

# Data Types

*By Tanya Hoskin, a statistician in the Mayo Clinic Department of Health Sciences Research who provides consultations through the Mayo Clinic CTSA BERD Resource.*

Don't let the title scare you. I know it sounds like one of those topics that only statisticians care about – the kind of topic that makes the eyes of most non-statisticians glaze over. In many ways, data types are very intuitive. However, when you need to collect, record or analyze your data, you can only accomplish these tasks successfully by thinking carefully about what type of data you have.

Suppose an investigator wants to determine if a specific lab value or test result is associated with patient outcome. In most cases, my first question to this investigator would be “What does the variable look like?” In other words, what possible values can the variable take and how will the variable be recorded?

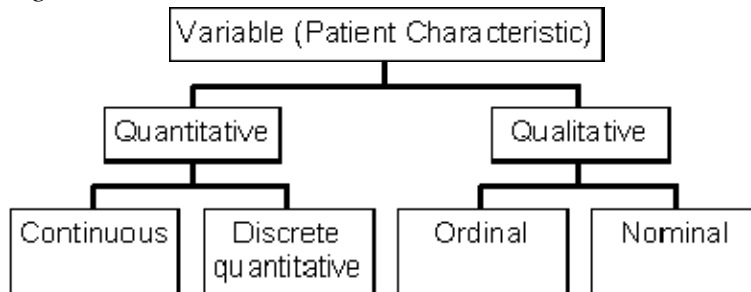
## The basics

Typically, a variable can describe either a quantitative or qualitative characteristic of an individual. Examples of quantitative characteristics are age, BMI, creatinine, and time from birth to death. Examples of qualitative characteristics are gender, race, genotype and vital status. Qualitative variables are also called categorical variables. Unfortunately, it gets a little more complicated.

Quantitative and qualitative data types can each be divided into two main categories, as depicted in Figure 1. This means that there are four basic data types that we might need to analyze:

1. Continuous
2. Discrete quantitative
3. Ordinal
4. Nominal

*Figure 1*



## Quantitative variables

You might think of a quantitative variable as one that can only be recorded using a number. These variables describe some quantity about the individual and are often

measured (e.g., body mass is measured with a scale) or counted (e.g., the number of needle punctures required to obtain the biopsy specimen is counted).

A quantitative variable can be either continuous or discrete. A continuous variable is one that in theory could take any value in an interval. We say “in theory” simply because we are limited by the precision of the measuring instrument (e.g., a patient’s true creatinine value might be 1.21345615 but we might only be able to measure it as 1.213). Examples of continuous variables are body mass, height, blood pressure and cholesterol.

A discrete quantitative variable is one that can only take specific numeric values (rather than any value in an interval), but those numeric values have a clear quantitative interpretation. Examples of discrete quantitative variables are number of needle punctures, number of pregnancies and number of hospitalizations. For these examples, positive whole numbers are the only possible values (i.e., it is not possible to have 1.5 pregnancies).

### **Qualitative variables**

Qualitative or categorical variables describe a quality or attribute of the individual. Categorical data can be either nominal or ordinal. Sex is an example of a nominal variable, and histologic stage is an example of an ordinal variable. What is the difference between these two variables? The values for one of these variables have a specific order; for the other variable, they do not. If one patient has histologic stage 4 and another patient has histologic stage 1, you know that the stage 4 patient has more severe disease. Although the histologic stages are categories, the categories have an inherent order. The same cannot be said for the variable sex. Qualitative data with unordered categories is referred to as nominal; qualitative data with ordered categories is referred to as ordinal.

### **Why do we care about data types during the data collection phase?**

The answer here seems pretty obvious – you must understand the data type of each variable in order to record its values in a consistent manner. This probably won’t require much thought in most cases, but consider the following example.

Suppose you are interested in the variable creatinine but plan to analyze it as a binary variable by classifying patients as creatinine  $< 1.8$  or creatinine  $\geq 1.8$ . You could simply collect which of these categories each individual falls into, but this probably isn’t the best choice. If a categorical variable is based on the value of a continuous variable, it is generally a good idea to collect the continuous variable. A continuous variable provides more information than a binary variable, which usually translates into more statistical power to detect differences among patients. If, in the analysis phase, you decide that you really do want to use the binary version of the variable, you can easily use a formula in a spreadsheet or statistical software package to create the binary variable from the continuous one you collected. On the other hand, if you only collect the binary variable, you do not have the source measurement recorded to go back to if necessary.

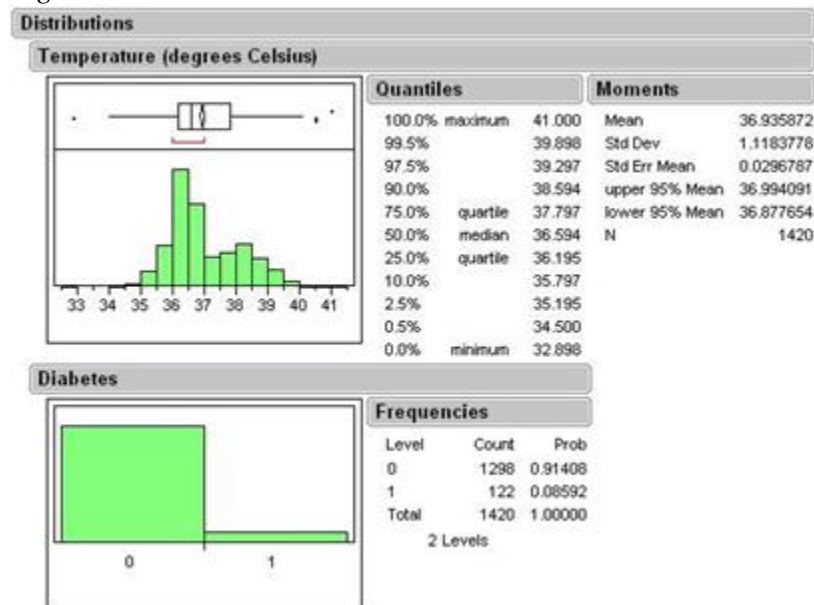
### **Why do we care about data types during the analysis phase?**

You are probably frequently exposed to terms such as mean, median, frequency,

proportion, two-sample t-test, chi-square test, regression, correlation, logistic regression, etc. These are all statistical calculations or procedures, but which ones do you use – and when? The appropriate statistical calculation or procedure is driven in large part by the data types.

The most basic example of data types driving statistical calculations is illustrated in Figure 2, which shows the distributions of the variables body temperature (°C) and diabetes (0 = No diabetes, 1 = Yes diabetes) among 1420 hospitalized cancer patients. Diabetes is a nominal variable with only two possible values. Thus, we want to know the number (frequency) of patients with diabetes and what proportion of the total sample they represent. Because body temperature is a continuous variable with many possible values, we summarize its distribution by reporting statistics such as the median, minimum, maximum, mean and standard deviation. Clearly it would not be feasible or helpful to summarize the number and proportion of patients who had each specific body temperature value, just as it would make no sense to calculate the mean of the diabetes variable.

Figure 2



### An analysis example

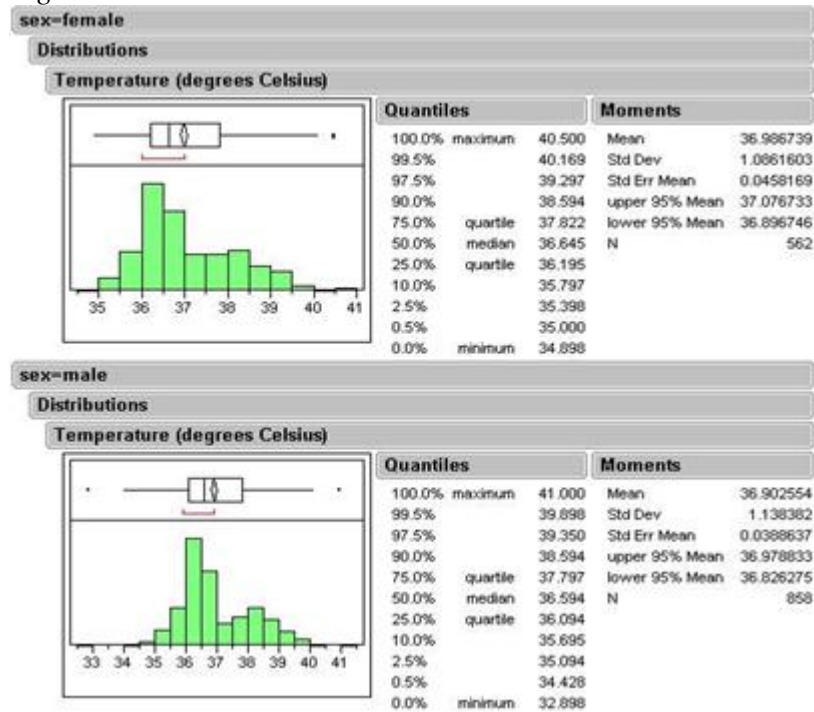
When you begin to pursue analyses more complex than descriptive statistics, data types are just as important and will lead you to the appropriate statistical procedures. Consider two questions in a hypothetical study involving these 1420 patients:

1. Is gender associated with body temperature?
2. Is gender associated with diabetes?

First, we must consider the data types. We know that body temperature is quantitative and more specifically continuous; diabetes is qualitative and more specifically nominal. Okay, now what? What information might be helpful in addressing the first question?

Would it be helpful to know the distribution of body temperature separately for males and females?

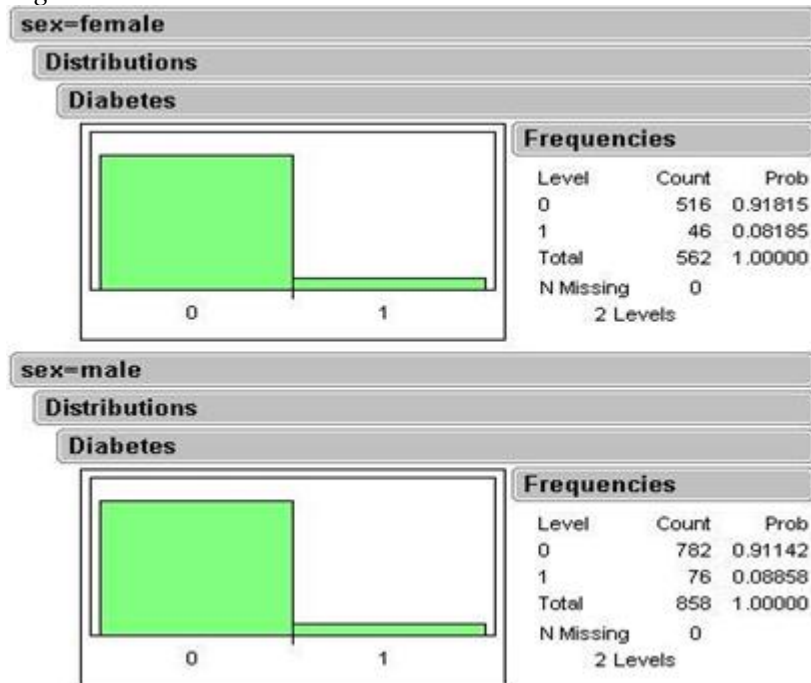
Figure 3



The distributions shown in Figure 3 summarize a continuous variable (body temperature) for each of two groups (females and males). A statistical quantity used to summarize the distribution of a continuous variable is the mean. We see that the mean body temperature for males was  $36.90^{\circ}$ , compared to  $36.99^{\circ}$  for females. Just as we compare means in the two groups in our descriptive statistical analysis, we need a procedure that will statistically compare the mean among males to the mean among females. One statistical test for comparing means between two groups is a two-sample t-test.

To answer question 2, we might start by summarizing the distribution of the diabetes variable separately for males and females (Figure 4).

Figure 4



The statistical quantity used to summarize the distribution of a nominal variable such as diabetes is a proportion. From Figure 4 we see that 46/562 (8.2 percent) females have diabetes, compared to 76/858 (8.9 percent) males. Because of the data types, we know that we would need a statistical procedure to compare proportions. The appropriate procedure to statistically compare proportions between two groups is a chi-square or Fisher's exact test.

## Conclusion

It's not always easy to classify the data type of a variable or to decide how it should be analyzed. Continuous and nominal variables are usually straightforward, but discrete quantitative and ordinal variables can be more challenging. For example, if you are interested in reporting the number of pregnancies among women in your study group, is it meaningful to treat this as a continuous variable and provide the mean number of pregnancies? Or would it be more meaningful to treat it as an ordinal variable and summarize the number of women with one pregnancy, two pregnancies, etc.? Or would it be more meaningful to report the number of women who had one or more pregnancies?

The answers to questions like these often depend on many factors such as the reason that you are summarizing a particular variable and what you believe will be the most meaningful and useful statistic for your audience.

## More information

The Mayo Clinic CTSA provides a biostatistical consulting service through its BERD Resource. More information can be found on the [BERD home page](#).