Introduction to Statistics

1.1 Introduction

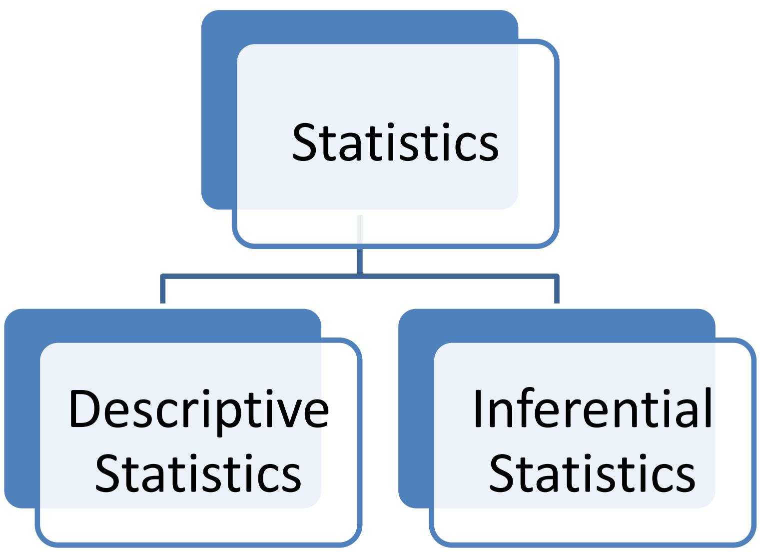
Statistics is a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting and drawing conclusions based on the data. It is the science of data.

By the end of this course, statistical thinking should become part of your thought process.

Statistical thinking involves applying rational thought and the science of statistics to critically assess information (data) and inferences.

In this course we will divide our study of statistics into two categories:

Descriptive statistics, where we will organize and summarize data and **inferential statistics,** where we use data to make predictions and decisions about a population based on information from a sample.



- **Descriptive statistics** utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form.

- **Inferential statistics** utilizes sample data to make estimates, decisions, predictions or other generalizations about a larger set of data.

Above, we mentioned the word population. Let's define two important terms used in this course:

Population vs. Sample

The **population** is the set of *all* measurements of interest to the investigator. Typically, there are too many experimental units in a population to consider *every* one. However, if we can examine every single one, we conduct what is called a **census**.

A **sample** is a subset of measurements selected from the population of interest.

Example 1: In 1936, Literary Digest magazine attempted to predict the winner of the contest for President of the United States between Alf Landon (R) and Franklin D. Roosevelt (D). They attempted to poll ten million people by sending out ballots to magazine subscribers, automobile owners, and telephone users. Over two million ballots were returned, which is a large number of respondents. The polls predicted that Alf Landon would capture 57% of the vote, which would have made him the winner by a landslide.

This poll result carried a lot of weight since Literary Digest magazine had used similar methods to accurately predict the winner in the 1916, 1920, 1924, 1928, and 1932 elections, so it was a surprise when Alf Landon only received 37% of the 44,431,180 votes cast in the 1936 election.

In this example, what was the target population and what was the sample? (Answer: Population = the collection of voting preferences of the entire set of eligible adults in the USA in 1936. Sample = the over two million poll respondents. The poll had a major selection bias—i.e., it was not a random selection of the population, not everyone had a car or a phone or a magazine membership in 1936, so not everyone member of the population had an equal chance of being selected.)

1.2 Parameters vs. Statistics

A **parameter** is a numerical measurement describing some characteristic of a *population* and computed from all of the population measurements. For example, a population average (mean), the average obtained from every item in the population, is a parameter.

A **statistic** is a numerical measurement describing some characteristic of a *sample* drawn from the population.

The example below will illustrate these ideas. One way to remember where parameters and statistics come from is to notice that the letter **P is the first letter of population and parameter** and **S is the first letter of sample and statistic.**

Example 2: In a study of household incomes in a small town of 1000 households, one might conceivably obtain the income of every household. However, it is probably very expensive and time consuming to do this. Therefore, a better approach might be to obtain the data from a portion of the households (let's say 125 households). In this scenario, the incomes of the 1000 households are referred to as the population and the incomes of the randomly selected 125 households are referred to as a sample.

In this household incomes example, is the average (mean) income of all 1000 households a parameter or statistic? Is the average (mean) income of the 125 households a parameter or a statistic? (Answer: The average (or mean) of the 1000 households is a parameter, whereas the average (mean) income of the 125 households is a statistic.)

Example 3 (parameter vs. statistic) - Determine whether the given value is a statistic or a parameter.

In July 2011, 89.6% of all Florida International University Law School graduates passed the bar exam, which incidentally was the best passage rate in Florida. Is this passage rate an example of a population parameter or a sample statistic? (Answer: Parameter)

A sample of law students attending the evening program is selected from FIU and their average age in years is 28.7. Is this average age an example of a population parameter or a sample statistic? (Answer: Statistic)

Example 4 Fill in the blank with the proper term (population vs. sample).

The ________ is the set of *all* measurements of interest to the investigator. (Answer: **population**)

A ______ is a subset of measurements selected from the population of interest. (Answer: **sample**)

Here are some more terms often used in this class:

A **variable** is a characteristic that changes or varies over time or varies across different individual subjects.

An **experimental unit** is the individual or object on which a variable is measured, or about which we collect data.

- Person
- Place
- Thing
- Event

A **measure of reliability** is a statement about the degree of uncertainty of a statistical inference.

For example, the plus and minus 5% in the below statement indicates the real number of people who prefer Pepsi might be as low as 51% or as high as 61%.

> Based on our analysis, we think 56% of soda drinkers prefer Pepsi to Coke, ± 5%.

Problems in statistics vary widely, but there is a broad outline that most of them fit into. That outline is listed below:

Five Elements of a Statistical Problem:

1. Specify the question to be answered and the population of interest
2. Design the experiment or sampling procedure to be used
3. Analyze the sample data
4. Make an inference
5. Measure the goodness (reliability) of the inference

1.3 Classifying Types of Data

Since Statistics is the science of data, we should attempt to classify the kinds of data we will face in the world around us. Very broadly, we can say there are numerical data and non-numerical data. Either the measurement you are taking results in a number or it doesn't (For example, I can put you on a scale and weigh you, or perhaps, I could ask where you were born. One is a number, and one is a characteristic that is not a number). Let's call the numerical data **Quantitative** (think quantity, it will help you remember these data are numbers), and let's call the other type **Qualitative** data because it captures qualities or characteristics that aren't normally numerical in nature.

- **Quantitative Data** are measurements that are recorded on a naturally occurring numerical scale.
  - Age
  - GPA
  - Salary
  - Cost of books this semester

- **Qualitative Data** are measurements that cannot be recorded on a natural numerical scale, but are recorded in categories.
  - Year in school
  - Live on/off campus
  - Major
  - Gender

Now that we have split data into two broad categories (numerical and other), let's further split the Quantitative (numerical) category into two groups:

Continuous numerical data result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

Example: The finishing times of a marathon

Discrete numerical data result when the number of possible values is either a finite number or a countable number.

(That is, the number of possible values is 0 or 1 or 2 and so on.)

Example: The numbers of fatal automobile accidents last month in the 10 largest US cities

To help you recognize continuous data and discrete data, try to think of how the data was obtained. Did it come from a **count** or from a **measurement**? Count data is discrete in nature, and measurement data is continuous. If you use a ruler, a speedometer, a scale, a stop watch to take a measurement, the data is probably continuous.

Since Statistics is a branch of mathematics, it will not surprise you that we think of the results of our measurements as variables (since the measurements vary from subject to subject or over time in a given subject). This leads to the following corresponding definitions regarding the types of variables we will deal with in this class:

Quantitative variables are numerical observations.

We divide the above variable type into two separate types again:

Continuous Variables can assume all of the infinitely many values corresponding to a line interval.

Example: Y = The amount of milk that a cow produces; e.g. 2.343115 gallons per day

Discrete Variables can assume only a countable number of values. **(i.e. the number of possible values is 0, 1, 2, 3, . . .)**

Example: X = The number of eggs that a hen lays

Then we have the non-numerical variables:

Qualitative variables are non-numerical or categorical observations.

Example 5 (discrete/continuous) - Determine whether the given value is from a discrete or continuous data set.

1. A research poll of 1015 people shows that 752 of them have internet access at work. (answer: Discrete—it is not possible to have a fraction of a person having internet access)

2. Josh Becket's fastball was clocked at 98 mph during the World Series. (answer: Continuous--You can have any fraction of a mile per hour as the speed)

3. A student spent $86.53 on her calculator for class. (answer: Discrete—You cannot pay any decimal amount for the calculator. For example, we can't pay $86.532. Since this is not possible it is technically discrete. However, currency is typically treated as continuous data since it is virtually continuous (allows us to use values out to the hundredths place.)

Levels of Measurement

Nominal Level of Measurement

Observations of a qualitative variable that can only be named and counted are conducted at the nominal level of measurement. The names or labels given to these observations function only as a means of classification. These names or labels do not have an implied order. As a result, the nominal level of measurement involves counting the number of occurrences only. For example, a university could track the home state of its students. The states have no inherent order, so it does not matter which state is listed first in a report. The university might also recorded the race of each student; this is another example of data collected at the nominal level of measurement.

The Ordinal Level of Measurement

Observations of a qualitative variable are at the ordinal level of measurement if they can be arranged in some order. For example, consider the following tabulated results from a faculty evaluation:

Excellent	25
Above Average	30
Average	25
Below Average	2
Poor	1

There is a natural order to these classifications. A rating of "Excellent" is better than one of "Above Average." A rating of "Above Average" is better than "Average," so on and so forth. We can certainly order data, but it is not possible to determine how much of a difference actually exists between any two classifications. For example, we cannot say a rating of "Average" is twice as good as a rating of "Poor." There is no way to subtract two ratings to find a meaningful difference between them. While it is true that "Average" teaching is better than "Below Average" teaching, we do not have a meaningful way to say exactly how much better.

The Interval Level of Measurement

Like the ordinal level of measurement, data measured at the interval level can be ranked or ordered, but the interval level of measurement has an additional feature. Differences between values are meaningful. The classic example of data at the interval level of measurement is temperature. Temperatures can certainly be placed in an order, for example, from coolest to warmest, but we can go further than simply ordering temperatures. We can subtract two temperatures and find a meaningful difference. It is correct to say that 75 degrees Fahrenheit is ten degrees warmer than 65 degrees Fahrenheit. However, ratios are not valid when working with interval level measurements. For example, 100 degrees Fahrenheit is not twice as warm as 50 degrees Fahrenheit. The ratio 100 to 50 is not meaningful because data measured at the interval level does not have a true zero starting point. Zero degrees Fahrenheit does not represent the complete absence of heat. It is because of this lack of a true zero that ratios at the interval level of measurement are meaningless.

The Ratio Level of Measurement

The ratio level of measurement combines the properties of the interval level of measurement (meaningful differences) with the additional feature that there is a true zero starting point. This makes meaningful ratios possible. The data at the ratio level of measurement is quantitative and may be measured on a discrete or continuous scale. For example, distance, weight, volume, or length are all examples of the ratio scale of measurement involving data on a continuous scale. Number of customers, number of workers, and number of computers are examples of the ratio scale of measurement using data measured on a discrete scale.

Profit is an example of the ratio level of measurement. If a business has zero profit for a quarter, it means they have no profit. If the same business profits $80,000 in one quarter and $40,000 in the quarter after that, not only is it valid to say over the second quarter the business made $40,000 less profit, but it is also valid to state the first quarter was twice as profitable as the second.

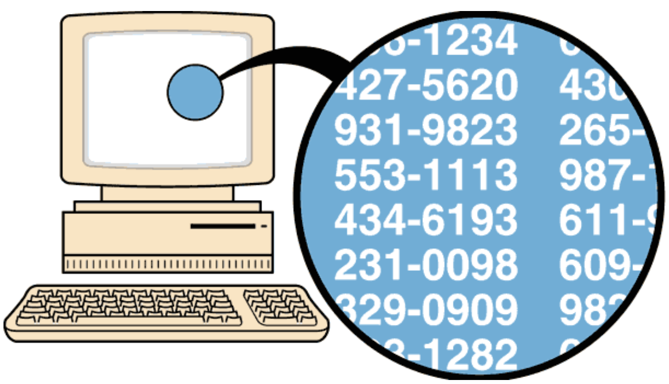
There are just two short definitions left to learn in this section. When we try to learn something about a particular group, we need to draw a sample that is similar to the overall population. This idea is conveyed in the next definition.

A **representative sample** exhibits characteristics typical of the target population.

In order to ensure that we get a good sample that is representative, we often employ a random sampling approach.

A **random sample** is selected in such a way that every different sample of size n has an equal chance of selection.

For example, if everyone at FIU has a Panther ID, we may draw a random selection of 36 ID numbers from the Panthersoft system. This selection of 36 Panther ID's would constitute a random sample from the larger FIU population.



In a book published in 1987, Shere Hite revealed the results of a poll she conducted of women. The poll was sent to 100,000 women who were subscribers to a selection of magazines, but only 4,500 of the women polled responded. The results were shocking. Over 90% of the women reported marital dissatisfaction, and more than half reported having extramarital affairs. These results are often used as an example of self-selection bias. Often people who volunteer to complete a poll or survey are very different from the general population being targeted. This becomes particularly problematic when only a very small percent of those surveyed choose to respond. In this case, it might be that the women who responded were a vocal few who were unhappy with their marriage and wanted to voice their dissatisfaction.