Print

Exploring Continuous Data: Enhanced Tools

You've seen many graphical tools for exploring continuous data. You can use distributions and box plots to describe one variable at a time, comparative box plots to explore the relationship between a continuous variable and a categorical variable, scatterplots and scatterplot matrices to explore pairs of continuous variables and run charts to visualize data over time.

In this video, you revisit some of these tools, and see how to enhance your data exploration using colors, markers, legends, and column switching.

For illustration, we use the Chemical Manufacturing example, which has a continuous variable, Yield, along with several other process measures.

To accept a batch, the yield must be greater than 80%.

The categorical variable, Performance, indicates whether a batch was accepted or rejected.

Yield is the output, or response variable, that we're interested in understanding.

The other variables are inputs, or factors.

The goal of our exploratory data analysis is to use these data to identify potential causes of poor yield.

You've seen that you can use linked histograms as an initial step in understanding your data.

But with 16 variables, this is a bit overwhelming.

As you have learned, you can use box plots to compare the yield across levels of a categorical variable.

From this box plot, there appears to be a difference in yield for the different vessel sizes.

To see this better, you can color the points and apply different markers, based on whether or not the batch met the yield specification of 80%.

You can see that most of the rejected batches were produced with the largest vessel size.

Only three of the rejected batches were from the two smaller vessel sizes.

You might not need to use both different colors and different markers, but the markers can help when it is difficult to distinguish the colors. You learn more about the use of colors and markers in the next lesson.

There are seven categorical input variables in this example.

To facilitate exploring all of these variables, without creating a new box plot each time, you can use the Column Switcher in JMP.

This enables you to scroll through each of the input variables to identify variables that might be related to poor yield.

Here, for instance, there doesn't appear to be much of a difference between the two amine suppliers.

But there might be a difference in Base Particle Size and Base Supplier.

There are eight continuous input variables. You can use scatterplots to explore the relationship between Yield and each of the variables, again using colored markers and the Column Switcher.

Here, it looks like there is a relationship between Yield and Carbamate Amount most of the rejected batches were at lower values of Carbamate Amount. You can add a reference line to better differentiate the accepted from the rejected batches.

There doesn't appear to be a relationship between Yield and Vacuum. The rejected batches don't have particularly low or high values of Vacuum, relative to the accepted batches.

Instead of looking at each scatterplot one at a time, you can use a scatterplot matrix. It doesn't look like there are particularly strong relationships between any of the variables and Yield.

But with a total of nine variables (the response, Yield, plus the eight input variables), this graph is a bit overwhelming. For this example, a scatterplot matrix isn't the most informative graph.

We learn more from the individual graphs with the Column Switcher.

Let's take a look at a different scenario, the Impurity example.

Remember that you are studying the impurity in a polymer, and an acceptable batch of polymer has less than 7% impurity.

Using linked histograms, you can see the relationships between Impurity and the three continuous process variables: Temp, Catalyst Conc, and Reaction Time.

What can we learn from a scatterplot matrix? Here's a scatterplot matrix for the four variables.

This enables you to see the relationship between all pairs of variables.

Look at how much more informative the scatterplot matrix is when you add colors and markers. Here, we've colored and marked the points by Outcome, which measures whether the batches passed or failed.

In this scatterplot matrix, you also see a row legend, which enables you to toggle between the two outcomes. This makes it easier for you to see points that might be hidden or obscured by other points.

It's easy to see that batches that failed to meet the impurity specification of 7% tend to have high values for Temp and Catalyst Conc.

If you want to see where observations with high Impurity values fall in the other scatterplots, you can select these points, and they are selected everywhere.

For example, the two observations with the highest values for Impurity had among the highest values for Temp and Catalyst Conc.

Let's summarize what you've learned in this video. Adding colors and different markers can help you see the patterns in your data.

Column switching enables you to easily explore many variables without having to re-create your analysis each time.

A scatterplot matrix, and any graphical display for that matter, can be enhanced with the addition of a row legend. You learn how to add colors, change markers, add a legend, and use the Column Switcher in JMP demonstration videos.

For additional information on how to enhance your analyses with colors, markers, legends, and other customizations, see the Read About It for this module.

In upcoming videos, you see that a data filter can also be used to select values in your graph. This is an effective tool for stratifying your data based on values of selected variables. You also learn additional graphical tools for visualizing categorical and continuous data.

Close