Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 1a: Introduction

Happiness and the Tea Party movement

- A Brooks *New York Times* op-ed:

  > *People at the extremes are happier than political moderates . . . . none, it seems, are happier than the Tea Partiers . . .*

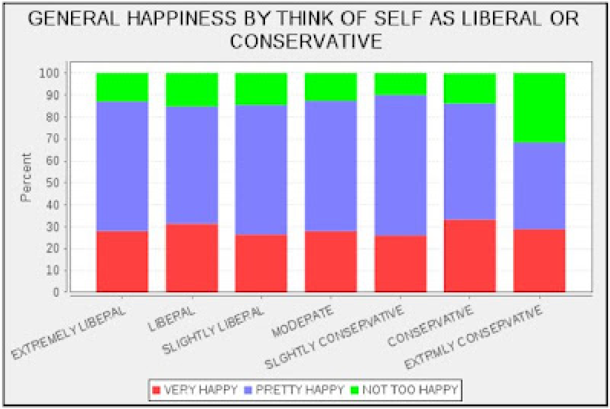
- But sociologist Jay Livingston writes:

  > *The GSS does not offer "bitter" or "Tea Party" as choices, but extreme conservatives are nearly three times as likely as others to be "not too happy."*

- Let's look at the data!

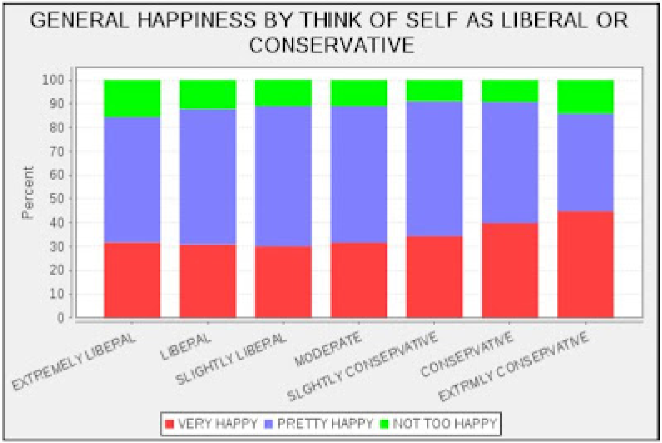
Data from General Social Survey

Chart for YEAR = 4(2009-2010)



- Is this just sampling variation?
  - Sample size for “Extremely Conservative” here is 80
  - Thus the standard error for that green bar on the right is approx $\sqrt{0.3 \cdot 0.7/80} = 0.05$

How did Brooks get this wrong?



- Averaging over all the years, conservatives seem pretty happy!
- The importance of descriptive inference
  - Be careful about explaining patterns that aren't real!

This course

- Statistical theory and methods
- Political science
- Computing

Statistical theory and methods

- Estimates and standard errors
- Weighted averages
- Regression
- Sampling probabilities

Political science

- U.S. public opinion and voting
- Sampling of records
- Other countries and other topics

Computing

- Stata
- R
  - Working with data
  - Simple calculations and regressions
  - Simple sampling
  - Simulation
  - Multilevel regression and poststratification
  - Survey package

Manipulating data in R

- Pulling in data
- Displaying and checking data
- Breaking up a survey question into multiple variables
- Combining several survey questions into a single variable
- Fitting models
- Graphs

Simple calculations and regressions in R

- Mean, standard deviation
- Linear regression
- Logistic regression

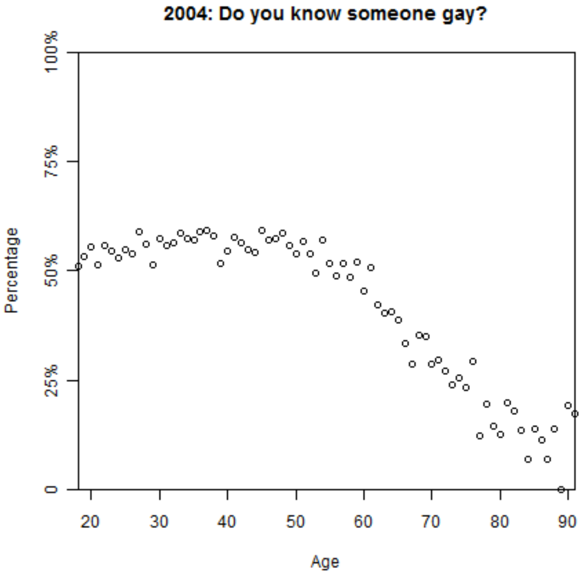
Simple sampling in R

- Simulation from a distribution
- Random sampling
- Stratified sampling, cluster sampling

More in R

- Multilevel regression and poststratification
- Survey package

How many people were in this survey?



Basics for yes/no survey responses

- $y = 1$ if yes, 0 if no
- Estimate is $\hat{p} = y/n$
- Standard error is s.e. $= \sqrt{\hat{p}(1 - \hat{p})/n}$
- 95% interval $[\hat{p} \pm 2\,\text{s.e.}]$
- How do you deal with "don't know" responses?
  - Party identification
  - Vote choice
  - Death penalty

The 95% confidence interval in R

- 1000 people surveyed: 700 support the death penalty and 300 oppose it

```
y <- 700
n <- 1000
estimate <- y/n
se <- sqrt (estimate*(1-estimate)/n)
int.95 <- estimate + qnorm(c(.025,.975))*se
```

The secret weapon

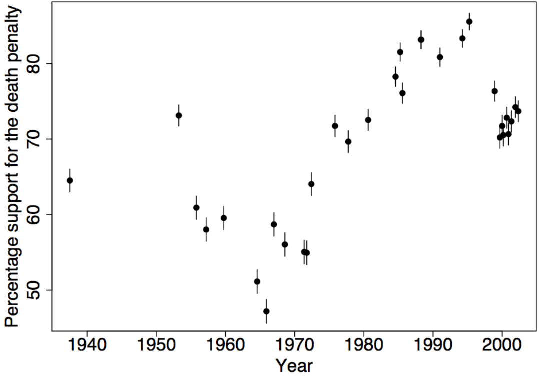


Figure 2.3 *Illustration of visual comparison of confidence intervals. Graph displays the proportion of respondents supporting the death penalty (estimates ±1 standard error—that is, 68% confidence intervals—under the simplifying assumption that each poll was a simple random sample of size 1000), from Gallup polls over time.*

The conservative upper bound on s.e

- $y = 1$ if yes, 0 if no
- Estimate is $\hat{p} = y/n$
- Standard error is s.e. $= \sqrt{\hat{p}(1-\hat{p})/n}$
- Maximum value of s.e. is when $\hat{p} = 0.5$
- Conservative s.e. is $\sqrt{0.5 \cdot 0.5/n} = 0.5/\sqrt{n}$
- When is this a bad idea?

Sample size calculations

- How large a survey do we need to estimate the president's approval rating so that the 95% confidence interval is $\pm 3\%$?
  - s.e. must be $1.5\% = 0.015$
  - $\sqrt{\hat{p}(1-\hat{p})/n} = 0.015$
  - Use the conservative guess $\hat{p} = 0.5$, then $0.5/\sqrt{n} = 0.015$
  - $n = (0.5/0.015)^2 = 1100$
- Assumes simple random sampling with no nonresponse

Sample size calculations—alternative solution

- How large a survey do we need to estimate the president's approval rating so that the 95% confidence interval is $\pm 3\%$?
  - Start with a guess, for example $n = 2000$
  - Work out the s.e., in this case $0.5/\sqrt{2000} = 0.011$
  - That's overkill, all we need is 0.015
  - Adjust sample size by factor $(0.011/0.015)^2 = 0.54$
  - Solution: $n = 0.54 \cdot 2000 = 1100$
  - Check: $0.5/\sqrt{1100} = 0.015$
- The $1/\sqrt{n}$ rule

Complications: $y = 0$ or $y = n$

- Example from recent consulting project: 75 out of 75 files had problems
- Problems with simple estimate:
  - $\hat{p} = 75/75 = 1$, complete certainty??
  - s.e. $= \sqrt{1 \cdot 0/75} = 0$??
- Agresti and Coull interval:
  - $\hat{p} = (y+2)/(n+4)$
  - s.e. $= \sqrt{\hat{p}(1-\hat{p})/n}$
  - 95% interval $[\hat{p} \pm 2\,\text{s.e.}]$
- When does this not make sense?

Complications: Finite-population correction

- Sample size n, population size N
- Simple formula: s.e. $= \sqrt{\hat{p}(1-\hat{p})/n}$
- Correct formula: s.e. $= \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}-\frac{1}{N})}$
- Consider special cases:
    - $N \to \infty$
    - $n = N$

Numerical survey responses

- Analysis
  - Compute average of data, $\bar{y}$, and standard deviation of data, s_y
  - s.e. $= s_y/\sqrt{n}$
  - 95% interval $[\bar{y} \pm 2\,\text{s.e.}]$
- Examples
  - Continuous (height, weight, age)
  - Continuous-like (feeling thermometer)
  - Counts (how many political events did you participate in during the past year?)
  - Discrete and finite (are you unhappy, somewhat happy, or very happy?)

Some examples from my work

- National public opinion polls
- Home radon surveys
- Post office surveys
- New York City telephone surveys
- Traffic exposure of Australian schoolchildren
- Alcoholics Anonymous membership survey

Schedule

- Weeks 1–2: Statistical background
- Weeks 3–4: Missing data and survey adjustments
- Weeks 5–6: Sampling and estimation
- Weeks 7–8: Measurement
- Weeks 9–10: Surveys in political science
- Weeks 11–12: More elaborate statistical modeling
- Weeks 13–14: Hard-to-reach populations
- Regular homeworks, final exam
- All course material at `http://www.stat.columbia.edu/~gelman/surveys`
- Course plan at `surveyscourse.pdf`

Textbooks

- Groves et al.: practical issues in surveys
- Lumley: the “survey” package in R
- Gelman and Hill: statistical methods
- Also, lots of readings (see syllabus)

Section meetings

- T.A. will set these up
- Key part of the course
- R
- Help with data

Jitts

- Due an hour before every class
- You don't have to get the questions right, but you do have to try them

Homework

- Two-week problem sets
- First homework due beginning of class 3a and class 4a
  1. Sample size calculation
  2. Linear regression in R
  3. Logistic regression in R
  4. Working with survey data and making graphs in R