Print

Identifying Issues in the Data Table

Let's consider a new scenario involving small components.

You are studying potential causes of high scrap rates. You compile data on 369 batches of components, produced over a three-month period. The data are in the file Components Raw.jmp. The data set includes information about the batch number, the part number, the customer number, the batch size, the number of parts scrapped, the yield, and various process variables.

Before you formally analyze these data, you need to evaluate data quality. Remember that you are looking for these common problems: incorrect formatting, incomplete data, missing data, and dirty or messy data.

You start by looking at the data table to make sure it is properly formatted.

This data table has 14 variables and 369 rows. Each row includes information about the variables for a given batch.

Three of the variables are coded as nominal, and the others are coded as continuous.

When you import data into JMP, numeric variables are automatically coded as continuous, and variables containing text or symbols are coded as nominal.

Let's look at the first variable, batch number, which is coded as continuous. This is a unique identifier for each batch. You can see that the data are sorted in order of batch number.

Part number is also coded as continuous.

Both the batch number and the part number are really just labels. These are categorical data, so the variables should be coded as nominal.

Now, we check to see whether the data are complete.

Remember that you are studying the scrap rate. You have data on the batch size, the number of parts scrapped per batch, and the yield. But you don't have a variable directly measuring the scrap rate. You need to calculate this variable from the data.

You know that you can calculate scrap rate from the yield.

But you don't know how this yield variable was calculated, and you don't fully trust the data.

Instead, you can create a new variable, Scrap Rate, using this formula.

Now you check to see whether you are missing any values.

You quickly see that you might have a problem with missing data.

For example, the variable temp is missing a lot of values. Because temp is numeric data, the missing values are displayed as black dots.

You look at the variable supplier and see that some of these cells are empty. Missing values for character data are displayed as blank cells. So, for example, the supplier for the third observation is missing.

When you look at the supplier names, you also see evidence of messy data. For example, you see both the name Cox and Cox Inc. The supplier name was entered two ways. You need to clean this up if you are planning on using supplier in an analysis.

In looking at your data table, you can identify and correct some obvious issues. For our example, this is what we have learned thus far: Batch Number is a unique identifier, so it should be coded as nominal. If you know you won't use batch number in any analysis, you can change the modeling type to none. Part Number is a label, so it should have a nominal modeling type. You are missing a variable for scrap rate. You should use a formula to create this variable.

These issues are easy to address. The file Components.jmp has been saved with these changes.

You also learned that you are missing values for some of your variables, and that the variable supplier needs to be cleaned up.

You could continue to scan the data table for potential data quality issues, but this isn't very efficient.

Instead, you use the exploratory tools that you have learned about in this module to evaluate data quality.

In the next video, you see how to diagnose common data quality issues by exploring data one variable at a time.

Close