Lecture I

Definition 1. Statistics is the science of collecting, organizing, summarizing and analyzing the information in order to draw conclusions.

It is a process consisting of 3 parts.

First part: collecting data

- Identify the research objective: the group that it is to be study is called **population**. A member of the population is called **individual**.
- Collect the information needed to answer the questions posed: typically look at a subset of the population called **sample**.

Second part: Organize and summarize the information

This step is called **descriptive statistics**. Uses tables, charts, graphs, etc to describe the data collected.

Third part: Draw conclusions from the information

This part is called **inferential statistics**.
Example:
We can not learn everything about the population just by looking at a sample!!! But we might be able to say something with a certain level of confidence.

1 Types of data

Definition 2. The characteristics, that we decided we are interested to study, of the individual within the population are called **variables**.

Variables can be classified into two groups:

Definition 3. **Qualitative** or categorial variables allow for classification of individuals based on some attribute or characteristics. **Quantitative variables** provide numerical measures of individuals. Arithmetic operations can be performed on the values of a quantitative variable and provide meaningful results.

Examples:

Quantitative variables can be classified into two types:

Definition 4. A **discrete variable** is a quantitative variable whose possible values could be counted: 0,1,2,3,4,5.

Examples:

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable.

Examples:

The list of observations a variable assumes is called data. Data could be classified in the same categories as variables.

Example:

Data can be obtained from four sources:

1. A census
2. Existing sources
3. Survey sampling
4. Designed experiments

Definition 5. A **census** is a list of all individuals in a population along with certain characteristics of each individual.

Existing data: Don't collect data that have already been collected.

Survey sampling is used when no attempt to influence the value of the variable of interest.

Examples: Polling, ...

Data obtained from a survey sample lead to an **observational study**. Sometimes it is referred to as *expost facto* (after the fact) studies because the value of the variable of interest has already been established.

A **designed** experiment applies a treatment to individuals (referred to as **experimental units**) and attempts to isolate the effects of the treatment on a **response variable**.

Observational studies are very useful tools for determining whether there is a relation between two variables, but it requires a design experiment to isolate the cause of the relation.

If control is possible, an experiment should be performed. If control is not possible or necessary, then observational studies are appropriate.

Sampling

How can a researcher obtain accurate information about the population through the sample while minimizing the costs?

There are 4 types of sampling:

- simple random sampling
- stratified sampling
- systematic sampling
- cluster sampling

All these sampling methods are based on the planned randomness techniques. The surveyor does not have a choice as to who is in the study.

Simple random sampling

Definition 6. If the population is of size N and we want a sample of size n $(n < N)$, a simple random sampling is obtained if every possible sample of size n has an equally likely chance of occurring. The sample is then called a **simple random sample**.

Examples:

What is sample with replacement or sample without replacement? When do we use them?

How do we obtain such a sample?

1. using a hat if the population is small!
2. using random number if the population is large:
   - (a) number the individuals in the population, from 1 to N. (that means that we have to have the **frame**-the list of all individuals in the population!
   - (b) select n random numbers from this list using a table of random numbers or using your calculator.

Using the table:

- Select a starting point.
- Look for numbers that have as many digits as N has.
- If a number is repeated, discard it.
- If a number is larger than N discard it.
- Stop when you obtain n numbers.

Stratified Sampling

Definition 7. A **stratified sample** is obtained by separating the population into nonoverlapping groups called strata and then obtaining a simple random sample from each stratum. The individuals within stratum should be homogeneous (or similar) in some way.

How do we obtain a stratified sample?

- Find how many individuals you need from each stratum by computing the percentage of the stratum in the population
- Perform a simple random sample in each of the stratum to find the desired number of individuals

Definition 8. A **systematic sample** is obtained by selecting every kth individual from the population. The first individual selected is a random number between 1 and k.

- Does not require a claim!
- How do we obtain a systematic sample without a frame? How do we establish k?

How to obtain a systematic sample when the population size is known to be N:

1. Determine the sample size n
2. Compute N/n and round down to the nearest integer. This value is k.
3. Randomly select a number between 1 and k. Call this number p.
4. The sample will consist of the following individuals:

$$p, p+k, p+2k, \cdots, p+(n-1)k$$

Cluster sampling

Definition 9. A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.

How do we obtain a cluster sampling?
• randomly select the cluster (using random sampling for example)
• survey all the individuals in the clusters.
Other questions: • How do I cluster the population?
• How many individuals in a cluster?
• How many clusters do I sample?