Print

Visualizing Continuous Data

In previous videos, you learned how to graph continuous data using histograms and box plots.

Consider the White Polymer case study. You are on a team that is trying to improve the yield of a molding process.

You learned that the historical crisis team data used earlier has some problems, and you collected new data. These data are in the course file VSSNewData.jmp.

The maximum possible value for Yield is 100%. This is the ideal. The distribution is bounded by 100% on the right and has a long tail to the left. So this is left-skewed data.

The median of Yield is 91.75%, and the mean for Yield is 87.21%. There's a lot of room for improvement.

Let's take a look at the distributions of the key output characteristics, MFI (the melt flow index) and CI (the color index). Remember that these two characteristics should be closely related to Yield.

Using linked histograms, you can see the relationship between Yield, MFI, and CI. Batches with high Yield tend to have high CI. The lower spec for CI is 80.

These batches also tend to have values close to the target range for MFI of 192-198. This is what you would hope to see. If you can improve MFI and CI, you can improve Yield.

Let's look at some other ways of visualizing these relationships.

When you have two continuous variables, a scatterplot shows you the relationship between these two variables.

This scatterplot, which has a smoother drawn through it, shows that the relationship between Yield and MFI appears to have curvature.

A scatterplot also shows some unusual observations. For example, observation 14, highlighted here, has an MFI of 197 and a Yield of 40%. And four points have both extremely high MFI values and extremely low Yield values.

Was there a serious issue with MFI that caused these values to be so high relative to the other MFI values? This warrants some investigation.

What does the relationship between Yield and MFI look like if you don't include or show these five observations?

The curved relationship is now much clearer. The target for MFI is 195. All of the batches with Yield over 90% have MFI above the target of 195.

You can also see that when the upper spec of 198 for MFI is met, many of the batches have high Yield values.

The data table, with the five extreme values hidden and excluded, is saved in your course folder as VSSTeamData.jmp. We'll continue the analysis with this file.

With a scatterplot, you can explore the relationship between two variables. When you have many variables that you want to explore at the same time, you can create a matrix of scatterplots.

Here is a scatterplot matrix for Yield, CI, and MFI. There is one scatterplot for each pair of variables: Yield and CI, Yield and MFI, and CI and MFI.

This is a triangular scatterplot matrix. That is, the scatterplots are in a triangular arrangement.

You can also create a square scatterplot matrix. This enables you to see the relationship between pairs of variables from different perspectives.

Here, you see two scatterplots for CI and Yield. With scatterplots, you have two continuous variables.

What if one of the variables is continuous and the other is categorical?

In the White Polymer case study, data were also collected on the quarry and the shift. You might want to see whether there is a difference in Yield across the different quarries or shifts.

You can create a tabular summary of the data, using descriptive statistics like the mean and the standard deviation.

In this summary, we don't see big differences in the means and standard deviations for the different quarries.

You can also summarize the data graphically. In a previous video, you learned about using box plots to explore the shape, centering, and spread of a distribution.

You can also use box plots for comparative purposes. Here, you see an outlier box plot for Yield for each of the three quarries, with the individual Yield values plotted as points for each quarry.

As you saw with the numeric summaries, there doesn't appear to be a big difference in Yield for the three quarries.

Before we close this video, let's talk about one final graphical tool for continuous data. Most of the data you collect will have a time dimension.

The Yield data are stored in the data table in the order that the batches were produced.

You might be interested in examining the performance of the process, with regard to Yield, over time. In this situation, you use a run chart, also known as a line graph.

With a line graph, the variable of interest is plotted on the Y axis and the time-ordered variable is plotted on the X axis.

Here, you see the Yield values plotted in the order that the batches were produced. The Yield values bounce around for the most part.

But you do see one period, starting at around batch 4080, that had several batches with very low yields.

Let's take a look at MFI. Here, you see a spike in MFI values, starting around batch 4070.

When you plot these two variables on the same graph, you can see that many of the batches with the highest MFI values also tend to have low values for Yield.

Clearly, if you want to improve Yield, you need to develop a better understanding of MFI!

You continue to explore the relationship between Yield, MFI and CI in a practice.

In these videos, you learned the building blocks of exploratory data analysis, or EDA. You explored data one variable at a time, then two variables at a time, and then more than two variables at a time.

You explored distributions, developed an understanding of the patterns in the data, identified some outliers, and found potential relationships that might bring you one step closer to solving the problem.

You learn more about exploratory data analysis and see additional tools and techniques for visualizing data in future videos.

In these videos, you revisit the White Polymer case study. Rather than focusing on the KPI, Yield, we focus on understanding the relationship between the two output variables, MFI and CI, and the various input variables.

In the next JMP demonstration videos, you learn how to create tabular summaries, scatterplots, comparative box plots, and line graphs in JMP.

Close