Problema 2 - Validación de resultados

Carlos Sierra Guzman, Camilo Vega Rámirez

Contenido

Introducción

El presente documento es la respuesta al problema 2 de la Unidad 2 del curso Métodos y Simulación Estidística.

Cada seccion esta compuesta por el puto a resolver, metodologia y resultado.

Al final del documuento se encuentra como anexos los códigos usados para la creación de las metodologias de las secciones, y se cuentan con links en el cuerpo del documento para navegar a travez del mismo.

Problema 2

> *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).*

Punto A

> *a. Realice una simulación en la cual genere dos poblaciones de $N_1 = 1000$ (Lote 1) y $N_2 = 1500$ (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).*

Metodología Punto A

Se crea la función `sim_plantas_enfermas` para simular la una proporción de plantas enfermas dada una población.

Se generan simulacines de N = 1000 con 10% de plantas enfermas y N = 1500 con 10% de plantas enfermas. Se genera tabla para comprobar que las cantidades sean las correctas.

Ir a código sección a

Resultado A

grupo	no_enferma	eneferma
grupo_1	900	100
grupo_2	1350	150

Punto B

> *b. Genere una función que permita obtener una muestra aleatoria de los lotes y calcule el estimador de la proporción muestral para cada lote ($\hat{p}1$ y $\hat{p}2$) para un tamaño de muestra dado $n_1 = n_2$. Calcule la diferencia entre los estimadores ($\hat{p}1$ - $\hat{p}2$).*

Metodología Punto B

Se crea la función `sample_prop_multi` para extraer `n` muestras de dos vector `x` y `y`, luego cacular por separado sus estimadores de proporción muestral y finalmente restar al estimador de `x` el estimador de `y`.

Se verifica el funcionamiento de la función para un `n = 500` sobre las poblacions simuladas.

Ir a código sección b

Resultado B

```
## [1] "Diferencia estimadores de prueba = 0.016"
```

Punto C

> *c. Repita el escenario anterior (b) 500 veces y analice los resultados en cuanto al comportamiento de los 500 estimadores (diferencias $\hat{p}1$ - $\hat{p}2$). ¿Qué tan simétricos son los resultados?, ¿Son siempre cero las diferencias?*

Metodología Punto C

Se crea la función `rep_sample_prop_multi` que nos permite repetir la función `sample_prop_multi` un numero `rep` de veces.

Se realiza la simulación de 500 veces las diferencia de los estimadores con muestras de `n = 500` sobre las poblaciónes simuladas.

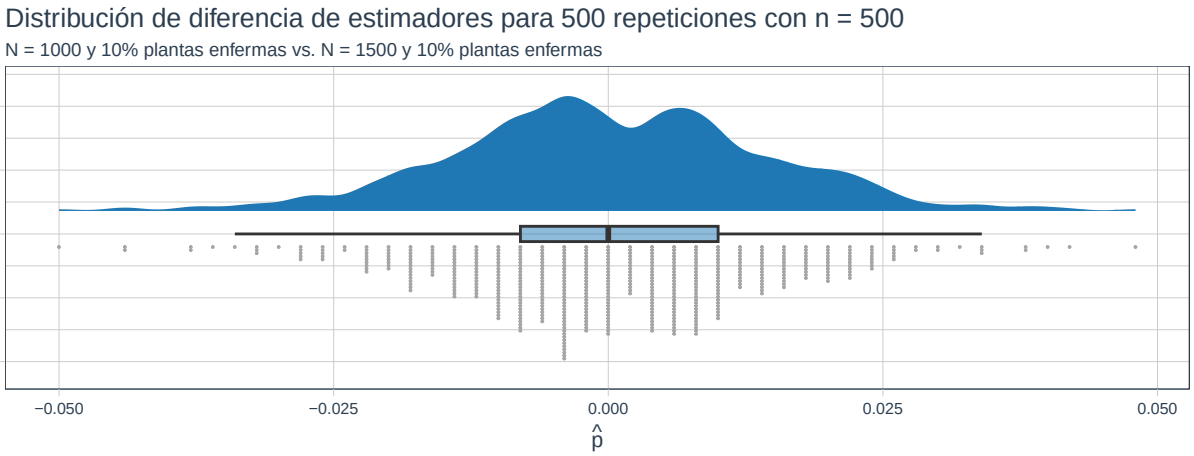
Se crea la función `gg_rain_cloud`, que toma un vector y genera un grafico de rain cloud.

Se crea la función `medidas_resumen`, que toma un vector y muestra en forma de tabla medidas de resumen respecto a simetria, sesgo y variabilidad.

Se usan `gg_rain_cloud` y `medidas_resumen` sobre las 500 simulaciones de las diferencias en los estimadores para su analisis.

Ir a código sección c

Resultado C



mean	median	sd	min	max	skewness	kurtosis
0.0002	0	0.0144	-0.05	0.048	-0.0346	0.4444

Con un tamaño de muestra n = 500 y 500 repeticiones, se observa que las diferencias de los estimadores presentan indicadores de skewness y kurtosis, bajos, que sumados a la grafica nos muestran que los datos pueden considerarse simetricos, igualmente tanto la mediana como el promedio del estimador se aproximan a 0 lo que nos indica que la distribución de las diferencias de los estimadores es in-sesgadas.

Vemos que las diferencias no siempre son 0 pero la tendencia media es que las diferencia se aproxima a ese valor, notando la presencia de alguns outlies con diferencias de $\pm$ 5%.

Punto D

> *d. Realice los puntos b y c para tamaños de muestra $n_1 = n_2 = 5$, 10, 15, 20, 30, 50, 60, 100, 200, 500. Compare los resultados de los estimadores ($\hat{p}1$ - $\hat{p}2$) 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?*

Metodología Punto D

Se realiza la simulación de 500 veces el calculo de las diferecias de los estmadores con multiles tamaños de muestra `n` (5, 10, 15, 20, 30, 50, 60, 100, 200 y 500) sobre las poblaciones simuladas, y se colocan en un data frame.

Se crea la función `medidas_resumen_multiple`, que toma un data frame y muestra en forma de tabla medidas de resumen y tests de normalidad de una columna seleccionada agupados por otra columna seleccionada.

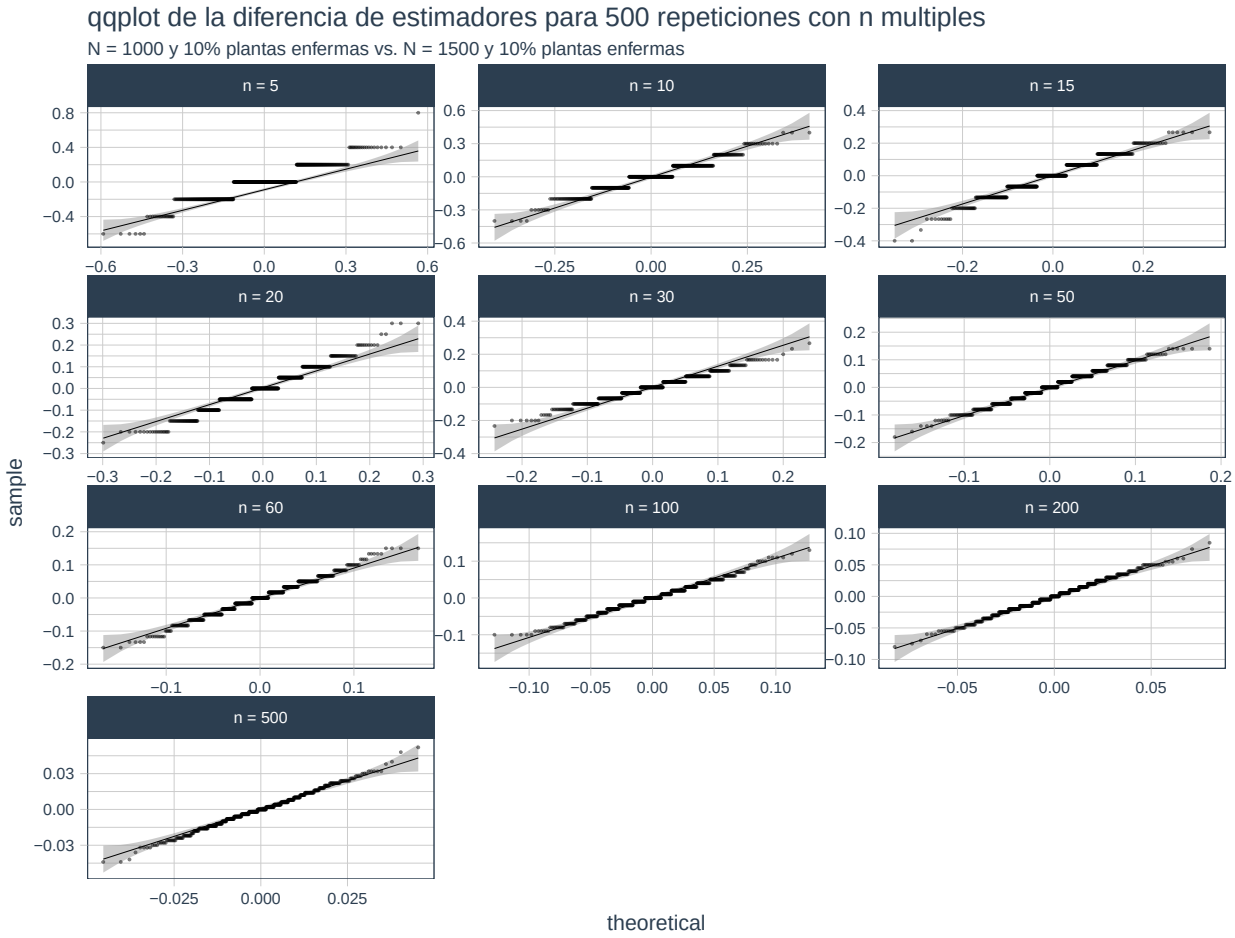
Se crea la función `gg_qq_plot`, que toma un data frame y realiza graficos de normalidad tipo `qqnor` de una columna seleccionada, agrupados por otra columna seleccionada.

Se usan `medidas_resumen_multiple` y `gg_qq_plot` sobre el data frame con las 500 simulaciones de diferencias de estimadores para distintos tamaños de nuestra `n` para su analisis.

Ir a código sección d

Resultado D

n	mean	median	sd	Shapiro-Wilk test P-Value
5	-0.0128	0	0.1874	0.0000
10	0.0024	0	0.1325	0.0000
15	-0.0017	0	0.1129	0.0000
20	-0.0040	0	0.0955	0.0000
30	-0.0009	0	0.0781	0.0000
50	0.0027	0	0.0596	0.0001
60	0.0008	0	0.0544	0.0012
100	-0.0004	0	0.0414	0.0024
200	-0.0011	0	0.0263	0.1494
500	0.0000	0	0.0147	0.1694



Podemos ver que a medida que aumentan los tamaños de muestas el promedio de las diferencias de los estimadores se aproxima cada vez más a 0, sin embargo la reducción no es constante y preseta variaciones

por arriba o abajo de este valor y e incluso en algunos casos aumenta como podemos ver en n = 200 donde su promedio esta (en valores absulutos) por arriba incluso de la diferencia de n = 60. Respecto a la desviación estardar esta disminuye cosntante mente a mediada que aumenta n. Así mismo se observa en las graficas de qqnorm que a mayor n, la distribución de las diferencias de los estimadores se parece más a una distrivución normal lo cual se comprueba con el test de Shapiro-Wilk el cual es positivo para normalidad a partir de n = 200.

Vemos que a mayor tamaño de muestra disminuye la probabilidad de concluir que hay diferencias entre la comparación de proporciones para los tratamientios, ya que los promedios se aproximan a 0, el pricipal factor para determinar esto son las desviaciones son cada vez menores pasando de un n = 5 cons desviaciones muy altas cercanas al 19% a desviaciones de solo 1.5% aproximadamente para n = 500.

Punto E

> *e. Ahora realice nuevamente los puntos a-d bajo un escenario con dos lotes, pero de proporciones de enfermos diferentes ($\hat{p}1 = 0.1$ y $\hat{p}2 = 0.15$), 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 ($\hat{p}1$ - $\hat{p}2$) 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 $\hat{p}1$ - $\hat{p}2$ bajo ambos escenarios (escenario 1: sin diferencias entre $\hat{p}1$ y $\hat{p}2$, escenario 2: diferencia de 5%)?*

Metodología Punto E

Se genera la simulación de N = 1000 con 15% de plantas enfermas, uniendola con la simulación ya generadad de N = 1000 con 10% de plantas enfermas y se genera tabla para comprobar que las cantidades sean las correctas.

Se realiza la simulación de 500 veces el calculo de las diferencias de los estimadores, con una muestra de `n = 500` sobre la poblaciónes simuladas.

Se usan `gg_rain_cloud` y `medidas_resumen` sobre las 500 simulaciones de diferencias de estimadores para su analisis.

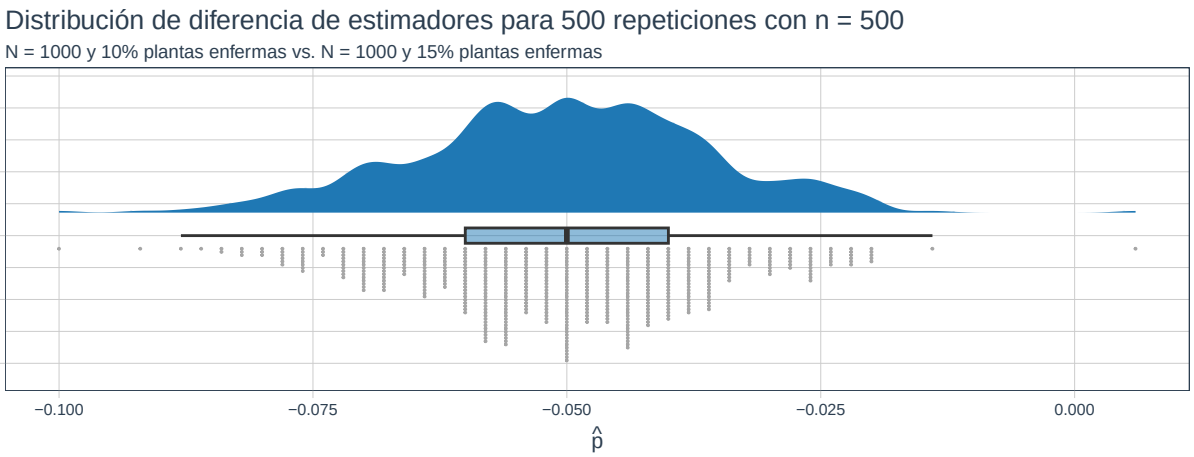
Se realiza la simulación de 500 veces el calculo de las diferencias de los estimadoes, con multiples tamaños de muiestra `n` (5, 10, 15, 20, 30, 50, 60, 100, 200 y 500) sobre las poblaciones simuladas, y se colocan en un data frame.

Se usan `medidas_resumen_multiple` y `gg_qq_plot` sobre el data frame con las 500 simulaciones de las diferencias de los estimadoes para distintos tamaños de nuestra `n` sobre las poblaciónes simuladad para su analisis,

Ir a código sección e

Resultado E

grupo	no_enferma	eneferma
grupo_1	900	100
grupo_3	850	150

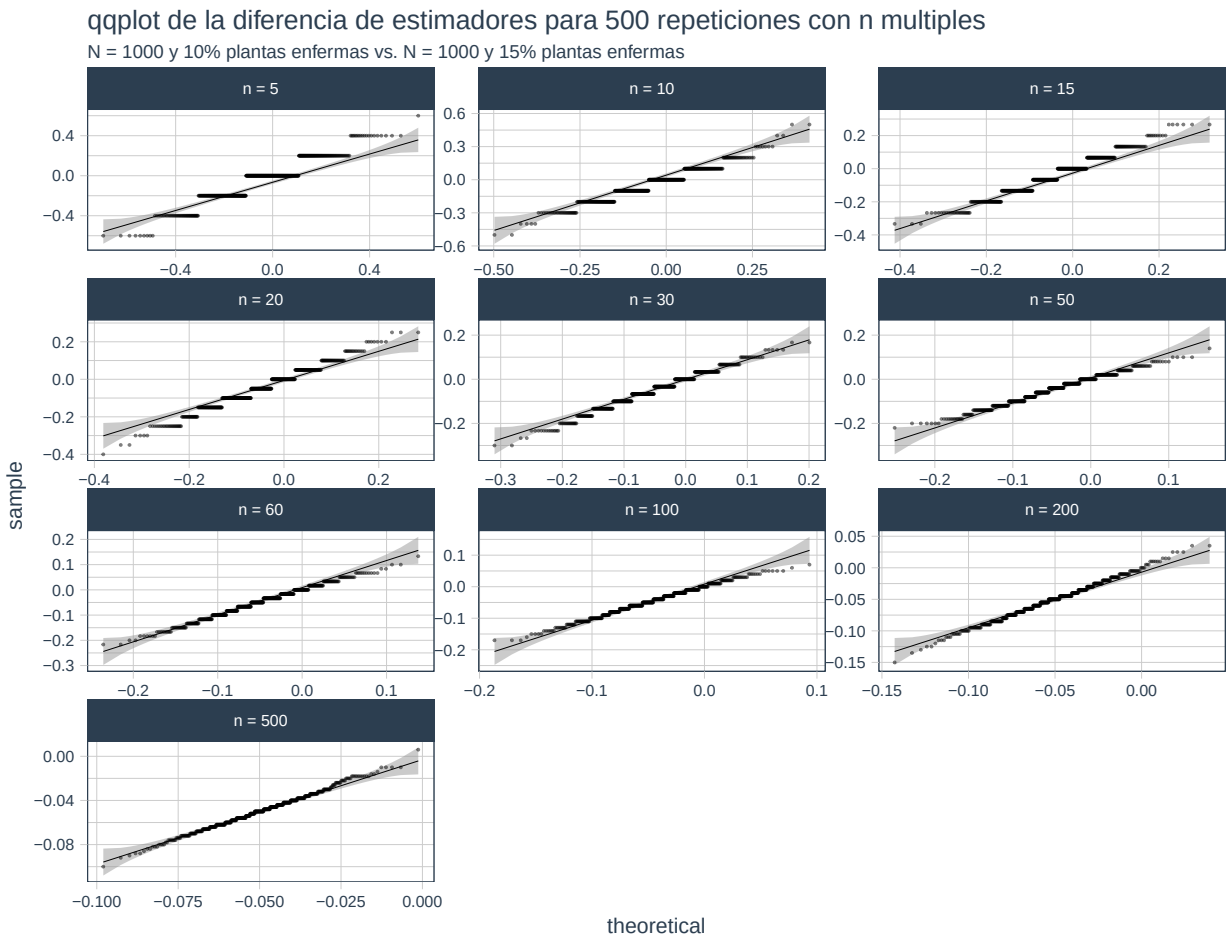


mean	median	sd	min	max	skewness	kurtosis
-0.0503	-0.05	0.0144	-0.1	0.006	-0.0735	0.2432

Con un tamaño de muestra n = 500 y 500 repeticiones, se observa que las diferencias de los estimadores de plantas enfermas al 10% vs 15% presentan indicadores de skewness y kurtosis, bajos, que sumados a la grafica nos muestran que los datos pueden considerarse simetricos, igualmente tanto la mediana como el promedio del estimador se aproximan a -5% lo que nos indica que la distribución de las diferencias de los estimadores es in-sesgadas.

El resultado en lo referente a normalidad es similar al de la comparación de plantas enfermas al 10% para tamaños de población distintos, mostrando en los dos casos que con un tamaño de muestra n = 50 las distribuciones se asemejan a la normalidad.

n	mean	median	sd	Shapiro-Wilk test P-Value
5	-0.0488	0.0000	0.2104	0.0000
10	-0.0412	0.0000	0.1479	0.0000
15	-0.0476	-0.0667	0.1182	0.0000
20	-0.0489	-0.0500	0.1076	0.0000
30	-0.0546	-0.0667	0.0827	0.0000
50	-0.0496	-0.0400	0.0654	0.0001
60	-0.0490	-0.0500	0.0604	0.0030
100	-0.0466	-0.0500	0.0454	0.0088
200	-0.0517	-0.0500	0.0295	0.2128
500	-0.0496	-0.0500	0.0157	0.2803



Podemos ver que a medida que aumentan los tamaños de muestas el promedio de las diferencias de los estimadores se aproxima cada vez más a -5%, sin embargo nunca se llega al valor exacto a diferencia de lo ocurrido en elescenario de diferencia de dos poblaciones con igual número de plantas enfermas donde se obtuvo una diferncias estimada igual a la poblacional en n = 500. Respecto a la desviación estardar esta disminuye cosntantemente a mediada que aumenta n, sin embago sus valores son muy altos en tamaños de muestra altos llegando hasta 21% en n = 5.

En las graficas de qqnorm de nota que a mayor n, la distribución de las diferencias de los estimadores se parece más a una distribución normal lo cual se comprueba con el test de Shapiro-Wilk el cual es positivo para normalidad a partir de n = 200.

Comparando los dos escenarios *escenario 1: sin diferencias entre p^1 y p^2, escenario 2: diferencia de 5%*, vemos que a tamaños de muestras bajos las desviaciones son muy altas a pesar de que los valores medios de las diferencias son cercanos al real, lo cual dependiendo de las muestras podria llevar a escenarios donde erroneamente se prodria indicar una diferencia en los tratamientos para el escenario 1, o sobrestimar el tamaño de la diferencia en el escenario 2. Estas altas desviaciones van disminuyendo a medida que se aumenta el tamaño de dandonos para el escenario 1 que no hay diferencia entre los tratamientos y para el escenario 2 que hay una diferencia aproximada del 5% en los tratamiendos siendo el tratamiento del la Población 1 mejor.

Conclusión

Se concluye que para el caso de la comparación de tratamientos entre 2 poblaciones comparados por la diferecnia entre las proporciones de exitos, es recomendable el contar con metodos rigurosos en lo referente a las repeticiones del experimiento y el tamaño de muestra para el analisis de resultados con el objetivo de contar con la mayor veracidad en la conclusión de los mismos.

Anexos

Código Librerias

```
library(tidyverse)      # Transformación de datos
library(normtest)       # Pruebas de normalidad
library(knitr)          # Renderizar tablas
library(ggdist)         # Expanción de graficas de ggplot
library(tidyquant)      # Tema de graficas de ggplot
library(nortest)        # Pruebas de normalidad
library(rapportools)    # Pruebas de normalidad
library(qqplotr)        # QQplot usando ggplot
options(scipen = 999)   # Anulación de notación cientifica
```

Código Punto A

Volver a metodología sección a

```
# Función para generar población n, con una propoporción prop de plantas enfermas
sim_plantas_enfermas <- function(n, prop){
    p <- round(n*prop)
    q <- n-p
    c(rep(TRUE,p), rep(FALSE,q))
}
# Simulando para N = 1000 y 0.1 de plantas enfermas
plantas_enfermas_10_1 <- sim_plantas_enfermas(1000, 0.1)

# Simulando para N = 1500 y 0.1 de plantas enfermas
plantas_enfermas_10_2 <- sim_plantas_enfermas(1500, 0.1)

# Tabla para visualizar simulación
grupo1 <- table(plantas_enfermas_10_1) %>%
    as_tibble() %>%
    mutate(grupo = "grupo_1") %>%
    rename(enferma = 1)

grupo2 <- table(plantas_enfermas_10_2) %>%
    as_tibble() %>%
    mutate(grupo = "grupo_2") %>%
    rename(enferma = 1)

bind_rows(grupo1, grupo2) %>%
```

```
    pivot_wider(names_from = enferma, values_from = n) %>%
    set_names(c("grupo", "no_enferma", "eneferma")) %>%
    kable()
```

Código Punto B

Volver a metodología sección b

```
# Función para tomar una muestra de tamaño n para los vectores x,y  calcular la
# proporcion del estimador para cada uno y restal al estimadoe de x el estimador
# de y
sample_prop_multi <- function(x, y, n){
    x_sample <- sample(x, n) %>%
        sum()/n
    y_sample <- sample(y, n) %>%
        sum()/n
    x_sample-y_sample
}

# Reproducibilidad
set.seed(4321)

# Test de función con n = 500, sobre los vectores plantas_enfermas_10_1 y
# plantas_enfermas_10_2

str_c("Diferencia estimadores de prueba = ",
sample_prop_multi(plantas_enfermas_10_1, plantas_enfermas_10_2,500))
```

Código Punto C

Volver a metodología sección c

```
# Función para repetir la función rep_sample_prop_multi un rep número de veces
rep_sample_prop_multi <- function(x, y, n, rep){
    map_dbl(1:rep, ~ sample_prop_multi(x, y, n))
}

# Reproducibilidad
set.seed(4321)

# Creación de 500  simulaciones de diferenciaas de estimadores, para un n = 500
# de muestras de plantas_enfermas_10_1 y plantas_enfermas_10_2
muesta_repetida_10_multi <-  rep_sample_prop_multi(plantas_enfermas_10_1,
                                                   plantas_enfermas_10_2,500, 500)

# Función para crear grafico de rain cloud sobre un vector
gg_rain_cloud <- function(x, title, subtitle){
    ggplot(x %>% as_tibble() , aes( y = value)) +
    stat_halfeye(adjust = 0.5, justification = -0.2, .width = 0, point_colour = NA,
                 fill = "#1F78B4") +
    geom_boxplot(width = 0.12, outlier.color = NA, alpha = 0.5, fill = "#1F78B4") +
    stat_dots(side = "left", justification = 1.1, fill = "#1F78B4") +
```

```
    coord_flip() +
    theme_tq() +
    scale_fill_tq(theme = "light") +
    theme(axis.text.y = element_blank(),
          text = element_text(size = 8)) +
    xlab("") +
    ylab(expression(hat("p"))) +
    ggtitle(label = title,
            subtitle = subtitle)
}

# Función para crear tabla con medidas de resumen sobre un vector
medidas_resumen <-  function(x){
    x %>%
    as_tibble() %>%
    summarise(mean = mean(value), median = median(value), sd = sd(value),
              min = min(value), max = max(value),
              skewness = skewness(value), kurtosis= kurtosis(value))  %>%
    mutate(across(where(is.numeric), ~ round(.,4))) %>%
    kable()
}

# Creación rain cloud y tabla de resumen sobre las 500 rerpeticiones con muestra
# n = 500, sobre las diferencias de estimadores simuladas.
gg_rain_cloud(
    muesta_repetida_10_multi,
    "Distribución de diferencia de estimadores para 500 repeticiones con n = 500",
    "N = 1000 y 10% plantas enfermas vs. N = 1500 y 10% plantas enfermas")
medidas_resumen(muesta_repetida_10_multi)
```

Código Punto D

Volver a metodología sección d

```
# Reproducibilidad
set.seed(4321)

# Creación de 500 estimadores, para multiples n de muestras de plantas_enfermas_10_1
# y plantas_enfermas_10_2
muesta_repetida_multiple_10 <-  map_df(c(5, 10, 15, 20, 30, 50, 60, 100, 200, 500),
                                       ~ tibble(
                                           diff_p_hat = rep_sample_prop_multi(
                                               plantas_enfermas_10_1,
                                               plantas_enfermas_10_2,
                                               ., 500),
                                           n = as_factor(.)))

# Función para crear tabla con medidas de resumen sobre un data frame, para una
# columna agrupada por otra columna
medidas_resumen_multiple <- function(df, value, group){
  value <- enquo(value)
  group <- enquo(group)
  df %>%
```

```
    group_by(!!group) %>%
    summarise(mean = mean(!!value), median = median(!!value),
              sd = sd(!!value),
              `Shapiro-Wilk test P-Value` = shapiro.test(!!value)$p.value) %>%
    ungroup() %>%
    mutate(across(where(is.numeric), ~ round(.,4))) %>%
    kable()
}

# Función para crear grafico de qqplot sobre un data frame, para una columna
# agrupada por otra columna
gg_qq_plot <- function(df, value, group, title, subtitle){
  ggplot(df,  aes(sample = {{value}})) +
    stat_qq_band(alpha = 0.5) +
    stat_qq_line(linewidth = 0.1) +
    stat_qq_point(alpha = 0.5, size = 0) +
    facet_wrap(vars(factor(str_c("n = ",{{group}}),
                           levels = c(str_c("n = ",c(5, 10, 15, 20, 30, 50, 60,
                                                     100, 200, 500))))),
               nrow =  4, scales = "free") +
    theme_tq() +
    theme(panel.spacing = unit(0, "lines"),
          text = element_text(size = 8)) +
    ggtitle(label = title,
            subtitle = subtitle)
}


# Creación de tabla de resumen y qqplots sobre las 500 rerpeticiones con muestra
# n multiples, sobre la población simulada
medidas_resumen_multiple(muesta_repetida_multiple_10, diff_p_hat, n)
gg_qq_plot(muesta_repetida_multiple_10, diff_p_hat, n,
           "qqplot de la diferencia de estimadores para 500 repeticiones con n multiples",
           "N = 1000 y 10% plantas enfermas vs. N = 1500 y 10% plantas enfermas")
```

Código Punto E

Volver a metodología sección e

```
#Simulando para N = 1000 y 0.15 de plantas enfermas
plantas_enfermas_15_1 <- sim_plantas_enfermas(1000, 0.15)

# Tabla para visualizar simulación
grupo3 <- table(plantas_enfermas_15_1) %>%
    as_tibble() %>%
    mutate(grupo = "grupo_3") %>%
    rename(enferma = 1)

bind_rows(grupo1, grupo3) %>%
    pivot_wider(names_from = enferma, values_from = n) %>%
    set_names(c("grupo", "no_enferma", "eneferma")) %>%
    kable()
```

```r
# Reproducibilidad
set.seed(4321)

# Creación de 500 estimadores, para un n = 500 de muestras de plantas_enfermas_10_1
# y plantas_enfermas_15_1
muesta_repetida_10_15_multi <-  rep_sample_prop_multi(plantas_enfermas_10_1,
                                                      plantas_enfermas_15_1,500, 500)

# Creación rain cloud y tabla de resumen sobre las 500 rerpeticiones con muestra
# n = 500, sobre la población simulada al 10% de plantas enfermas
gg_rain_cloud(
muesta_repetida_10_15_multi,
"Distribución de diferencia de estimadores para 500 repeticiones con n = 500",
    "N = 1000 y 10% plantas enfermas vs. N = 1000 y 15% plantas enfermas")
medidas_resumen(muesta_repetida_10_15_multi)
```

```r
# Reproducibilidad
set.seed(4321)

# Creación de 500 estimadores, para multiples n de muestras de plantas_enfermas_10_1
# y plantas_enfermas_15_1
muesta_repetida_multiple_10_15 <-  map_df(c(5, 10, 15, 20, 30, 50, 60, 100, 200, 500),
                                    ~ tibble(
                                        diff_p_hat = rep_sample_prop_multi(
                                            plantas_enfermas_10_1,
                                            plantas_enfermas_15_1,
                                            ., 500),
                                        n = as_factor(.)))

# Creación de tabla de resumen y qqplots sobre las 500 rerpeticiones con muestra
# n multiples, sobre las poblaciónes simuladas
medidas_resumen_multiple(muesta_repetida_multiple_10_15, diff_p_hat, n)
gg_qq_plot(muesta_repetida_multiple_10_15, diff_p_hat, n,
"qqplot de la diferencia de estimadores para 500 repeticiones con n multiples",
          "N = 1000 y 10% plantas enfermas vs. N = 1000 y 15% plantas enfermas")
```