A. Sampling

A.1 Random Sampling
Sampling Designs
Sample Statistics
Exercises

Introduction

Let's say you're a scientist, (an *actuarial* scientist?) and you want to perform a study with some goal in mind:

- Determine if eating fewer eggs reduces prevalence of heart disease
- Determine if people who purchase above-the-minimum car insurance have increased odds of accident
- Determine the number of people who still watch movies on VHS

In each case, you are trying to learn something about a **population**, the entire group of interest.

Often, you are trying to determine a **parameter** of the population - a numerical characteristic defined for some specific variable.

Introduction

To determine the true answer to our question or the true value of a parameter, it would be necessary to observe the entire population.

This is not usually feasible as populations are often too large or difficult to access.

Instead we attempt to collect data from a **representative sample** of the population - a sample that does not differ in any systematic or important way from the population.

A numerical function of the sample data is called a **statistic**.

Forming a representative sample is not always trivial. Poor sampling may result in biased statistics.

To address sampling difficulties, numerous **sampling designs** have been developed.

Types of Sampling

How can we sample a population?

Haphazard/Convenience Sampling - choosing a sample that is easily available

- Example: asking a question to random people on the street
- Pros: easy, fast, cost-effective
- Cons: typically subject to bias, non-representative of most populations

Judgment Sampling - sample is selected based on the judgment of the data-collector

- Example: asking questions to a group of people selected because of their diversity
- Pros: less biased than convenience sampling, still easy, fast
- Cons: subject to bias of data-collector, may give too much weight to certain groups within population

Types of Sampling

Quota Sampling -sample is based on quotas for specific characteristics, based on population

- Example: If a population is known to be 50% Male and 50% Female, the sample is also selected as 50/50 M/F
- Pros: decreases bias if selected characteristics explain population parameters
- Cons: population characteristics may be unknown/difficult to determine, selected characteristics cannot include all explanatory characteristics

Types of Sampling

Simple Random Sampling - a sample of size n taken from population N without replacement such that each possible sample of size n is equally likely to be chosen

- Example: assign each item of the population a unique number, then select a sample using a random number generator, ignoring duplicates
- Pros: No bias involved, provides a good representation of population characteristics provided n is sufficiently large
- Cons: May be difficult or impossible to implement, due to access or willingness of population, not good for large or highly diverse populations that would require very large n to be representative

Types of Sampling

Stratified Random Sampling - divide a population into homogeneous sub-populations having similar units, then take a simple random sample from each. These sub-populations are called **strata**

- Example: divide a population into buckets based on income, then randomly sample each
- Pros: can obtain information on sub-populations without additional work, sampling sub-populations may be more administratively feasible than sampling the full population as n decreases for groups with less impact
- Cons: may still require unfeasible n for very large populations, certain strata may be impossible to observe.

Types of Sampling

Multistage cluster sampling - perform a series of tiered simple random samples, based on specified characteristics

- Example: suppose the population is all US citizens. The you would draw a simple random sample of states, from the selected states a simple random sample of counties, from those a simple random sample of towns, and from those a simple random sample of citizens.
- Pros: flexible, allows sampling of very large populations
- Cons: high variance - repeating the process twice may result in very different samples, sample may ignore key elements of population

Types of Sampling

Systematic Sampling - Select one random unit from the first k units, and then every kth unit after.

- Example: we want to study cars coming off an assembly line, so we randomly select one of the first 20 cars off the line, and then every 20 cars after that. If we selected the $5th$ car, then our sample would be cars $5, 25, 45, 65, \ldots$
- Pros: practical for very large populations that have been previously ordered (ie. people in a telephone directory), gives representative samples of large populations, widely used and understood
- Cons: does not work if there are cyclical variations, requires population to have been previously ordered

Sample Statistics

A **sample statistic** can be any numerical function of observed sample data

Of particular interest are sample statistics correspond to parameters of the underlying population.

Examples:

- Sample mean, sample median, sample mode
- Sample variance
- Sample min/max
- $\bar{x} \pm 1.645 \frac{s}{\sqrt{n}}$

$\bar{X}$ and S^2

Two sample statistics are of great importance in many applications and are critical to this course.

Let $x_1, x_2, \ldots, x_n$ be n observations that make up the sample. The **sample mean** and **sample variance** are defined as follows.

Sample Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sample mean corresponds to the population mean, typically denoted μ.

$\bar{X}$ and S^2

Sample Variance:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

If the population mean, μ is known. If it is not know, then the sample variance is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

We divide by $n-1$ in the latter case because we lose one **degree of freedom** when we use our data to estimate μ.
Typically, μ is not known, so we use the latter formula.
The sample variance corresponds to the population variance, typically denoted σ^2,

$\bar{X}$ and S^2

Example:
The following sample data is observed during a study on actuarial exam scores:

$$3 \quad 3 \quad 4 \quad 5 \quad 5 \quad 6 \quad 8 \quad 10$$

Calculate the sample mean and variance of the sample data.

Sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(3+3+4+5+5+6+8+10) = 5.5$$

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{8-1}\left[(3-5.5)^2 + \ldots + (10-5.5)^2\right] = 6.00$$

Exercise 1

A researcher is studying the impact of egg consumption on heart disease. The population for the study is all Canadian citizens over the age of 40.

To collect sample data, the researcher divides the population into decennial age groups (40 to 49, 50 to 59, etc.), assigns a number to each citizen within each group, and then randomly selects n citizens from each group.

This is known as which of the following?

A. Quota Sampling
B. Multistage Cluster Sampling
C. Simple Random Sampling
D. Stratified Random Sampling

Exercise 1

A researcher is studying the impact of egg consumption on heart disease. The population for the study is all Canadian citizens over the age of 40.

To collect sample data, the researcher divides the population into decennial age groups (40 to 49, 50 to 59, etc.), asigns a number to each citizen within each group, and then randomly selects n citizens from each group.

This is known as which of the following?

A. Quota Sampling
B. Multistage Cluster Sampling
C. Simple Random Sampling
D. Stratified Random Sampling

Exercise 2

The following sample data are collected during a study of hours spent annually watching YouTube videos:

$$0, 17, 55, 134, 175, 244, 278, 299, 367, 822, 901, 977, 999$$

Calculate the mean and variance of the sample.

(Hint, try the 'data' function on the TI-30XS MultiView if you have one)

Exercise 2

Sample Mean:

$$\bar{x} = \frac{1}{13}(0 + 17 + ... + 977 + 999) = 405.23$$

Sample Variance:

$$s^2 = \frac{1}{13-1}\left[(0 - 405.23)^2 + ... + (999 - 405.23)^2\right] = 143,172$$