Print

# Histograms

In the previous video, you saw a histogram and summary statistics for the Impurity data. Here, the percent impurity was measured for 100 batches of polymer.

A histogram is a picture of the distribution of the data. A distribution is the pattern that is formed by your data.

The height of each of the bars indicates the number of impurity values that fall within each interval.

Here, you can see that 28% of the observations in the data set fall in the interval from 5 to 6.

You can also see that 26% of the observations have an unacceptable value for Impurity. That is, 26% of the observations fall above the specification for impurity of 7.

You can use the histogram to understand three characteristics of the distribution: the centering (or location), the spread, and the shape.

Histograms are also useful for identifying unusual patterns in the data. Let's take a look at some possible distribution shapes to understand these characteristics.

This first distribution has more or less a bell shape.

It is centered at zero, and the spread of the distribution is from approximately −3 to +3.

These data are from a very special distribution in statistics, known as the standard normal distribution. You learn more about the normal distribution, and the standard normal distribution, in the Probability Concepts lesson in this module.

This next distribution has a long right tail.

This is a right-skewed distribution. Most of the values are between zero and 1, with a few values falling in the tail of the distribution.

When the tail of the distribution points to the left, the distribution is called left-skewed.

In this distribution, the variable takes on values between zero and 1. The values are distributed across this range.

This is called a uniform distribution. For a uniform distribution, the shape is flat, and the data span the range of possible values for the variable. You might also see a histogram that looks like this.

Do you notice that there are two peaks? This is a bimodal distribution. The second peak is lower than the first peak, and the data around the second peak are more spread out than the data around the first peak.

When you see a histogram with this shape, it usually means that your data come from two different distributions.

In this case, the two distributions have both different centering and different spread.

Let's return to the Impurity data. We collected impurity values for 100 batches, but we also collected information on several process variables.

Here, you see histograms for Temp, Catalyst Conc, and Reaction Time. From these histograms, you can easily see the centering, shape, and spread of the data for the three variables.

You can also see how Impurity is related to the process variables. Here, we show vertical histograms for Impurity and the three process variables so that we can view all of the distributions on one screen.

We have selected the Impurity values that are greater than 7. The histogram bars for these observations are shaded.

The histograms are linked, so observations are also shaded in the histograms for the process variables.

Notice that batches that were unacceptable for Impurity also tend to have high values for Temp and Catalyst Conc. This could be very useful information in identifying potential causes for high impurity levels.

In the next JMP demonstration video, you see how to create histograms, and how to interact with linked histograms to explore potential relationships between variables. You also learn how to save the steps used to create an analysis to the data table so that you can easily re-run the analysis later.

Then you practice what you've learned using two case studies, which will be revisited throughout this module.

---




Close