Probability Sampling (Part 1)

Brady T. West

Research Associate Professor, Survey Research Center, Institute for Social Research

UNIVERSITY OF MICHIGAN

© 2018 The Regents of the University of Michigan
Except where otherwise noted, this work is licensed under
http://creativecommons.org/licenses/by-nc/3.0/

Share

0:01 / 10:57

Share

Lecture Overview

- Simple random sampling (SRS), and links to i.i.d. data

  Example: Email response times

- Complex sampling for larger populations: stratification, cluster sampling, and weighting

  Example: The NHANES

- Key benefits of probability sampling

we'll keep talking about key benefits of
probability sampling as we move forward.

1:25 / 10:57

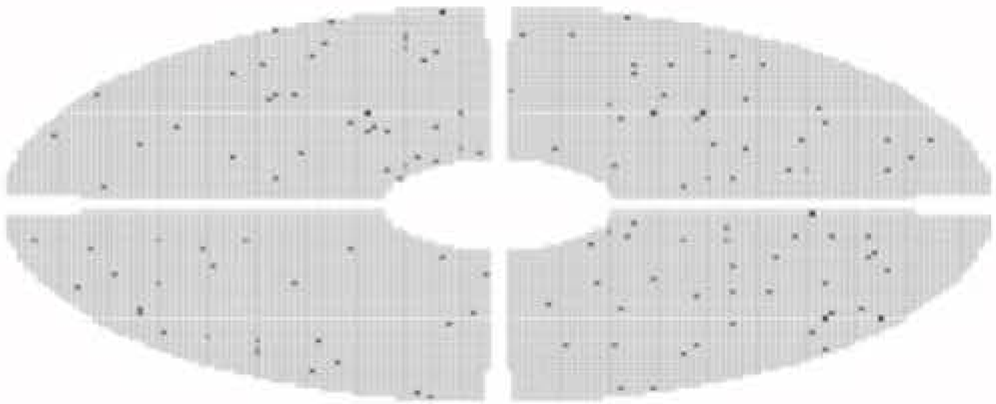
Simple Random Sampling (SRS)

- Start with known list of *N* population units, and randomly select *n* units from the list
- Every unit has **equal probability of selection = *n* / *N***
- All possible samples of size *n* are equally likely
- Estimates of means, proportions, and totals based on SRS are ***unbiased*** *(equal to the population values on average!)*

and other statistics of interest
based on the data that we collect
2:55 / 10:57
Share

Simple Random Sampling (SRS)

Consider this **stadium view** of a random sample of *n* = 134 people out of 10,000 people:



So we have a representative selection from
all the different areas of that particular

4:09 / 10:57

Share

Simple Random Sampling

- Can be **with replacement** or **without replacement**
- For both: probability of selection for each unit still n / N

- SRS rarely used in practice ~
  collecting data from n randomly sampled units
  in large population can be expensive $$$ *(more on this later!)*

Collecting data from n
randomly sampled units

5:32 / 10:57

SRS: Connection to i.i.d. Data

- Recall: i.i.d. observations are **independent** and **identically distributed**
- SRS will generate i.i.d. data for a given variable, *in theory*...

All randomly sampled units will yield observations that are independent (not correlated with each other) and identically distributed (representative, *in theory*)

Okay, so they're representative of some larger population of values, again,

Share
7:08 / 10:57

Share

SRS Example

- Customer service database: N = 2,500 email requests in 2018
- Director wants to estimate: **mean email response time**
- Exact calculations require manual review of each email thread
- Asks analytics team: sample, process and analyze n = 100 emails

8:39 / 10:57

Share

SRS Example

- **Naive Approach**: process the first 100 emails on the list
  - Estimated mean could be **biased** if customer service representatives learn or get better over time at responding more quickly
  - First 100 observations may come from a small group of staff

    → **not fully representative, independent, or identically distributed**!
  - **No random selection** according to **specific probabilities**!

probability sample, and that provides
us with important limitations.

9:52 / 10:57

Share

SRS Example

- **Better SRS Approach:** number emails 1 to 2,500 and randomly select 100 using a random number generator
  - **Every email has known probability of selection** = 100 / 2,500
  - Produces **random, representative sample** of 100 emails (*in theory*)
  - **Estimated mean response time** will be an ***unbiased*** estimate of the population mean

The estimated mean response time in this
case will also be an unbiased estimate

10:36 / 10:57

An ordered list of all students in a classroom has the following ages:

17, 21, 20, 21, 19, 18, 21, 20, 20, 17, 19, 20

A researcher wishes to select a simple random sample of size 5, and a random number generator calls for the sampling of elements 3, 8, 9, 2, and 5 from the ordered list. What is the probability of selection into this simple random sample, and what is the mean age based on the sample?

- 1/5, 19
- 1/5, 20
- 5/12, 19
- 5/12, 20

Correct

Of the 12 students, 5 are selected, making the probability of selection 5/12. The ages of the selected students based on the ordered list are 20, 20, 20, 21, and 19, making the average age 20.