

CHAPTER 1: Introduction to Statistics

Section 1.1: An Overview of Statistics

“Statistics is the grammar of science.”

---Karl Pearson

Statistics – the study of procedures for collecting, describing, and drawing conclusions from information.

“**Statistics** is the science of collecting, organizing, analyzing, and interpreting data to make decisions.”

Sources of Information:

A. Population – A **population** is the entire collection of individuals about whom information is sought.

- Census – the collection of info from the entire population
- Parameter – a numerical measurement that describes a characteristic of a population → the result of a census.

B. Sample – A **sample** is a subset of a population containing the individuals that are actually observed.

- Sample Data – the collection of info from a sample.
- Statistic – a numerical measurement that describes characteristics of a sample → the result of sample data.

*Ideally, we want the sample we choose to represent the population appropriately. Since this is rarely possible, we use methods of sampling (choosing a sample) that give a better chance of the sample reflecting the major characteristics of the population under consideration.

1. Distinguishing Between a Parameter and a Statistic:
2. In a survey of automobile owners, 6% said they had to change their engine control module at least once. Ans: Statistic
3. In a recent year, the average math score on the ACT for all graduates was 20.2.
Ans: **Parameter**
4. In a random check of several hundred retail stores, the Food and Drug Administration found that 34% of the stores were not storing fish at the proper temperature. Ans: Statistic
5. Last year, a small company spent a total of \$5,150,694 on employees' salaries.
Ans: **Population parameter, because the total spent on employees' salaries, \$5,150,694, is based on the entire company.**

Popular Methods of Sampling:

Simple Random Sample – A simple random sample of size n is a sample chosen by a method in which each collection of n population items is equally likely to make up the sample.

Examples: Suppose there are ten equal shapes and sizes of red marbles in a basket, and we want to draw three red marbles without looking at the basket. So, each red marble has an equal chance to be selected.

*For this course, a sample will refer to a simple random sample (SRS).

statistics has two major branches:

Descriptive statistics: **It involves the organization, summarization, and display of data.**

Example: Finding five-number summary, drawing histogram, RF bar graph, pie chart, etc.

Inferential statistics: **It involves using a sample to draw conclusions about a population.**

Example: Hypothesis tests, confidence intervals, prediction intervals, and regression analysis

Exercise: A study of 513 respondents to an Internet-wide survey found that 97% of the respondents' said music is important to them, and 83% of the respondents said they actively look for new music.

- a) Identify the population and the sample.

Population- **The population consists of the responses of all Internet users.**

Sample- **the sample consists of the responses of the 513 respondents in the study.**

- b) Determine which part of the study represents the descriptive branch of statistics.

The part of this study that represents the descriptive branch of statistics involves the statement, "97% of the respondents' said music is important to them, and 83% of the respondents said they actively look for new music."

- c) What conclusions might be drawn from the study using inferential statistics?

A possible inference drawn from the study is that a large majority of Internet users find music to be important to them and actively look for new music.

For this course, sample will refer to a simple random sample.

Section 1.2 : Types of Data

When collecting information, we get a ____ **data set** _____. The items or subjects in a sample or a population are called the ____ **individuals** _____. Examples: **people, student, animals etc.**

The characteristics of the individuals about which we collect information are called ____ **variables** _____. Examples: **height, weight, #of credit hours etc.**

The ____ **values of the variables** _____ that we obtain are called ____ **data** _____.

Types of Variables:

1. **Qualitative or Categorical Variables** – **Qualitative variables classify individuals into categories.** Names or labels that are not numerical in nature
 - a. **Ordinal Variables** – **categories that have natural order.**
 Example: Letter grade A, B, C, D, E, F; size of a shirt XS, S, M, L, XL, XXL;
 - b. **Nominal Variables** – **name only, no order**
 Example: Brand name, gender, textbook titles etc.
2. **Quantitative or Numerical Data** – **Numerical values representing counts or measurements. Quantitative variables tell how much or how many of something there is.**

Example: # of pets, distance between home and school, #of correct answer,
height, weight etc.

- a. **Discrete Variables** – **# of values is finite and countable-all possible values can be listed. Example: #of pets, #of siblings, # of pairs of shoes.**
- b. **Continuous Variables** – **result from infinitely many possible values, too many values to list**
 Example: height, weight,

STA 2023 SECTION 1.3 DATA COLLECTION AND EXPERIMENTAL DESIGN

Learning Outcomes:

- 1) Distinguish between observational and experimental study
- 2) Analyze experimental designs.
- 3) Identify and analyze different sampling techniques
- 4) Analyze survey questions and identify bias

When we conduct an experiment or study:

- The individuals studied are called ____ **experimental units** _____ or ____ **subjects** _____. **example: people, animals etc.**
- The ____ **outcome or response** _____ is the variable being observed or measured for the subjects.
- The ____ **treatments** _____ are the procedures applied to each subject. There are always 2 or more ____ **treatments** _____. The purpose is to determine whether the treatment affects the outcome (placebo, different doses, etc.)

The goal of every statistical study is to collect data and then use the data to make a decision.

A statistical study can usually be categorized as an observational study or an experiment.

1. **Observational study:** a researcher observes and measures characteristics of interest of part of a population but does not change existing conditions.

Example: Cancer rate due to smoking, brain problems after a head injury, automobile crashes due to Speed, etc.

2. **Randomized Experiment** – the investigator assigns treatments to the experimental unit at random. If there are large differences in outcomes among the treatment groups, we can conclude that the differences are due to the treatments.

Example: medication, teaching methods (before and after)

***Confounding Variables** – a variable that is related to both the treatment and outcome

Example: In a weight loss program, if the treatment is diet, then exercise, sleep, and genetics are also variables that can influence the outcome and the treatment.

Data Collection

There are several ways to collect data.

- A **simulation** is the use of a mathematical or physical model to reproduce the conditions of a situation or process. Collecting data often involves the use of computers.
- A **survey** is an investigation of one or more characteristics of a population. Most often, surveys are carried out on *people* by asking them questions.

Experimental Design

The goal of sampling is to get a subset of the population that appropriately represents the population's characteristics.

If a procedure produces this result, it is said to be _____ **unbiased** _____.

If the sample overestimates or underestimates a population value, it is said to be _____ **biased** _____.

Blinding

It

double-blind experiment,

****Sampling types:**

Most of the time, it's tough and costly to collect data from the entire population. why? as because of the size of a population is much bigger than we can handle properly. Ideally, we want the sample we choose to represent the population appropriately. Since this is rarely possible, we use methods of sampling (choosing a sample) that give a better chance of the sample reflecting the major characteristics of the population under consideration.

Methods of Sampling:

1. **Simple Random Sample** – A simple random sample of size n is a sample chosen by a method in which each collection of n population items is **equally likely** to make up the sample, just as in a lottery.

Examples: Suppose there are ten equal shapes and sizes of red marbles in a basket, and we want to draw three red marbles without looking at the basket. So, each red marble has an equal chance to be selected.

2. **Samples of Convenience** – A sample that is not drawn by a well-defined random method.

Examples:

I want to choose five people from this class, and I can pick up the first five names from the attendance paper.

(It excludes a great proportion of the population.)

* A sample of convenience may be okay if the sample does reflect the population appropriately.

3. **Stratified Sampling** – A sample in which the population is divided into groups, called strata, where the members of each stratum have similar characteristics. Then a random sample is drawn from each stratum.

****each stratum may differ from one another, but members within each stratum tend to be alike.** Example: Freshmen, Sophomore, Junior, Senior...then choose five random students from each group.

4. **Cluster Sampling** – A sample in which items are drawn from the population in ____clusters____ or ____groups____. This type of sample is useful when a straight forward simple random sample is difficult to take. Examples: *Imagine drawing a simple random sample of households from a city and interviewing every member of each household. This would be a cluster sample, with the households as the clusters.*

The difference between cluster sampling and stratified sampling:

In both cluster sampling and stratified sampling, the population is divided into groups. In stratified sampling, a simple random sample is chosen from each group. In cluster sampling, a random sample of groups is chosen, and every member of the chosen groups is sampled

5. **Systematic Sampling** – A sample in which population items are ordered in some way and then it is decided how frequently to take items.
- Examples: *Imagine walking alongside a line of people and choosing every third one. One could sample every third item, or every fifth item, or every hundredth item.*

Automobiles are coming off an assembly line. It is decided to draw a systematic sample for a detailed check. The starting point will be the third car, then **every fifth** car after that will be sampled. We start with the third car, then count by fives to determine which cars will be sampled. The sample will consist of cars numbered 3, 8, 13, 18, and so on.

6. **Voluntary Response Sample** – A sample in which the respondents themselves decide to be involved in a study or process. Examples: **online survey, amazon reviews**.