

# Análisis de Normalidad

Paúl Arévalo, Esteban Vizhñay

2024-07-03

El análisis de normalidad tiene como objetivo analizar si la distribución de frecuencias relativas de una variable cuantitativa en un población se aproxima a una Distribución Normal con la misma media y desviación estándar que la variable.

## Métodos Gráficos

Para estas demostraciones se creará una distribución randomica:

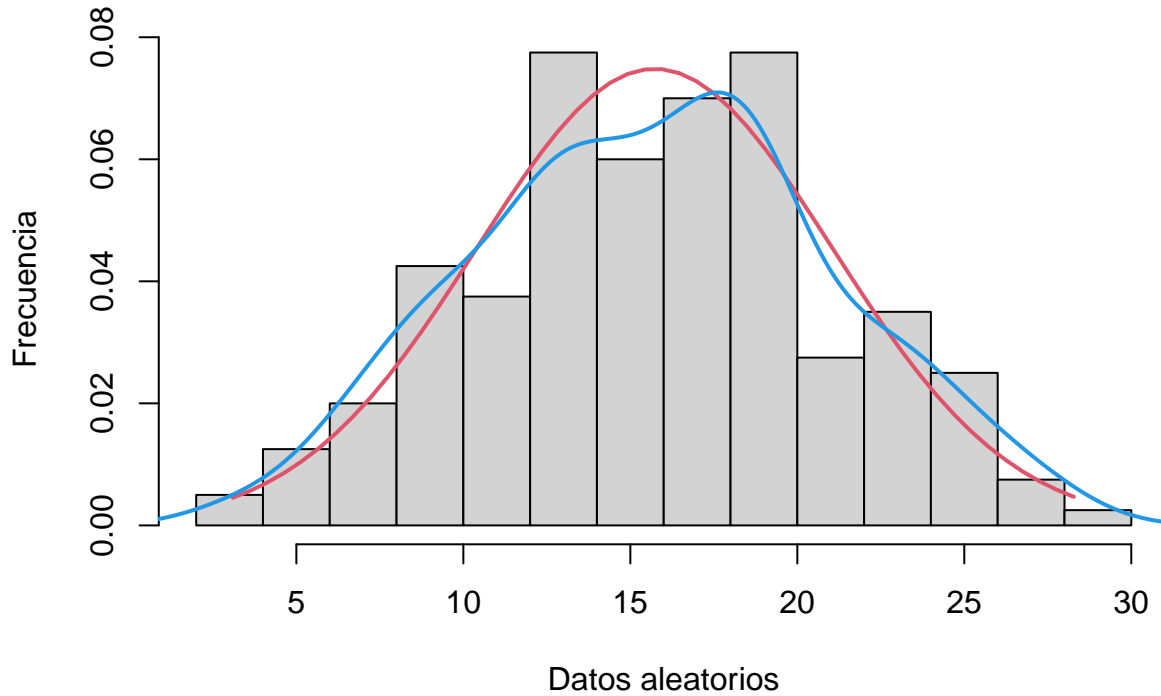
```
set.seed(12345)
x <- rnorm(n = 200, mean = 15, sd = 5)
x
```

```
## [1] 17.927644 18.547330 14.453483 12.732514 18.029437 5.910220 18.150493
## [8] 13.619079 13.579201 10.403390 14.418761 24.086560 16.853139 17.601082
## [15] 11.247340 19.084499 10.568212 13.342112 20.603563 16.493618 18.898110
## [22] 22.278925 11.778358 7.234313 7.011452 24.025488 12.591763 18.101899
## [29] 18.060617 14.188445 19.059366 25.984168 25.245952 23.162228 16.271356
## [36] 17.455941 13.379567 6.689749 23.838669 15.129005 20.642554 3.098210
## [43] 9.698672 19.685703 19.272259 22.303647 7.934506 17.837016 17.915938
## [50] 8.466006 12.298070 24.738463 15.267951 16.758314 11.645117 16.389768
## [57] 18.455856 19.118977 25.725325 3.265280 15.747960 8.287343 17.766515
## [64] 22.949814 12.065602 5.838113 19.440697 22.967442 17.584273 8.521642
## [71] 15.273078 11.076753 9.753236 26.652560 22.013527 19.713004 19.131291
## [78] 10.942298 17.381241 20.106292 18.226915 20.215718 13.478154 27.385555
## [85] 19.856103 24.335496 18.360212 13.460233 17.682619 19.124350 10.180493
## [92] 10.724587 24.434735 13.040903 10.096835 18.436661 12.474782 25.788599
## [99] 12.001012 11.527267 16.119627 9.218883 17.112093 8.376224 15.705422
## [106] 12.319760 13.441970 22.780548 12.759834 16.605618 8.849139 8.379707
## [113] 21.306211 21.596159 14.596231 12.474551 14.739232 18.144303 25.900012
## [120] 14.654913 22.724318 21.607260 16.610758 22.654776 12.893802 9.205895
## [127] 5.773159 20.786626 4.382251 9.019842 23.210960 19.418274 17.624379
## [134] 9.076705 28.278941 9.760431 9.944387 18.344608 15.645886 12.887116
## [141] 9.298679 8.531424 12.026506 7.495930 15.079278 17.700848 7.263540
## [148] 19.248265 19.480066 15.693455 6.903358 17.741990 15.976411 10.967510
## [155] 14.456879 13.745267 23.496733 13.278506 15.338860 11.747151 12.561807
## [162] 16.515756 13.790130 12.591332 10.040986 13.596754 18.165087 8.800908
## [169] 23.821570 14.881601 15.999602 21.735964 15.180367 19.122906 6.486641
## [176] 17.404751 27.417750 17.006825 16.075886 5.921438 10.441303 14.754777
## [183] 12.973063 20.651909 19.077324 15.382088 22.268737 16.870605 14.145480
## [190] 12.488936 17.717611 12.474070 18.933979 16.504747 21.551120 18.992169
## [197] 19.254302 12.782160 12.766126 15.066525
```

## Histograma de frecuencias

```
h <- hist(x, main = "Histograma de datos con distribución normal", xlab = "Datos  
↪ aleatorios", ylab = "Frecuencia", probability = TRUE)  
x2 <- seq(min(x), max(x), length = 50)  
normal <- dnorm(x2, mean = mean(x), sd = sd(x))  
lines(x2, normal, col = 2, lwd = 2)  
lines(density(x), col = 4, lwd = 2)
```

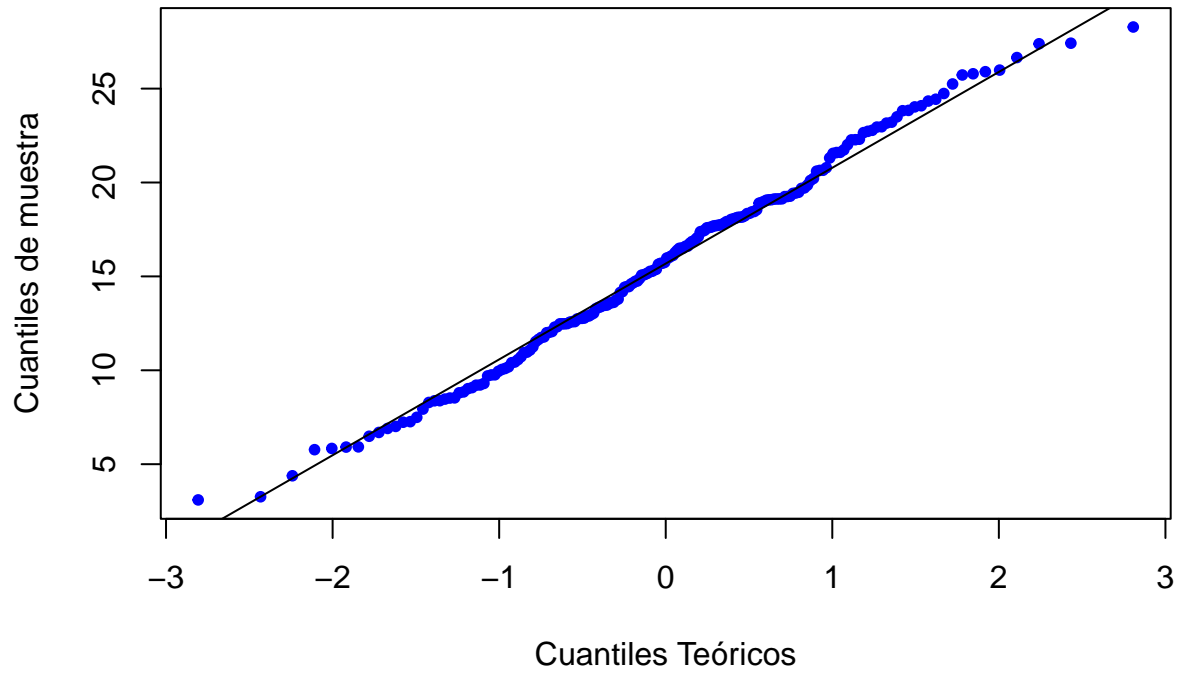
### Histograma de datos con distribución normal



## Gráfico de cuantiles teóricos (Gráfico Q-Q)

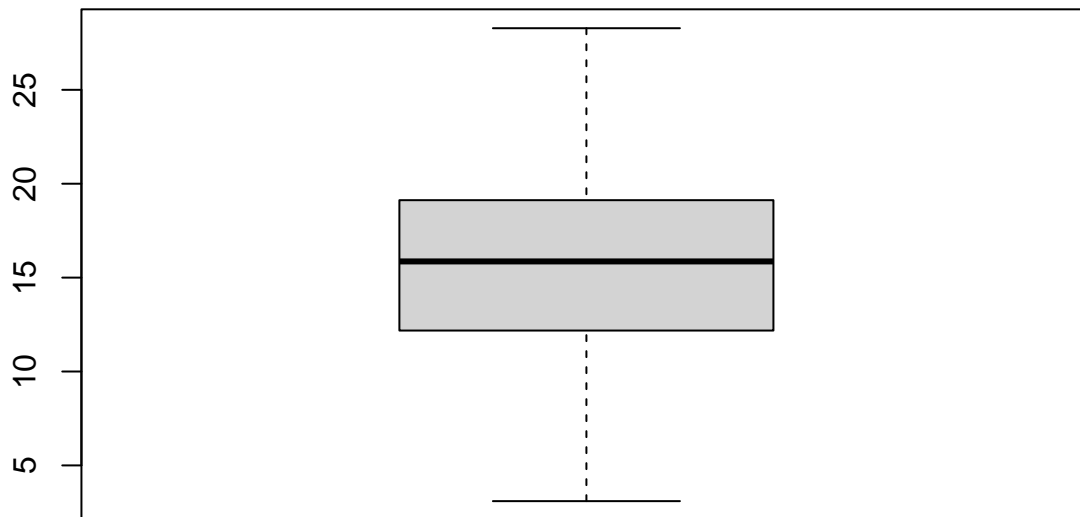
```
qqnorm(x, main = "Gráfico Q-Q Normal", xlab = "Cuantiles Teóricos", ylab = "Cuantiles de  
↪ muestra", pch = 20, col = "blue")  
qqline(x)
```

## Gráfico Q-Q Normal



## Diagrama de Caja

```
boxplot(x)
```



## Métodos Analíticos

### Asimetria

Existen tres tipos de asimetría

Para el analisis del mismo vamos a definir los siguientes rangos:

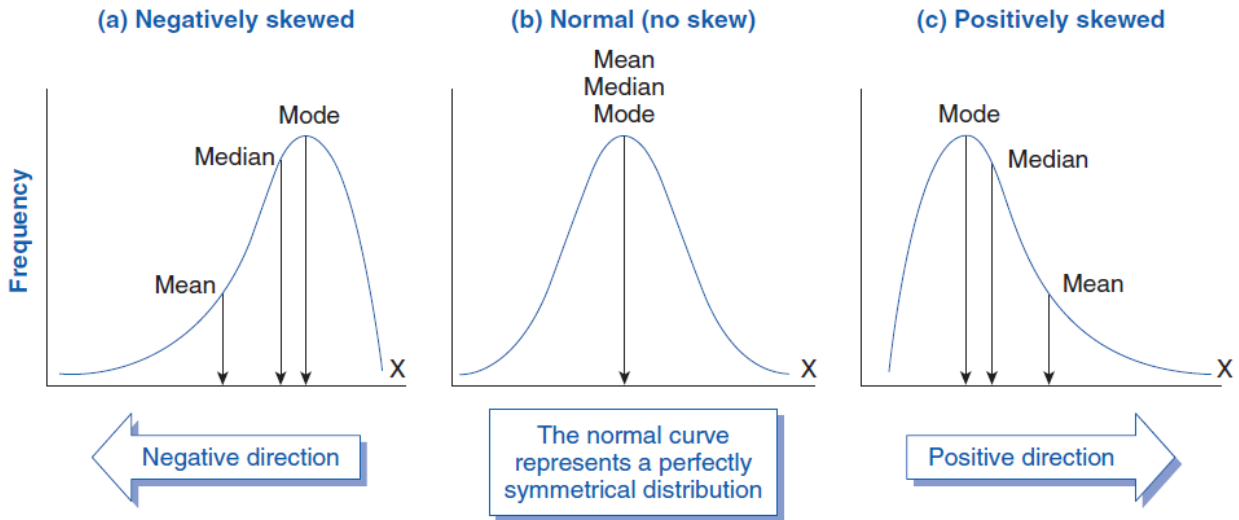


Figure 1: Comparativa de Normalidad

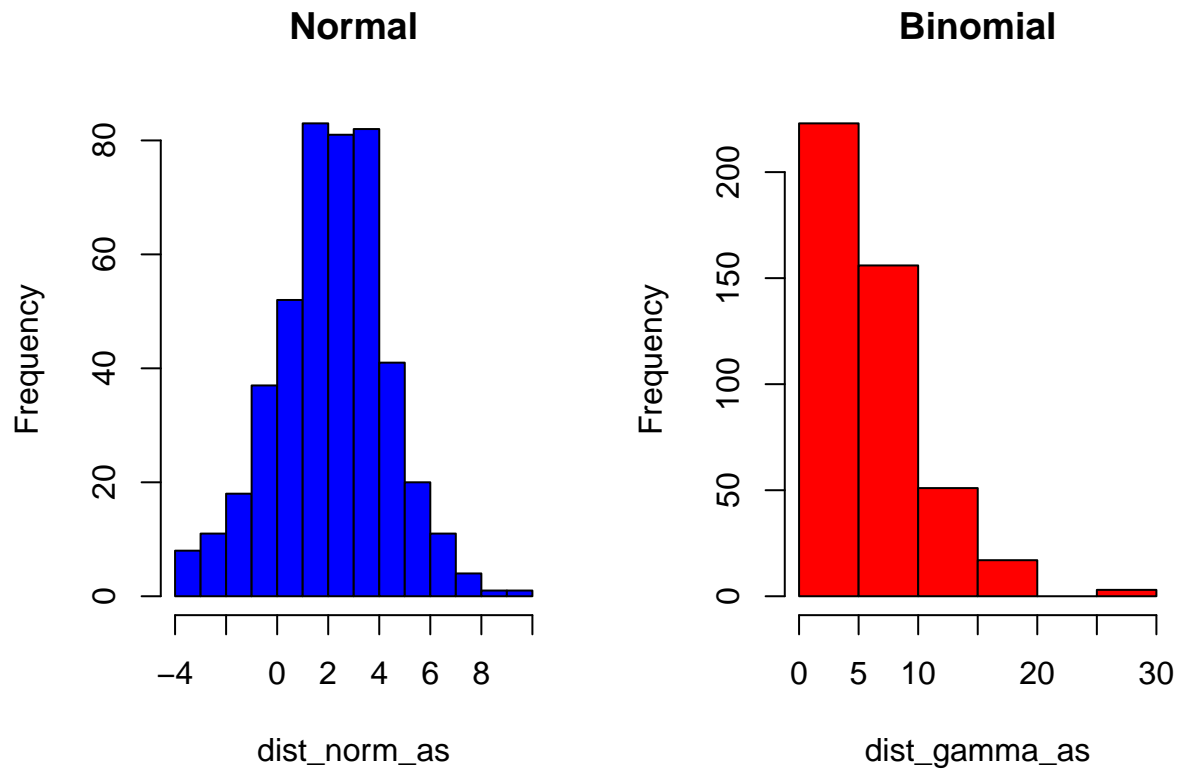
- El valor se encuentra entre -0.5 y 0.5. La distribución es aproximadamente sesgada.
- El valor se encuentra entre -1 a -5 o 1 a 5. La distribución es ligeramente sesgada.
- El valor es menor que -1 o mayor a 1. La distribución es moderadamente sesgada pero aceptable.
- Si el valor supera -2 o 2. La distribución no es normal

```
dist_norm_as <- rnorm(n = 450, mean = 2, sd = 2.3)
dist_gamma_as <- rgamma(n = 450, shape = 2, scale = 3)
```

```
# Configurar la disposición de los gráficos
par(mfrow = c(1, 2))

# Dibujar el primer histograma
hist(dist_norm_as, main = "Normal", col = "blue")

# Dibujar el segundo histograma
hist(dist_gamma_as, main = "Binomial", col = "red")
```



```
library(moments)

print(skewness(dist_norm_as))

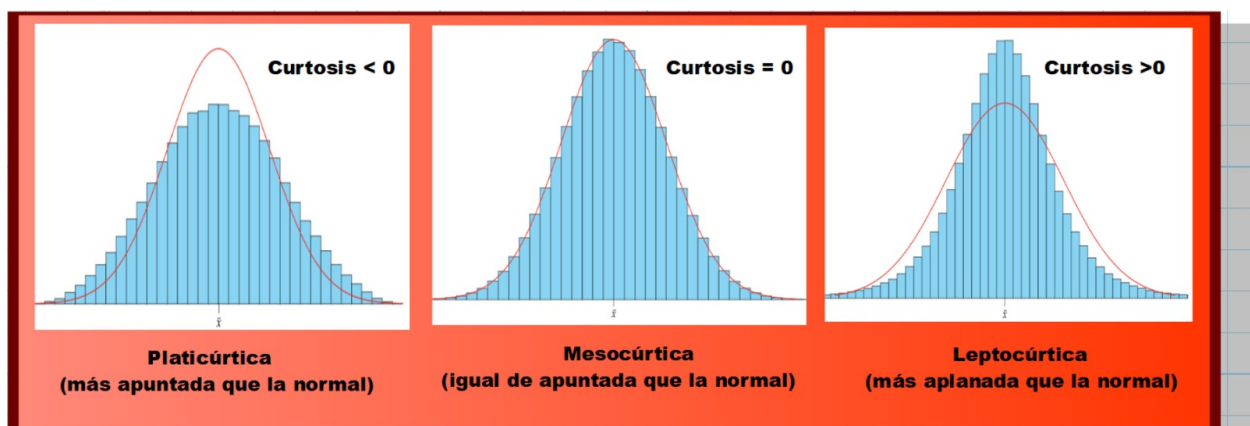
## [1] -0.08919541

print(skewness(dist_gamma_as))

## [1] 1.392248
```

## Curtosis

Permite medir el grado de apuntamiento o achatamiento de la distribución de frecuencia, respecto a la curva de la distribución normal que tiene coeficiente igual a 0.



```
print(kurtosis(dist_norm_as))  
  
## [1] 3.27443  
print(kurtosis(dist_gamma_as))  
  
## [1] 5.905284
```

## Contraste de hipótesis

```
library("nortest")
```

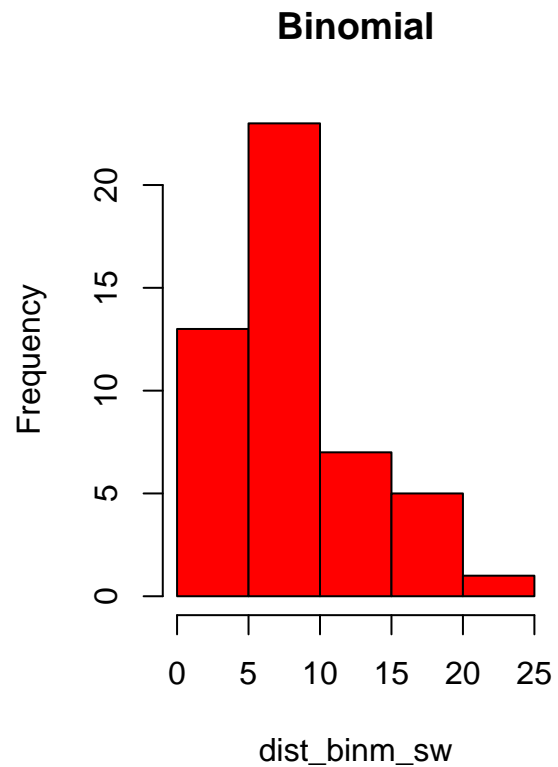
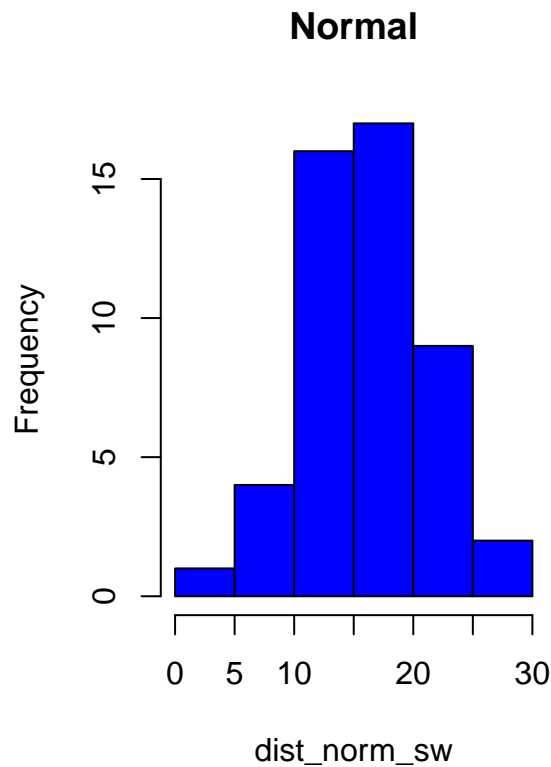
En todas las pruebas a realizar se considera como hipótesis nula que los datos sí proceden de una distribución normal y como hipótesis alternativa que no lo hacen. El *p-value* de estos test indica la probabilidad de obtener una distribución como la observada si los datos proceden realmente de una población con una distribución normal.

*El nivel de significancia es 0.05*

### Prueba de Shapiro-Wilk

La muestra debe ser menor a 50 datos.

```
dist_norm_sw <- rnorm(n = 49, mean = 15, sd = 5)  
dist_binm_sw <- rbinom(n = 49, size = 5, prob = 0.35)  
  
# Configurar la disposición de los gráficos  
par(mfrow = c(1, 2))  
  
# Dibujar el primer histograma  
hist(dist_norm_sw, main = "Normal", col = "blue")  
  
# Dibujar el segundo histograma  
hist(dist_binm_sw, main = "Binomial", col = "red")
```



```
print(shapiro.test(x = dist_norm_sw))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dist_norm_sw
## W = 0.99103, p-value = 0.9695
```

```
print(shapiro.test(x = dist_binm_sw))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dist_binm_sw
## W = 0.94026, p-value = 0.01513
```

## Prueba de Anderson-Darling

El número de muestras tiene que ser mayor a 7

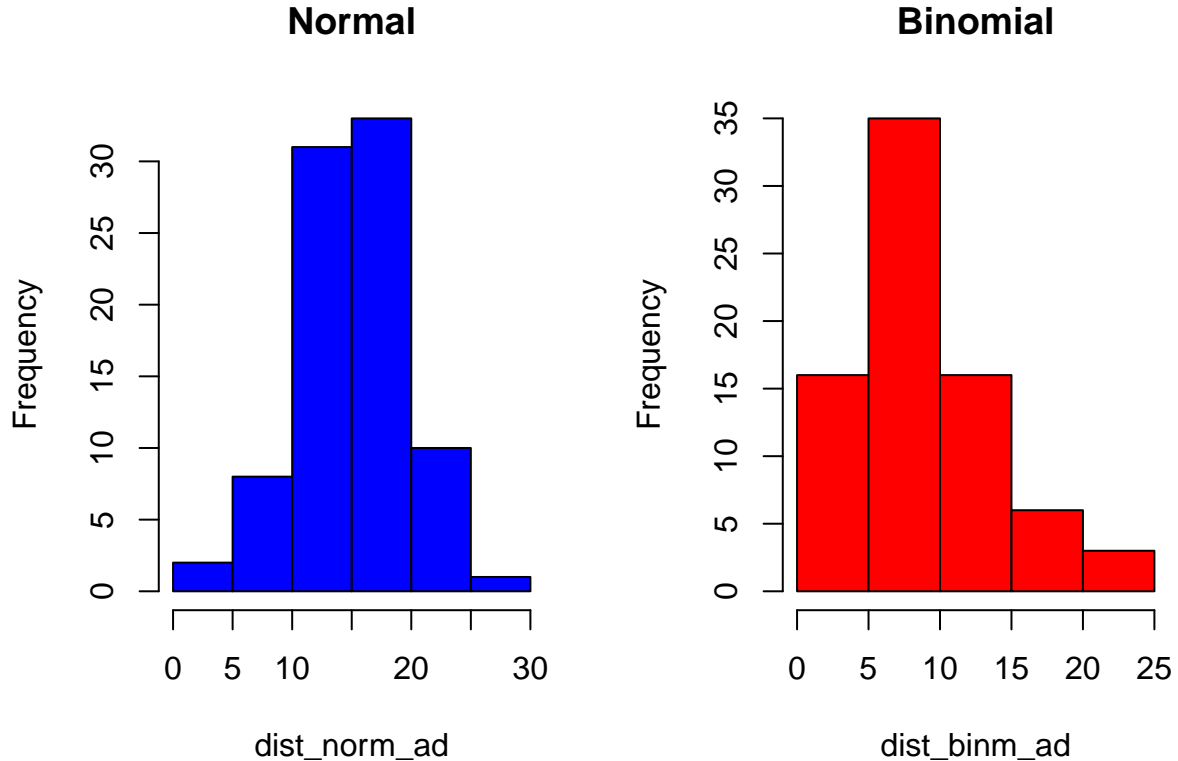
```
dist_norm_ad <- rnorm(n = 85, mean = 15, sd = 5)
dist_binm_ad <- rbinom(n = 76, size = 5, prob = 0.35)
```

```
# Configurar la disposición de los gráficos
par(mfrow = c(1, 2))
```

```
# Dibujar el primer histograma
hist(dist_norm_ad, main = "Normal", col = "blue")
```

```
# Dibujar el segundo histograma
```

```
hist(dist_binm_ad, main = "Binomial", col = "red")
```



```
print(ad.test(x = dist_norm_ad))
```

```
##
## Anderson-Darling normality test
##
## data: dist_norm_ad
## A = 0.22851, p-value = 0.8053
```

```
print(ad.test(x = dist_binm_ad))
```

```
##
## Anderson-Darling normality test
##
## data: dist_binm_ad
## A = 1.329, p-value = 0.001768
```

## Prueba de Cramer-von Mises

Útil para muestras pequeñas

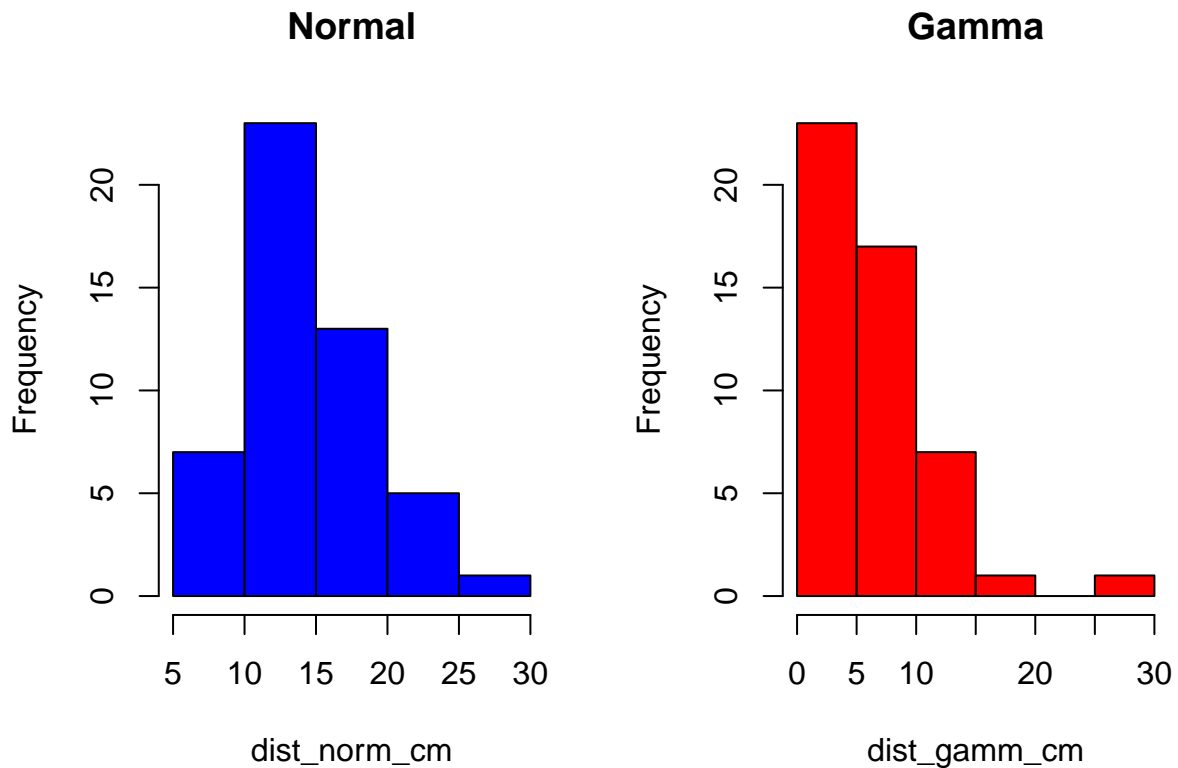
```
dist_norm_cm <- rnorm(n = 49, mean = 15, sd = 5)
dist_gamm_cm <- rgamma(n = 49, shape = 2, scale = 3)
```

```
# Configurar la disposición de los gráficos
par(mfrow = c(1, 2))
```

```
# Dibujar el primer histograma
hist(dist_norm_cm, main = "Normal", col = "blue")
```



```
# Dibujar el segundo histograma
hist(dist_gamm_cm, main = "Gamma", col = "red")
```



```
print(cvm.test(x = dist_norm_cm))

##
##  Cramer-von Mises normality test
##
## data:  dist_norm_cm
## W = 0.078312, p-value = 0.2128

print(cvm.test(x = dist_gamm_cm))

##
##  Cramer-von Mises normality test
##
## data:  dist_gamm_cm
## W = 0.23438, p-value = 0.001879
```

## Prueba Lilliefors

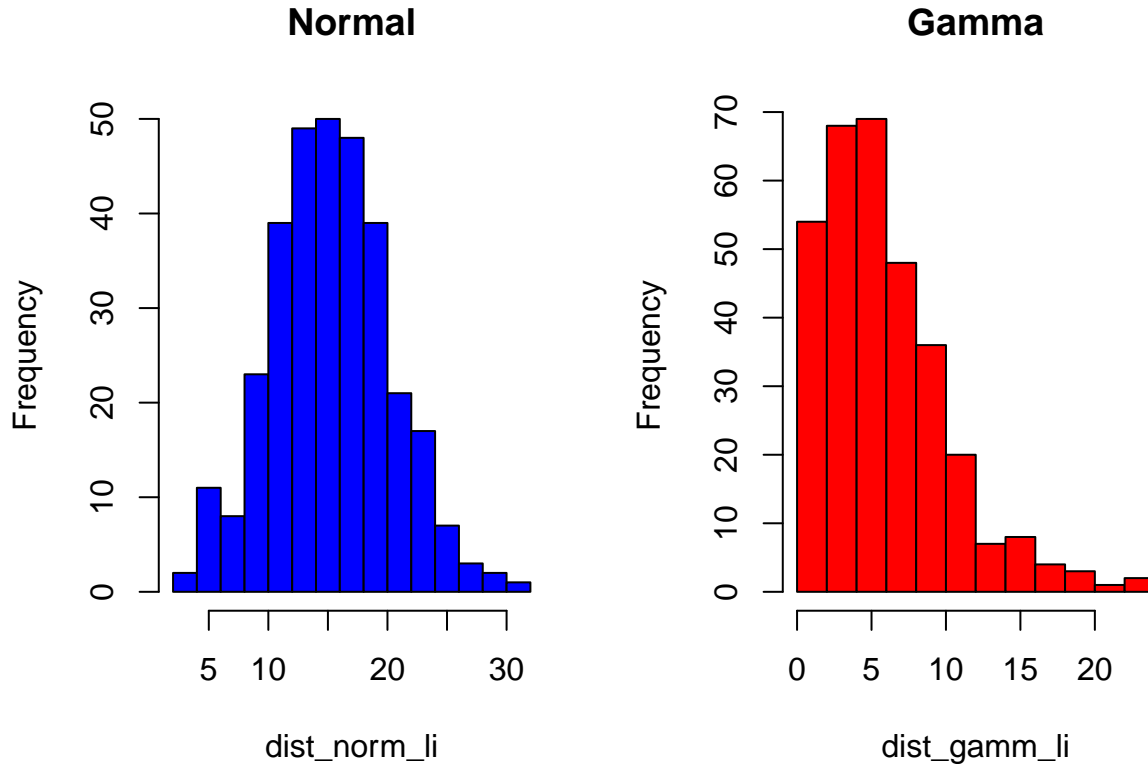
Número de observaciones mayor a 50

```
dist_norm_li <- rnorm(n = 320, mean = 15, sd = 5)
dist_gamm_li <- rgamma(n = 320, shape = 2, scale = 3)

# Configurar la disposición de los gráficos
par(mfrow = c(1, 2))
```

```
# Dibujar el primer histograma
hist(dist_norm_li, main = "Normal", col = "blue")

# Dibujar el segundo histograma
hist(dist_gamm_li, main = "Gamma", col = "red")
```



```
print(lillie.test(x = dist_norm_li))

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dist_norm_li
## D = 0.021506, p-value = 0.9744

print(lillie.test(x = dist_gamm_li))

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dist_gamm_li
## D = 0.097141, p-value = 1.066e-07
```

## Prueba Pearson chi-square

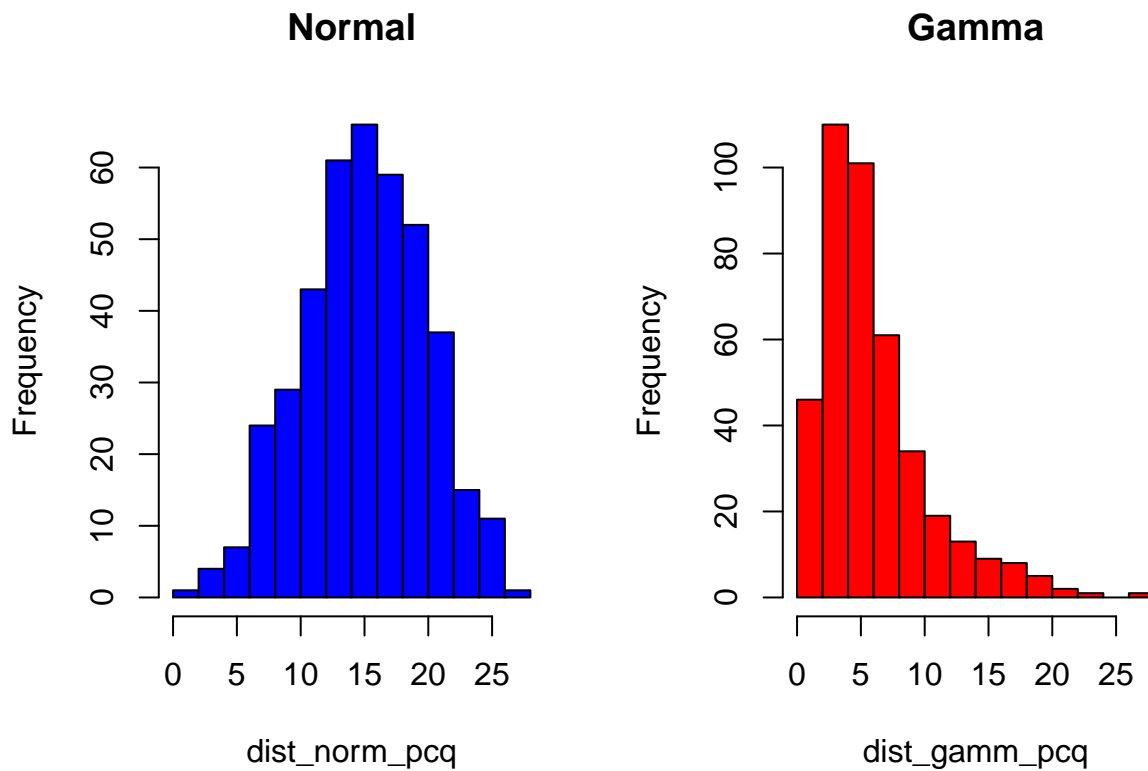
Basa en una distribución Chi Cuadrada

```
dist_norm_pcq <- rnorm(n = 410, mean = 15, sd = 5)
dist_gamm_pcq <- rgamma(n = 410, shape = 2, scale = 3)
```

```
# Configurar la disposición de los gráficos
par(mfrow = c(1, 2))

# Dibujar el primer histograma
hist(dist_norm_pcq, main = "Normal", col = "blue")

# Dibujar el segundo histograma
hist(dist_gamm_pcq, main = "Gamma", col = "red")
```



```
print(pearson.test(x = dist_norm_pcq))
```

```
##
## Pearson chi-square normality test
##
## data: dist_norm_pcq
## P = 24.868, p-value = 0.2065
```

```
print(pearson.test(x = dist_gamm_pcq))
```

```
##
## Pearson chi-square normality test
##
## data: dist_gamm_pcq
## P = 133.92, p-value < 2.2e-16
```