Análisis de Normalidad: gráficos y contrastes de hipótesis

Joaquín Amat Rodrigo j.amatrodrigo@gmail.com

Enero, 2016

Tabla de contenidos

Versión PDF: Github

Introducción

Los análisis de normalidad, también llamados contrastes de normalidad, tienen como objetivo analizar cuánto difiere la distribución de los datos observados respecto a lo esperado si procediesen de una distribución normal con la misma media y desviación típica. Pueden diferenciarse tres estrategias: las basadas en representaciones gráficas, en métodos analíticos y en test de hipótesis.

Métodos gráficos

Histograma y curva normal

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

```
library(ggplot2)
ggplot(data = mtcars, aes(x = mpg)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
                args = list(mean = mean(mtcars$mpg),
                            sd = sd(mtcars$mpg))) +
  ggtitle("Histograma + curva normal teórica") +
  theme_bw()
```

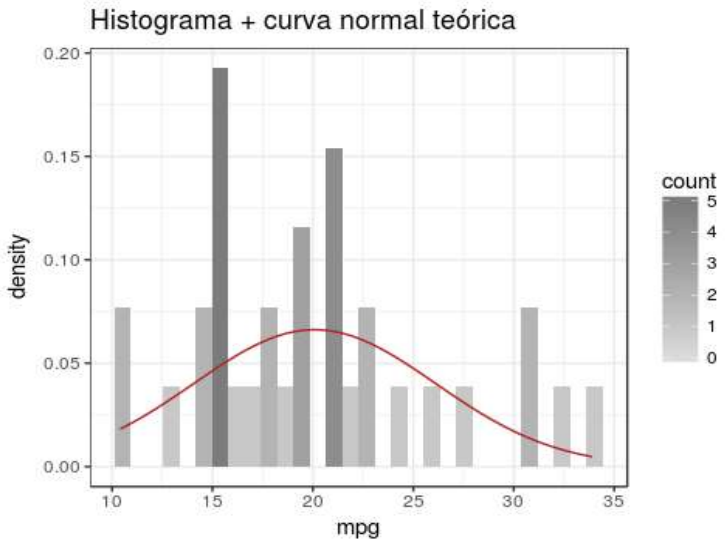
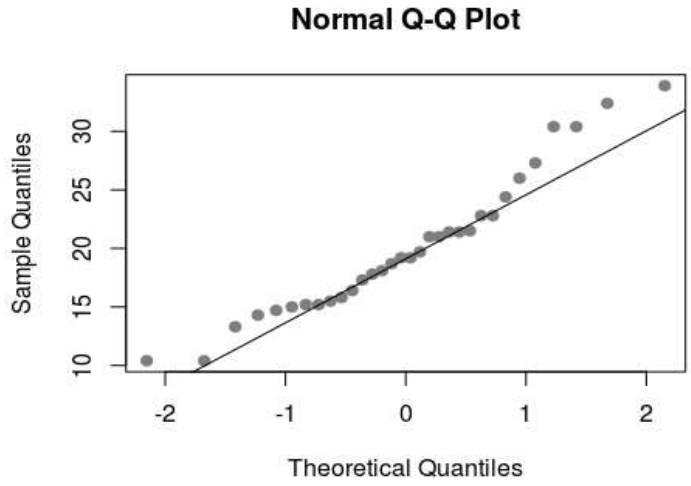


Gráfico de cuantiles teóricos (Gráficos Q-Q)

Consiste en comparar los cuantiles de la distribución observada con los cuantiles teóricos de una distribución normal con la misma media y desviación estándar que los datos. Cuanto más se aproximen los datos a una normal, más alineados están los puntos entorno a la recta.

```
qqnorm(mtcars$mpg, pch = 19, col = "gray50")
qqline(mtcars$mpg)
```



Métodos analíticos

Asimetría y curtosis

Un valor de curtosis y/o coeficiente de asimetría entre -1 y 1, es generalmente considerada una 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 y como 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 proceden realmente de una población con una distribución normal.

Cuando estos test se emplean con la finalidad de verificar las condiciones de métodos paramétricos, por ejemplo un t-test o un ANOVA, es importante tener en cuenta que, al tratarse de *p-values*, cuanto mayor sea el tamaño de la muestra más poder estadístico tienen y más fácil es encontrar evidencias en contra de la H_0 (normalidad). Al mismo tiempo, cuanto mayor sea el tamaño de la muestra, menos sensibles son los métodos paramétricos a la falta de normalidad. Por esta razón, es importante no basar las conclusiones únicamente en el *p-value* del test, sino también considerar la representación gráfica y el tamaño de la muestra.

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 procede 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 conoce 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 contrastar la normalidad. Es la alternativa al test de *Shapiro-Wilk* cuando el número de observaciones es mayor de 50. La función lillie.test() del paquete nortest permite aplicarlo.

```
library("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 calcularse mediante la función jarque.bera.test() del paquete tseries.

```
library("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. Las principales consecuencias de la no normalidad son:

- 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.