Testing for Normality - Performing Tests with R and R Commander

Dr. James. Ready

Updated: December 23, 2019

Many statistical tests (e.g., Independent Samples t-test, Analysis of Variance [ANOVA], calculation of the Pearson Product-Moment Correlation Coefficient [r]) are based on the assumption that the continuous variable being tested is normally distributed. One way of examining a variable's distribution is by reviewing a histogram of the variable with a normal distribution overlay or a Quantile-Quantile (QQ) plot. However, interpretation of a normal distribution is subjective. One researcher's view of a distribution may be different from another. An alternative way of examining a continuous variable's distribution is by performing a test of normality.

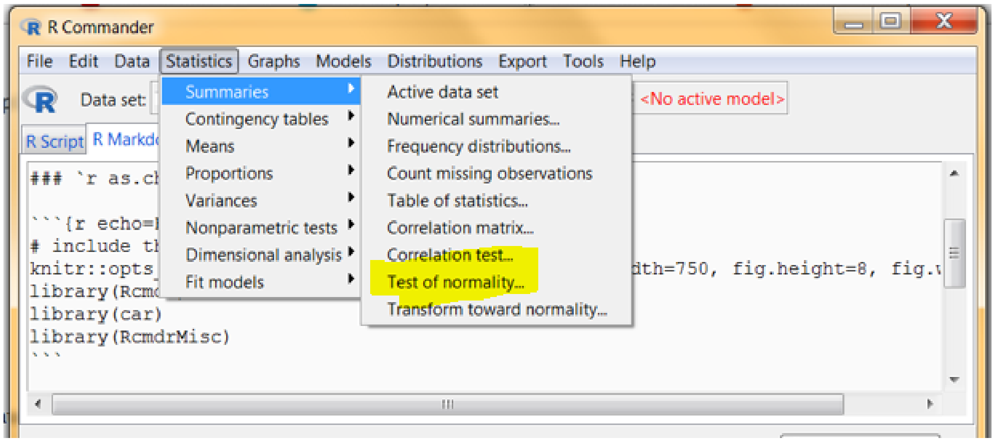
The purpose of this paper is to demonstrate how to perform tests of normality using R Commander and R, interpreting and reporting the test results, and connecting these results to graphical representations of data.

Performing Tests of Normality using R Commander

Several tests of normality can be accessed using R Commander by navigating to Statistics/Summaries menu tree and selecting Test of Normality (Figure 1).

Figure 1

R Commander Menu Tree



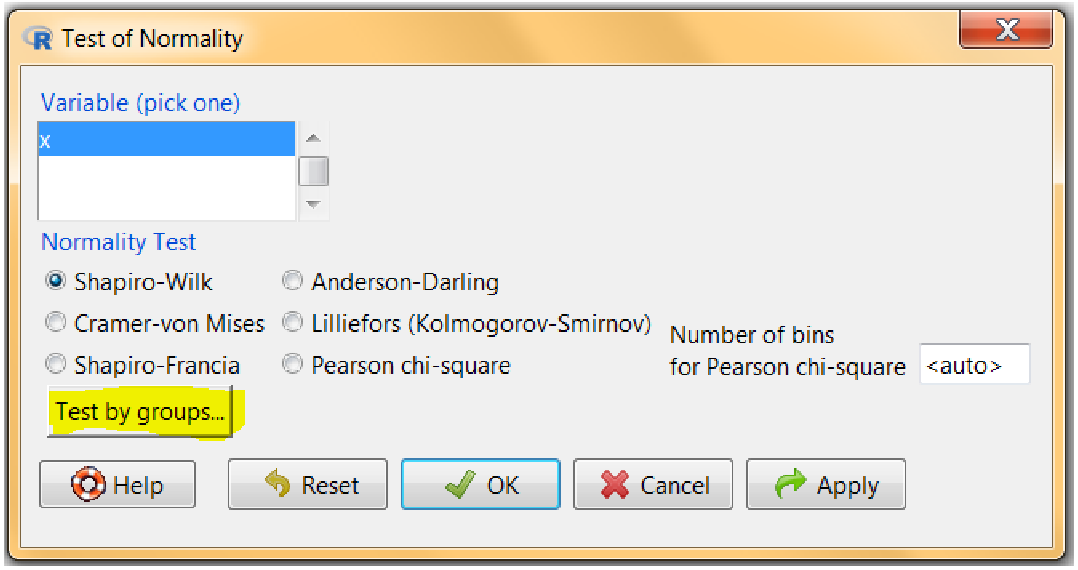
~~~
### `r as.cl

```{r echo=I
# include tl
knitr::opts_                dth=750, fig.height=8, fig.w
library(Rcmd
library(car)
library(RcmdrMisc)
```
~~~

Once this menu option is selected, a user can select from six tests: Shapiro-Wilk, Cramer-von Mises, Shapiro-Francia, Anderson-Darling, Lilliefors (Kolmogorov-Smirnov), and Pearson chi-square (not discussed in this paper). Pressing the Test by Group button, a user can perform tests of normality of a continuous variable by groups; a required step supporting an Independent Samples t-test or Analysis of Variances (ANOVA; Figure 2)

Figure 2

Test of Normality Option Box



Test of Normality

Variable (pick one)

x

Normality Test

● Shapiro-Wilk ○ Anderson-Darling

○ Cramer-von Mises ○ Lilliefors (Kolmogorov-Smirnov)

○ Shapiro-Francia ○ Pearson chi-square

Number of bins for Pearson chi-square <auto>

Performing Tests of Normality using R

The first exploration into examining departures from normality began with Karl Pearson in 1895. Since then, tests of normality available to today's researcher stem from the work of Cramer-von Mises (1928-1930) and Kolmogorov-Smirnov (1933). With the introduction of computers, enhancements have been made that have improved upon the original formulas for detecting deviations from normality. In the following pages, I'll illustrate how to implement some of these tests in R. Understand that many of these tests have been developed to focus on aspects of a specific distribution (e.g., skewness, kurtosis). A researcher should review the assumptions of each test before using.

Before testing, let's create some test data -

```
# set the seed value for replication purposes
set.seed(11022019)
# create a continuous variable  (N = 75) that has a normal distributed mean
# of 3.0 and a standard deviation of .25
normx <- rnorm(75, 3.0, .25)
# create a continuous variables (N = 75) that does not have a normal
# distribution (it follows a Poisson distribution with a Lambda of 10)
# Refer to your favorate statistics textbook to review a Poisson distribution
# and the definition of the Lambda
nonnormx <- rpois(75, 10)
```

Next, rather than show repetitive menu selections and output, I will execute the tests found in the R Commander menu on our two test variables using the *normalityTest* function found in the **RcmdrMisc** package from the R console. The following code checks to see if the **RcmdrMisc** package is installed in your system. If not, it loads it from the Cloud and imports it into your curren session.

```
if(!require(RcmdrMisc)){
    install.packages("RcmdrMisc")
}
```

```
## Loading required package: RcmdrMisc
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: sandwich
```

```
library(RcmdrMisc)
```

Shapiro-Wilk test

The Shapiro-Wilk (S-W) test is one of the most commonly used tests of normality. It has shown to be the most powerful test to identify deviations from normality in small sample sizes ($N < 30$). With larger samples though, this test is equal in power to the Anderson-Darling test and the Kolmogorov-Smirnov test with Lilliefors correction. An S-W W statistic range can be 0 to 1, with a score closer to 1 being associated with a normal distribution.

First, let's test the *normx* variable -

```
RcmdrMisc::normalityTest(normx, test = "shapiro.test")
```

```
##
##  Shapiro-Wilk normality test
##
## data:  normx
## W = 0.99185, p-value = 0.9167
```

With the p-value > .05, the implied null hypothesis that the variable is approximately normally distributed should not be rejected. Now, let's try the *nonnormx* variable -

```
RcmdrMisc::normalityTest(nonnormx, test = "shapiro.test")
```

```
##
##  Shapiro-Wilk normality test
##
## data:  nonnormx
## W = 0.96554, p-value = 0.03886
```

With the p-value < .05, the implied null hypothesis of an approximate normal distribution can be rejected.

Cramer-von Mises test

The Cramer-von Mises (CVM) test is based on a family of statistics using the empirical distribution function. The W statistic ranges from 0 to 1. The closer W is to 0, the closer the referenced distribution is to a normal distribution. Let's test both variables -

```
RcmdrMisc::normalityTest(normx, test = "cvm.test")
```

```
##
##  Cramer-von Mises normality test
##
## data:  normx
## W = 0.019451, p-value = 0.972
```

```
RcmdrMisc::normalityTest(nonnormx, test = "cvm.test")
```

```
##
##  Cramer-von Mises normality test
##
## data:  nonnormx
## W = 0.14042, p-value = 0.03102
```

The CVM test assessed the *norm* variable as approximately normally distributed and the *nonnormx* variables as not approximately normally distributed.

Shapiro-Francia test

The Shapiro-Francia (S-F) test was introduced in 1972, is a simplified S-W test. This simplification was developed because the S-W test was labor intensive before computers. However, the S-F test has been shown to be as equally powerful as the S-W test with large samples. Similar to the S-W test, the W statistic range can be 0 to 1, with a score closer to 1 being associated with a normal distribution.

```
RcmdrMisc::normalityTest(normx, test = "sf.test")
```

```
##
##  Shapiro-Francia normality test
##
## data:  normx
## W = 0.9918, p-value = 0.8472
```

```
RcmdrMisc::normalityTest(nonnormx, test = "sf.test")
```

```
##
##  Shapiro-Francia normality test
##
## data:  nonnormx
## W = 0.97216, p-value = 0.08882
```

Note that the S-F test results show that both *normx* and *nonnormx* are approximately normally distributed ($p = .089$ vs. $p > .05$). Could this be because of the simplications, or is this caused when a distribution is close to normal. A researcher in this situation should not rely solely on a test of normality but also on the review of graphs to make a judgement about a variable's distribution.

Anderson-Darling test

The Anderson-Darling (A-D) test, created in 1954, is a modification of the original K-S test. The A-D test is considered more powerful than the original K-S test in identifying deviations from a normal distribution; specifically in the tails. The A-D test can be used on any sample > 7. Similar to the CVM test, the A-D test range is 0 to 1, with a score close to 0 being indicative of a normal distribution. Let's try the same two variables again -

```
RcmdrMisc::normalityTest(normx, test = "ad.test")
```

```
##
##  Anderson-Darling normality test
##
## data:  normx
## A = 0.14082, p-value = 0.972
```

```
RcmdrMisc::normalityTest(nonnormx, test = "ad.test")
```

```
##
##  Anderson-Darling normality test
##
## data:  nonnormx
## A = 0.85387, p-value = 0.02684
```

The A-D test identified the *normx* variable as approximately normallys distributed and *nonnormx* as not approximately normally distributed.

Kolmogorov-Smirnov test with Lilliefors correction

The K-S test with Lilliefors correction, referred to as *Lilliefors* in R, R Commander, and by the author, is based on a 1967 modification of the K-S test. Lilliefors modified the K-S test by using the sample's distribution properties as part of the test, rather than specifying parameters before the test was performed. Similar to the CVM test and A-D test, the range of the test statistic D is 0 to 1, with a number closer to 0 being interpreted as a normal distribution.

Let's see how the Lilliefors test performs with the two test variables -

```
RcmdrMisc::normalityTest(normx, test = "lillie.test")
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  normx
## D = 0.043257, p-value = 0.9794
```

```
RcmdrMisc::normalityTest(nonnormx, test = "lillie.test")
```

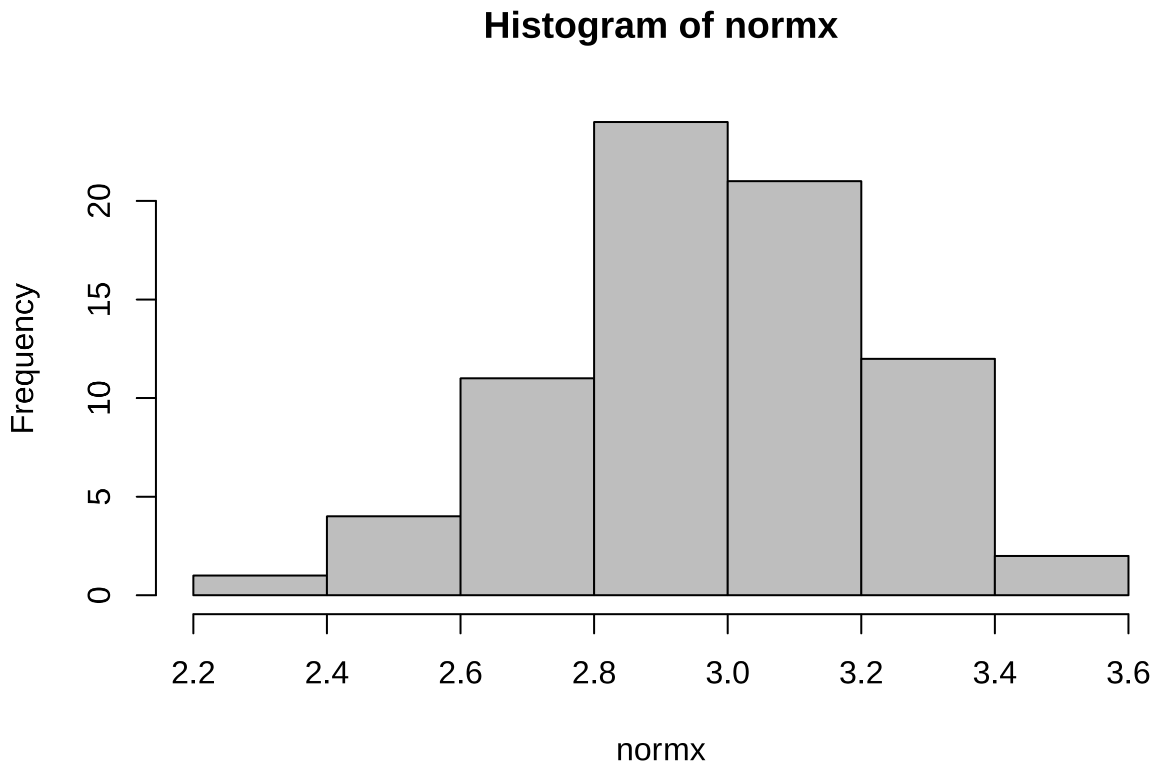
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  nonnormx
## D = 0.11523, p-value = 0.01522
```

Similar to the S-W test, CVM test, and A-D test, the Lilliefors test identified the *normx* variable as approximately normallys distributed and *nonnormx* as not approximately normally distributed..

Combining Tests of Normality with Graphs

Regardless of which test of normality is selected, researchers usually combine the results of the test with an analysis of a histogram with a normal distribution overlay and a review of a QQ Plot. First, let's examine *normx* variable from a graphical perspective via a simple histogram from the built-in R **graphics** library

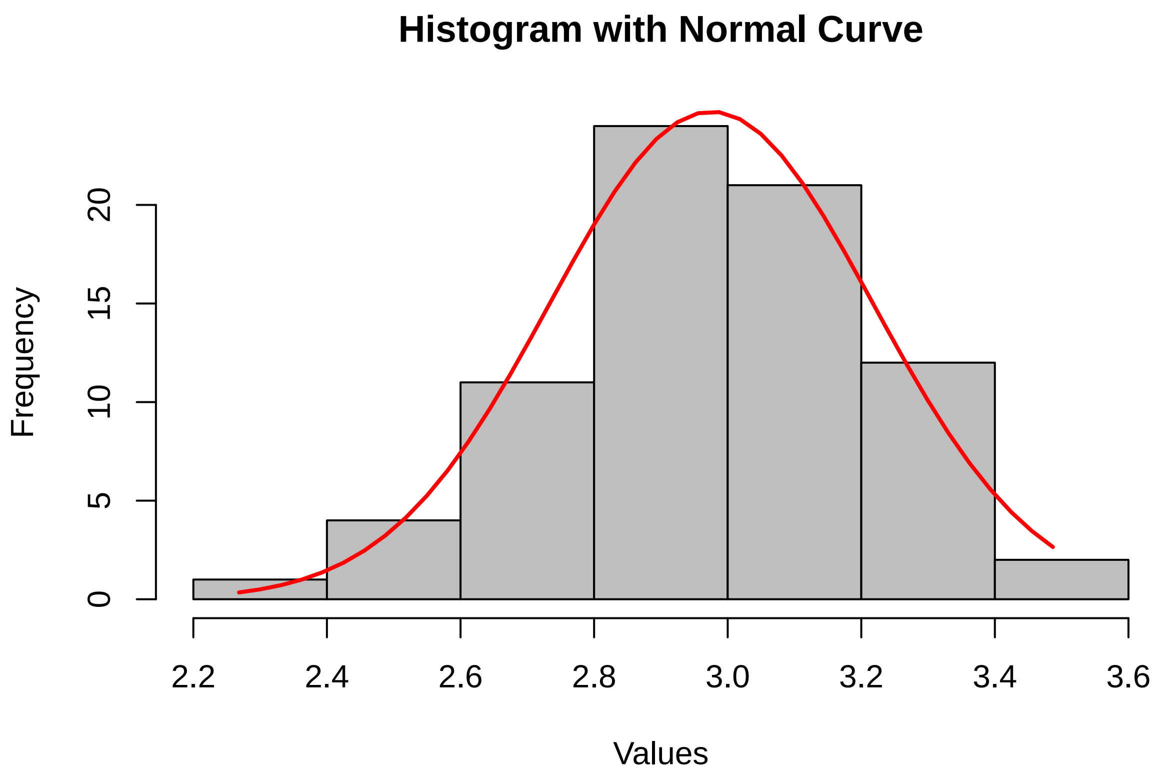
```r
# The bars in the histogram were colored for illustration purposes via the
# 'col=' option
graphics::hist(normx, col = "grey")
```



One can see that the *normx* variable is a bit high-peaked, but does the distribution approximate a normal distribution? Let's overlay a normal curve across the histogram -

```r
# Histogram was created as before only an x-axis label and a chart
# title was added
h<-hist(normx,
        col="grey",
        xlab="Values",
        main="Histogram with Normal Curve")
# store the minimum and maximum values of *normx* to the xfit variable
xfit<-seq(min(normx),max(normx),length=40)
# store the density function of *normx* variable to the yfit variable
yfit<-dnorm(xfit,mean=mean(normx),sd=sd(normx))
# overwrite the yfit variable with the difference between *xnorm* and a
# theoretical normal distribution
yfit <- yfit*diff(h$mids[1:2])*length(normx)
```

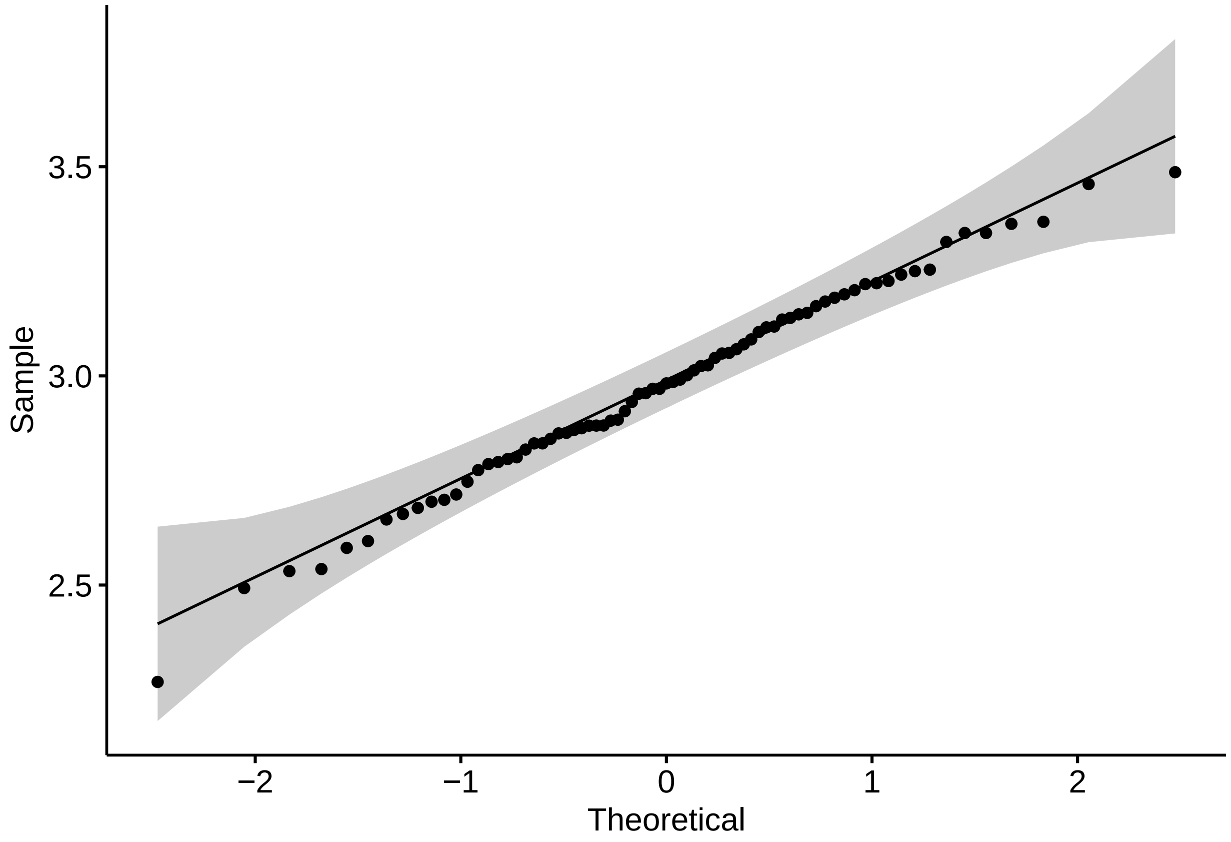
```
# add lines to the histogram
lines(xfit, yfit, col="red", lwd=2)
```



By adding a normal curve to a histogram, differences between values of the *normx* variable and a theoretical normal distribution can be assessed.

Another relevant graphical representation of data is a QQ plot. Let's examine the *normx* variable by reviewing a QQ Plot with 95% Confidence Intervals (CIs) using the ggqqplot() function found in the **ggpubr** package -

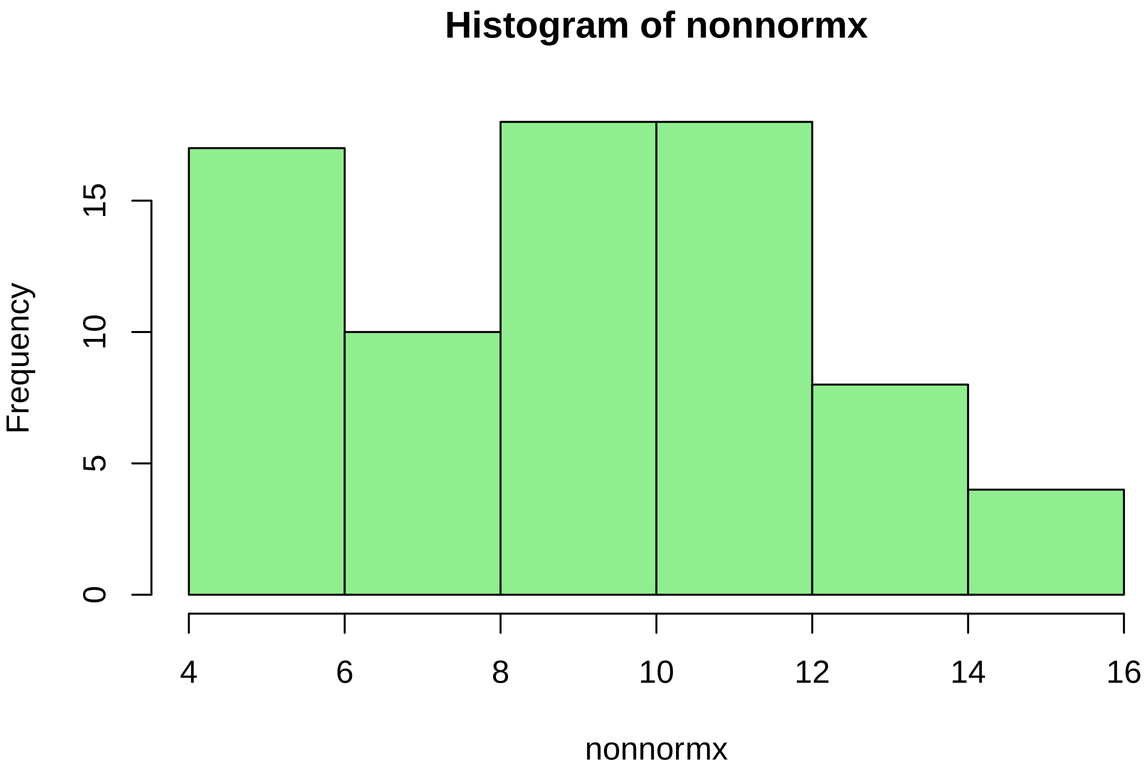
```
ggpubr::ggqqplot(normx)
```



Note that while all data points do not lay on a normal distribution line, those that vary are within 95% CI of a normal distribution.

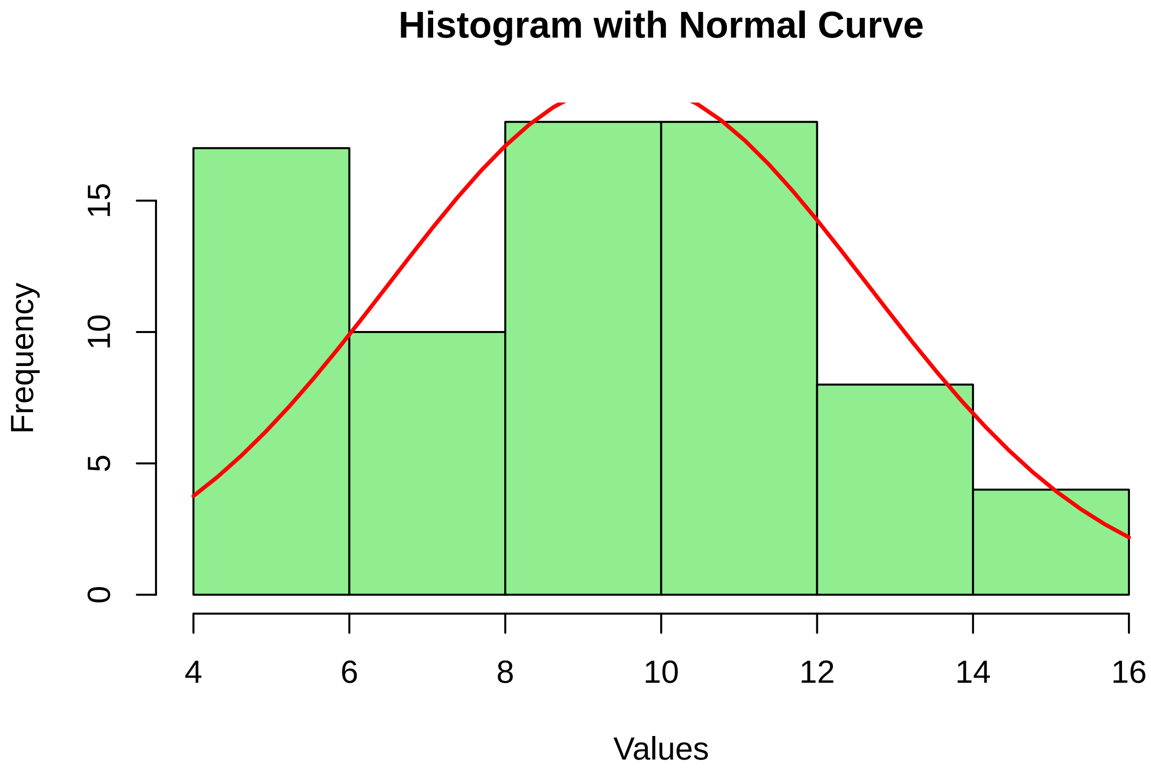
Now that we've viewed a normally distributed variable, let's follow the same approach with the *nonnormx* variable. First, a histogram with

```
# The bars in the histogram were colored for illustration purposes via the
# 'col=' option
graphics::hist(nonnormx, col = "lightgreen")
```



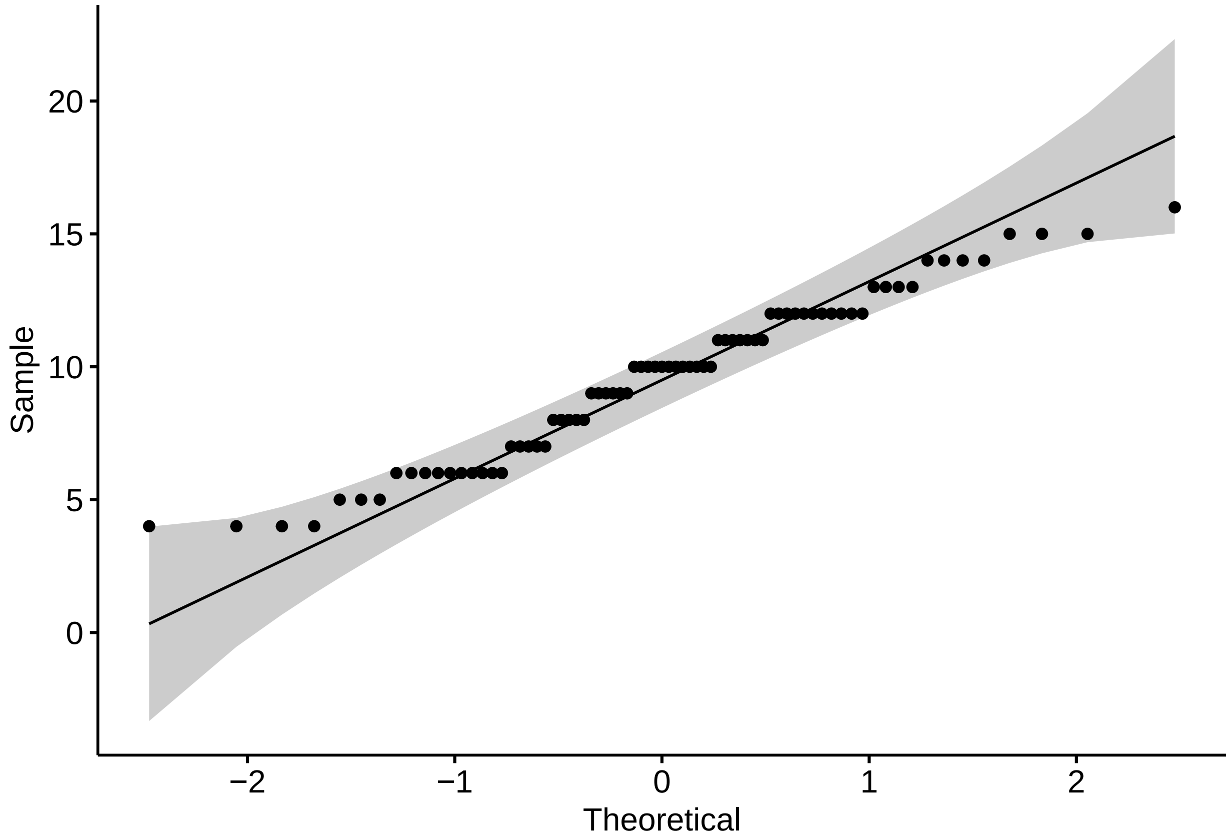
Note that *nonnormx* doesn't follow a normal distribution. In fact, it's right-skewed and potentially bi-modal. Let's overlay a normal curve -

```
h<-hist(nonnormx,
        col="lightgreen",
        xlab="Values",
        main="Histogram with Normal Curve")
xfit<-seq(min(nonnormx),max(nonnormx),length=40)
yfit<-dnorm(xfit,mean=mean(nonnormx),sd=sd(nonnormx))
yfit <- yfit*diff(h$mids[1:2])*length(nonnormx)
lines(xfit, yfit, col="red", lwd=2)
```



Note the difference between the bars and the normal curve in this graph versus the graph of a normal distribution; specifically in the left-tail. Let's examine a Q-Q plot of the *nonnormx* variable -

```
ggpubr::ggqqplot(nonnormx)
```



While nearly all points fall within the 95% CI of a normal distribution, note how very few data points are on the line. In fact, a learned researcher would conclude that the variable has 'levels;' 13 to be precise. A level-like variable can create errors when used in a test that requires a normally distributed variable. Thus, this is the contributing factor for *nonnormx* failing the series of tests of normality.

Reporting Normality Tests in APA format

Testing the distribution of a variable generally occurs in the Exploratory Data Analysis phase. Often, researchers make a simple statement about a variable being *approximately normally distributed* or *following a theoretical normal distribution*. In some cases, a researcher will display a graph for the reader to assess. In other cases, a statement is made with statistical results. In other cases, both graphs and statements or tables are used to report test results. If the statement approach is selected, reporting a test of normality could be written as -

> The Anderson-Darling test was used to assess deviations from normality. The result of the test on the *normx* variable was not significant, $A^2(30) = 0.168$, $p = .929$. Thus, the *normx* variable was considered normally distributed.