What is the statistical science?

The science of data.

It involves collecting, classifying, summarizing, analyzing, and interpreting data.

American Heritage Dictionary defines statistics as:

The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.

The Merriam-Webster’s Collegiate Dictionary definition is:

A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.

Most of what we will be doing in this course centers around trying to understand a set of information (data).

This set of information is from ALL objects in a population. Often, this set of information is so big that obtaining all of the information about these objects is extremely difficult. Instead, this information may be hard to obtain making it impossible to obtain all of it.

To understand the population, we usually will take a representative subset of information from the population.

The subset is a smaller set of information that is usually more manageable. The objects from the population used in this subset form the sample. Typically, we want the objects in the sample to be very similar (representative) to objects in the population.

Example: Suppose we are interested in estimating the average GPA of all students at this university. How would we do this? Assume we do not have access to any student records.

* + - 1. Define the population of interest

The population is all students here

Problems obtaining information on all of the students:

* Thousands of students!
* Students drop out and enroll late
	+ - 1. Define a characteristic or variable of interest

GPA; this is often denoted as a “X” or “Y”

* + - 1. Define the parameter of interest

Parameter: Numerical summary measure used to describe a population characteristic.

The parameter is the population average here. The average is often called the “population mean” and denoted by the Greek letter “mu”, μ.



To find the average GPA, we could add up ALL of the GPAs of students in the population and divide by the total number of students.

Why would this be difficult to do?

Instead, suppose a representative set of students in the population was found and their average GPA is found.

* + - 1. Define the sample

Suppose a representative sample of 20 students is taken from the population. Each student’s GPA is called an observation because we are “observing” a particular student’s GPA.

* + - 1. Define the statistic

Statistic: Numerical summary measure used to describe a sample characteristic.

Note: A statistic estimates a parameter

The statistic is the sample mean and it is typically denoted by  or .

* + - 1. Random sample

How do we take the sample?

Random sample: Select n items from a population where each has an equal chance of being chosen.

There are other ways to take the sample, but each of them is interested in being representative of the population.



In the above diagram, I could not fit the thousands of GPAs in the population and all 20 GPAs in the sample!

* + - 1. Inference

Inference: A deduction or conclusion about the population based on the sample

Based upon the statistic in the sample, we will make inferences about the parameter in the population with a certain level of accuracy. This level of accuracy can be made through the use of probability. We will start to discuss this in more detail in a few weeks. For now, we can say that the sample mean GPA estimates the population mean GPA.



Questions:

* The sample mean GPA *estimates* the population mean GPA. Is the sample mean GPA equal to the population mean GPA?
* How accurate is the sample mean GPA in estimating the population mean GPA?

The statistical science allows us to measure the accuracy. We will learn how to later in the semester.

* Suppose another random sample is taken. Is the sample mean GPA going to be the same?
* What would happen if a random sample is not taken? Suppose only students from a particular part of the university are sampled.

Where is statistics used?

* Pharmaceutical industry

Clinical trials – Does a drug cure a person of disease?

The FDA mandates that all new drugs go through testing to see if a drug works. New drugs are tried out on humans through conducting many clinical trials (or tests). Measurements are taken on humans included in a sample to make inferences to the entire human population.

Shelf-life or expiration date determination – What is the expiration date on a bottle of medicine?

The potency of a drug may be lost over time. Statistical models can be created to estimate the potency lost. A sample of the drug (suppose in tablet form) can be put into storage and measurements can be taken upon each tablet in the sample. Inferences about potency are then made to the entire population of drug tablets ever made or to be made.

* Agriculture

What corn variety produces the highest yield? A variety of corn varieties can be planted in test farms (sample) and yield can be observed for a number of years. Inferences about the best corn varieties can be made to the population of all farms.

* Marketing

Who should a mail advertisement be sent to? Generally, one would want to send it to people who have the largest chance of buying a product. How do you determine this?

Which Hallmark cards should be marketed to the public?

* Forecasting

Student enrollment at a university – How many students are going to enroll next year? Why is this important to know?

Models can be created based upon past enrollments to predict future enrollments. The sample could include enrollments for each semester over the past 10 years. The population would include these past enrollments and future enrollments.

Profit for a company – How much profit should a company expect for the next year?

Interesting quotes about statistics:

* Jon Kettenring, former American Statistical Association president, says:

I like to think of statistics as the science of learning from data. It presents exciting opportunities for those who work as professional statisticians. Statistics is essential for the proper running of government, central to decision making in industry, and a core component of modern educational curricula at all levels.

* Florence Nightingale says

The most important science in the whole world: for upon it depends the practical application of every other science and of every art: the one science essential to all political and social administration, all education, all organization based on experience, for it only gives results of our experience.

If you are interested in knowing more about the statistics profession, go to the American Statistical Association’s website at [www.amstat.org](http://www.amstat.org). The link, <https://www.amstat.org/ASA/Your-Career/home.aspx>, contains specific information about being a statistician.

Why are you here?

Hopefully, it is not only because your advisor made you take the course!

* Work with statisticians
* Perform some statistical calculations on your own
* Take additional courses that have some statistical content

The purpose of this course is to give you a background in the basics of statistics so that you can do one or more of these items above.

What are we going to do in our course?

Below is the diagram used earlier in the GPA example:



Sections

1. Introduction - Current material; R

2.Summarize Data – Initial look at data, later relate it to more sophisticated analysis methods

3.-6. Probability – Discuss how we formally define it in mathematical ways; this will be useful then for inference

7.-End – Inferences that put all parts of the diagram together for a number of scenarios: One mean, two means, proportion, other parameters, relationships among variables