Identifying Issues One Variable at a Time

After scanning the data table for potential issues, you look at the variables one at a time. You use summary statistics and graphical summaries to get familiar with the data. At the same time, you look for potential data quality issues.

For continuous variables, you're interested in summary statistics, such as the mean, the standard deviation, the minimum and maximum values, and the number of missing values.

You're also interested in the shape of the distribution.

Is the distribution more or less symmetric? Is it skewed? Are there clusters of data or severe outliers? Are there values that aren't physically possible?

For categorical data, you're interested in the number of categories (or levels), the number of observations in each category, and the number of missing values.

We'll continue with the components example. The file Components.jmp has the new variable scrap rate. The variable yield has been hidden and excluded, and the modeling types for batch number and part number have been changed to nominal.

First, let's look at summary statistics for all of the variables.

The column N Categories tells you how many levels you have for your categorical variables. You see that there are 369 batch numbers, three part numbers, and 20 customers. You can also see that there are two levels for vacuum and 10 suppliers.

When you look at the N Missing column, you can see that many of the variables are missing values. The biggest issue is temp, which is missing 265 values. This tells you that more than 2/3 of the values for temp are missing.

For the continuous variables, you can see the minimum and maximum values, the mean, and the standard deviation.

You see an interesting problem with number scrapped and scrap rate. The minimum value for both of these variables is negative.

Using bivariate (or two-variable) plots can help you understand the negative values for these two variables.

When you plot the number scrapped against the scrap rate using a scatterplot, you see that there are two negative values.

Does this make sense? Can you have a negative number of scrapped parts? This is something you need to investigate.

You also see something interesting with the statistics for process, which is coded as continuous data. The minimum value is 1, and the maximum value is 2. Is this just a categorical variable with two numeric values, 1 and 2?

This is also something you need to check.

Next, you look at distributions for all of the variables (except for batch number).

You have 20 customer numbers, but most of the orders are for only 6 of the customers. Very few of the orders are from the other 14 customers.

You see that scrap rate is slightly skewed, and you can see the two negative values. You noticed this earlier when you looked at the summary statistics.

You see that there are only four batch sizes, but batch is coded as continuous. You might want to code batch size as nominal if you use it in an analysis.

The distribution for scrapped parts is highly skewed. But this makes sense, given that different batch sizes have a different number of scrapped parts, and most of the components were made in the smaller batch sizes.

You also see a few other issues.

You learn that process does indeed have only two levels, 1 and 2.

You see that most of the values for speed are between 60 and 140. But there is one value that is close to zero. Is this a typo?

You also see some issues for supplier that you noted earlier. Some of the supplier names are spelled differently. And some use different capitalization. If you are going to use supplier in an analysis, you need to fix these issues.

Statistical Thinking for Industrial Problem Solving

Close