Problema No. 1

2023-03-13

## Elaborado por:

Mario Felipe Aristizabal Trujillo

Juan Carlos Villalba Acevedo

## Presentado a:

Profesora Delia Ortega Lenis

## Teorema del Límite Central

El Teorema del Límite Central es uno de los más importantes en la inferencia estadística y habla sobre la convergencia de los estimadores como la proporción muestral a la distribución normal. Algunos autores afirman que esta aproximación es bastante buena a partir del umbral n>30.

A continuación se describen los siguientes pasos para su verificación:

### a. Realice una simulación en la cual genere una población de N=1000 (Lote), donde el porcentaje de individuos (supongamos plantas) enfermas sea del 50%.

#### Rta.

Poblacion = rep(c(0,1), each = 500)  
Poblacion

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [260] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [297] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [519] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [556] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [593] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [630] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [667] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [704] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [741] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [778] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [815] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [852] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [889] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [926] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [963] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1000] 1

### b. Genere una función que permita:

* Obtener una muestra aleatoria de la población y
* Calcule el estimador de la proporción muestral pˆ para un tamaño de muestra dado n.

#### Rta.

n = 30  
Muestra = sample(Poblacion, n)  
Muestra

## [1] 0 0 1 1 1 1 1 0 1 1 1 0 0 0 1 1 1 0 1 0 1 1 1 1 1 0 0 0 0 0

Propor = sum(Muestra)/n  
Propor

## [1] 0.5666667

### c. Repita el escenario anterior (b) n=500 veces y analice los resultados en cuanto al comportamiento de los 500 resultados del estimador pˆ. ¿Qué tan simétricos o sesgados son los resultados obtenidos? y ¿qué se puede observar en cuanto a la variabilidad?. Realice en su informe un comentario sobre los resultados obtenidos.

#### Rta.

prop <- function(n) {  
 Muestra = sample(Poblacion, n)  
 Propor = sum(Muestra)/n  
 return(Propor)  
}  
  
Proporcion <- sapply(rep(30,500), prop)

Histograma:

hist(Proporcion, col="#BC2B6A")

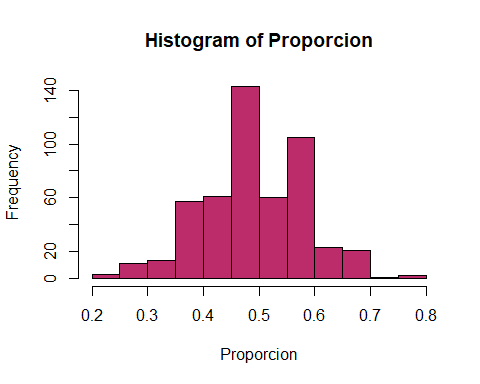
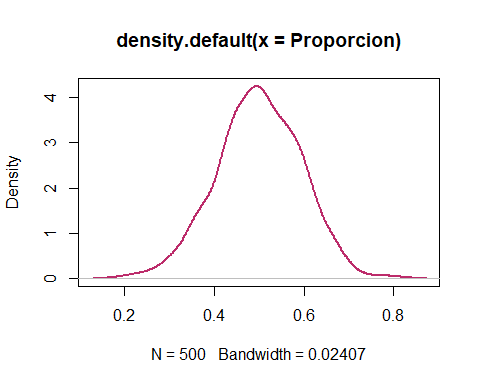


Gráfico de Densidad:

plot(density(Proporcion), col="#BC2B6A", lwd=2)



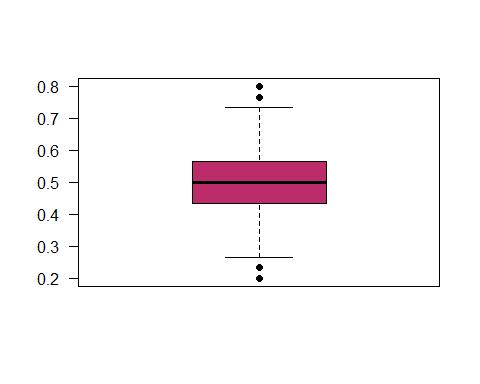
Resumen:

psych::describe(Proporcion)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 500 0.5 0.09 0.5 0.5 0.1 0.2 0.8 0.6 -0.12 0.06 0

Boxplot:

boxplot(Proporcion, pch=19, col="#BC2B6A", las=1)



Análisis de resultados:

En primera instancia podemos decir que la estimación de la proporción se comporta como una variable aleatoria ya que cada vez que la calculamos obtenemos un resultado diferente. Tanto en el histograma como en la gráfica de la función de densidad de la proporción podemos ver que los datos tienen una forma simétrica, muy parecida a la gráfica de una distribución normal. Otro dato importante es el valor de la desviación estándar, medida de dispersión que nos indica cuanto varían los valores de la muestra con respecto a su media, en este caso, el valor de sd es de 0.09, valor muy pequeño que nos permite inferir que los datos están muy cercanos de la media. Finalmente, si observamos el valor de la media y la mediana nos damos cuenta que son iguales, característica de una distribución normal.

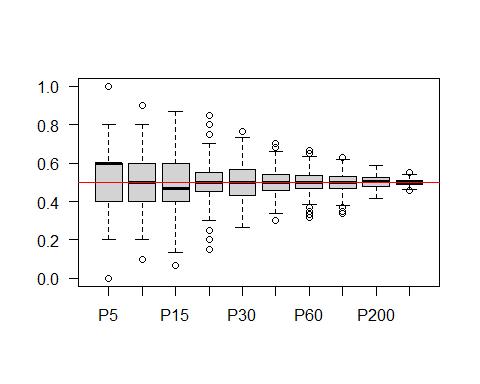
### d. Repita los puntos b y c para tamaños de muestra n=5, 10, 15, 20, 30, 50, 60, 100, 200, 500. Compare los resultados obtenidos para los diferentes tamaños de muestra en cuanto a la normalidad. Utilice pruebas de bondad y ajuste (shapiro wilks :shspiro.test()) y métodos gráficos (grafico de normalidad: qqnorm()). Comente en su informe los resultados obtenidos

#### Rta.

P5 <- sapply(rep(5,500), prop)  
P10 <- sapply(rep(10,500), prop)  
P15 <- sapply(rep(15,500), prop)  
P20 <- sapply(rep(20,500), prop)  
P30 <- sapply(rep(30,500), prop)  
P50 <- sapply(rep(50,500), prop)  
P60 <- sapply(rep(60,500), prop)  
P100 <- sapply(rep(100,500), prop)  
P200 <- sapply(rep(200,500), prop)  
P500 <- sapply(rep(500,500), prop)  
PT = data.frame(P5,P10,P15,P20,P30,P50,P60,P100,P200,P500)

Boxplot:

boxplot(PT, las=1)  
abline(h=0.5, col="red")



Prueba de normalidad:

shapiro.test(P5)

##   
## Shapiro-Wilk normality test  
##   
## data: P5  
## W = 0.92419, p-value = 3.475e-15

shapiro.test(P50)

##   
## Shapiro-Wilk normality test  
##   
## data: P50  
## W = 0.99133, p-value = 0.005025

shapiro.test(P100)

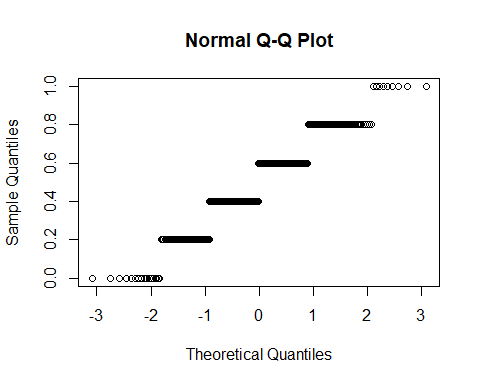
##   
## Shapiro-Wilk normality test  
##   
## data: P100  
## W = 0.99224, p-value = 0.01058

shapiro.test(P500)

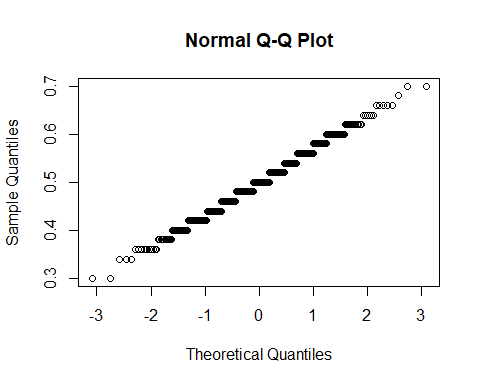
##   
## Shapiro-Wilk normality test  
##   
## data: P500  
## W = 0.99448, p-value = 0.06827

Gráfico de normalidad:

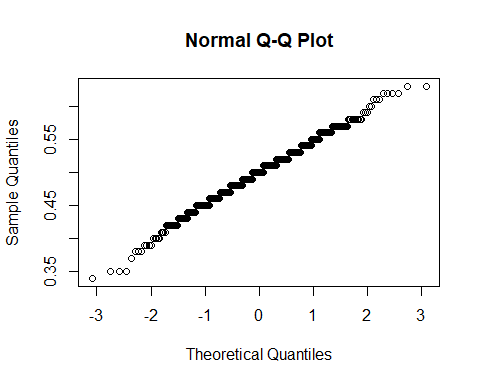
qqnorm(P5)



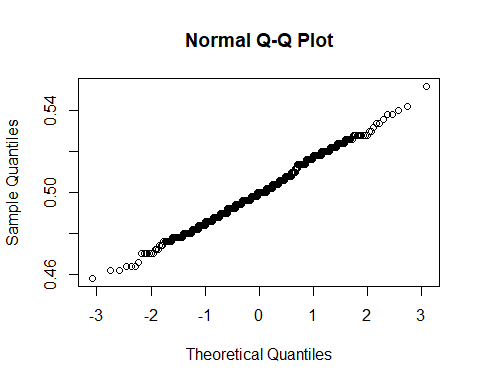
qqnorm(P50)



qqnorm(P100)



qqnorm(P500)



Análisis de resultados:

En primera instancia, si observamos el grafico boxplot con la comparación del comportamiento de los datos para todas las muestras tomadas podemos observar que a medida que el tamaño de la muestra (n) aumenta, la dispersión del 50% de los datos ubicados entre el primer cuartil y el tercer cuartil disminuye. Así mismo, podemos ver como en la medida que el tamaño de n aumenta la mediana toma un valor más cercano a la media. En cuanto a la forma podemos deducir que con un tamaño de n menor a 30 se observa asimetría, por ejemplo, con un tamaño de n=5 los datos presentan una asimetría positiva mientras que a partir de n=30 la forma tiene una tendencia simétrica. En cuanto a los valores atípicos, podemos decir que en la medida que n aumenta el límite de detección de los datos atípicos disminuye considerablemente.

Con la realización de la prueba de Shapiro-Wilk realizada a las muestras de tamaño n = 5, 50 y 500 buscamos determinar si los datos para cada muestra presentan una distribución normal. La hipótesis nula (H0) es que los datos provienen de una distribución normal, mientras que la hipótesis alternativa (Ha) es que los datos no provienen de una distribución normal. Para poder evaluar las hipótesis nos apoyamos en el valor de P, si el valor de P es mayor que el nivel de significancia elegido (generalmente 0.05), no se puede rechazar la hipótesis nula, si el valor de P es menor que el nivel de significancia elegido, se rechaza la hipótesis nula. Analizando los resultados obtenidos en las pruebas realizadas podemos concluir que: para un tamaño de muestra n=5 el valor de P (3.475e-15) es menor al valor de significancia (0.05), por lo tanto, se rechaza la hipótesis nula (H0), concluyendo que los datos tomados con un tamaño de muestra n=5 no presentan un comportamiento normal. Para un tamaño de muestra n=50 el valor de P (0.005025) continúa siendo menor que el valor de significancia (0.05), por ende, podemos decir que los datos no presentan un comportamiento normal. A partir de un tamaño de muestra n=100 el valor de P (0.01058) empieza a ser cercano al valor de significancia, sin embargo, no podemos aceptar la hipótesis nula. Finalmente, si analizamos el valor de P (0.06827) para una muestra de tamaño n=500, es mayor que el valor de significancia (0.05), esto implica no rechazar la hipótesis nula, con lo cual podemos concluir que a medida que aumenta el tamaño de la muestra los datos presentaran un comportamiento normal.

El gráfico QQ se utiliza comúnmente para evaluar si una muestra de datos sigue una distribución teórica (en este caso normal), y si no es así, cómo difiere la distribución evaluada de la distribución teórica. Los cuartiles de la muestra se trazan en el eje x y los cuartiles teóricos en el eje y. Si los datos siguen la distribución teórica (normal), los puntos en el grafico QQ seguirán una línea recta. Por el contrario, si los datos difieren de una distribución normal, entonces los puntos en el grafico QQ se alejarán de la línea recta. Si observamos los tres gráficos generados para una muestra de n = 5, 50, 500, podemos ver que a medida que el tamaño de la muestra aumenta los puntos se asemejan más a una línea recta, confirmado que los datos presentaran una tendencia a una distribución normal en la medida que el tamaño de la muestra supera un tamaño aproximado de 100.