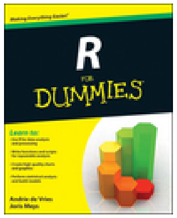
dummies
A Wiley Brand

🏠 > **PROGRAMMING** > **R** > HOW TO TEST DATA NORMALITY IN A FORMAL WAY IN R

HOW TO TEST DATA NORMALITY IN A FORMAL WAY IN R



RELATED BOOK

R For Dummies

By **Andrie de Vries, Joris Meys**

The graphical methods for checking data normality in R still leave much to your own interpretation. There's much discussion in the statistical world about the meaning of these plots and what can be seen as normal.

If you show any of these plots to ten different statisticians, you can get ten different answers. That's quite an achievement when you expect a simple yes or no, but statisticians don't do simple answers.

On the contrary, everything in statistics revolves around measuring uncertainty. This uncertainty is summarized in a probability — often called a *p-value* — and to calculate this probability, you need a formal test.

Probably the most widely used test for normality is the Shapiro-Wilks test. The function to perform this test, conveniently called `shapiro.test()` , couldn't be easier to use. You give the sample as the one and only argument, as in the following example:

```
> shapiro.test(beaver2$tem
 Shapiro–Wilks normality test
data: beaver2$temp
W = 0.9334, p-value = 7.764e-05
```

REMEMBER

This function returns a list object, and the p-value is contained in a element called `p.value` . So, for example, you can extract the p-value simply by using the following code:

```
> result <- shapiro.test(beaver2$temp)
> result$p.value
[1] 7.763782e-05
```

This p-value tells you what the chances are that the sample comes from a normal distribution. The lower this value, the smaller the chance. Statisticians typically use a value of 0.05 as a cutoff, so when the p-value is lower than 0.05, you can conclude that the sample deviates from normality.

In the preceding example, the p-value is clearly lower than 0.05 — and that shouldn't come as a surprise; the distribution of the temperature shows two separate peaks. This is nothing like the bell curve of a normal distribution.

TIP

When you choose a test, you may be more interested in the normality in each sample. You can test both samples in one line using the `tapply()` function, like this:

```
> with(beaver, tapply(temp, activ, shapiro.test)
```

This code returns the results of a Shapiro-Wilks test on the temperature for every group specified by the variable `activ`.

WARNING

People often refer to the Kolmogorov-Smirnov test for testing normality. You carry out the test by using the `ks.test()` function in base R. But this R function is not suited to test deviation from normality; you can use it only to compare different distributions.