of 5, and a standard deviation of 5 of the effects across classes.

   (a) Simulate fake data from this world.

   (b) Fit a hierarchical linear model to your fake data and check that you approximately recover the assumed parameter values.

In class:

1. Discuss Jitts
2. Play with multilevel models in R
3. Discuss readings and next class

Class 11b: Multilevel logistic regression

Readings before class:

1. Andrew Gelman, "Multilevel modeling: What it can and cannot do": http://www.stat.columbia.edu/~gelman/surveys.course/Gelman2006.pdf

2. Andrew Gelman, "Two-stage regression and multilevel modeling: a commentary": http://www.stat.columbia.edu/~gelman/surveys.course/Gelman2005.pdf

3. *Multilevel Regression*, chapter 4

Homework due at beginning of class:

1. *Multilevel modeling.* From the Pollster data, estimate a time series of support for Obama and Romney, adjusting for house effects and then smoothing the curve using some function such as lowess. Compare to the smoothed average of the unadjusted approval numbers from this series and comment on any differences.

In class:

1. Discuss Jitts

2. Play with multilevel models in R

3. Discuss readings and next class

Class 12a: Item-response and ideal-point modeling

Readings before class:

1. Joseph Bafumi, Andrew Gelman, David Park, and Noah Kaplan, "Practical issues in implementing and understanding Bayesian ideal point estimation": http://www.stat.columbia.edu/~gelman/research/published/171.pdf

2. Joseph Bafumi and Michael Herron, "Leapfrog representation and extremism: A study of American voters and their members of Congress" http://mc-stan.org

3. Valen Johnson, Robert Deaner, and Carel van Schaik, "Bayesian analysis of rank data with application to primate intelligence experiments": http://www.stat.columbia.edu/~gelman/surveys.course/JohnsonDeanerSchaik2002.pdf

Homework due at beginning of class:

1. *Ideal-point modeling.* You will create a measure of economic ideology using the following questions from the 2000 Annenberg survey: Are tax rates a problem (CBB01), Favor cutting taxes or strengthening social security (CBB05), Federal government should reduce the top tax rate (CBB10), Federal government should adopt flat tax (CBB13), Federal government should spend more on social security (CBC01), Favor investing social security in stock market (CBC05), Is poverty a problem (CBP01), Federal government should reduce income differences (CBP02), Federal government should spend more on aid to mothers with young children (CBP03), Federal government should expend effort to eliminate many business regulations (CBT01).

   Fit a hierarchical logistic regression to estimate ideal points for individuals and survey questions.

   (a) Display the estimated ideal points and standard errors of the survey questions (listing the questions in order of their estimated ideal points)

   (b) Display the distribution of estimated ideal points of the survey respondents. On this same graph, display the distributions for Democrats, independents, and Republicans.

In class:

1. Discuss Jitts

2. Play with item-response models in Stan

3. Discuss readings and next class

Class 12b: Multilevel regression and poststratification

Readings before class:

1. *Multilevel Regression*, chapter **

2. Jeffrey Lax and Justin Phillips, "How should we estimate public opinion in the states?": `http://www.stat.columbia.edu/~gelman/surveys.course/LaxPhillips2009a.pdf`

3. Jeffrey Lax and Justin Phillips, "Gay rights in the states: Public opinion and policy responsiveness": `http://www.stat.columbia.edu/~gelman/surveys.course/LaxPhillips2009a.pdf`

4. Yair Ghitza and Andrew Gelman, "Deep interactions with MRP: Presidential turnout and voting patterns among small electoral subgroups": `http://www.stat.columbia.edu/~gelman/research/published/misterp.pdf`

Homework due at beginning of class:

1. *Multilevel regression and poststratification.* Download the 2012 National Election Study.

   (a) Fit a multilevel logistic regression estimating support for gun control given state, sex, and ethnicity (white/black/hispanic/other). Use the `display()` function in R to display the fitted model. Explain the output in a brief paragraph.

   (b) Using your model, get estimates of the proportion of people who support gun control, for all 8 demographic groups in each state (excluding Alaska and Hawaii) for the year 2012. Using the 2010 census, poststratify to get an estimate for each state.

   (c) Make the following two graphs: (i) a plot of estimated gun control support vs. Obama vote share in 2012 (indicating each state by its two-letter abbreviation); (ii) a plot of estimated gun control support in 2012 vs. the raw proportion of respondents in the state from 2012 who supported gun control.

In class:

1. Discuss Jitts

2. Play with MRP in R

3. Discuss readings and next class

Class 13a: Constructing survey weights

Readings before class:

1. Andrew Gelman, "Struggles with survey weighting and regression modeling": http://www.stat.columbia.edu/~gelman/surveys.course/Gelman2007a.pdf

2. Sharon Lohr, "Comment: Struggles with survey weighting and regression modeling": http://www.stat.columbia.edu/~gelman/surveys.course/Lohr2007.pdf

3. Andrew Gelman, "Rejoinder: Struggles with survey weighting and regression modeling": http://www.stat.columbia.edu/~gelman/surveys.course/Gelman2007b.pdf

Homework due at beginning of class:

1. Take the Pew 2008 survey, ignore the existing weights, and construct your own:

   (a) Construct weights based on sex, education (less than high school, high school, some college, college, graduate school), and ethnicity (white, black, hispanic, other), adjusting for the variables one at a time, matching to the 2010 census numbers on the population of U.S. adults.

   (b) Make a scatterplot of your weights vs. the Pew weights. How do they differ? If there are points on the plot that are far from the rest, take a look and find out who are they.

In class:

1. Discuss Jitts

2. Construct survey weights in R

3. Discuss readings and next class

Class 13b: Missing-data imputation

Readings before class:

1. Christopher Ingraham, "Kansas is the nation's porn capital, according to Pornhub": http://wonkviz.tumblr.com/post/82488570278/kansas-is-the-nations-porn-capital-according-to

2. Groves et al., chapter 6

3. *Regression and Other Stories*, chapter 17

Homework due at beginning of class:

1. *Missing-data imputation.* Create a miniature version of the 2010 General Social Survey (http://www.thearda.com/Archive/Files/Codebooks/GSS10PAN_CB.asp), including the following variables: sex, age, ethnicity (use four categories), urban/suburban/rural, education (use five categories), political ideology (on a 7-point scale from "extremely liberal" to "extremely conservative"), and general happiness.

(a) Fit a logistic regression on whether respondents feel “not too happy,” given the other variables in the dataset. Display (using `display()`) the results for the logistic regression fit to the complete cases (this is the result if you just feed the data including NA’s into R).

(b) Impute the missing values using `mi()` in the `mi` package in R. Then take one of the completed datasets and fit and display a logistic regression as above.

(c) Repeat, this time imputing using `aregImpute()` in the `Hmisc` package.

(d) Briefly discuss the differences between the three inferences above.

In class:

1. Discuss Jitts
2. Play with missing-data imputation in R
3. Discuss readings and next class

Class 14a: Open problems in analysis of survey data

Readings before class:

1. Seth Stephens-Davidowitz, “How many American men are gay?”: `http://www.nytimes.com/2013/12/08/opinion/sunday/how-many-american-men-are-gay.html`
2. Ken Shirley and Andrew Gelman, “Hierarchical models for estimating state and demographic trends in U.S. death penalty public opinion”: `http://www.stat.columbia.edu/~gelman/research/published/A12052_Shirley.pdf`
3. Yair Ghitza and Andrew Gelman, “The Great Society, Reagan’s revolution, and generations of presidential voting”: `http://www.stat.columbia.edu/~gelman/research/unpublished/cohort_voting_20140605.pdf`

Homework due at beginning of class:

1. To be announced

In class:

1. Discuss Jitts
2. Set up models in R and Stan for open problems
3. Discuss readings and next class

Class 14b: Summary of the course

Readings before class:

1. Groves et al., chapters 11 and 12

Homework due at beginning of class:

1. To be announced

In class:

1. Discuss Jitts
2. Go over the semester