STAT22000 Autumn 2013 Lecture 9

Yibi Huang

October 18, 2013

3.2 Sampling Designs

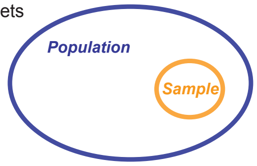
Four Keywords in Sampling

- □ **Population:** The entire group of individuals in which we are interested but can't usually assess directly.

  Example: All humans, all working-age people in California, all crickets

- □ **Sample:** The part of the population we actually examine and for which we do have data.

  How well the sample represents the population depends on the sample design.



- □ A **parameter** is a number describing a characteristic of the population.
- □ A **statistic** is a number describing a characteristic of a sample.

Example

Suppose we want to predict the result of an election in a state.
A sample of 3000 citizens are interviewed:
45% support candidate A, and 40% support candidate B.

- Population: citizens in the state that is going to vote
- Parameter(s): the percentage of votes for candidate A, and also the percentage of votes for candidate B
- Sample: the 3,000 citizens interviewed
- Statistic: the percentage in the sample that support candidate A, and that for candidate B

Sample Survey v.s. Experiments and Observational Studies

- Experiments and observational studies are for comparison, or to explore relationships between variables (association or causation)
- Sampling is for making inference or conclusions about a population from a sample. Whether results found in a sample can be extended to the whole population depends on whether the sample is properly selected the population.

Some Bad Sampling Methods

- **Convenience Sampling** — Just ask whoever is around.
  - E.g. "Man on the street" survey (cheap, convenient, now very popular with TV "journalism")
  - Problem: results may vary greatly with "when and where" the survey is done, lack of representation
- **Voluntary Response Sampling**
  - e.g., internet polls, call-in surveys
  - Only people visiting the website/watching the program will be sampled
  - People with strong opinions are more likely to participate

Better Sampling Designs

- Simple Random Sampling
- Stratified Sampling
- Cluster Sampling
- Multistage Clustered Sampling

Simple Random Sampling

Basic idea: put the names in a box and make draws from the box

- need a list of names of all subjects in the population, called **sampling frame**
- all subjects have the same chance to be chosen
- the Law of Large Number ensures that the makeup of a simple random sample will mimic the makeup of the population (age/gender/race/income...)
- impractical for large population

Stratified Sampling

The population is divided into groups, called **strata**, and then a separate simple random sample is chosen in each stratum.

- e.g. divide by school grade/sex/geographical region
- after division, subpopulations are smaller, easier to conduct simple random sampling
- (works better for population with large strata-to-strata variation but small within-strata variation)

Clustered Sampling

The population is divided into groups, called **clusters**.

- A sample of clusters is chosen. All subjects in the selected clusters are interviewed.
- Example 1: Suppose Walmart wants to survey its employees. It can choose a number of stores, and interview all employees in the selected stores. Here a cluster is a store.
- Example 2: Suppose a biologist wants to access the percentage of pine trees affected by some tree disease. He may divide forests into small regions, randomly pick a few regions, then examine every pine tree in the selected regions. Here a region is a cluster.
- (Works better for population with small cluster-to-cluster variation but large variation within clusters)

Multistage Cluster Sampling

- First stage: the population is divided into groups, called **clusters**, and a sample of clusters is chosen.
- Second stage: the selected clusters is further divided into sub-clusters, and a sample of sub-clusters is chosen in each selected cluster.
- (Third stage: ...)
- (Fourth stage: ...)

Most nationwide surveys (like GSS) use this method

- towns → wards → precincts → households

Advantage:

- reducing traveling cost of interviewers,
- no need to make sampling frame for unselected sub-clusters

Common Problems in Sample Surveys

- **Undercoverage** – some groups of the population are left out of sampling frame
  - e.g., U.S. Census goes "house to house", homeless people are not represented
  - More and more people use cell phone only, having no land lines. Telephone surveys that sample from land lines will miss these cell-phone-only people
- **Non-response bias** – non-respondents can be very different from respondents.
  - Solution: call back, double sampling scheme
- **Response bias** – the answers by respondents are influenced to some extent by the phrasing of the questions, and even the tone or attitude of the interviewer.
  - Solution: interviewer control, proper design of questionnaires

Example 1: The *Literary Digest* Poll

Literary Digest

- well-known magazine in U.S. from 1890 to 1936
- old issues at Regenstein
- had run presidential polls since 1920; always right
- bankrupt in 1938

The 1936 election

- 10 million postcard were sent (20% of voters in the country)
- Names from phone lists, auto registrations, and club registers
- 2,376,523 postcard replies, response rate ≈ 24%

	FDR	Landon	Lemke	Sample Size
Literary Digest	41%	55%	4%	2.4 million
Gallup	56%	?	?	50,000
Result	61%	37%	2%	

Why failed?

- Undercoverage: in 1936, poor people were less likely to have cars, phones or join clubs. They were under-represented
- Low response rate