**Collecting Data Notes**

*Statistics is a systematic and objective process of Collecting, describing and analyzing information to aid in drawing conclusions or making decisions.*

There are three parts of this definition to be aware of.

1st – statistics is a systematic and objective process.

2nd – statistics is about collecting, describing, and analyzing information.

3rd – statistics helps us draw conclusions or make decisions based on collected data.

**Data/Variables**

Typically, data are in a *data set.* An Excel spreadsheet is a good example of a data set. You can see here that a data set has information arranged in rows and columns.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Race | Gender | Age | Major | Q1 | Q 2 | Q3 | Q4 | Q5 | Q6 | Q7 |
| 01 | B | F | 20 | BUS | 2 | 4 | 1 | 4 | 3 | 5 | 2 |
| 02 | H | F | 21 | EDU | 1 | 4 | 5 | 1 | 3 | 2 | 5 |
| 03 | W | M | 18 | PHY | 3 | 3 | 4 | 5 | 2 | 1 | 3 |
| 04 | H | F | 18 | ANT | 2 | 4 | 1 | 4 | 3 | 4 | 2 |
| 05 | B | F | 19 | PSY | 1 | 2 | 3 | 1 | 1 | 2 | 4 |
| 06 | H | M | 18 | SOC | 4 | 5 | 5 | 2 | 4 | 1 | 3 |
| 07 | W | M | 19 | EDU | 3 | 1 | 4 | 3 | 1 | 5 | 2 |
| 08 | A | F | 20 | BUS | 3 | 3 | 4 | 5 | 4 | 1 | 1 |
| 09 | W | M | 18 | MTH | 1 | 4 | 5 | 1 | 3 | 2 | 5 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

In a data set, rows and columns have specific functions. *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* represent the objects we collect information about. In a data set, *rows are referred to as cases or units*. In this data set, it looks like we collected information about people, because we collected things such as race, gender, age, and major.

As it turns out, race, gender, age, and major – the information in columns – has a name also. They are referred to as *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*. *A variable is any characteristic that is collected/recorded for each case*.

Notice that cases are in rows and variables are in columns.

For our purposes, *there are two kinds of variables – \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_and \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*.

*Categorical* variables help us group cases and *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* variables help us measure some aspect of our cases. It is important to remember that cases can belong to one and only one group in a categorical variable. So, we must make sure that our groups are mutually exclusive.

*Continuous* variables have numbers and mathematical operations make sense. That is, let’s say I collected the distance run by each runner in a race. Let’s say I called the variable “Distance”. If I added together the value of “Distance” for two runners, I would have the total distance run by the two runners.

Sometimes, however, it isn’t so obvious. Let’s say I have a variable called “Section” and in that variable I recorded which of my 3 sections of Elementary Statistics each student is enrolled. Let’s say one student is enrolled in “1” and another student is enrolled in “2”. If I add “1” and “2”, does that mean the two students together are enrolled in section “3”? No; not at all. So, in this case, mathematical operations do NOT make sense. Although the variables have numbers, they are NOT quantitative. They are categorical. In *categorical variables,* numbers represent *Categories.*

Whether a variable is categorical or quantitative, they can be used for different purposes. For our discussions, they can be used for two purposes – as *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* variables or as *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* variables. ALL variables are either, but NO variable is both. It’s an either/or situation.

An *explanatory* variable is used to help understand – or explain – another variable. A *response* variable, on the other hand, is the variable that we are trying to understand. Another way to think about it is that the *response* variable is explained or predicted by the *explanatory* variable. It may be easier to think about it with an example.

Suppose I was interested in differences in scores on an exam. And, suppose I was wondering if exam scores were different by academic level – such as freshman, sophomore, junior, or senior. In this case, the exam score would be my response variable, because that’s the variable I’m trying to understand, and classification level would be my explanatory variable, because I’m wondering if classification level explains or predicts the grade on an exam.

**Data Collection**

Garbage *in* = Garbage *out*

Data collection is the foundation of statistics. And, the only way to conduct sound statistical analysis is grounded in proper data collection.

If we are going to collect data, the first consideration is in whom we are interested? The answer to this question is different depending on the question we’re attempting to answer.

As an example, suppose we want to increase attendance of adults living in the Jackson metro area at our home football games. So, all adults in the Jackson metro area would be the individuals we are interested in. This group is referred to as a *population.*

In a perfect world, we’d like to ask every adult in the Jackson metro area why they don’t attend Millsaps home games, but that probably isn’t feasible. So, instead, we ask a smaller group of adults from the Jackson metro area. This subset of our population is referred to as a *sample.* Instead of trying to reach all adults in the metro area, we’ll pose our question to this sample.

In statistics, we denote the size of a population with a capital *N* and the size of the sample with a lower case *n.*

Once we have collected data from our sample, we want to use that information to understand the population. This process of using data from a sample to understand a population is referred to as *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* statistics.

In our sample, we’re hoping to use the information we gain about why our sample doesn’t attend Millsaps home games to help us understand why the general metro area doesn’t attend.

But, we can run into a lot of problems if we aren’t careful. One potential problem is if our sample data inaccurately reflects the population. Anything that causes sample data to inaccurately reflect the population is known as sampling bias. *Sampling bias* occurs when the method of selecting a sample causes the sample to differ from a population in some relevant way.

Suppose I stand outside the grocery store in my neighborhood and ask people why they don’t attend Millsaps games. Might those people inaccurately reflect the population of the metro area? Probably not. If there is sampling bias, we cannot generalize what we gather from our sample to the population. But, what exactly is generalization? Formally, *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* is using statistical inferences from a sample to make conclusions about a sample.

There are other ways bias can occur (other than sampling). People may choose to not participate. Typically, participation is voluntary and if people choose not to participate, you are missing their input. If I stand outside a grocery store and ask people about home games, some people may walk past without responding or go in a different door to avoid me altogether.

Another source of bias is the way questions are worded. Suppose I asked people to answer the following question with true or false: Using a toboggan in winter often results in injury.

Without further explanation, that question can lead to useless results.

It all depends on how you define ‘toboggan.’ Some people define toboggan as a sled, whereas other people define toboggan as a cap. Depending on your definition of toboggan, you may respond to that question very differently.

Another example is “grits”. If you aren’t from the south, you may not even know what grits are.

Fortunately, we have a way to address sampling bias. We address it through a process known as simple random *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*. With simple random sampling, everyone in the population has the same chance of being selected to be a part of the sample. Consequently, *collecting random samples avoids sampling bias*.

**Research Design**

Let’s start by looking at a real example from the 20th century. During the 1st half of the century, polio was an epidemic throughout the industrialized world. A tremendous amount of research was dedicated to finding the cause of polio in hopes of finding a vaccine and a cure.

Researches noticed that polio cases would spike during the summer with a smaller spike around Christmas. As a result, they linked polio to eating ice cream, which led them to believe that increased levels of sugar consumption led to polio.

In 1940, there was a journal article in the American Journal of Pathology linking polio to sugar consumption. In reality, however, polio has nothing to do with sugar consumption. Instead, the increase in polio was indirectly brought about by the invention of flushable toilets. Prior to flushable toilets, children were exposed to low levels of infections that helped build immunities, and with flushable toilets, they were no longer subjected to those immunizing infections, which led to a higher susceptibility to polio.

The major lesson here is that two things can be related without one causing the other and we need to be careful in our research to realize the difference.

When the values of one variable tend to be related to the values of another variable, this is known as association, or *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*.

When changing the value of one variable influences the value of another variable, this is referred to as *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*.

Sugar consumption and polio cases were related, but increasing sugar consumption would have no effect on the number of polio cases.

Flushable toilets and polio cases were also related. But, increasing the number of flushable toilets had an effect on the number of polio cases. So, sugar consumption and polio were only associated, whereas flushable toilets and polio had a causal relationship – having a flushable toilet was the explanatory variable and contracting polio was the response variable. Flushable toilets helped explain the increase in polio cases.

The situation can be further complicated when a 3rd variable is associated with the explanatory and response variables. While researching polio, researchers also noticed that family income as associated with polio. They noticed that polio was more prevalent in families with higher income. Family income was also associated with the amount of ice cream eaten and, in the end, family income was associated with have a flushable toilet – the wealthier the family, the more likely they were to have a flushable toilet.

In this case, family income is what we call a *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*variable – a third variable associated with both the explanatory and response variables.

Because ice cream consumption was associated with having a flushable toilet (wealthier families who could afford a flushable toilet could also afford ice cream) and with the likelihood of contracting polio, ice cream consumption was a confounding variable. And, ultimately, it was the confounding variable that caused all the problems in deciphering polio.

Luckily, we have methods to help us determine whether variables are just associated or have a causal relationship. How we approach our explanatory variable through research design helps us establish whether there is a causal relationship.

There are two general ways to design research. We can conduct an *observational study* or we can conduct an *experiment*. The difference between the two rests in how much control we, as the researcher, exert on the process.

In an *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*study, the researcher doesn’t actively control the value of any variable. Instead, the researcher just observes or looks on. An observational study is often referred to as a correlational study.

In an *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*, on the other hand, the researcher actively controls some aspect of the study. In an experiment, we have two groups – a control group and an experimental group. (More on them shortly).

The major difference in the outcomes of the two research designs is that *observational studies \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_lead to conclusions about causality*. They may show an association but we CANNOT conclude that the association is causal. *Conclusions about causality can ONLY come from \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ experiments*.

In a *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*, participants are randomly assigned to the control group or experimental group. Now, let’s talk about control groups and experimental groups.

In a *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* group, the explanatory variable is not manipulated (that is, there is no experimental treatment). On the other hand, in an *\_\_\_\_\_\_\_\_\_\_\_\_\_\_* group, the explanatory variable is manipulated (there is experimental treatment).

A common example is a trial for a new drug. Participants in the study are randomly divided between a control group and an experimental group. The experimental group is given the drug being studied, whereas the control group is not given the drug. Instead, they are given a fake drug, referred to as a placebo. A *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* is just a fake treatment. The researchers can then study the responses of people in the experimental group, who actually took the drug, with those in the control group, who took a placebo.

Placebos can cause their own problem, however. Sometimes, administering a placebo to the control group can lead to a phenomenon where members of the control group experience an effect even though they didn’t receive any treatment. This is referred to as the *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*. We’ll look at how to address the placebo effect shortly.

Depending on what we’re trying to find out, we have options for designing randomized experiments. Among them are the randomized comparative experiment and the matched pairs experiment. The situation and our goals as a researcher dictate the one we use.

In a *\_\_\_\_\_\_\_\_\_\_\_\_\_\_* experiment, the researcher randomly assigns cases to experimental /control group and compares results on the response variable.

A *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* experiment works a little differently. In a matched pairs experiment, each case gets both treatments in a random order and differences between the response variable are compared.

Example – suppose you have been tasked with recommending whether nitrile or latex gloves should be used in the chemistry lab. To determine your recommendation, you decide to conduct an experiment with 30 nitrile gloves and 30 latex gloves.

Your first step is to draw a random sample of 30 students (you have 60 gloves but each person has 2 hands – so 30 people) enrolled in chemistry lab. Keep in mind that we want random selection so that we can generalize to ALL chemistry students.

You could do a randomized comparative experiment. You can randomly assign 15 people to wear latex gloves and 15 people to wear nitrile gloves.

But this introduces a confounding variable. Each student has a dominant hand. What if left handed people ended up wearing most of the nitrile gloves and right handed people ended up wearing most of the latex gloves? Also, some students may just be naturally harder on gloves than other students and may naturally go through gloves faster. In this case, the student is the confounding variable.

To help address this confounding variable, we can make sure that EACH student gets both a latex glove and a nitrile glove. Students can be randomly assigned to wear the nitrile on their right hand and latex on their left hand. Now we can compare the length of time the gloves last for each student. This is a matched pairs experiment.

Now let’s go back and talk about how we can address the placebo effect. We can do it through a process called *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*. There are two versions of blinding that you’ve probably heard of.

In a *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* experiment, participants are not aware of which group they have been assigned to.

Sometimes we have a concern about allowing the researcher to know which groups is the control group and which is the experimental group. To address this concern, we can conduct a double-blind study, where neither the participants nor the researchers know which groups the participants are assigned to. Instead, some 3rd party controls membership in each group.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* experiments are especially useful in high-stakes studies such as drug trials. Pharmaceutical companies have an extreme interest in the success of drugs. To reduce the likelihood of inappropriate influence, double-blind studies are prevalent in drug trials.

Regardless of whether an experiment is randomized comparative or matched pairs, it can be single-blind or double-blind. Or, there may be no blinding at all.

**Randomization**

Randomization can occur at two levels – *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* and *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*.

We can have random selection AND random assignment.

We can have random selection OR random assignment.

We can NOT have random selection nor random assignment.

So, we can have both, either, or neither.

As you can see, random selection and random assignment each serves a *specific but different* purpose.

Personally, it helps me to visualize it in this chart.

So, if we have random selection, we can *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* to our population. Otherwise we CANNOT generalize.

Likewise, if we have random assignment, we can infer *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*. Otherwise, we CANNOT infer causality.

If we have both random selection and random assignment, we can generalize our causal inferences to the population of interest.