

Análisis de Normalidad: gráficos y contrastes de hipótesis

Joaquín Amat Rodrigo

Enero, 2016

Índice

Métodos gráficos	2
Histograma superpuesto con una distribución normal con la misma media y sd que los datos	2
Gráfico de cuantiles teóricos (Gráficos Q-Q).....	3
Métodos analíticos	4
Asimetría y curtosis.....	4
Contraste de hipótesis	4
Test de Shapiro-Wilk.....	4
Test de Kolmogorov-Smirnov y modificación de Lillefors.....	5
Test de normalidad de Jarque-Bera	6
Consecuencias de la falta de normalidad	6

Los análisis de normalidad, también llamados contrastes de normalidad, analizan cuanto difiere la distribución de los datos observados respecto a una distribución normal con la misma media y desviación típica.

Métodos gráficos

Histograma superpuesto con una distribución normal con la misma media y sd que los datos

Consiste en representar los datos mediante un histograma y superponer la curva que describe una distribución normal con la misma media y sd que muestran los datos.

```
library(ggplot2)
p <- ggplot(data = mtcars, aes(x = mpg))
p <- p + geom_histogram(aes(y = ..density..), binwidth = 0.9, colour = "white")
p <- p + stat_function(fun = dnorm, colour = "blue", args = list(mean = mean(mtcars$mpg), sd = sd(mtcars$mpg)))
p <- p + scale_x_continuous(name = "mpg")
p <- p + ggtitle("Histograma + curva normal teórica")
p <- p + theme_bw()
p
```

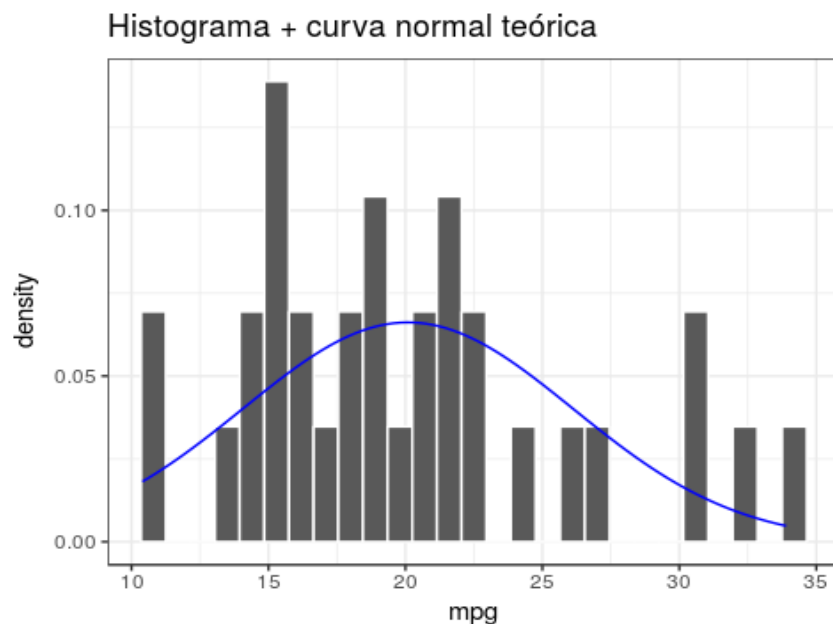
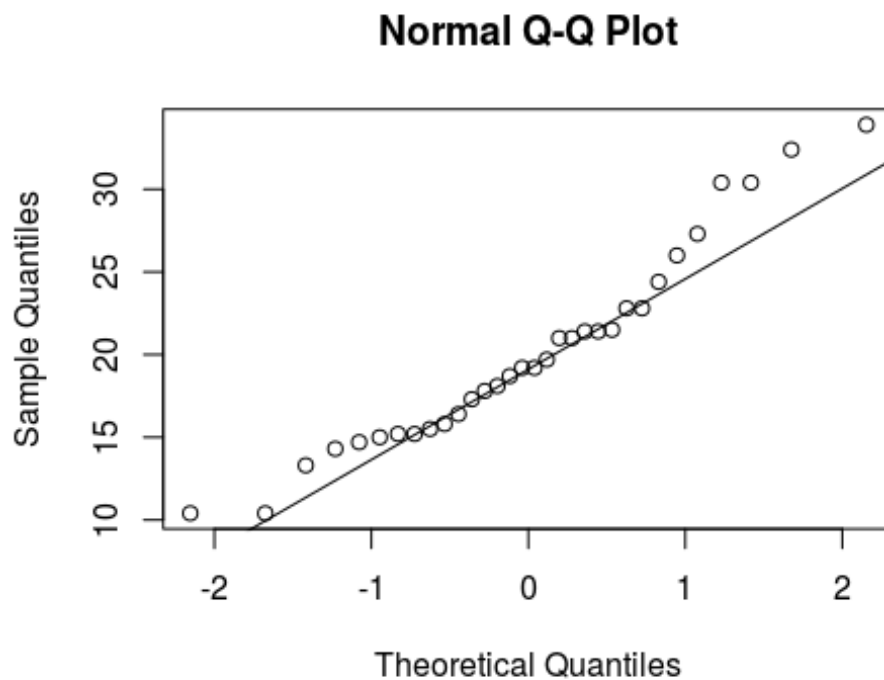


Gráfico de cuantiles teóricos (Gráficos Q-Q)

Consiste en comparar los cuantiles de la distribución observada con los cuantiles teóricos de una distribución normal. Cuanto más se asemejen los datos a una normal, más alineados están los puntos entorno a una recta.

```
qqnorm(mtcars$mpg)  
qqline(mtcars$mpg)
```



Métodos analíticos

Asimetría y curtosis

Un valor de curtosis y/o coeficiente de asimetría entre -1 y 1, es generalmente considerada una muy ligera desviación de la normalidad (Bulmer, 1979), (Brown, n.d.). Entre -2 y 2 hay una evidente desviación de la normal pero no extrema.

Contraste de hipótesis

A continuación se muestran los test de hipótesis más empleados para analizar la normalidad de los datos. En todos ellos se considera como hipótesis nula que los datos sí proceden de una distribución normal, la hipótesis alternativa que no lo hacen. El *p-value* de estos test indica la probabilidad de obtener una distribución como la observada si los datos realmente proceden de una población con una distribución normal. Es importante tener en cuenta que al tratarse de *p-values*, cuando mayor sea el tamaño de la muestra más estrictos son los test y es más fácil encontrar evidencias en contra de H_0 . Pero al mismo tiempo, cuanto mayor sea la muestra menos sensibles son los test paramétricos a la falta de normalidad.

Test de Shapiro-Wilk

Este test se emplea para contrastar normalidad cuando el tamaño de la muestra es menor de 50. Para muestras grandes es equivalente al test de kolmogorov-Smirnov.

```
shapiro.test(x = mtcars$mpg)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  mtcars$mpg  
## W = 0.94756, p-value = 0.1229
```

Test de Kolmogorov-Smirnov y modificación de Lillefors

El test de Kolmogorov-Smirnov permite estudiar si una muestra proviene de una población con una determinada distribución (media y desviación típica), no está limitado únicamente a la distribución normal. Se ejecuta con la función `ks.test()`.

```
ks.test(x = mtcars$mpg, "pnorm", mean(mtcars$mpg), sd(mtcars$mpg))

## Warning in ks.test(x = mtcars$mpg, "pnorm", mean(mtcars$mpg), sd(mtcars
## $mpg)): ties should not be present for the Kolmogorov-Smirnov test

##
## One-sample Kolmogorov-Smirnov test
##
## data:  mtcars$mpg
## D = 0.1263, p-value = 0.687
## alternative hypothesis: two-sided
```

A pesar de que continuamente se alude al test Kolmogorov-Smirnov como un test válido para contrastar la normalidad, esto no es del todo cierto. El Kolmogorov-Smirnov asume que se conocen la media y varianza poblacional, lo que en la mayoría de los casos no es posible. Esto hace que el test sea muy conservador y poco potente. Para solventar este problema se desarrolló una modificación del Kolmogorov-Smirnov conocida como test Lilliefors. El test Lilliefors asume que la media y varianza son desconocidas, estando especialmente desarrollado para testar la normalidad. Es la alternativa al test de Shapiro-Wilk cuando el número de observaciones es mayor de 50.

En R existe la función `lillie.test()` dentro del paquete *nortest* que contiene la modificación.

```
require("nortest")
lillie.test(x = mtcars$mpg)

##
## Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  mtcars$mpg
## D = 0.1263, p-value = 0.2171
```

Test de normalidad de Jarque-Bera

El test de Jarque-Bera no requiere estimaciones de los parámetros que caracterizan la normal. El estadístico de Jarque-Bera cuantifica que tanto se desvían los coeficientes de asimetría y curtosis de los esperados en una distribución normal.

Puede realizarse mediante la función `jarque.bera.test()` del paquete *tseries*.

```
require("tseries")
jarque.bera.test(x = mtcars$mpg)
```

```
##
##  Jarque Bera Test
##
## data:  mtcars$mpg
## X-squared = 2.2412, df = 2, p-value = 0.3261
```

Consecuencias de la falta de normalidad

El hecho de no poder asumir la normalidad influye principalmente en los test de hipótesis paramétricos (t-test, anova,...) y en los modelos de regresión provocando que:

- Los estimadores mínimo-cuadráticos no son eficientes (de mínima varianza).
- Los intervalos de confianza de los parámetros del modelo y los contrastes de significancia son solamente aproximados y no exactos.

El teorema del límite central requiere que la población o poblaciones de las que proceden las muestras tengan una distribución normal, no que la tengan las muestras. Si las muestras se distribuyen de forma normal, se puede aceptar que así lo hacen las poblaciones de origen. En el caso de que las muestras no se distribuyan de forma normal pero se tenga certeza de que las poblaciones de origen sí lo hacen, entonces los resultados obtenidos por los contrastes paramétricos sí son válidos. EL teorema del límite central, permite reducir los requerimientos de normalidad cuando las muestras suficientemente grandes.