MATH 1150 Chapter 1 Terminology

Consider the following **dataset** concerning students in a college introductory biology course, who recently completed an exam:

Student	Year	Hours Studied	Test Score
Alice	Freshman	0	37
Bob	Junior	3.5	64
Eve	Sophomore	5	79
Li	Freshman	6	86
Jason	Senior	2	78
Eleanor	Senior	9	94
Ron	Junior	1	97
Tariq	Sophomore	4	83

The rows of the table, in this case the students taking the exam, are called the **cases** of the dataset, while the columns of the table, the pieces of information being collected, in this case the year of the students in school, the number of hours they studied for the exam, and their score on the exam, are called the **variables** of the dataset.

If a variable assigns a numerical value to each case, then we refer to it as a **quantitative variable**. In this case, two of the three variables, specifically the number of hours studied and the exam score, are quantitative variables. If, alternatively, a variables divides the cases into groups without assigning a numerical value, then we refer to it as a **categorical variable**. In this case, the year in school is a categorical variable.

When collecting data, it is often the case that we are attempting to investigate the relationship between two or more of the variables, in an effort to answer a question or come to a conclusion. More specifically, if we are trying to use one variable to explain or predict another variables, we refer to the former as the **explanatory variable** and the latter as the **response variable**. For example, in this case we might be asking "Does more studying lead to higher test scores?", in which case the explanatory variable is the hours studied and the response variable is the test score. Carefully note that the role a variable is playing depends on the question we are asking. Alternatively, if we were asking "Do underclassmen study more than upperclassmen?", then the explanatory variable is the year in school and the response variable is the hours studied.

As discussed in class, we need to be careful about how we use data to come to conclusions. If the question we are asking is truly as broad as "Does more studying lead to higher test scores?", that would imply that we are interested in forming a conclusion about ALL students on ALL tests. The collection of all individuals or objects of interest for a question is called the **population**, while the subset of the population from which data is actually collected is called the **sample**. In this case, if we don't refine our question, then the population is quite vague (perhaps every person who has ever taken a test ever?), while our sample is simply these eight specific students in this one introductory biology class.

The process, and associated shortcomings, of using data from a sample to form a conclusion about a population is called **statistical interference**. In order to form better conclusions, we want to minimize this interference. In particular, we would like to avoid **sampling bias**, which occurs when the sample differs from the population as a whole in a relevant way. In this case, the fact that all of the students attend the same school, the fact that we are restricting to one specific subject, etc., are all potential sources of sampling bias. Perhaps a more realistic goal would be to address the question "Does more studying lead to higher test scores in Dr. Guerrier's Millsaps College introductory biology class?"

Even with this more refined question, the risk of bias still exists. How were these eight students chosen for data collection? Maybe they volunteered, maybe the professor picked his favorites, etc., any of which could skew the data. One way to avoid sampling bias is to use a **simple random sample**, in which the cases are chosen completely randomly amongst all members of the population. For example, perhaps Dr. Guerrier put every one of his students' names on individual slips of paper, then drew eight of them from a hat, and collected data from those eight students.

When asking questions like this, we are searching for connections between different variables. If the values of one variable tend to be related to the values of another variable, we say that the two variables are **associated**. However, it is crucial to note that just because two variables are associated, it does NOT mean that they are actually influencing each other. If changing the value of one variable (while keeping everything else fixed) DOES influence the value of another variable, we say that the two variables are **causally associated**.

In our example, it may be a reasonable hypothesis that hours studied is causally associated with test score, but let's explore another example. Collected data would likely indicate that, amongst all children under 10, weight is associated to language skills in a very compelling way. Does that mean that if you want your child to improve their speech and vocabulary, you should feed them french fries and ice cream all the time? No way! Rather, both of these variables, weight and language skills, are being influenced by a hidden third variable, the AGE of the child. Age is causally associated with weight, and age is causally associated with language skills, but weight and language skills are not causally associated with one another. In this example, we refer to the age of the child as a **confounding variable**, or **lurking variable**.

Assuming the students weren't told what to do ahead of time, our original example with the biology students is an example of an **observational study**, because the researcher did not actively control the value of any variable, but rather simply observed them as they naturally exist. Alternatively, data can be collected using an **experiment**, in which case the researcher DOES actively control one or more variables. For example, if each biology student were told exactly how many hours to study in order to observe the impact on his or her test score, that would be an experiment. If the number of hours each student was told to study was determined at random, say using a computer or drawing numbers out of a hat, then that would be a **randomized experiment**.