What are the four methods used to test for normality in R?

Authored by

stats writer

May 12, 2024

ARABPSYCHOLOGY.COM

RECOMMENDED CITATION

stats writer (2024). *What are the four methods used to test for normality in R?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=143930

There are four commonly used methods in R to test for normality, which is the assumption that a dataset follows a Gaussian distribution. The first method is the Shapiro-Wilk test, which uses a statistic to measure the deviation from normality and a p-value to determine the significance of the deviation. The second method is the Kolmogorov-Smirnov test, which also calculates a statistic and p-value, but does not assume a specific distribution. The third method is the Anderson-Darling test, which uses a critical value to determine if the data significantly deviates from normality. The final method is the Chi-Square test, which compares the observed frequency distribution to the expected frequency distribution under the assumption of normality. These methods are commonly used to assess the normality of a dataset and inform further statistical analysis.

Test for Normality in R (4 Methods)

Many statistical tests make that datasets are normally distributed.

There are four common ways to check this assumption in R:

1. (Visual Method) Create a histogram.

If the histogram is roughly "bell-shaped", then the data is assumed to be normally distributed.

2. (Visual Method) Create a Q-Q plot.

If the points in the plot roughly fall along a straight diagonal line, then the data is assumed to be normally distributed.

ARABPSYCHOLOGY.COM

3. (Formal Statistical Test) Perform a Shapiro-Wilk Test.

If the p-value of the test is greater than α = .05, then the data is assumed to be normally distributed.

4. (Formal Statistical Test) Perform a Kolmogorov-Smirnov Test.

If the p-value of the test is greater than α = .05, then the data is assumed to be normally distributed.

The following examples show how to use each of these methods in practice.

Method 1: Create a Histogram

The following code shows how to create a histogram for a normally distributed and non-normally distributed dataset in R:

```
#make this example reproducible
set.seed(0)

#create data that follows a normal distribution
normal_data <- rnorm(200)

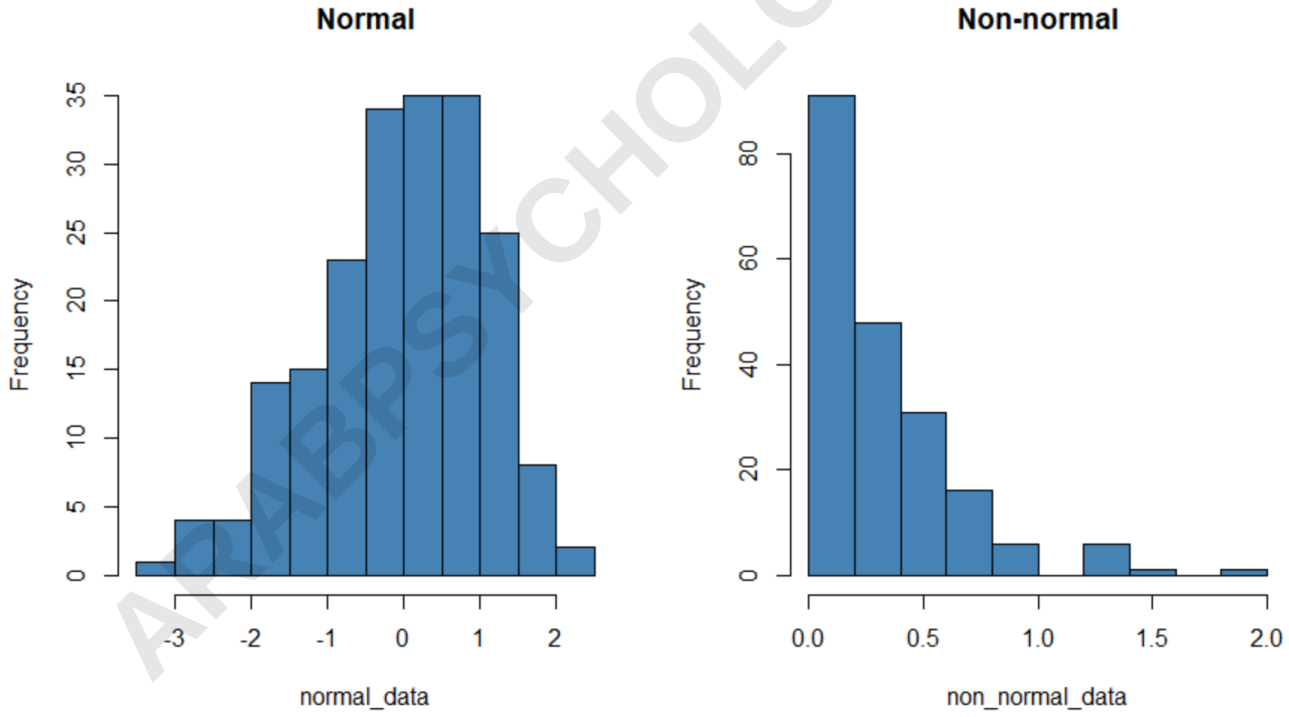
#create data that follows an exponential distribution
```

ARABPSYCHOLOGY.COM

```
non_normal_data <- rexp(200, rate=3)

#define plotting region
par(mfrow=c(1,2))

#create histogram for both datasets
hist(normal_data, col='steelblue', main='Normal')
hist(non_normal_data, col='steelblue', main='Non-normal')
```



ARABPSYCHOLOGY.COM

The histogram on the left exhibits a dataset that is normally distributed (roughly a "bell-shape") and the one on the right exhibits a dataset that is not normally

distributed.

Method 2: Create a Q-Q plot

```
#make this example reproducible
set.seed(0)

#create data that follows a normal distribution
normal_data <- rnorm(200)

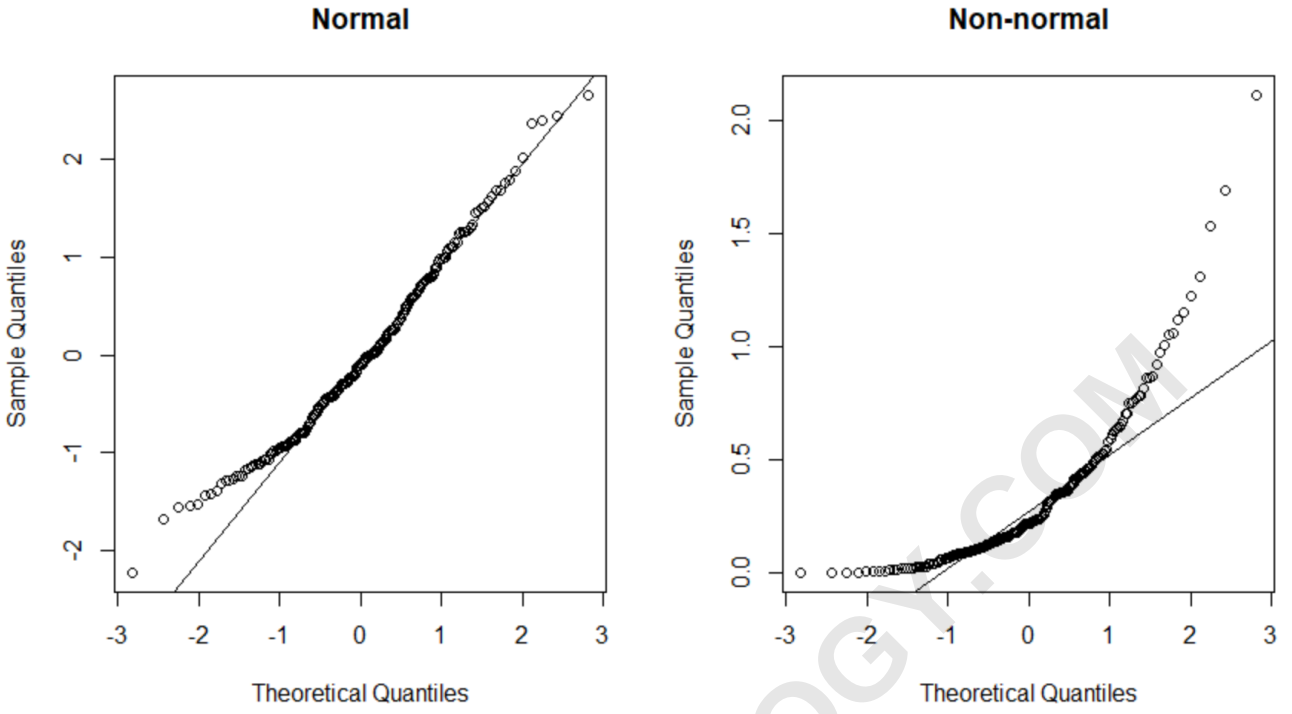
#create data that follows an exponential distribution
non_normal_data <- rexp(200, rate=3)

#define plotting region
par(mfrow=c(1,2))

#create Q-Q plot for both datasets
qqnorm(normal_data, main='Normal')
qqline(normal_data)

qqnorm(non_normal_data, main='Non-normal')
qqline(non_normal_data)
```

ARABPSYCHOLOGY.COM



The Q-Q plot on the left exhibits a dataset that is normally distributed (the points fall along a straight diagonal line) and the Q-Q plot on the right exhibits a dataset that is not normally distributed.

Method 3: Perform a Shapiro-Wilk Test

The following code shows how to perform a Shapiro-Wilk test on a normally distributed and non-normally distributed dataset in R:

```
#make this example reproducible
set.seed(0)

#create data that follows a normal distribution
```

ARABPSYCHOLOGY.COM

```
normal_data <- rnorm(200)

#perform shapiro-wilk test
shapiro.test(normal_data)

Shapiro-Wilk normality test

data: normal_data
W = 0.99248, p-value = 0.3952

#create data that follows an exponential distribution
non_normal_data <- rexp(200, rate=3)

#perform shapiro-wilk test
shapiro.test(non_normal_data)

Shapiro-Wilk normality test

data: non_normal_data
W = 0.84153, p-value = 1.698e-13
```

The p-value of the first test is not less than .05, which indicates that the data is normally distributed.

The p-value of the second test *is* less than .05, which indicates that the data is not normally distributed.

ARABPSYCHOLOGY.COM

Method 4: Perform a Kolmogorov-Smirnov Test

The following code shows how to perform a Kolmogorov-Smirnov test on a normally distributed and non-normally distributed dataset in R:

```
#make this example reproducible
set.seed(0)

#create data that follows a normal distribution
normal_data <- rnorm(200)

#perform kolmogorov-smirnov test
ks.test(normal_data, 'pnorm')

One-sample Kolmogorov-Smirnov test

data: normal_data
D = 0.073535, p-value = 0.2296
alternative hypothesis: two-sided

#create data that follows an exponential distribution
non_normal_data <- rexp(200, rate=3)

#perform kolmogorov-smirnov test
ks.test(non_normal_data, 'pnorm')
```

ARABPSYCHOLOGY.COM

One-sample Kolmogorov-Smirnov test

data: non_normal_data
D = 0.50115, p-value < 2.2e-16
alternative hypothesis: two-sided

The p-value of the first test is not less than .05, which indicates that the data is normally distributed.

The p-value of the second test *is* less than .05, which indicates that the data is not normally distributed.

How to Handle Non-Normal Data

If a given dataset is *not* normally distributed, we can often perform one of the following transformations to make it more normally distributed:

1. Log Transformation: Transform the values from x to log(x).

2. Square Root Transformation: Transform the values from x to √x.

3. Cube Root Transformation: Transform the values from x to x1/3.

ARABPSYCHOLOGY.COM

By performing these transformations, the dataset typically becomes more normally distributed.

Read this tutorial to see how to perform these transformations in R.

ARABPSYCHOLOGY.COM