15.4.4

Test for Normality

Jeremy had forgotten how important distribution is! After brushing up on distribution, he asks Colleen to give him a quick recap on how to actually test for normality.

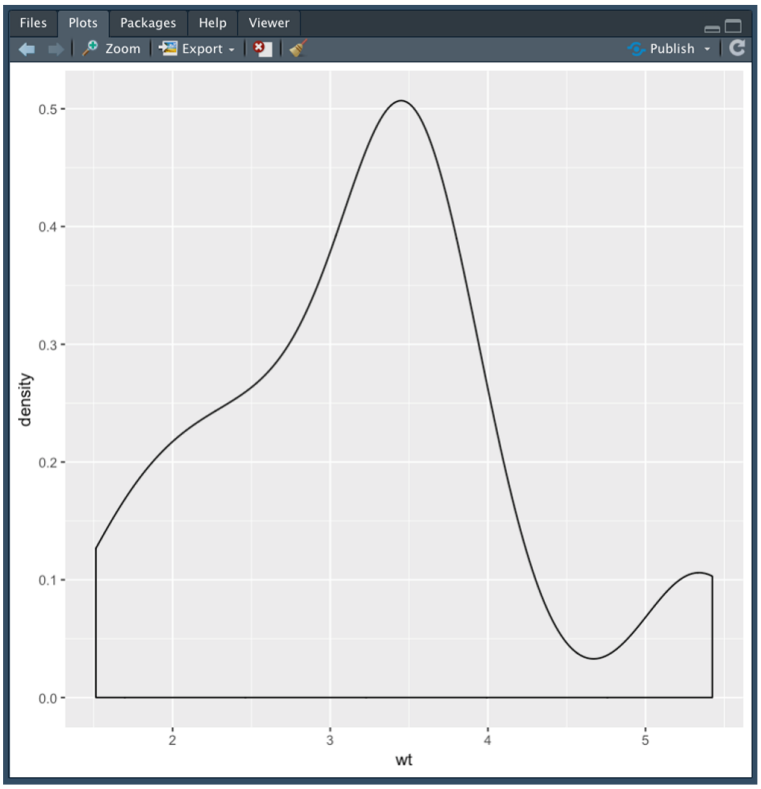
You can test for normality during data analysis by performing a qualitative test or a quantitative test.

Qualitative Test for Normality

The **qualitative test for normality** is a visual assessment of the distribution of data, which looks for the characteristic bell curve shape across the distribution. In R, we would use ggplot2 to plot the distribution using the `geom_density()` function.

For example, if we want to test the distribution of vehicle weights from the built-in `mtcars` dataset, our R code would be as follows:

```
> ggplot(mtcars,aes(x=wt)) + geom_density() #visualize distribution using density plot
```



The `geom_density()` function plots a numerical vector by creating buckets of similar values and calculating the density (number of bucket data points/total number of data points) for each bucket.

The results of each bucket density calculation are plotted, connected, and smoothed out to create our distribution plot. Although our data distribution does not perfectly match the normal bell curve shape, the distribution does approximate a normal distribution and could be used for further analysis.

But what if our data distribution is noisy—meaning that the dataset contains uncharacteristically large or small values at high frequency—or we need to make more informed, quantitative decisions? In these cases, we would want to perform our quantitative test.

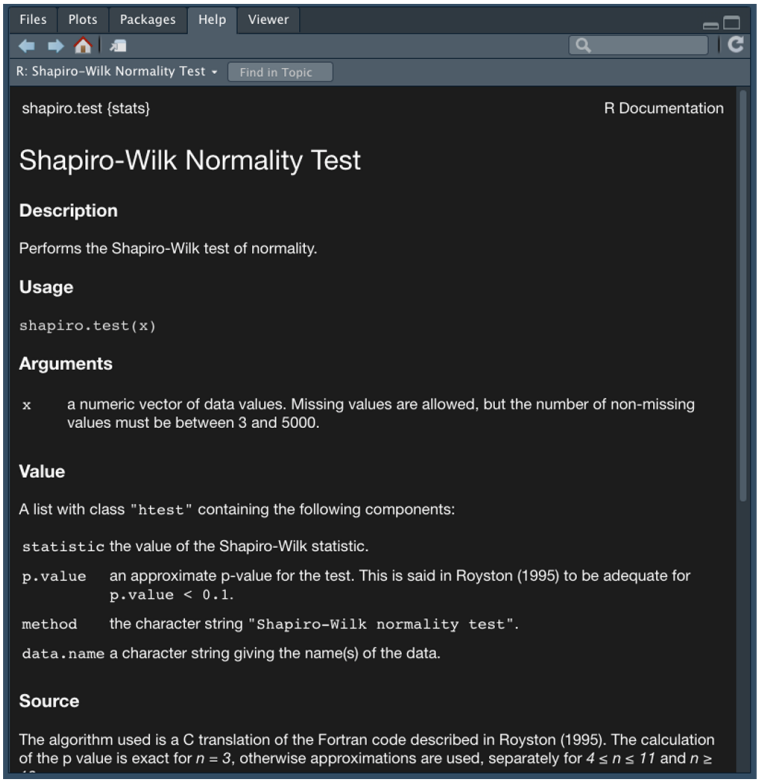
Quantitative Test for Normality

The **quantitative test for normality** uses a statistical test to quantify the probability of whether or not the test data came from a normally distributed dataset.

In most cases, data scientists will use the Shapiro-Wilk test for normality, though there are many other statistical tests available. In R, we can use the built-in stats library to perform our quantitative test with the `shapiro.test()` function.

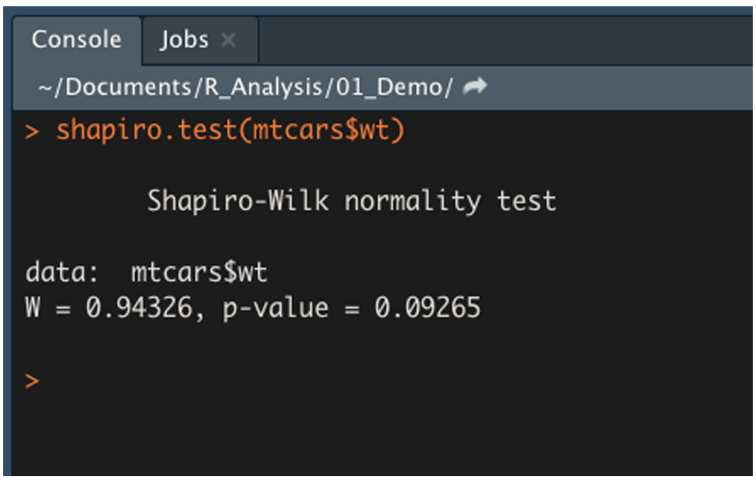
Type the following code into the R console to look at the `shapiro.test()` documentation in the Help pane:

```
>?shapiro.test()
```



The `shapiro.test()` function only requires the numeric vector of values you wish to test. Therefore, if we want to perform a quantitative Shapiro-Wilk test on our previous example, our R code would look as follows:

```
> shapiro.test(mtcars$wt)
```



Later we'll discuss what a p-value is and how it is used in statistics. For our purposes, you just need to know that if the p-value is greater than 0.05, the data is considered normally distributed.

Remember that most basic statistical tests assume an **approximate normal distribution**. Therefore, if our p-value is around 0.05 or more, we would say that our input data meets this assumption. But what happens if our data distribution does not look like a bell curve, or the p-value of the Shapiro-Wilk tests is too small?

© 2020 - 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.