

2.5: Working with Data

What are data?

The first important point about data is that data *are* – meaning that the word “data” is plural (though some people disagree with me on this). You might also wonder how to pronounce “data” – I say “day-tah”, but I know many people who say “dah-tah”, and I have been able to remain friends with them in spite of this. Now, if I heard them say “the data is” then that would be a bigger issue...

Operationalizing Variables

We need to have specifically defined how we are measuring our construct or our variable. The act of defining how to measure your data is to **operationalize**. Some variables are easier to define, like height or weight. I can measure height in inches tall or weight in pounds. Some other variables can be more open to measurement, like happiness or love. How would I measure happiness? Would I simply ask are you happy (yes or no)? Would I use a questionnaire for a self-report measure? Would I rate individuals from observing them for happiness? Would I ask their partner, teacher, parent, best friend about the person’s happiness? Researchers’ decisions on how to measure data is an important factor and helps to determine what kind of data is being used.



How would you measure happiness in a research study? [Image Source](#)

Qualitative and Quantitative Variables

Data are composed of *variables*, where a variable reflects a unique measurement or quantity. An important distinction between variables is between qualitative variables and quantitative variables. **Qualitative variables** are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. Qualitative means that they describe a quality rather than a numeric quantity. *Qualitative variables are sometimes referred to as categorical variables*. For qualitative variables, response options are usually limited or fixed to a set of possible values. Assigning a person, animal or event to a category is done on the basis of some qualitative property. For example, in my stats course, I generally give an introductory survey, both to obtain data to use in class and to learn more about the students. One of the questions that I ask is “What is your favorite food?”, to which some of the answers have been: blueberries, chocolate, tamales, pasta, pizza, and mango. Those data are not intrinsically numerical; we could assign numbers to each one (1=blueberries, 2=chocolate, etc), but we would just be using the numbers as labels *rather than as real numbers*.

Personality type, gender, and shirt sizes are all categorical, or qualitative, variables. The values of a qualitative variable do not necessarily imply order and do not produce numerical responses or use real numbers. For example, there is an order to shirt size but shirt size is categorical and not number based. Another example is postal Zip Code data. Those numbers are represented as integers, but they don’t actually refer to a numeric scale; each zip code basically serves as a label or category representing a different region. Because this data is not using real numbers, what we do with those numbers is constrained; for example, it wouldn’t make sense to compute the average of those numbers.

More commonly in statistics we will work with *quantitative* data, meaning data that are numerical. For example, here Table 1 shows the results from another question that I ask in my introductory class, which is “Why are you taking this class?”

Table 1: Counts of the prevalence of different responses to the question “Why are you taking this class?”

Why are you taking this class?	Number of students
It fulfills a degree plan requirement	105
It fulfills a General Education Breadth Requirement	32
It is not required but I am interested in the topic	11
Other	4

Note that the students' answers were qualitative, but we generated a quantitative summary of them by counting how many students gave each response. **Quantitative variables** are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

Experimental studies can involve qualitative and quantitative data. In the study on the effect of diet discussed previously, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

Discrete and Continuous Variables

Variables such as number of children in a household are called **discrete variables** since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as "time to respond to a question" are **continuous variables** since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

2.5: Working with Data is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.