Probability Sampling

Brady T. West

Lecture Overview

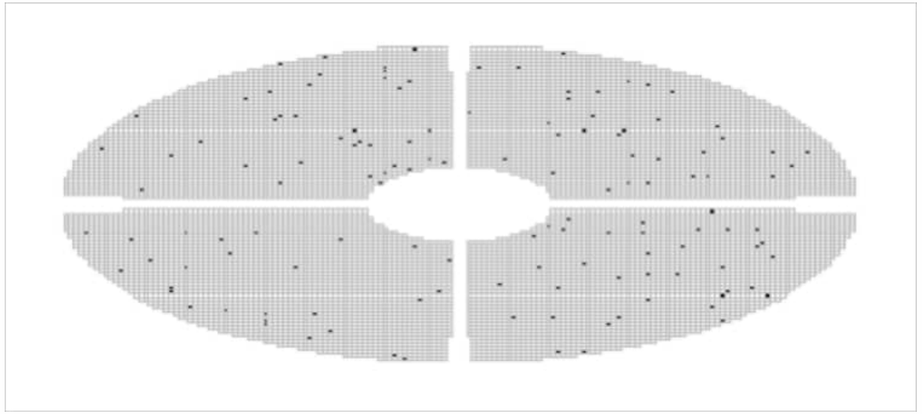
- Simple random sampling (SRS), and links to i.i.d. data
- Example: Email response times
- Complex sampling for larger populations: stratification, cluster sampling, and weighting
- Example: The NHANES
- Key benefits of probability sampling

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!)

Simple Random Sampling (SRS)

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



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!)*

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*)

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

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**!

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