Gov 50: 8. Measurement: Survey Sampling

Matthew Blackwell

Harvard University

Fall 2018

1. Today's agenda

2. The role of randomization

3. The power of randomization

4. Missing data in R

1/ Today's agenda

Logistics

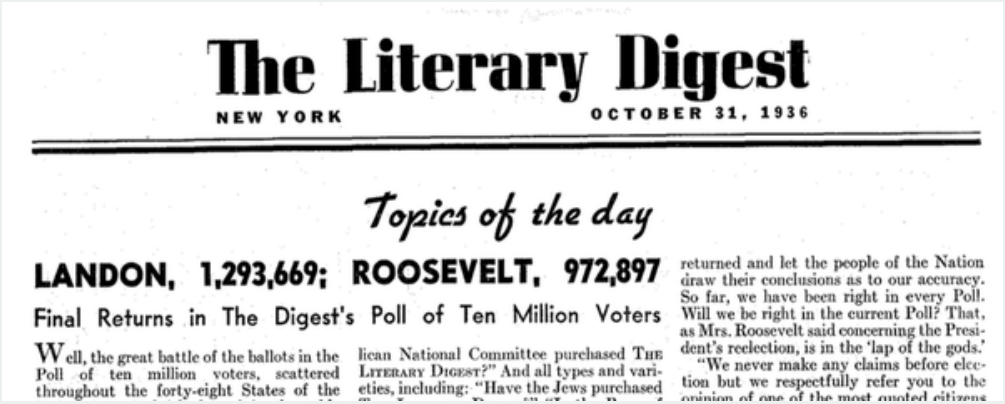
- HW 2: out on Canvas/Rstudio
- DataCamp 3: due tonight.
- Midterm 1: Week from Tuesday.
  - Mostly conceptual question.
  - Mix of multiple choice and short answer.
  - Practice exam coming soon.
  - Will cover up through next lecture.
- Next Thursday is a midterm review session run by TFs.

Where are we going?

1. Review Sections 3.1–3.4 of Imai
   - Role of randomization in survey sampling
   - Non-response and other sources of bias
   - Missing data

2/ The role of randomization

1936 Literary Digest Poll



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

- Literary Digest predicted elections using mail-in polls.
- Source of addresses: automobile registrations, phone books, etc.
- In 1936, sent out 10 million ballots, over 2.3 million returned.
- George Gallup used only 50,000 respondents.

	FDR's vote share
Literary Digest	43
George Gallup	56

Poll fail

	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

- **Selection bias**: ballots skewed toward the wealthy (with cars, phones)
  - Only 1 in 4 households had a phone in 1936.
- **Nonresponse bias**: people who respond are different than those who don't.
- Lesson: when selection procedure is biased, adding more observations doesn't help!

1948 Election

The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

- **Quota sampling**
- fixed quota of certain respondents for each interviewer
- sample resembles the population on these characteristics
- potential unobserved confounding $\rightsquigarrow$ **selection bias**
- Republicans easier to interview within quotas (phones, listed addresses, etc)

2020????


PRICE $7.99
THE NEW YORKER
SEPT. 14, 2015
2020
TRUMP DEFEATS KANYE

Sample surveys

- **Probability sampling** to ensure representativeness
  - Definition: every unit in the population has a known, non-zero probability of being selected into sample.
- **Simple random sampling**: every unit has an **equal** selection probability.
- Random digit dialing:
  - Take a particular area code + exchange: 617-495-XXXX.
  - Randomly choose each digit in XXXX to call a particular phone.
  - Every phone number in America has an equal chance of being included in sample.

Sampling lingo

- **Target population**: set of people we want to learn about
  - Ex: people who will vote in the next election.
- **Sampling frame**: list of people who are going to vote.
  - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample**: set of people contacted.
- **Respondents**: subset of the sample that actually picks up the phone.
  - Unit non-response: sample $\neq$ respondents
- **Completed items**: subset of questions that respondents answer.
  - Item non-response

Difficulties of sampling

- Problems of telephone survey
  - Cell phones (double counting for the wealthy)
  - Caller ID screening (unit non-response)
  - Response rates down to 9%!
- An alternative: Internet surveys
  - Opt-in panels, respondent-driven sampling $\rightsquigarrow$ **non-probability sampling**
  - Cheaper, but non-representative
  - Digital divide: rich vs. poor, young vs. old
  - Correct for potential sampling bias via statistical methods.

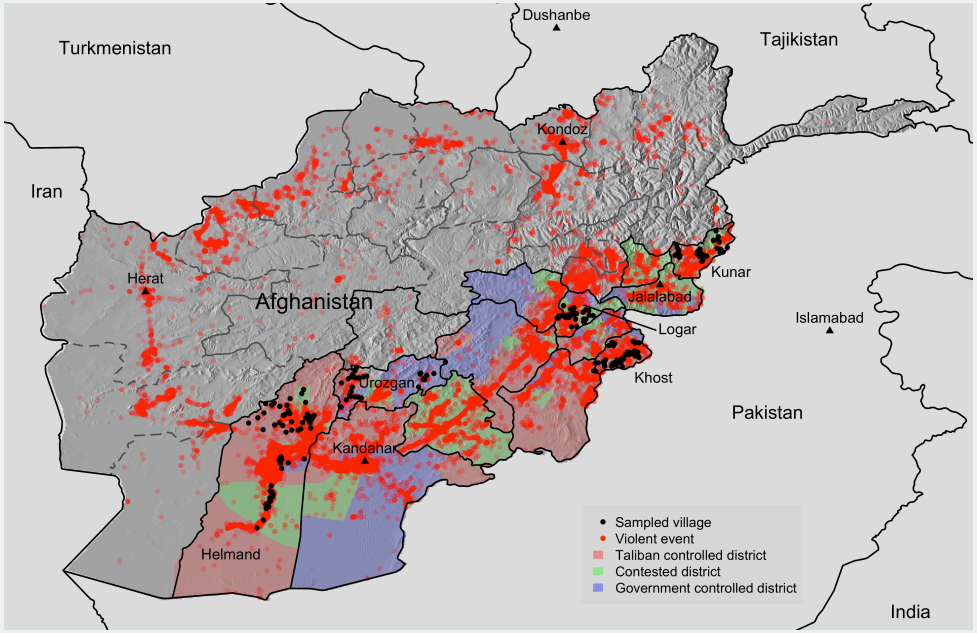
3/ The power of randomization

Why randomization works

- Randomization of surveys creates two groups: the sampled and the unsampled.
  - Just like RCTs creating two groups: treatment and control.
- If coin flips decide who gets in the sample, then the **sampled** and **unsampled** groups should be identical, at least on average.

Civilian attitudes and war against insurgency

- Conventional war: one military against another
- Counter-insurgency war: military against insurgents
  - From Vietnam to Iraq/Afghanistan
  - Key to victory: winning hearts and minds of civilians
  - aid provision, information campaign, minimizing civilian casualties
- Afghanistan study: sample civilians on their exposure to violence and support for Taliban, coalition forces



Cluster sampling

- One problem with randomization: need a list to sample from.
  - Random digit dialing: all phone numbers.
  - Other polls are using voter files.
  - No comprehensive list of citizens in Afghanistan to use
- Alternative: **multi-stage cluster sampling**
  - Randomly choose villages from a list of all villages
  - Go to each village and randomly choose households.
- Question: do the sampled villages look representative?

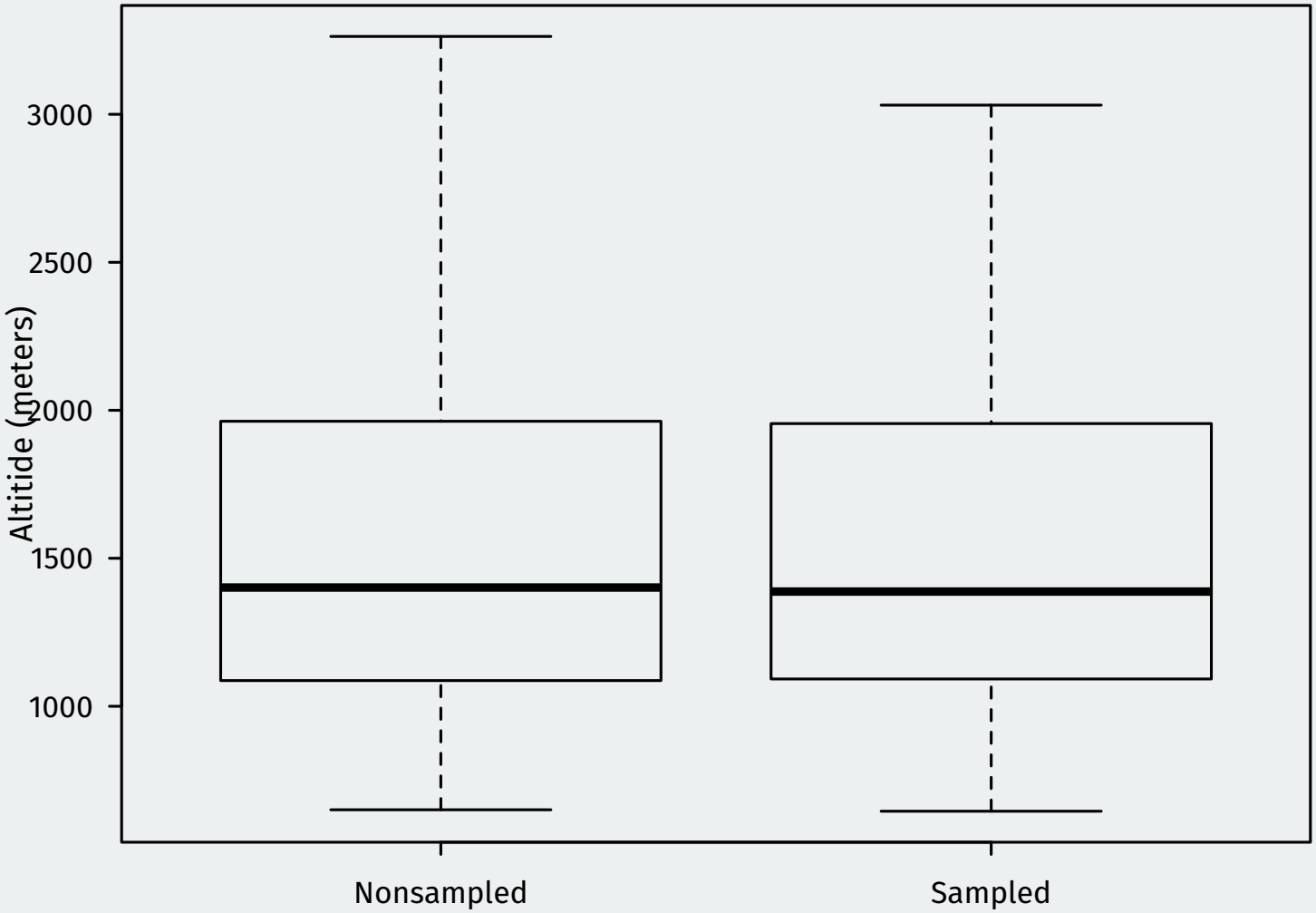
Are the sampled villages representative?

```
afghan.village <- read.csv("data/afghan-village.csv")
head(afghan.village)
```

```
##   altitude population village.surveyed
## 1     1959        197                 1
## 2     2426        744                 0
## 3     2237        179                 1
## 4     1692        225                 0
## 5     1928        379                 0
## 6     1195        617                 0
```

```
boxplot(altitude ~ village.surveyed, data = afghan.village,
        ylab = "Altitide (meters)",
        names = c("Nonsampled", "Sampled"))
```

Altitude distributions



4/ Missing data in R

Handling missing data in R

- Missing data in R: a special value `NA`
- Adding `na.rm = TRUE` to some functions removes missing data.

```
afghan <- read.csv("data/afghan.csv")
## prop. of those who got hurt by ISAF
mean(afghan$violent.exp.ISAF)
```

```
## [1] NA
```

```
mean(afghan$violent.exp.ISAF, na.rm = TRUE)
```

```
## [1] 0.375
```

- Or, you can explicitly remove missing values using `na.omit()` function:

```
mean(na.omit(afghan$violent.exp.ISAF))
```

```
## [1] 0.375
```

Available-case vs complete-case analysis

- **available-case analysis**: use the data you have for that variable:

```
sum(!is.na(afghan$violent.exp.ISAF))
```

```
## [1] 2729
```

```
mean(afghan$violent.exp.ISAF, na.rm = TRUE)
```

```
## [1] 0.375
```

- **complete-case analysis**: only use units for which you have data on all variables (**listwise deletion**)

```
dim(na.omit(afghan))
```

```
## [1] 2554   11
```

```
mean(na.omit(afghan)$violent.exp.ISAF)
```

```
## [1] 0.372
```

Cross-tabs with missing data

- Add `NA` to `table()` with `exclude = NULL`:

```
table(ISAF = afghan$violent.exp.ISAF, exclude = NULL)
```

```
## ISAF
##    0    1 <NA>
## 1706 1023   25
```

- **Contingency table**: distribution cases are spread across two variables.

```
table(ISAF = afghan$violent.exp.ISAF,
      Taliban = afghan$violent.exp.taliban, exclude = NULL)
```

```
##        Taliban
## ISAF       0    1 <NA>
##   0     1330  354   22
##   1      475  526   22
##   <NA>     7    8   10
```

Non-response and other biases

- Item non-response, like unit non-response, can create bias.
- More violent areas $\rightsquigarrow$ more non-response:

```
tapply(is.na(afghan$violent.exp.taliban), afghan$province,
       mean)
```

```
## Helmand   Khost   Kunar   Logar Uruzgan
## 0.03041 0.00635 0.00000 0.00000 0.06202
```

```
tapply(is.na(afghan$violent.exp.ISAF), afghan$province,
       mean)
```

```
## Helmand   Khost   Kunar   Logar Uruzgan
## 0.01637 0.00476 0.00000 0.00000 0.02067
```

- Sensitive questions $\rightsquigarrow$ non-response, **social desirability bias**
- racial prejudice, corruption, even turnout
- Do you support ISAF? What about Taliban?

Public nature of interviews

List experiments

- Script for the **control group**:

  I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

  Karzai Government; National Solidarity Program; Local Farmers

List experiments

- Script for the **treatment group**:

  I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

  Karzai Government; National Solidarity Program; Local Farmers; ISAF (Taliban)

Analysis of List Experiments

- Proportion of those who support ISAF:

```
mean(afghan$list.response[afghan$list.group == "ISAF"]) -
  mean(afghan$list.response[afghan$list.group == "control"])
```

```
## [1] 0.049
```

- Why does this work?
  - Control group mean: avg number of control items
  - Treatment group mean: avg number of control items + proportion of people supporting ISAF.

Next time

- Summarizing the relationships between two variables.
- Make sure to have read QSS 3.5–3.6