1.2 Statistical Sampling

Dr. Lauren Perry

Goals

2. Describe sampling techniques.
   - Understand key terms
   - Describe how to collect data using a random sample
   - Understand the differences between simple random, stratified, cluster, and systematic sampling.

Statistical Sampling

How do we get samples?

- We want a sample that represents our population.
- **Representative samples** reflect the relevant characteristics of our population.
- In general, we get representative samples by selecting our samples *at random* and with an adequate sample size.

A non-representative sample is said to be **biased**.

Ex: A sample of chihuahuas to represent all dogs.

These can be a result of **convenience sampling**, choosing a sample based on ease.

Common sources of bias in our daily lives:

- *Anecdotal evidence* is data based on personal experience or observation.
  - Typically this consists of only one or two observations and is NOT representative of the population.
- Availability bias is your brain's tendency to think that examples of things that come readily to mind are more representative than is actually the case.

Simple Random Samples

- We avoid bias by taking random samples.
- One type of random sample is a **simple random sample**.
  - We can think of this as "raffle sampling", like drawing names out of a hat.
  - Each case (or each possible sample) has an equal chance of being selected.
  - Knowing that A is selected doesn't tell us anything about whether B is selected.
- Instead of literally drawing from a hat, we usually use a **random number generator** from a computer.

Stratified Sampling

- In a **stratified sample**, we break the population down into groups called **strata**
  - Strata are based on characteristics we think might be relevant to our study.
  - Individuals or items within a strata should be fairly similar to each other.
  - We then take a random sample from each strata.
    - This ensures we have representation from each group.

Example: Stratified Sampling

- A local politician believes men and women will vote differently on an upcoming ballot measure.
- She goes into the community and randomly samples 50 men and 50 women to ask for their thoughts on the ballot measure.

Example: Stratified Sampling

We can also use stratified sampling to make sure that the proportion of items in each group in the population matches the proportions in our sample.

- A local high school is 29% freshmen, 27% sophomores, 24% juniors and 19% seniors.
- They want to collect a sample of 100 students using class level as strata.
- Since some class levels have more students than others, they set their strata to match:
  - 29 freshmen (29% of their sample), 27 sophomores, 24 juniors, and 19 seniors.

This approach to stratified sampling can also help us ensure that small strata are adequately represented in our study.

Example: Stratified Sampling

- Suppose we are doing some drug development research for a particular disease and know that a very small part of our population develops an especially severe form of the disease.
  - In order to make sure those individuals are represented in our sample, we could treat disease severity as strata.
  - Motivation: in a simple random sample, we might miss those individuals entirely!

Cluster Sampling

- In a **cluster sample**, we break the population into **clusters**.
  - Each cluster is similar to the population (and so the clusters are all similar to each other).
  - We then take a random sample of clusters and measure *all* items or individuals within each of those randomly selected clusters.

Example: Cluster Sampling

- An airline wants to survey people who take its international flights from the United States to Asia.
- They randomly select 10 of these flights and give the survey to every individual on each of those 10 flights.

Example: Cluster Sampling

- A farmer wants to know something about the plants in their fields.
- They randomly select 5 of their fields and examine all of the plants in each field.

Example: Cluster Sampling

The potential downside to cluster sampling is that there may be factors that make clusters meaningfully different from one other.

- Say our airline randomly samples flights to Asia.
  - It's likely the people going to Vietnam are different from the people going to the Philippines who are different from the people going to India.
  - That is, not every flight to Asia is representative of the *entire* population of individuals who fly to Asia.

Systematic Sampling

- In a **systematic sample**, we choose some starting point in our population and then collect every kth observation.

Examples: Systematic Sampling

1. If I have a list of student ID numbers for every student at Sacramento State, I could generate a sample of students by selecting every 100th number.
2. Suppose we want to examine some machine part coming off on an assembly line. We collect a sample by pulling every 20th part off of the assembly line for additional testing.

Examples: Systematic Sampling

One potential issue with systematic sampling is that there may be some pattern in the data.

- Suppose a machine on an assembly line is oiled after producing 20 components, and its performance degrades steadily after it is oiled.
- If we select every 20 components, we match this *periodicity* and fail to capture a representative sample of components.