Actividad 1 - Teorema del Límite Central

Keyner Martinez y Jairo Alba - Metodos y Similución Estadística

2023-02-21

1. Introducción

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.

Supongamos que la población donde se hace el muestreo es finita de tamaño N y sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño n, tomada de una población con media y varianza. Entonces, si n es lo suficientemente grande, X tiene una distribución normal aproximada, es decir:

- La media de la distribución muestral es igual a la media de la población en que se toma la muestra.
- La varianza de la distribución muestral es igual a la varianza de la población dividida por el tamaño de la muestra.

El potencial de este teorema está en que no importa la distribución de la variable X, la distribución de la media proveniente de la muestra tomada de esta población se distribuye aproximadamente normal.

2. Desarrollo del proyecto

Para verificar el teorema del límite central, descrito en la sección 1, generamos una población (supongamos plantas) de $\boldsymbol{N = 1000}$, donde analizamos tres simulaciones, la primera simulación corresponde a que el cincuenta por ciento (50%) de la población se encuentra enferma, la segunda simulación con una población enferma del noventa por ciento (90%) y la última simulación con un diez por ciento (10%) de la población enferma.

Para cada población se tomaron ***500 muestras*** de tamaños $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$, y analizamos los resultados obtenidos para cada población.

En razón de lo anterior, se realizó el siguiente código, para generar cada una de las simulaciones con las especificaciones anteriormente descritas:

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# especificando semilla
set.seed(138)

# Población
pob <- 1000

# número de simulaciones
nsim <- 500

# tamaños de las muestras
muestra_vect <- c(5,10,15,20,30,50,60,100,200,500)

muestras_chr <- c("n = 5","n = 10","n = 15","n = 20","n = 30","n = 50","n = 60",
                  "n = 100","n = 200","n = 500")


# porcentaje de plantas enfermas 500=0.5 ; 900=0.9 ; 100=0.1
p <- c(500,900,100)

# variables de almacenamiento de los resultados de cada simulación
sim_pob <- list()
med_sim <- vector(length=nsim)
lista_med <- list() # lista de estimadores con población del 50% de plantas enfermas
lista_alt <- list() # lista de estimadores con población del 90% de plantas enfermas
lista_baj <- list() # lista de estimadores con población del 10% de plantas enfermas
repetidor <- vector()
```

a) Se generan las poblaciones de plantas enfermas del 50%, 90% y 10%:

```r
# generando población de plantas enfermas 50% ; 90% ; 10%
for (i in 1:3) {
  repetidor <- rep(c(1,0),times=c(p[i],pob-p[i]))
  sim_pob[[i]] <- repetidor
}
```

b) Se genera una función que permite i) obtener muestras aleatorias dado la población y el tamaño de muestras; y ii) calcula los estimadores de la proporción muestral para un tamaño de muestra dado:

```r
# genera la muestra
muestra <- function(poblacion,n){
  return(sample(poblacion,n))
}

# calculó del estimador de cada una de las muestras

for (i in 1:3) {
  for (j in 1:10) {
    for (z in 1:nsim) {
      x <- muestra(sim_pob[[i]],muestra_vect[j])
      med_sim[z] <- mean(x)

    }
```

```
    if(i==1){
      lista_med[[j]] <- med_sim
    } else if (i==2) {
      lista_alt[[j]] <- med_sim
    } else {
      lista_baj[[j]] <- med_sim
    }
  }
}
```

3. Análisis de los Resultados

En esta sección, nos centraremos en analizar los resultados obtenidos en las distintas simulaciones y verificar el teorema del limite central.

3.1. Caso 1

El primer caso, se ha extraído una muestra aleatoria de tamaño $\boldsymbol{n = 500}$ de una población de plantas con tamaño $\boldsymbol{N = 1000}$ de cincuenta por ciento (50%) de plantas enfermas.

En el gráfico de distribución de la proporción de cada una de las muestras se observa que, **i)** mientras la población tiene una proporción del 50% de plantas enfermas, la media de la proporción muestral indica lo mismo, es decir, la muestra tomada tiene una proporción del 50% de plantas enfermas; **ii)** cuando n es grande, en este caso $n = 500$, algunos autores establecen que la aproximación es bastante buena y así se comprobó, se puede observar que la media de la proporción muestral tiene una distribución normal.

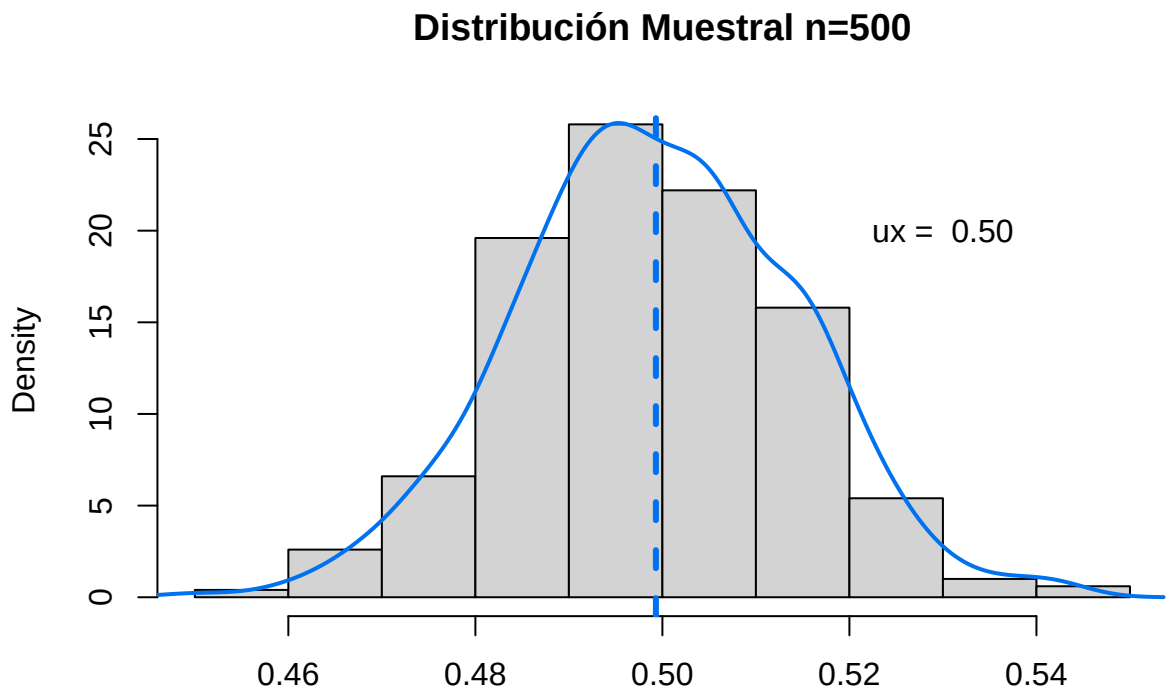


Figura 1. Histograma y curva de densidad de las distribución muestral

Los estimadores son variables aleatorias que toman su valor de los datos obtenidos en una muestra y que siguen una distribución conocida, pero además poseen propiedades deseables como son: insesgadez, eficiencia, consistencia, entre otros.

Se observa en la siguiente figura, donde los estimadores, para el presente caso, presentan una alta precisión (poca varianza), no presentan sesgo y es consistente.

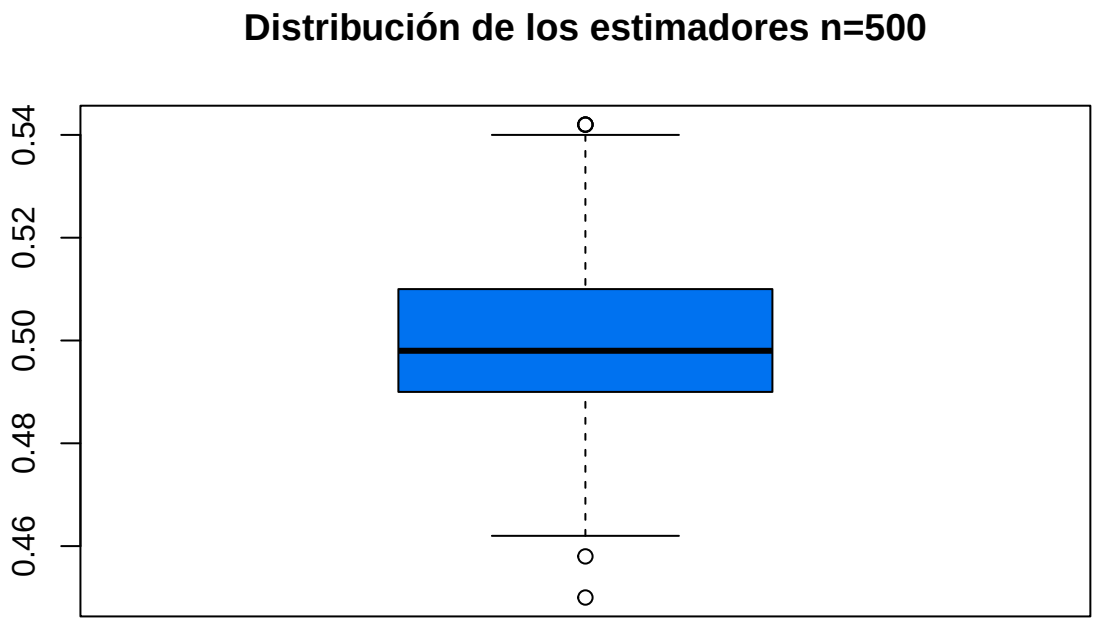


Figura 2. Boxplot de la distribución de los estimadores

```
##   muestras    media     varianza
## 1    n=500 0.499304 0.0002214144
```

3.2. Caso 2

El segundo caso, se ha extraído muestra aleatoria de tamaño ***n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500*** de una población de plantas con tamaño ***N = 1000*** de cincuenta por ciento (50%) de plantas enfermas.

A continuación, se muestra los datos y gráficos de distribución de la proporción de cada una de las muestras, se observa que, ***i)*** mientras la población tiene una proporción del 50% de plantas enfermas, la media de la proporción muestral indica lo mismo, es decir, la muestra tomada tiene una proporción del 50% de plantas enfermas; ***ii)*** mientras aumenta el tamaño de la muestra se observa que la distribución de las media proporcional se va transformando en una distribución simétrica, hasta convertirse en una normal; ***iii)*** mientras aumenta el tamaño de la muestra los estimadores presentan alta precisión, por lo cual, la varianza disminuye conforme aumenta los tamaños de la muestra.

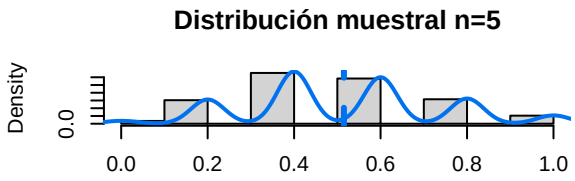


Fig 3. Media de las proporciones

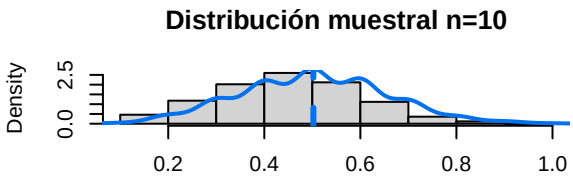


Fig 4. Media de las proporciones

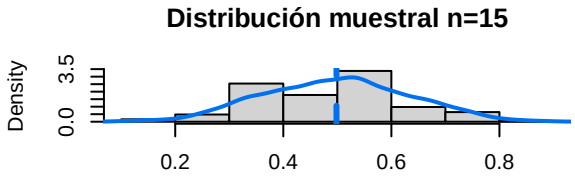


Fig 5. Media de las proporciones

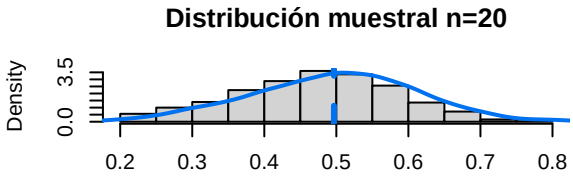


Fig 6. Media de las proporciones

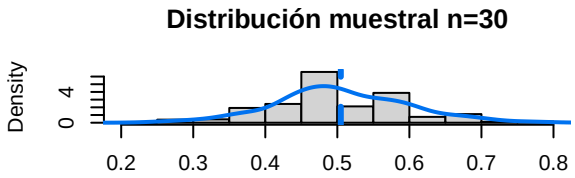


Fig 7. Media de las proporciones

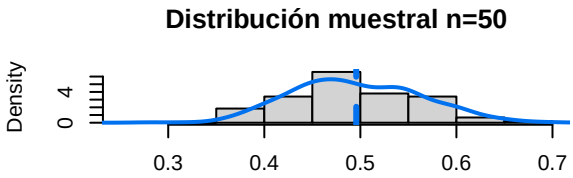


Fig 8. Media de las proporciones

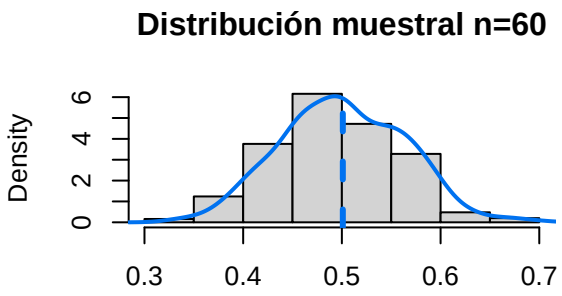


Fig 9. Media de las proporciones

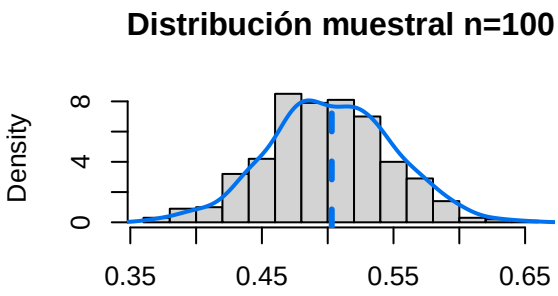


Fig 10. Media de las proporciones

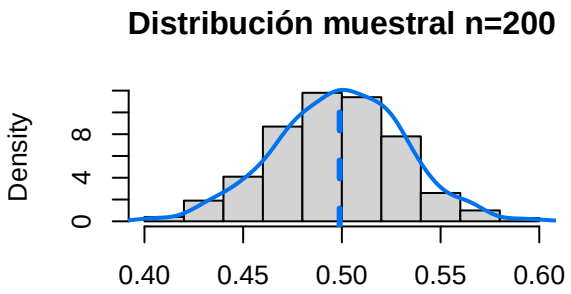


Fig 11. Media de las proporciones

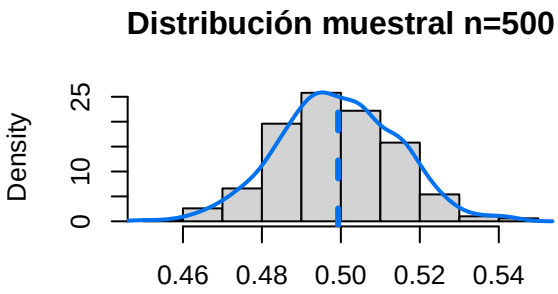


Fig 12. Media de las proporciones

```
##    muestras     media     varianza
## 1     n = 5 0.5152000 0.0514719038
## 2    n = 10 0.5022000 0.0234620842
## 3    n = 15 0.4984000 0.0178019328
## 4    n = 20 0.4967000 0.0129299699
## 5    n = 30 0.5048000 0.0086253218
## 6    n = 50 0.4957200 0.0044666148
## 7    n = 60 0.5009667 0.0039720096
## 8   n = 100 0.5030600 0.0022064493
## 9   n = 200 0.4988700 0.0010384500
## 10  n = 500 0.4993040 0.0002214144
```

Así mismo, se puede observar la comparación de los distintos tamaños de muestras ***n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500*** evidencia que al ir aumentando los tamaños muestrales los estimadores tienden a ser insesgado, eficiente y consistente.

Se puede observar que a medida que aumenta los tamaños de muestra, es decir, que el limite tiende a ***N = 1000*** hay menos variabilidad:

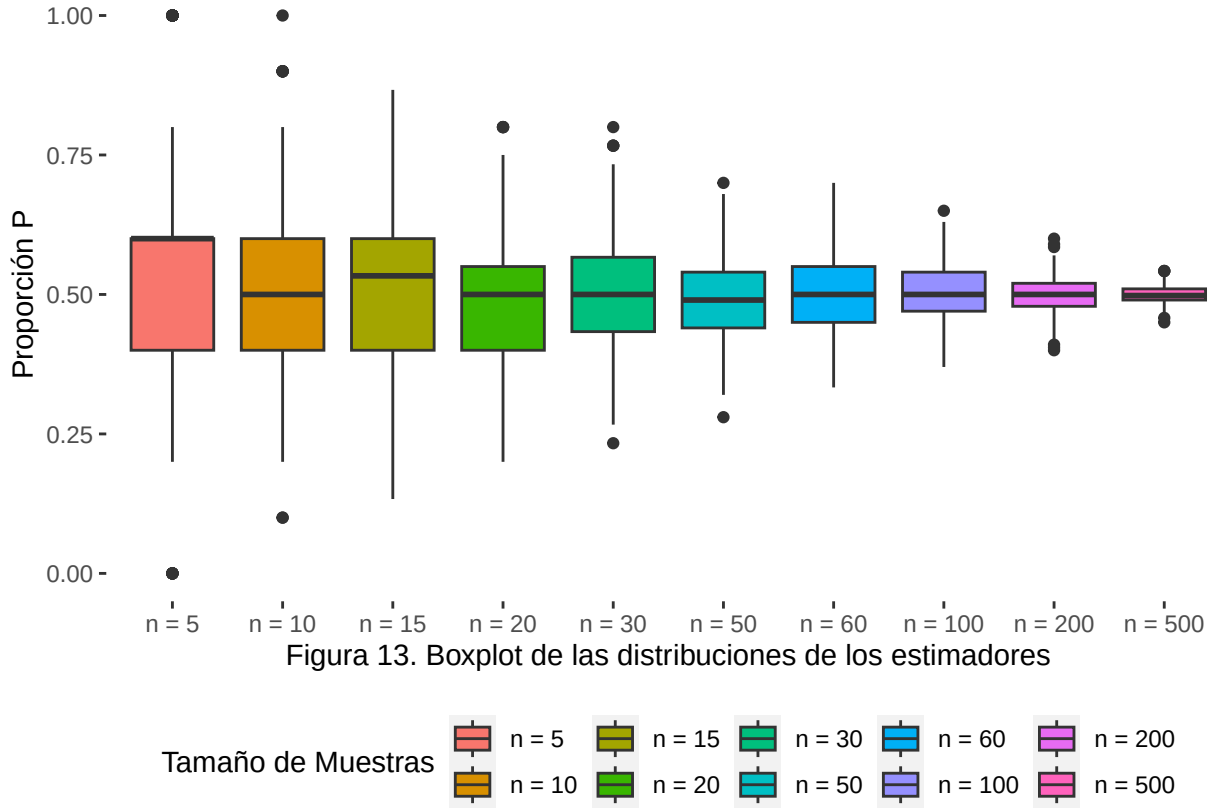


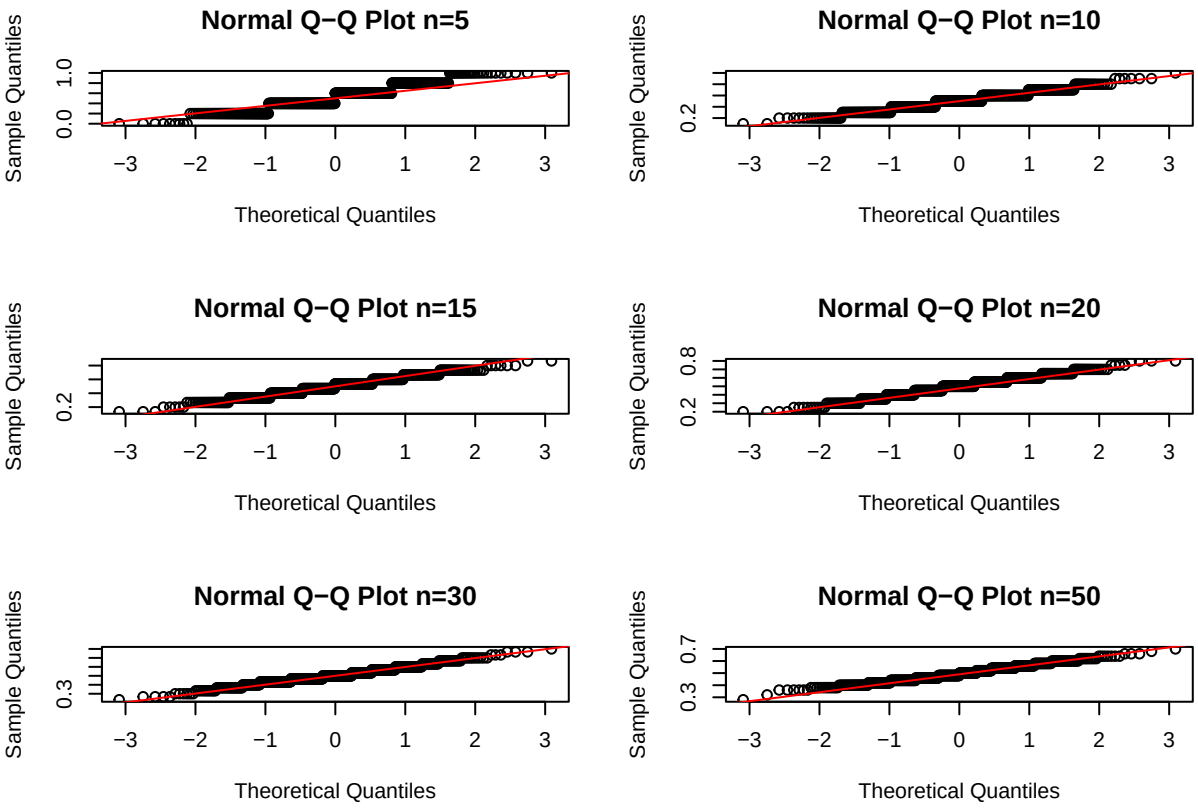
Figura 13. Boxplot de las distribuciones de los estimadores

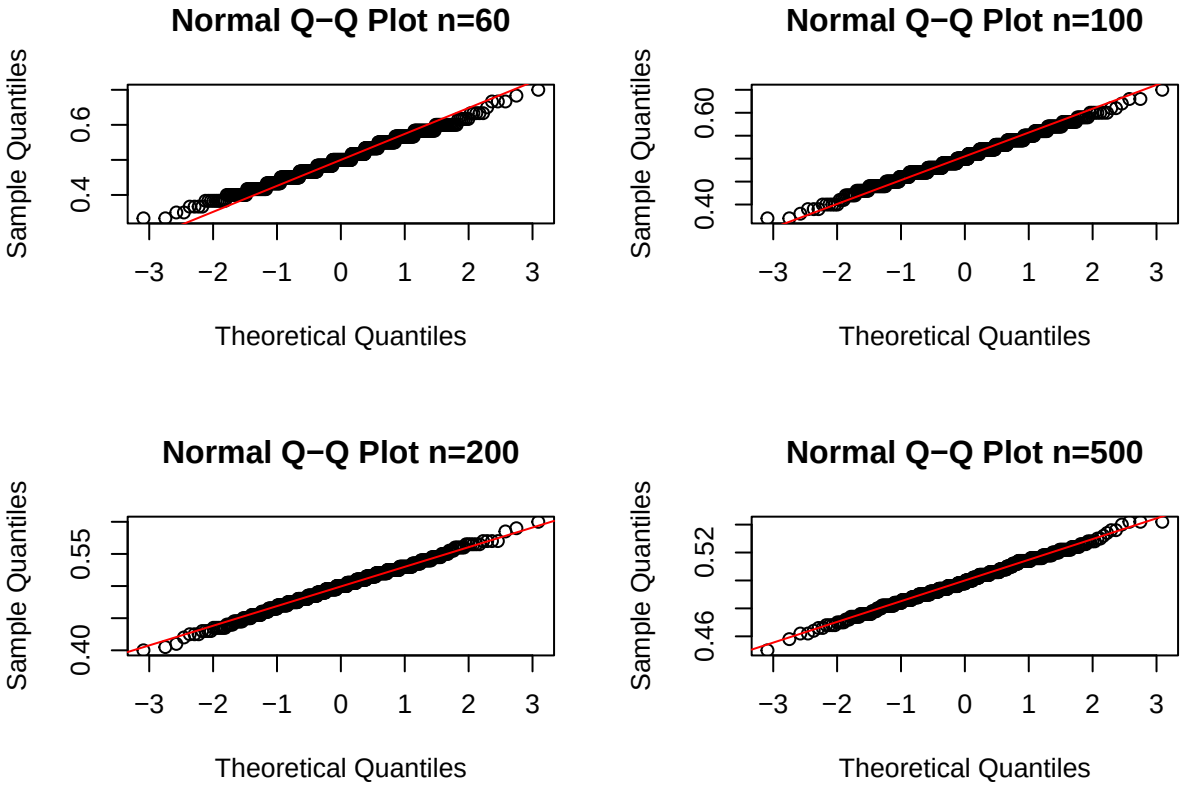
Posteriormente se realizó un test de normalidad ***SHAPIRO-WILK***, es una de las más utilizadas y eficiente para comprobar la normalidad de una variable. En este caso, la hipotesis nula del test es que la población representa una distribución normal. Por lo tanto, un valor menor a 0.05 indica que se debe rechazar la hipótesis nula, es decir, los datos no poseen distribución normal.

A continuación, se muestra una tabla de los resultados de la prueba para cada uno de los tamaños de muestras $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$, donde se puede observar la convergencia de la distribución de la media muestral a una distribución normal al aumentarse el tamaño de la muestra, más especificamente, a partir del tamaño de muestra $\boldsymbol{n = 100}$.

```
##    muestras shapiro
## 1     n = 5    0.00
## 2    n = 10    0.00
## 3    n = 15    0.00
## 4    n = 20    0.00
## 5    n = 30    0.00
## 6    n = 50    0.00
## 7    n = 60    0.00
## 8   n = 100    0.09
## 9   n = 200    0.33
## 10  n = 500    0.38
```

Así mismo, se muestra la prueba de normalidad de una linea recta, donde evidenciamos que a medida que aumenta los tamaños de muestras $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$ los puntos se acercan a la línea recta.


Normal Q−Q Plot n=5
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=10
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=15
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=20
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=30
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=50
Sample Quantiles
Theoretical Quantiles



3.3. Caso 3.

El tercer caso, se ha extraído muestra aleatoria de tamaño ***n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500*** de una población de plantas con tamaño ***N = 1000*** de noventa por ciento (90%) de plantas enfermas.

A continuación, se muestra los datos y gráficos de distribución de la proporción de cada una de las muestras, se observa que, ***i)*** mientras la población tiene una proporción del 90% de plantas enfermas, la media de la proporción muestral indica lo mismo, es decir, la muestra tomada tiene una proporción del 90% de plantas enfermas; ***ii)*** mientras aumenta el tamaño de la muestra se observa que la distribución de las media proporcional se va transformando en una distribución simétrica, hasta convertirse en una normal; ***iii)*** mientras aumenta el tamaño de la muestra los estimadores presentan alta precisión, por lo cual, la varianza disminuye conforme aumenta los tamaños de la muestra.

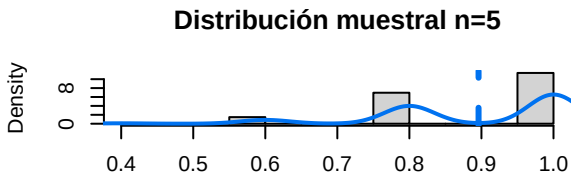


Fig 14. Media de las proporciones

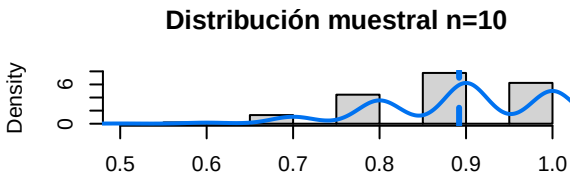


Fig 15. Media de las proporciones

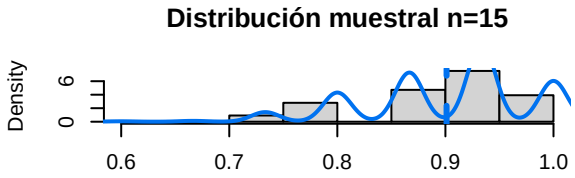


Fig 16. Media de las proporciones

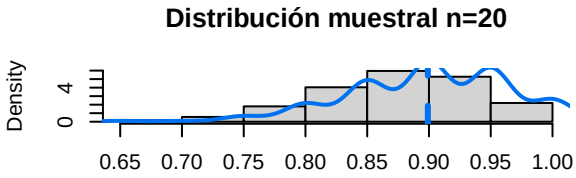


Fig 17. Media de las proporciones

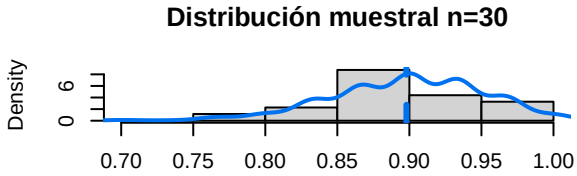


Fig 18. Media de las proporciones

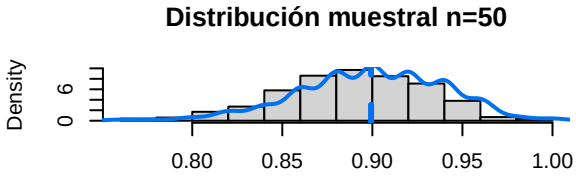


Fig 19. Media de las proporciones

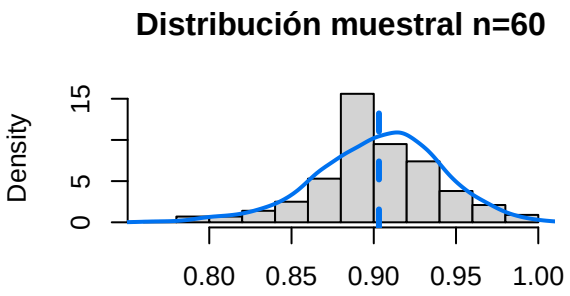


Fig 20. Media de las proporciones

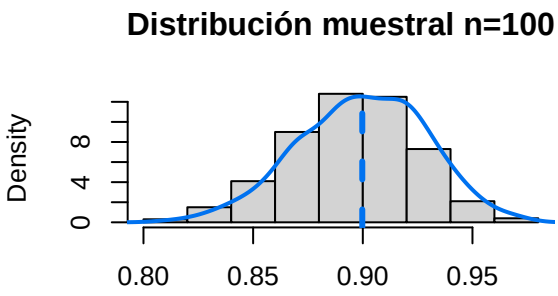


Fig 21. Media de las proporciones

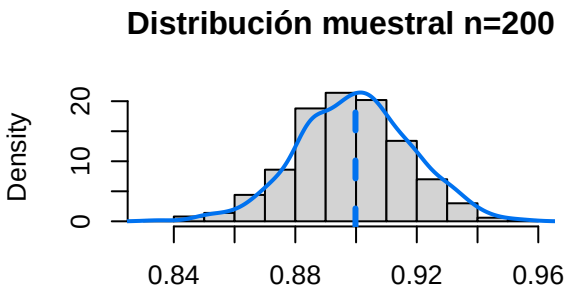


Fig 22. Media de las proporciones

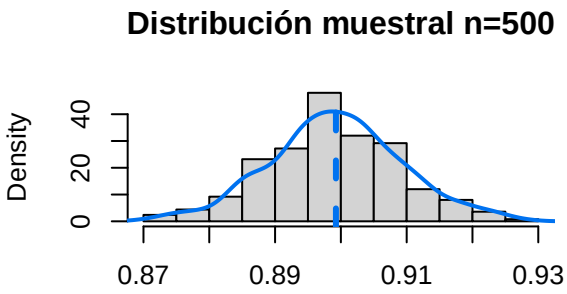


Fig 23. Media de las proporciones

```
##    muestras     media     varianza
## 1     n = 5 0.8960000 0.0178597194
## 2    n = 10 0.8920000 0.0091543086
## 3    n = 15 0.9010667 0.0057258695
## 4    n = 20 0.8991000 0.0042126152
## 5    n = 30 0.8979333 0.0028302783
## 6    n = 50 0.8992800 0.0016636088
## 7    n = 60 0.9031667 0.0014100145
## 8   n = 100 0.8997600 0.0008460345
## 9   n = 200 0.8998100 0.0003439017
## 10  n = 500 0.8992480 0.0001025516
```

Así mismo, se puede observar la comparación de los distintos tamaños de muestras $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$ evidencia que al ir aumentando los tamaños muestrales los estimadores tienden a ser insesgado, eficiente y consistente.

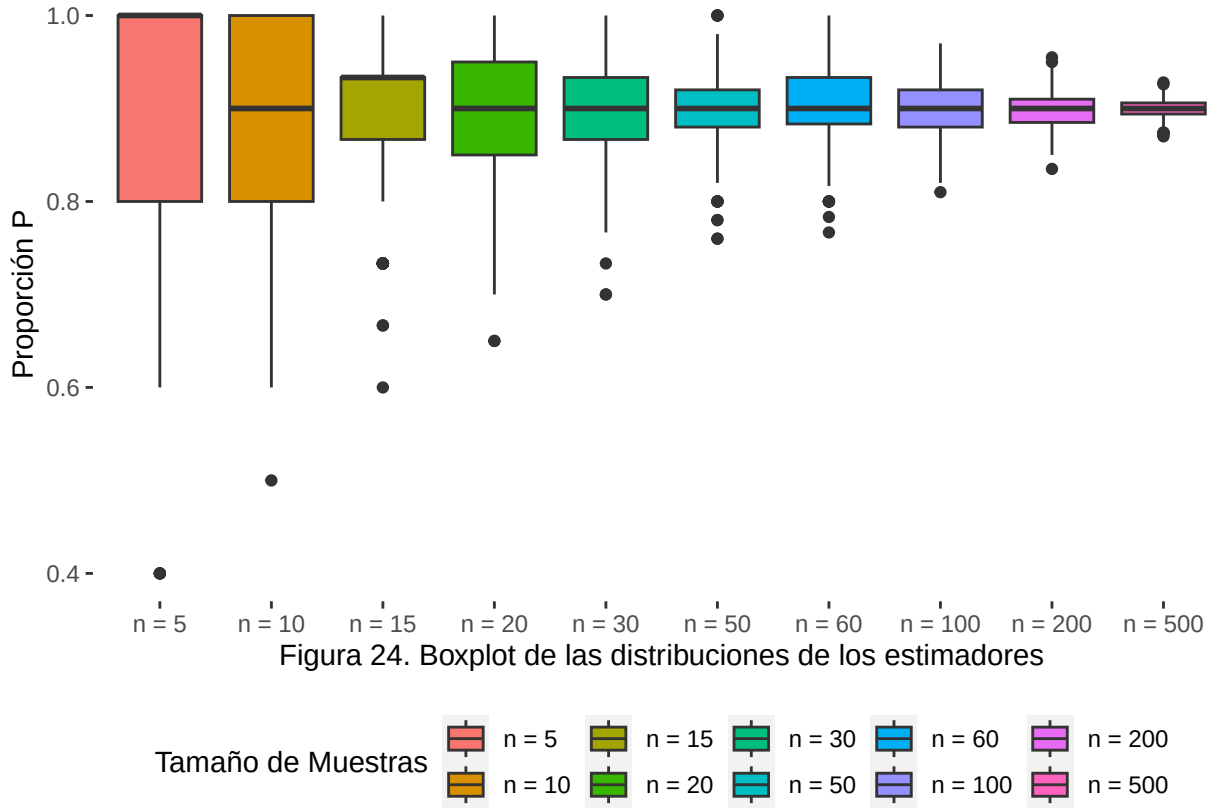
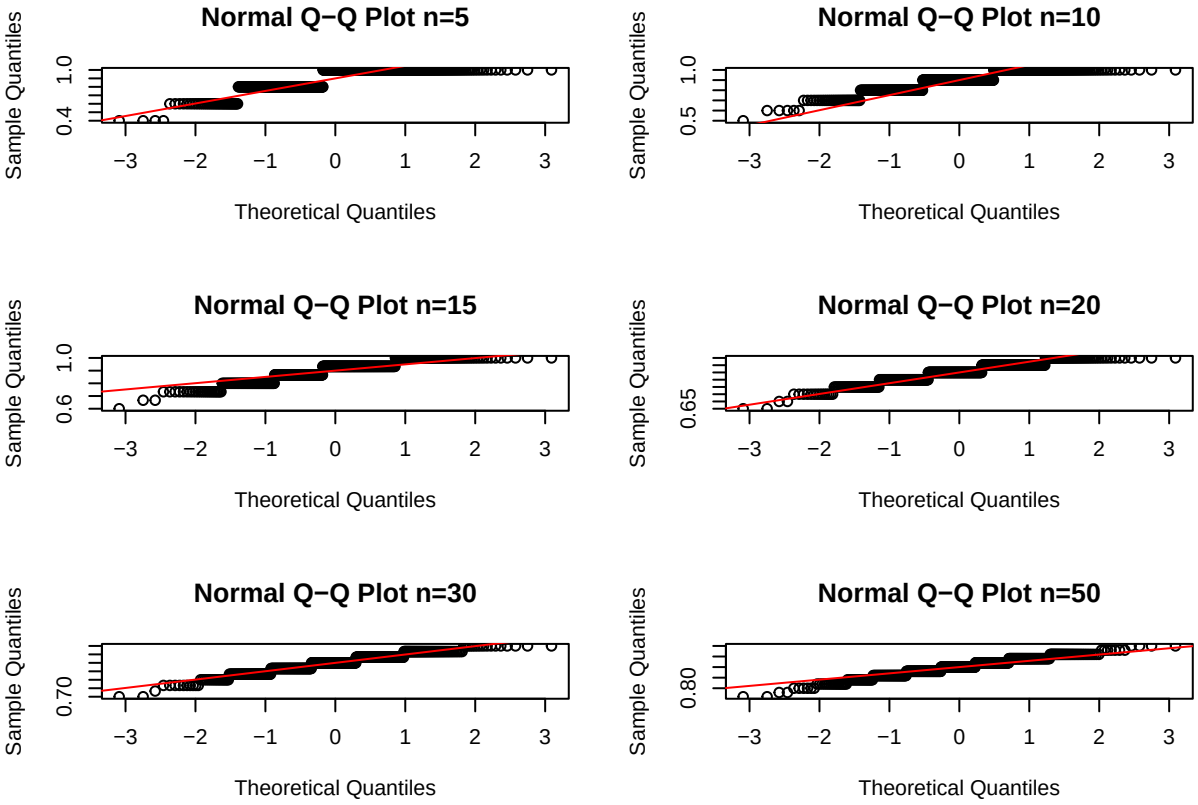


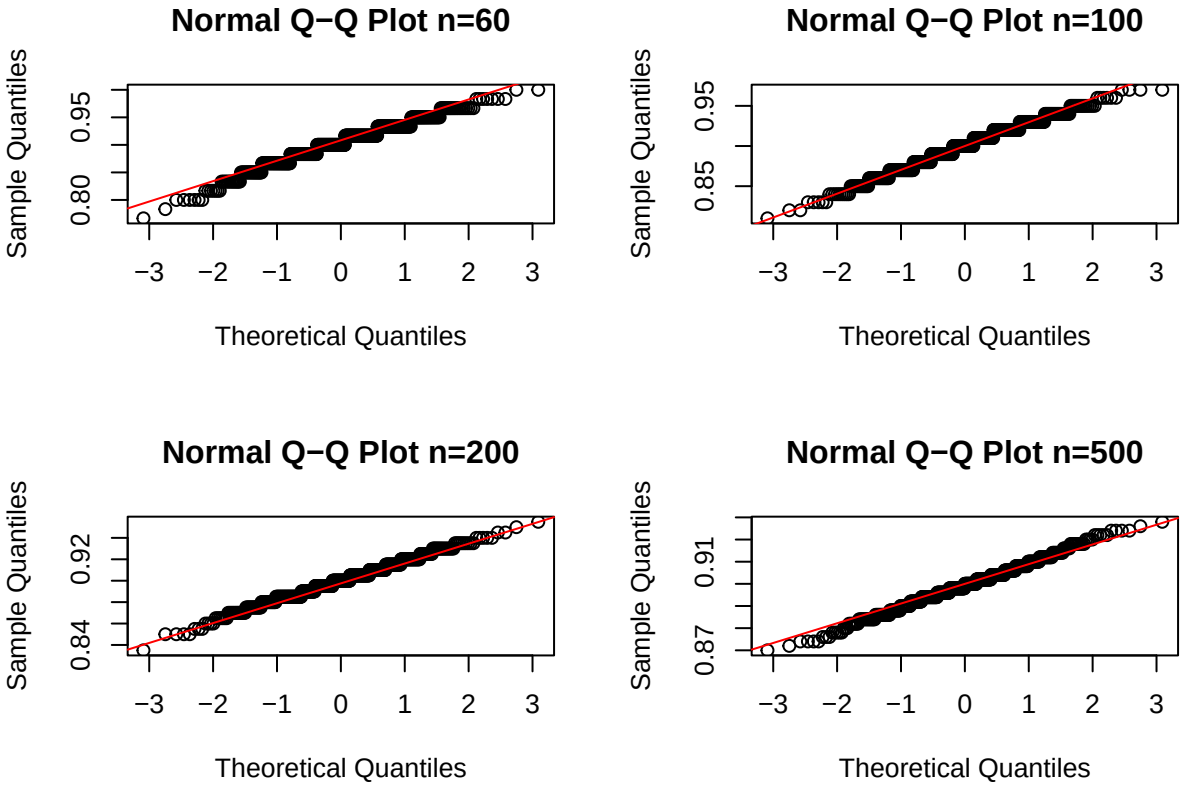
Figura 24. Boxplot de las distribuciones de los estimadores

A continuación, se muestra una tabla de los resultados de la prueba ***SHAPIRO-WILK***, descrita en la anterior caso, para cada uno de los tamaños de muestras $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$, donde se puede observar la convergencia de la distribución de la media muestral a una distribución normal al aumentarse el tamaño de la muestra, más especificamente, a partir del tamaño de muestra $\boldsymbol{n = 500}$.

```
##    muestras shapiro
## 1     n = 5    0.00
## 2    n = 10    0.00
## 3    n = 15    0.00
## 4    n = 20    0.00
## 5    n = 30    0.00
## 6    n = 50    0.00
## 7    n = 60    0.00
## 8   n = 100    0.00
## 9   n = 200    0.01
## 10  n = 500    0.06
```

Así mismo, se muestra la prueba de normalidad de una linea recta, donde evidenciamos que a medida que aumenta los tamaños de muestras $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$ los puntos se acercan a la línea recta.


Normal Q−Q Plot n=5
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=10
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=15
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=20
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=30
Sample Quantiles
Theoretical Quantiles
Normal Q−Q Plot n=50
Sample Quantiles
Theoretical Quantiles



3.4. Caso 4

El último caso, se ha extraído muestra aleatoria de tamaño $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$ de una población de plantas con tamaño $\boldsymbol{N = 1000}$ de diez por ciento (10%) de plantas enfermas.

A continuación, se muestra los datos y gráficos de distribución de la proporción de cada una de las muestras, se observa que, ***i)*** mientras la población tiene una proporción del 10% de plantas enfermas, la media de la proporción muestral indica lo mismo, es decir, la muestra tomada tiene una proporción del 10% de plantas enfermas; ***ii)*** mientras aumenta el tamaño de la muestra se observa que la distribución de las media proporcional se va transformando en una distribución simétrica, hasta convertirse en una normal; ***iii)*** mientras aumenta el tamaño de la muestra los estimadores presentan alta precisión, por lo cual, la varianza disminuye conforme aumenta los tamaños de la muestra.

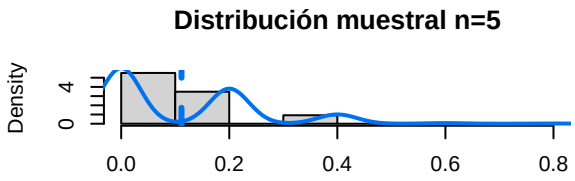


Fig 25. Media de las proporciones

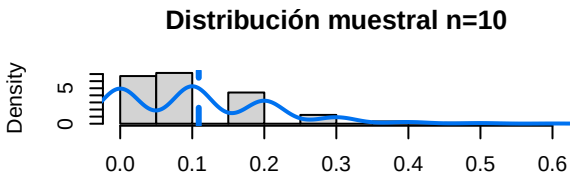


Fig 26. Media de las proporciones

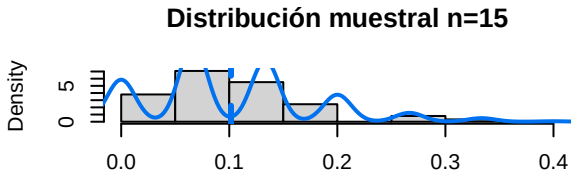


Fig 27. Media de las proporciones

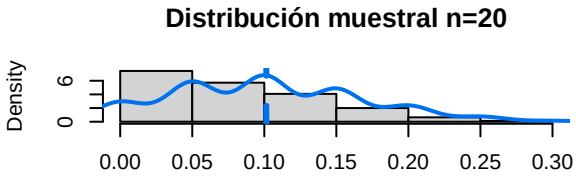


Fig 28. Media de las proporciones

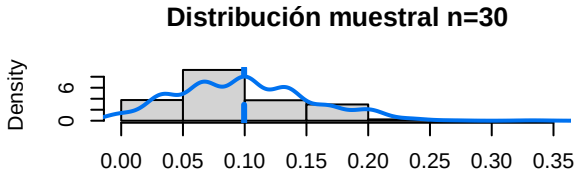


Fig 29. Media de las proporciones

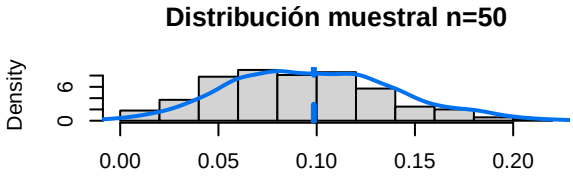


Fig 30. Media de las proporciones

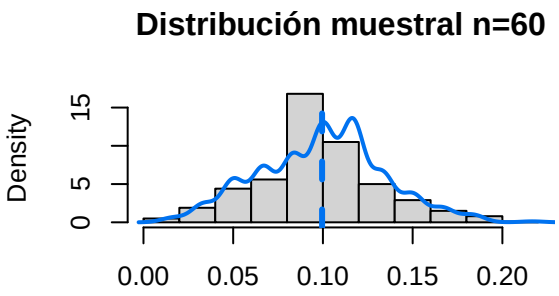


Fig 31. Media de las proporciones

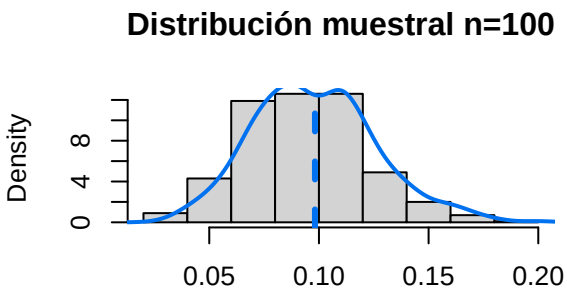


Fig 32. Media de las proporciones

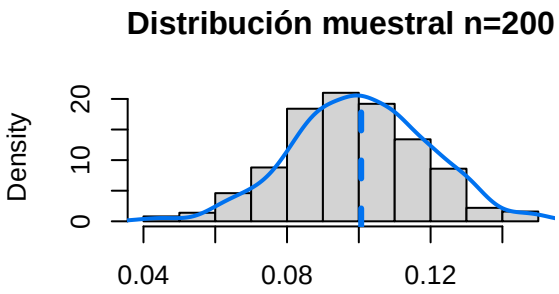


Fig 33. Media de las proporciones

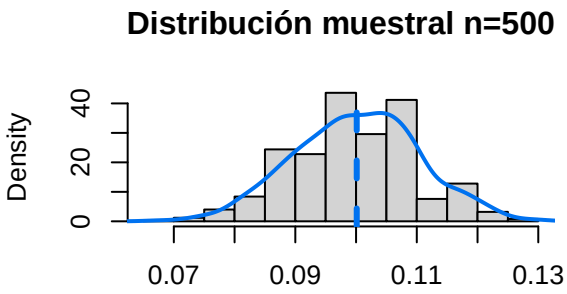


Fig 34. Media de las proporciones

```
##    muestras      media     varianza
## 1     n = 5 0.11160000 0.0196647695
## 2    n = 10 0.10900000 0.0108807615
## 3    n = 15 0.10160000 0.0059916455
## 4    n = 20 0.10140000 0.0043367134
## 5    n = 30 0.09953333 0.0029011401
## 6    n = 50 0.09856000 0.0017133531
## 7    n = 60 0.09953333 0.0012088666
## 8   n = 100 0.09818000 0.0007816509
## 9   n = 200 0.10066000 0.0003542729
## 10  n = 500 0.10011200 0.0001048211
```

Así mismo, se puede observar la comparación de los distintos tamaños de muestras ***n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500*** evidencia que al ir aumentando los tamaños muestrales los estimadores tienden a ser insesgado, eficiente y consistente.

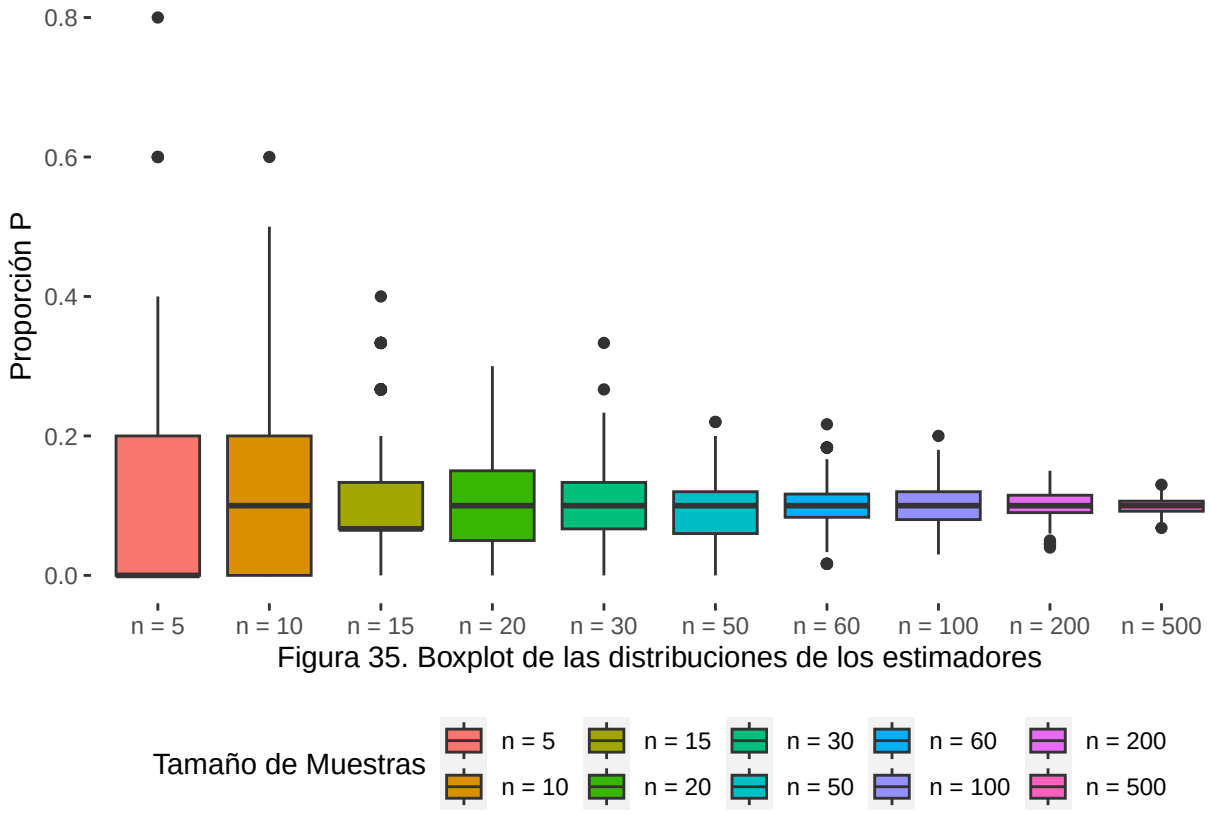
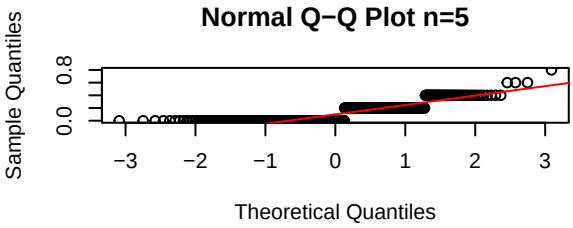


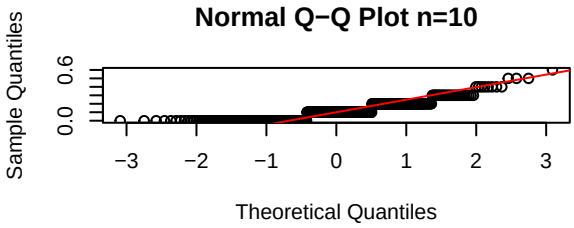
Figura 35. Boxplot de las distribuciones de los estimadores

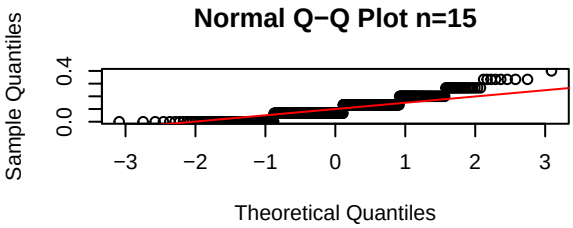
A continuación, se muestra una tabla de los resultados de la prueba ***SHAPIRO-WILK***, descrita en el primer caso, para cada uno de los tamaños de muestras $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$, donde se puede observar la convergencia de la distribución de la media muestral a una distribución normal al aumentarse el tamaño de la muestra, más especificamente, a partir del tamaño de muestra $\boldsymbol{n = 500}$.

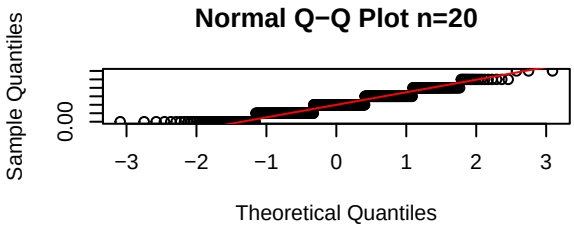
```
##    muestras shapiro
## 1     n = 5    0.00
## 2    n = 10    0.00
## 3    n = 15    0.00
## 4    n = 20    0.00
## 5    n = 30    0.00
## 6    n = 50    0.00
## 7    n = 60    0.00
## 8   n = 100    0.00
## 9   n = 200    0.01
## 10  n = 500    0.10
```

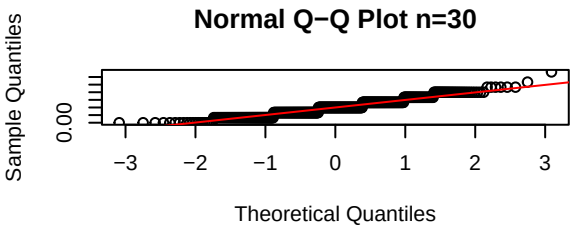
Así mismo, se muestra la prueba de normalidad de una linea recta, donde evidenciamos que a medida que aumenta los tamaños de muestras $\boldsymbol{n = 5, 10, 15, 20, 30, 50, 60, 100, 200, 500}$ los puntos se acercan a la línea recta.

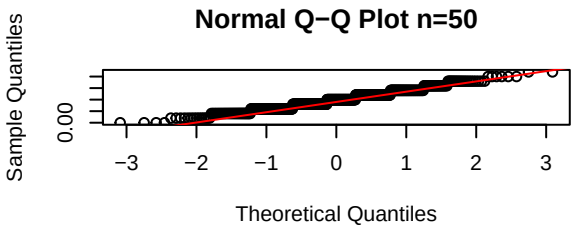

Normal Q−Q Plot n=5
Sample Quantiles
0.0
0.8
−3
−2
−1
0
1
2
3
Theoretical Quantiles

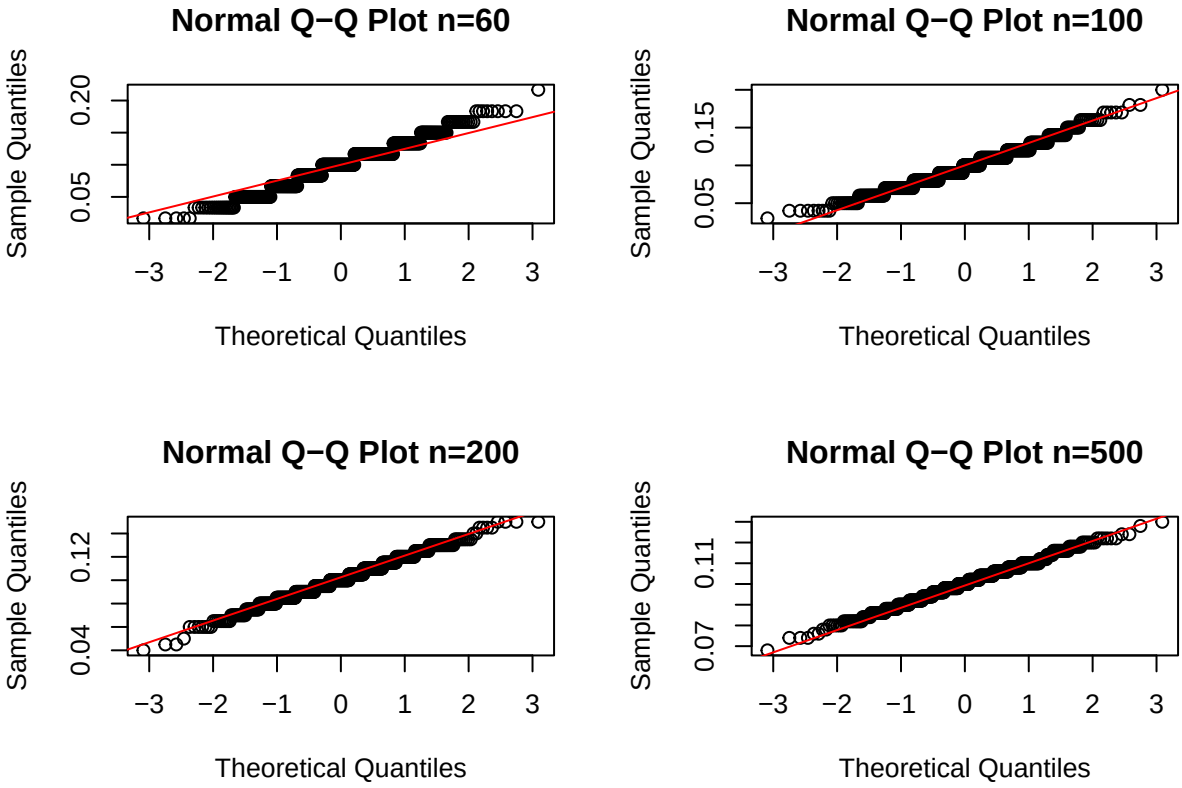

Normal Q−Q Plot n=10
Sample Quantiles
0.0
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot n=15
Sample Quantiles
0.0
0.4
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot n=20
Sample Quantiles
0.00
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot n=30
Sample Quantiles
0.00
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot n=50
Sample Quantiles
0.00
−3
−2
−1
0
1
2
3
Theoretical Quantiles



4. Conclusión

El teorema del limite central indica que, independientemente de la forma de la distribución de población, la distribución muestral de medias se aproximará a la distribución de probabilidad normal, así como se evidenció en los ***casos 2,3 y 4***, donde la proporción de plantas enfermas era distinta para cada simulación, para el caso 2 era del 50%, para el caso 3 era del 90% y para el caso 4 era del 10%, y se comprobó que la distribución muestral de medias se aproxima a la distribución normal. Así mismo, cuanto mayor sea el número de observaciones en cada muestra, mayor será la convergencia.