Survey Sampling: Predicting U.S. Elections

Zeyang (Arthur) Yu

Princeton University

October 30, 2024

Outline

1. Survey Sampling: An Intuitive Example

2. Survey Sampling: Motivation

3. Survey Sampling: Definitions and Basic Properties

4. Survey Sampling: Caveat

5. Survey Sampling: Predicting U.S. Elections

Do POL 345 Students Like Hawaiian Pizza?

Super Hawaiian Pizza, by Papa John's Pizza

Do POL 345 Students Like Hawaiian Pizza? (Conti.)

I Want to Use A Number To Answer the Question

- Question: do POL 345 students like Hawaiian pizza?
  - This is a descriptive question, but **NOT** a causal question
  - Restrict to **POL 345 students** (all units of interest, population)
- Intuitively, we can answer this question by:
  - The percentage of POL 345 students who like Hawaiian pizza

The Number We Have in Mind: P(Likes Hawaiian Pizza)

- Probability space defined on POL 345 students
  - Equal probability for each student being drawn
- Random variable: X_i
  - $X_i = 1$: like Hawaiian Pizza; $X_i = 0$: doesn't like Hawaiian Pizza
- Answer: $E(X_i)$, that measures P(Likes Hawaiian Pizza)

Do POL 345 Students Like Hawaiian Pizza? (Conti.)

How to Know $E(X_i)$ - Be a "Ruthless" Instructor

- I can simply abuse my power as the instructor
  - Create a PSet on telling me your preference for Hawaiian Pizza
  - Then, I can know all of your preferences
- This solution sounds creepy (to myself)...

How to Know $E(X_i)$ - A "Statistical" Solution

- Let me randomly survey 5 students and ask their preferences
- Then, let me use these 5 students to guess $E(X_i)$
- Now, let's do these two surveys and ...
  - See some randomness
  - See how we guess $E(X_i)$ based on these five students

Outline

1. Survey Sampling: An Intuitive Example
2. Survey Sampling: Motivation
3. Survey Sampling: Definitions and Basic Properties
4. Survey Sampling: Caveat
5. Survey Sampling: Predicting U.S. Elections

Outline

1. Survey Sampling: An Intuitive Example
2. Survey Sampling: Motivation
3. Survey Sampling: Definitions and Basic Properties
4. Survey Sampling: Caveat
5. Survey Sampling: Predicting U.S. Elections

Survey Sampling: Motivation

Three canonical problems in mathematical statistics

- Sample: $\{X_i\}_{i=1}^n$ distributed according to P (population *dist.*)

- E.g., X_i: whether voter i voted for Obama in 2008

- “Learn” some “features” of P (e.g., a *param.* $\theta(P)$) from the data

- E.g., $\theta(P) = E[X_i]$, vote share of Obama in 2008

- Provides a “best guess” $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ for $\theta(P)$

- E.g., $\hat{\theta}_n = \frac{\sum_{i=1}^n X_i}{n}$, use sample average to “guess” $E[X_i]$

- Test a hypothesis about $\theta(P)$

- E.g., construct a *fun.*, $\phi_n(X_1, \ldots, X_n)$, to decide reject or not

  Given the sample, can we reject statement that McCain will lose

  Since we’ll make mistakes, can we control these errors

- Construct a confidence region for $\theta(P)$

- E.g., a *rand.* set, $C_n(X_1, \ldots, X_n)$ covers $E[X_i]$ *w.* pre-specified *prob.*

Outline

1. Survey Sampling: An Intuitive Example

2. Survey Sampling: Motivation

3. Survey Sampling: Definitions and Basic Properties

4. Survey Sampling: Caveat

5. Survey Sampling: Predicting U.S. Elections

Survey Sampling: Population vs. Sample

Population: definition

A population is all the units (e.g., individuals, firms, etc.) of interest.

Population: remarks

- In the motivation example, what is the population of interest?
- - Population: all the U.S. voters in the 2008 presidential election
- Usually, no available data for all units in a population

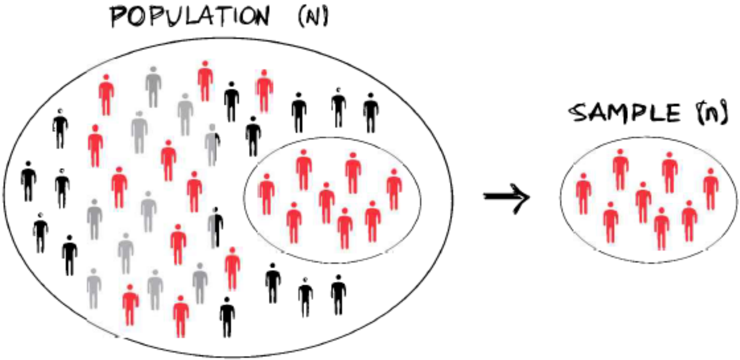
Sample: definition

A sample is a subset of the units in a population.

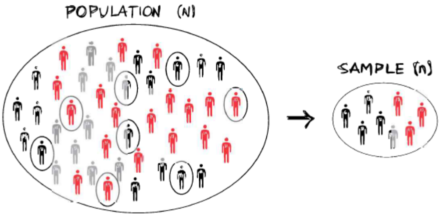
Sample: remarks

- Usually, there is available data for units in samples
- E.g., voters in the polls during 2008 presidential election

Survey Sampling: Population vs. Sample



Survey Sampling: Representativeness



Survey sampling: what is a “nice” property of survey data

- Natural to think the criteria: survey data ≈ population
- There is a jargon, representative sample, to describe this

Representative sample: definition

If we repeat the sampling procedure many times, the features of each resulting sample would on average equal to the population features.

Survey Sampling: Simple Random Sampling

Important sampling techniques

- Simple random sampling (SRS)
- (*Optional*) quota sampling, multistage cluster sampling

Simple random sampling (SRS): definition

SRS selects a predetermined number of respondents to be surveyed from a target population, with each potential respondent having an equal chance of being sampled into the survey.

Simple random sampling (SRS): example

- Each voter had a chance of $\frac{1}{N_{\text{total voter}}}$ being sampled in 2008

- $N_{\text{total voter}}$: total number of eligible voters in U.S. in 2008

SRS produces a sample that is representative of the population.

Outline

1. Survey Sampling: An Intuitive Example
2. Survey Sampling: Motivation
3. Survey Sampling: Definitions and Basic Properties
4. Survey Sampling: Caveat
5. Survey Sampling: Predicting U.S. Elections

Survey Sampling: Nonresponse Bias

Survey sampling: what can go wrong

- Naturally, we expect non-response in the sampling process
- China Employer-Employee Survey... (when I was young)

  Some firms refused survey because it disrupted production

Unit nonresponse (UNR): definition

Unit nonresponse (UNR) refers to a case in which a sampled respondent refuses to participate in the survey.

Unit nonresponse (UNR): what can go wrong

- It's fine if UNR is random (nonresponse doesn't depend on X_i)
- If nonresponse depends on X_i, then, will cause problems
- E.g., consider all McCain supporters refused to take your survey

Survey Sampling: An Example for Nonresponse Bias

i	inc	SRS	Nonresponse
1	10	$\frac{1}{5}$	$\frac{1}{4}$
2	25	$\frac{1}{5}$	$\frac{1}{4}$
3	10	$\frac{1}{5}$	$\frac{1}{4}$
4	25	$\frac{1}{5}$	$\frac{1}{4}$
5	500	$\frac{1}{5}$	0

Survey Sampling: An Example for Nonresponse Bias

Under SRS, randomly sample 1 individual *w. prob.* $\frac{1}{5}$

$$\underbrace{E(inc)}_{\text{Distribution is the 3rd col}} = \frac{10+25+10+25+500}{5} = 110$$

Under UNR, i.e., richest i does not respond

$$\begin{aligned}\underbrace{E(inc)}_{\text{Distribution is the last col}} &= \frac{10+25+10+25}{4} \\ &= \underbrace{15}_{\text{Under estimate } inc \text{ because miss the richest}}\end{aligned}$$

Outline

1. Survey Sampling: An Intuitive Example

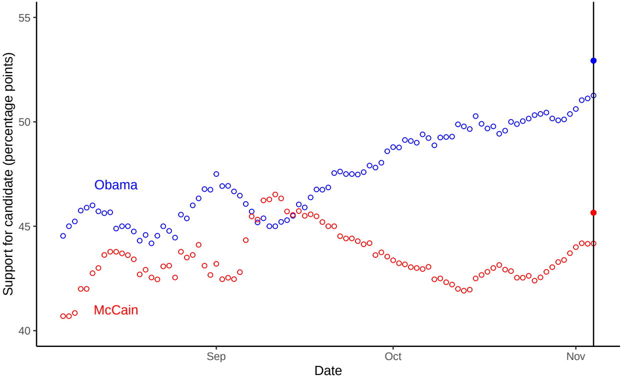
2. Survey Sampling: Motivation

3. Survey Sampling: Definitions and Basic Properties

4. Survey Sampling: Caveat

5. Survey Sampling: Predicting U.S. Elections

Survey Sampling: Predicting U.S. Elections



We are making errors with survey samples

- Discrepancies between samples and the actual outcome

  - prediction error = actual outcome - predicted outcome
- Wrongly claim that McCain will win during September

Survey Sampling: Errors Predicting U.S. Elections

	Parameter Space	
	Obama Won	McCain Won
Sample Tells Us		
Obama Won		Error
McCain Won	Error	

Learning Goals: Survey Sampling

Students will be able to:

- Know the following concepts in a survey

- Population, sample
- Representative sample
- Simple random sampling
- Nonresponse bias