Print

Checking for Normality

How can you tell if your data are approximately normally distributed?

You've learned that a histogram is the best way to see the distribution of sample data. Here, it looks like the distribution for Impurity is slightly right-skewed.

But histograms can be misleading when it comes to understanding the underlying distribution. Even if your data are known to come from a normal distribution, if you have a small sample size, the data might not appear to be normal. To illustrate, we use a simulation.

The first histogram here is for 1000 observations simulated using a random normal distribution.

The second is based on 100 observations, the third is based on 30 observations, and the last is based on 10.

These were all simulated using a random normal distribution with a mean of 0 and a standard deviation of 1.

Notice that this distribution based on 1000 observations looks normal.

The distribution based on 100 observations is mounded in shape but looks less normal.

The last two distributions don't really look normal at all.

If your data set is small, with fewer than 100 observations, a histogram isn't the best way to determine whether your data are approximately normal.

An alternative is to use a normal quantile plot.

A normal quantile plot is also called a QQ plot or a normal probability plot.

A normal quantile plot displays your data, which have been sorted in ascending order, plotted against the percentiles of the standard normal distribution.

Notice the scales for the normal quantile plot. There are two axes.

The first axis goes from around -3 to 3, which is the spread that you'd expect to see for a standard normal distribution.

The second scale shows the percentiles for the normal distribution, ranging from close to zero to close to 1.0.

The smallest data value is plotted at the smallest percentile, and the largest data value is plotted at the largest percentile.

Because we expect that 50% of the observations in a standard normal distribution fall below zero, the percentile for zero is 0.50.

If the distribution is approximately normal, the points in the normal quantile plot fall in a straight diagonal line, with no obvious nonlinear patterns.

Here are the normal quantile plots for our four distributions simulated from random normal data. Notice that, even for the distribution with only 10 observations, the points in the normal quantile plot fall more or less on a diagonal line.

Here is a normal quantile plot for data that are right-skewed. Notice the curve in the points.

The normal quantile plot for the uniform distribution is s-shaped.

Let's look at a normal quantile plot for some real data. Here's a histogram and a normal quantile plot for the Impurity data. The histogram is slightly right-skewed.

As a result, we see a slight curve in the normal quantile plot.

How do you decide whether the data are close enough to a straight (diagonal) line? There are formal tests for normality, which are beyond the scope of this course.

One rule of thumb is to use the red bands in the normal quantile plot, which are called confidence curves. If all of your observations fall within these bands, then it's safe to assume that the underlying distribution is approximately normal.

The width of these curves is based on the sample size, so when you have a smaller sample size the curves are wider.

Let's take another look at the normal quantile plots for the simulated random normal distributions.

For each plot, notice that all of the points fall within the red bands.

Also, notice how wide the bands are for the distribution with only 10 observations. This normal quantile plot is much less strict in detecting departures from normality, because we simply don't have a lot of information about the shape of the distribution.

What do you see in our Impurity example?

Some of the points fall outside the confidence curves. This indicates, as we've already learned, that the underlying distribution is likely not normal.

Although interpreting the normal quantile plot is somewhat subjective, most of the statistical techniques you learn in this course will still apply if the departure from normality is minor.

Throughout this video, the focus of the discussion has been the normal distribution. However, theres no rule that your data must be normally distributed in order to analyze it. Your data are what they are, and data can follow many different probability distributions.

Many continuous characteristics are known to be non-normal. For example, the time it takes to deliver a product follows a right-skewed distribution.

In fact, these data follow a lognormal distribution. You can see that the lognormal curve, which is right-skewed, fits the data well.

A distribution that is used widely in reliability applications is the Weibull distribution. This probability distribution is often used to model the time to failure for a product. In this example, the fitted Weibull distribution is right-skewed, but this is not always the case.

Earlier in this video, you learned that the area under a normal curve is 1.

The area under the curve for the lognormal, the Weibull, and every other probability distribution, is also 1.

This means that, if you understand the underlying distribution, you can make predictions and calculate probabilities, just like you can for normally distributed data.

You learn more about non-normal distributions in the Quality Methods module.

For information about fitting distributions and testing the fit of these distributions, search for Fit Distributions in the JMP Help or see the Read About It for this module.

In the next two JMP demos, you see how to check for normality and how to find the area under the curve for a distribution. Then you learn about the Central Limit Theorem, which enables you to use many statistical tools that assume normality, even though the characteristic might not be normally distributed.

Close