

## **Lesson 1**

### **Introduction**

#### Outline

Statistics

Descriptive versus inferential statistics

Population versus Sample

Statistic versus Parameter

Simple Notation

Summation Notation

#### **Statistics**

What are statistics? What do you think of when you think of statistics? Can you think of some examples where you have seen statistics used? You might think about where in the real world you see statistics being used, or think about how statistics are used in your major.

Statistics are divided into two main areas: descriptive and inferential statistics.

Descriptive statistics- These are numbers that are used to consolidate a large amount of information. Any average, for example, is a descriptive statistic. So, batting averages, average daily rainfall, or average daily temperature are good examples of descriptive statistics.

Inferential statistics- inferential statistics are used when we want to draw conclusions. For example when we want to determine if some treatment is better than another, or if there are differences in how two groups perform. A good book definition is using samples to draw inferences about populations. More on this once we define samples and populations.

Population- Any set of people or objects with something in common.

Anything could be a population. We could have a population of college students. We might be interested in the population of the elderly. Other examples include: single parent families, people with depression, or burn victims. For anything we might be interested in studying we could define a population.

Very often we would like to test something about a population. For example, we might want to test whether a new drug might be effective for a specific group. It is impossible most of the time to give everyone a new treatment to determine if it worked or not. Instead we commonly give it to a group of people from the population to see if it is effective. This subset of the population is called a sample.

When we measure something in a population it is called a parameter. When we measure something in a sample it is called a statistic. For example, if I got the average age of parents in single-family homes, the measure would be called a parameter. If I measured

the age of a sample of these same individuals it would be called a statistic. Thus, a population is to a parameter as a sample is to a statistic.

This distinction between samples and population is important because this course is about inferential statistics. With inferential statistics we want to draw inferences about populations from samples. Thus, this course is mainly concerned with the rules or logic of how a relatively small sample from a large population could be tested, and the results of those tests can be inferred to be true for everyone in the population. For example, if we want to test whether Bayer aspirin is better than Tylenol at relieving pain, we could not give these drugs to everyone in the population. It's not practical since the general population is so large. Instead we might give it to a couple of hundred people and see which one works better with them. With inferential statistics we can infer that what was true for a few hundred people is also true for a very large population of hundreds of thousands of people.

When we write symbols about populations and samples they differ too. With populations we will use Greek letters to symbolize parameters. When we symbolize a measure from a sample (a statistic) we will use the letters you are familiar with (Roman letters).

Thus, if I measure the average age of a population I'd indicate the value with the Greek letter "mu" ( $\mu = 24$ ). While if I were to measure the same value for a subset of the population or a sample then I would indicate the value with a roman letter ( $\bar{X} = 24$ ).

### Simple Notation

You might think about descriptive statistics as the vocabulary of the "language" of statistics. If this is true then summation notation can be thought of as the alphabet of that language. Notation and summation notation is just a short hand way of representing information we have collected and mathematical operation we want to perform. For example, if I collect data on a variable, say the amount of time (in minutes) several people spent waiting at a bus stop, I can represent that group of numbers with the variable X. The variable X represents all of the data that I collected.

Amount of Time

$\underline{X}$

5.0

11.1

8.9

3.5

12.3

15.6

With subscripts I can also represent an individual data point within the variable set we have labeled X. For example the third data point, 8.9, is the  $X_3$  data point. The fifth data point  $X_5$  is the number 12.3. Very often when we want to represent ALL of the data

points in a variable set we will use X by itself, but we may also add the subscript  $i$ . Whenever you the subscript  $i$ , you can assume that we are referring to all the numbers for the variable X. Thus,  $X_i$  is all of the numbers in the data set or: 5,11.1,8.9,3.5,12.3,15.6.

There are other common symbols we will use besides X. Sometimes we will have two data sets to deal with and refer to one distribution as X and the other distribution as Y. It is also necessary for many formulas to know how many data points are in a data set. The symbol for the number of data points in a set is N. For the data set above the number of data points or  $N = 6$ . In addition, we will use the average or mean value a good deal. We will indicate the mean, as noted above, differently for the population ( $\mu$ ) than for the sample ( $\bar{X}$ ).

### Summation Notation

Another common symbol we will use is the summation sign ( $\Sigma$ ). This symbol does not represent anything about our data itself, but instead is an operation we must perform. Whenever you see this symbol it means to add up whatever appears to the right of the sign. Thus,

$\Sigma X$  or  $\Sigma X_i$  tells us to add up all of the data points in our data set. For our example above it would be:  $5 + 11.1 + 8.9 + 3.5 + 12.3 + 15.6 = 56.4$ .

You will see the summation sign with other mathematical operations as well. For example  $\Sigma X^2$  tells us to add all the squared X values. Thus, for our example:

$$\Sigma X^2 = 5^2 + 11.1^2 + 8.9^2 + 3.5^2 + 12.3^2 + 15.6^2 \quad \text{-or-}$$

$$25 + 123.21 + 79.21 + 12.25 + 151.29 + 243.36 = 634.32.$$

A few more examples of summation notation are in order since the summation sign will be central to the formulas we write. The following examples should give you a better idea about how the summation sign is used. Be sure you recall the order of operations needed to solve mathematical expressions. You will find a review on the web page or you can click here: <http://faculty.uncfsu.edu/dwallace/sorder.html>

For the examples below we will use a new distribution.  $X = 1 \ 2 \ 3 \ 4$   
 $Y = 5 \ 6 \ 7 \ 8$

$$\Sigma X^2 \neq (\Sigma X)^2$$

For this expression we are saying that the sum of the squared X's is not equal to the sum of the X's squared. Notice here we want to perform the operation in parentheses first, and then the exponents, and then the addition. Thus:

$$\begin{aligned}\Sigma X^2 &\neq (\Sigma X)^2 \\ 1^2 + 2^2 + 3^2 + 4^2 &\neq (1+2+3+4)^2 \\ 1 + 4 + 9 + 16 &\neq (10)^2 \\ 30 &\neq 100\end{aligned}$$

For the next expression we show, like in algebra, that the law of distribution applies to the summation sign as well. Again, what is important is to get a feel for how the summation sign works in equations.

$$\begin{aligned}\Sigma(X + Y) &= \Sigma X + \Sigma Y \\ (1+5)+(2+6)+(3+7)+(4+8) &= (1+2+3+4)+(5+6+7+8) \\ 6 + 8 + 10 + 12 &= 10 + 26 \\ 36 &= 36\end{aligned}$$