Print

Types of Data

In the previous module, you learned that there are three modeling types used when analyzing data: nominal, ordinal, and continuous. These modeling types are used to guide you to the type of statistical method that makes sense, given the type of data and the number of variables you are analyzing.

Nominal and ordinal are both categorical data, where the data are grouped into categories or levels.

Nominal data are unordered categorical data. For example, the data might fall into two categories, defective or non-defective, or pass or fail.

Or the data might have multiple unordered categories, such as the type of defect or the reason for failure. We call this multinomial data.

Ordinal data consist of ordered categories, such as severity ratings or degree of agreement scales on a survey.

For example, if you take a survey and rate the degree of satisfaction with a product, the values might range from 1 (not at all) to 5 (completely). The numbers represent the ordered categories.

Continuous data consist of numerical data or counts. If it makes sense to calculate an average of the values for a variable, or apply operations like addition or subtraction, then the data are usually coded as continuous.

Continuous data are often numerical measurements, like size, time and temperature.

The term continuous is used because these measurements can, at least conceptually, take on an infinite number of values along a continuum.

For example, the weight of a part might be measured as 5 grams, 5.1 grams, or 5.1217 grams, depending on the resolution of the scale.

For the purpose of analysis, count data are often treated as continuous data. Examples of count data are the number of defects on a part or the number of family members in a household.

Count data are discrete numeric data, but the average count for a sample can be continuous.

For example, the number of defects on an individual part might be 0, 1, 2, and so on.

But the average number of defects on parts in a given sample might be 1.12.

As you see in the upcoming videos, there are more statistical tools available for continuous data than there are for categorical data. Let's take a look at a simple example.

Let's say you are studying the impurity in a polymer. An acceptable batch of polymer has less than 7% impurity.

This corresponds to an acceptable yield of 93%.

Consider these two possible measures of the quality of a batch of polymer. A batch of polymer is classified as pass or fail, based on whether the specification of 7% has been met, or the percent impurity in the polymer is measured. Which of these measures contains more information?

For the nominal pass/fail measure, Outcome, you have an individual classification for each batch. The descriptive methods you can use are based largely on frequencies.

You can calculate the number of batches that passed or failed, or you can calculate the percent of batches that fall in each category. The graphical methods available are largely limited to graphs showing the frequencies within each of the categories.

From this bar chart and frequency table for 100 batches of polymer, you can see that 26% of the batches failed to meet the specification.

For the continuous response, percent Impurity, the value for a batch can range from 0 to 100. You have a measured value for each batch, so you know how bad the problem is or how close you are to the acceptable level for each batch. You can also use descriptive statistics to tell you about the distribution of impurity values for the different batches.

You can summarize where the distribution is centered. Is it close to 7%? Is it much lower or higher than this value?

You can summarize the spread of the distribution. Are most of the values close to the center of the distribution? Or do the impurity values span a broad range?

You can also summarize the shape of the distribution. Are the values more or less evenly spread around the middle of the distribution, or does the distribution have some other shape, like a long tail of values?

Are there a handful of values that fall far away from the other values?

In this summary of the same 100 batches, you can see that the distribution of impurity values is centered at about 6 and that the impurity values range from around 3 to 10.

You can also see that the tail of the distribution is longer in the direction of the higher impurity levels.

We'll revisit this impurity scenario throughout this lesson as we introduce other graphical and statistical methods.

In upcoming videos, you learn about descriptive statistical methods for continuous and categorical data, and you see how to analyze these data in JMP.

Close