Problema No. 2

2023-03-13

## Elaborado por:

Mario Felipe Aristizabal Trujillo

Juan Carlos Villalba Acevedo

## Presentado a:

Profesora Delia Ortega Lenis

## Validación de resultados

La comparación de tratamientos es una práctica fundamental en las ciencias agropecuarias y para ello a nivel estadístico se cuenta con herramientas para apoyar el proceso de toma de decisiones y así poder lograr concluir con un alto grado de confianza sobre los resultados observados en una muestra. A través de la simulación es posible evaluar estimadores y sus propiedades, que nos permitan usarlos con toda tranquilidad.

Suponga un escenario en el cual se aplicó tratamientos diferentes a dos lotes de una misma plantas y se desea analizar si alguno de los dos tratamientos presenta un mejor desempeño en el control de una plaga presente en ambos al momento inicial. Para ello utilizará como criterio de desempeño el tratamiento, el menor porcentaje de plantas enfermas presente después de un tiempo de aplicación (es decir, si se presentan o no diferencias en las proporciones de enfermos p1 y p2 - proporciones poblacionales-).

### a. Realice una simulación en la cual genere dos poblaciones de N1=1000 (Lote 1) y N2=1500 (Lote 2), para los cuales se asume que el porcentaje de individuos (plantas) enfermas en ambos lotes es del 10% (es decir, sin diferencias entre los tratamientos).

#### Rta.

L1 = rep(c(0,1), times = c(900, 100))  
L2 = rep(c(0,1), times = c(1350, 150))

### b. Genere una función que permita obtener una muestra aleatoria de los lotes y calcule el estimador de la proporción muestral para cada lote (p1ˆ y p2ˆ) para un tamaño de muestra dado n1=n2. Calcule la diferencia entre los estimadores (p1ˆ−p2ˆ).

#### Rta.

n = 500  
M1 = sample(L1, n)  
M1

## [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 1 1 0 0 0 0 0 0 0 0 0 0  
## [38] 0 0 0 0 0 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  
## [75] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0  
## [112] 0 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  
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0  
## [186] 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0  
## [223] 0 0 0 0 0 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 1 0 0 0 0  
## [260] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  
## [297] 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0  
## [334] 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 1 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 1 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 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [445] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0  
## [482] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0

M2 = sample(L2, n)  
M2

## [1] 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0  
## [38] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 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 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [149] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 1 0 1 0 0 0 0 0  
## [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0  
## [223] 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0  
## [260] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0  
## [297] 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0  
## [334] 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0  
## [371] 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 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 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 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 1 0 0 0 0 0 1 0 0 1 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

Prop1 = sum(M1)/n  
Prop1

## [1] 0.084

Prop2 = sum(M2)/n  
Prop2

## [1] 0.106

Diferencia = Prop1 - Prop2  
Diferencia

## [1] -0.022

### c. Repita el escenario anterior (b) 500 veces y analice los resultados en cuanto al comportamiento de los 500 estimadores (diferencias p1ˆ−p2ˆ). ¿Qué tan simétricos son los resultados?, ¿Son siempre cero las diferencias?

#### Rta.

Diferencia <- function(n) {  
 M1 = sample(L1, n)  
 Prop1 = sum(M1)/n  
 M2 = sample(L2, n)  
 Prop2 = sum(M2)/n  
 Difer = Prop1 - Prop2  
 return(Difer)  
}  
  
Diferencias <- sapply(rep(500,500), Diferencia)

Histograma:

hist(Diferencias, col="#BC2B6A")

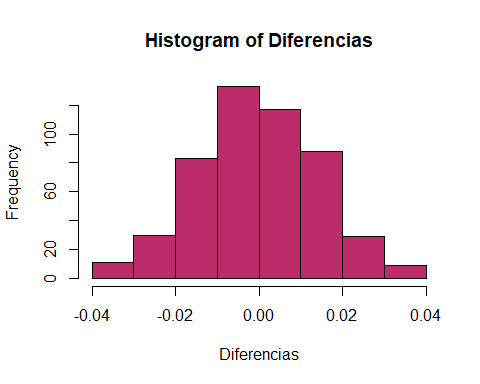
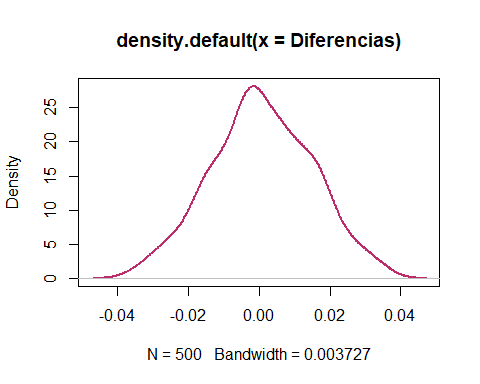


Gráfico de Densidad:

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



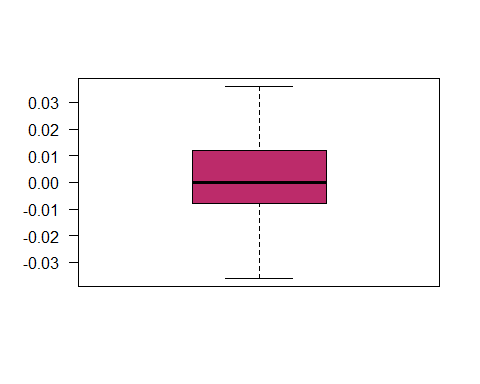
Resumen:

psych::describe(Diferencias)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 500 0 0.01 0 0 0.01 -0.04 0.04 0.07 -0.03 -0.34 0

Boxplot:

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



Análisis de resultados:

Se encontró con los resultados obtenidos con un tamaño de muestra n=500 y con la ayuda del histograma y el grafico de densidad que el comportamiento de los datos tiende a ser simétrico. Sin embargo, si analizamos los resultados de la diferencia de los estimadores observamos que la tendencia es a ser cero, pero no en todos los casos. Lo anterior, lo podemos ratificar si observamos el valor de la media y la mediana el cual es igual a 0. Así mismo, la desviación estándar presenta un valor muy pequeño (0.01), lo que nos indica que los resultados de la diferencia de los estimadores en su mayoría no se alejan mucho del valor 0.

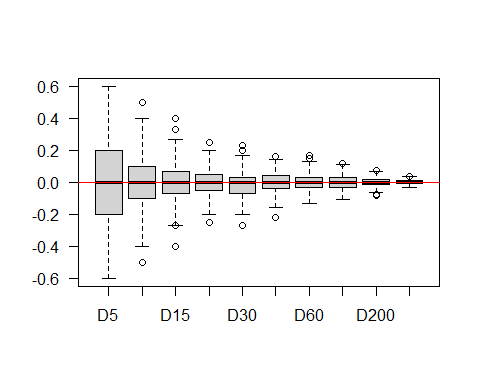
### d. Realice los puntos b y c para tamaños de muestra n1=n2=5, 10, 15, 20, 30, 50, 60, 100, 200, 500. Compare los resultados de los estimadores (p1ˆ−p2ˆ) en cuanto a la normalidad. También analice el comportamiento de las diferencias y evalúe. ¿Considera que es más probable concluir que existen diferencias entre los tratamientos con muestras grandes que pequeñas, es decir, cuál considera usted que es el efecto del tamaño de muestra en el caso de la comparación de proporciones?

#### Rta.

D5 <- sapply(rep(5,500), Diferencia)  
D10 <- sapply(rep(10,500), Diferencia)  
D15 <- sapply(rep(15,500), Diferencia)  
D20 <- sapply(rep(20,500), Diferencia)  
D30 <- sapply(rep(30,500), Diferencia)  
D50 <- sapply(rep(50,500), Diferencia)  
D60 <- sapply(rep(60,500), Diferencia)  
D100 <- sapply(rep(100,500), Diferencia)  
D200 <- sapply(rep(200,500), Diferencia)  
D500 <- sapply(rep(500,500), Diferencia)  
DT = data.frame(D5,D10,D15,D20,D30,D50,D60,D100,D200,D500)

Boxplot:

boxplot(DT, las=1)  
abline(h=0.0, col="red")



Prueba de normalidad:

shapiro.test(D5)

##   
## Shapiro-Wilk normality test  
##   
## data: D5  
## W = 0.90436, p-value < 2.2e-16

shapiro.test(D50)

##   
## Shapiro-Wilk normality test  
##   
## data: D50  
## W = 0.98819, p-value = 0.0004559

shapiro.test(D200)

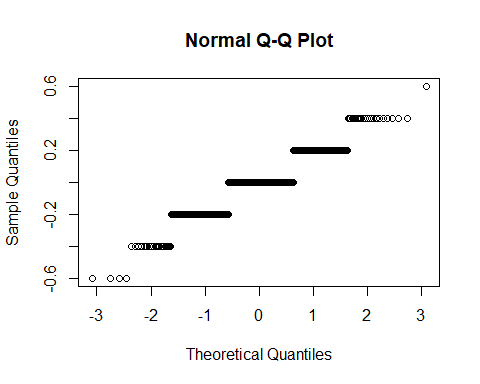
##   
## Shapiro-Wilk normality test  
##   
## data: D200  
## W = 0.99501, p-value = 0.1067

shapiro.test(D500)

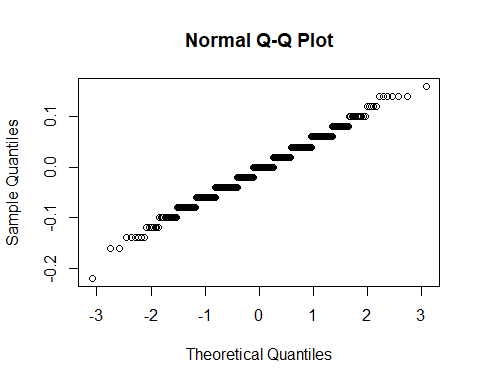
##   
## Shapiro-Wilk normality test  
##   
## data: D500  
## W = 0.99466, p-value = 0.07975

Gráfico de normalidad:

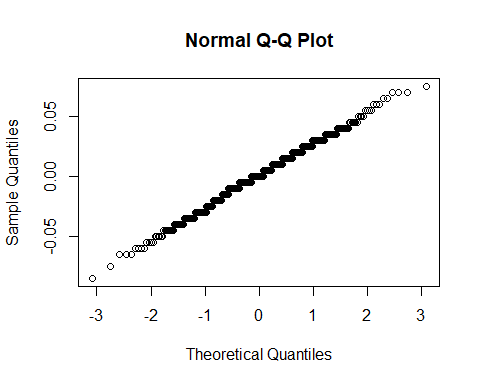
qqnorm(D5)



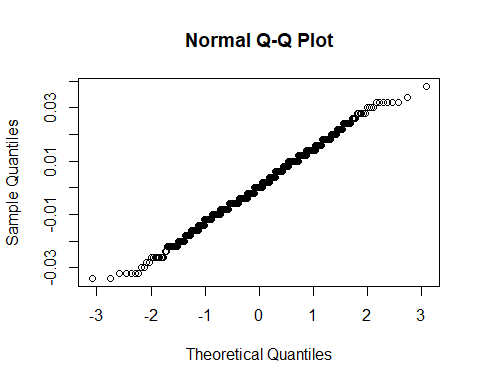
qqnorm(D50)



qqnorm(D200)



qqnorm(D500)



Análisis de resultados:

Con la ayuda de la prueba Shapiro-Wilk realizada al subconjunto de datos obtenidos para un tamaño de muestra n=5, 50, 200 y 500, observamos que a partir del tamaño de muestra n=200 el resultado del valor P es mayor que el valor de significancia (0.05), por lo tanto, no se puede rechazar la hipótesis nula, es decir, que los datos presentan un comportamiento normal. Lo anterior, lo podemos corroborar analizando los gráficos QQ-Plot, en los cuales observamos que con un tamaño de muestra superior a 200 la tendencia de los puntos es a formar una línea recta.

Analizando el grafico Boxplot podemos decir que el efecto que se obtiene con aumentar el tamaño de muestra se evidencia en el grado de dispersión del 50% de los datos centrales, ya que a medida que se aumenta el tamaño de muestra, la dispersión se hace más pequeña y los datos centrales tienden a estar más cerca de la mediana, que en este caso es 0.

### e. Ahora realice nuevamente los puntos a-d bajo un escenario con dos lotes, pero de proporciones de enfermos diferentes (p1=0.1 y p2=0.15), es decir, el tratamiento del lote 1 si presentó un mejor desempeño reduciendo en un 5% el porcentaje de enfermos. Bajo este nuevo escenario compare la distribución de estas diferencias (p1ˆ−p2ˆ) con las observadas bajo igualdad de condiciones en los lotes. ¿Qué puede concluir? ¿Existen puntos en los cuales es posible que se observen diferencias de p1−p2 bajo ambos escenarios (escenario 1: sin diferencias entre p1 y p2, escenario 2: diferencia de 5%)?

#### Rta.

##### Definición de la problación:

Le1 = rep(c(0,1), times = c(900, 100))  
Le2 = rep(c(0,1), times = c(1275, 225))

##### Cálculo de la diferencia de proporciones:

n = 500  
Me1 = sample(Le1, n)  
Me2 = sample(Le2, n)  
  
Prop1e = sum(Me1)/n  
Prop1e

## [1] 0.1

Prop2e = sum(Me2)/n  
Prop2e

## [1] 0.148

Diferencia\_e = Prop1e - Prop2e  
Diferencia\_e

## [1] -0.048

##### Definición función de diferencia de proporciones (para un tamaño de muestra n=500):

Diferencia\_e <- function(n) {  
 Me1 = sample(Le1, n)  
 Prop1e = sum(Me1)/n  
 M2 = sample(Le2, n)  
 Prop2e = sum(Me2)/n  
 Difer\_e = Prop1e - Prop2e  
 return(Difer\_e)  
}  
  
Diferencias\_e <- sapply(rep(500,500), Diferencia\_e)

Histograma:

hist(Diferencias\_e, col="#BC2B6A")

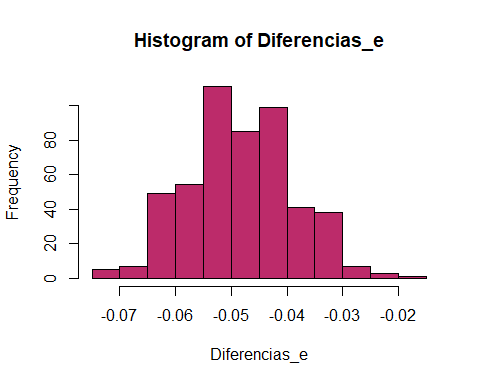
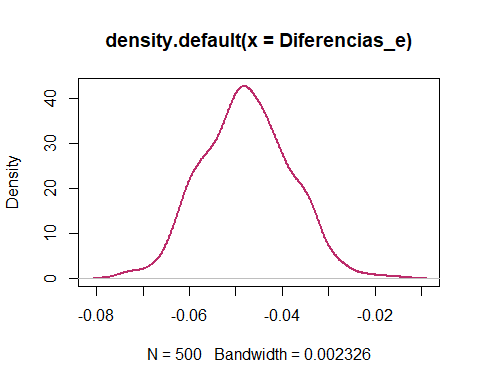


Gráfico de Densidad:

plot(density(Diferencias\_e), col="#BC2B6A", lwd=2)



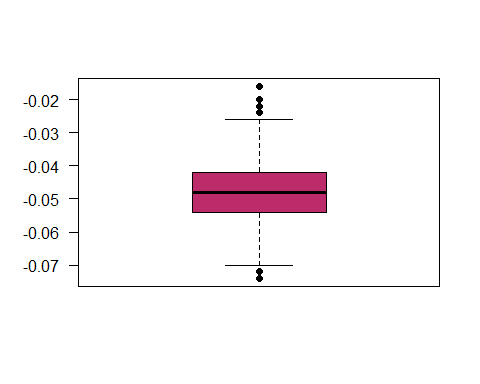
Resumen:

psych::describe(Diferencias\_e)

## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 500 -0.05 0.01 -0.05 -0.05 0.01 -0.07 -0.02 0.06 0.08 -0.07 0

Boxplot:

boxplot(Diferencias\_e, pch=19, col="#BC2B6A", las=1)



Análisis de resultados:

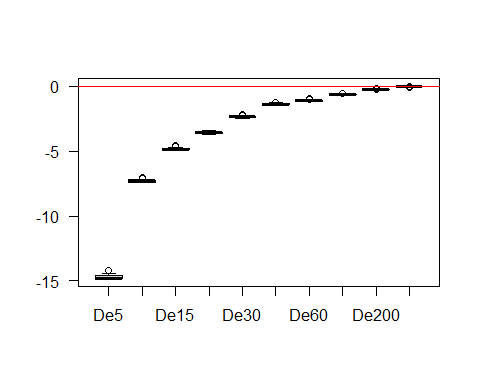
Como conclusión del escenario numero 2 (diferencia de 5%) podemos decir que la mediana y la media ya no presentan una tendencia a 0, en este caso, su valor es de -0.05. Si bien se observa una aparente simetría de los datos, no podemos concluir que su comportamiento se asemeje a una distribución normal, para esto sería necesario realizar una prueba de normalidad. Si analizamos el grafico Boxplot podemos decir que los datos centrales presentan una dispersión moderada con una asimetría positiva o sesgada a la derecha muy leve.

##### Diferencia de proporciones (para un tamaño de muestra n1=n2=5, 10, 15, 20, 30, 50, 60, 100, 200, 500):

De5 <- sapply(rep(5,500), Diferencia\_e)  
De10 <- sapply(rep(10,500), Diferencia\_e)  
De15 <- sapply(rep(15,500), Diferencia\_e)  
De20 <- sapply(rep(20,500), Diferencia\_e)  
De30 <- sapply(rep(30,500), Diferencia\_e)  
De50 <- sapply(rep(50,500), Diferencia\_e)  
De60 <- sapply(rep(60,500), Diferencia\_e)  
De100 <- sapply(rep(100,500), Diferencia\_e)  
De200 <- sapply(rep(200,500), Diferencia\_e)  
De500 <- sapply(rep(500,500), Diferencia\_e)  
DTe = data.frame(De5,De10,De15,De20,De30,De50,De60,De100,De200,De500)

Boxplot:

boxplot(DTe, las=1)  
abline(h=0.0, col="red")



Prueba de normalidad:

shapiro.test(De5)

##   
## Shapiro-Wilk normality test  
##   
## data: De5  
## W = 0.67134, p-value < 2.2e-16

shapiro.test(De50)

##   
## Shapiro-Wilk normality test  
##   
## data: De50  
## W = 0.97123, p-value = 2.426e-08

shapiro.test(De200)

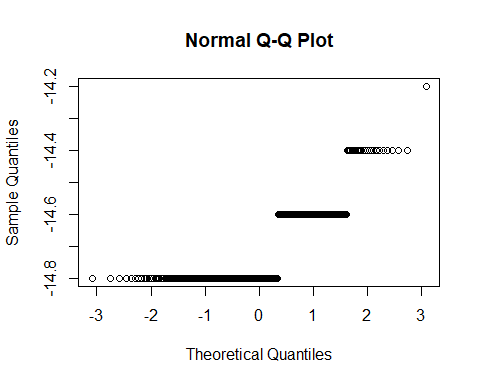
##   
## Shapiro-Wilk normality test  
##   
## data: De200  
## W = 0.98689, p-value = 0.00018

shapiro.test(De500)

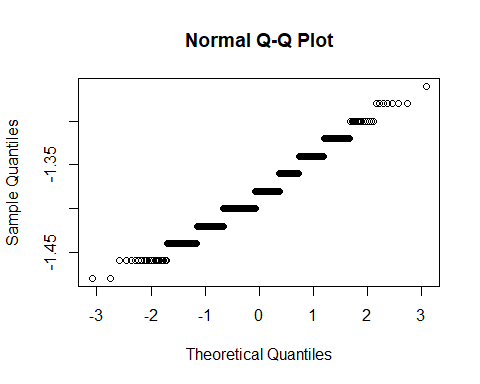
##   
## Shapiro-Wilk normality test  
##   
## data: De500  
## W = 0.99426, p-value = 0.05676

Gráfico de normalidad:

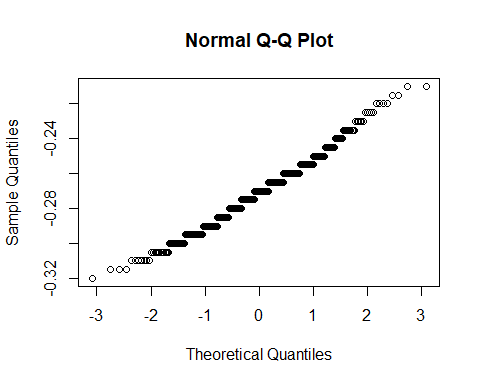
qqnorm(De5)



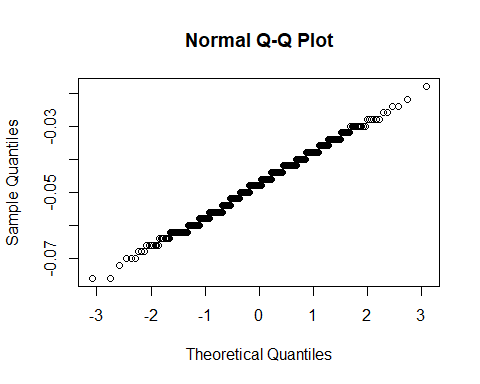
qqnorm(De50)



qqnorm(De200)



qqnorm(De500)



Análisis de resultados:

Después de realizar los puntos a, b, c y d nuevamente, esta vez en un escenario (escenario 2) con una diferencia de 5% en la proporción de individuos enfermos (plantas), podemos decir que cuando tenemos un tamaño de muestra pequeño la diferencia de proporciones aumenta, mientras que, si el tamaño de muestra es grande, observamos que la diferencia de las proporciones se disminuye y se acerca a 0. Esto lo podemos observar en la gráfica Boxplot donde se ve el comportamiento exponencial que tienen las cajas del conjunto de datos con un tamaño de muestra desde n=5 hasta n=500, donde la mediana tiende a cero en la medida que el tamaño de muestra aumenta.

En cuanto a la normalidad de los datos y con la ayuda de la prueba Shapiro-Wilk y el grafico QQ-Plot realizado a los subconjuntos de datos con un tamaño de muestra de n = 5, 50, 200 y 500, observamos que con excepción del tamaño de muestra n=500, el valor de P es menor al valor de significancia (0.05), por lo tanto, rechazamos la hipótesis nula, es decir, los datos no presentan un comportamiento similar a una distribución normal. A pesar que el valor de P en el tamaño de muestra n=500 es superior al valor de significancia, la diferencia no es muy grande entre los dos valores, por ende, no es muy preciso concluir que los datos presenten una distribución normal. Esta conclusión se puede comprobar si observamos los gráficos de QQ-Plot, donde no es claro que los datos se asemejen a una línea recta, incluso con un tamaño de muestra de n=500.

Finalmente, podríamos decir que, si tomamos muestras de dos poblaciones bajo condiciones diferentes, la diferencia de proporciones va tomar un valor diferente de cero, sin embargo, en la medida que se aumente el tamaño de la muestra, disminuirá exponencialmente la diferencia de proporciones tendiendo a ser cero.