Trabajo 1

Matias Bajac - Lucas Pescetto - Andres Vidal

2023-04-10

La variable que elegimos para trabajar es el total de hogares que no tienen acceso a computadoras XO. Para eso, usamos la variable "HOGCE09", que indica la cantidad de dispositivos que hay en el hogar.
En cuanto a la base de datos, creamos una variable indicadora para cada hogar, uniendo la variable identificadora de viviendas y el n° de hogar; para así quedarnos con una sola observación por hogar.
Luego creamos nuestras variables de interés: - **NBI** vale 0 si el hogar tiene 3 o menos NBI y 1 si tiene 4 o más. - **XO** vale 0 si el hogar tiene algún dispositivo, y 1 en caso de no contar con ninguno. -

```
datos <- load(here("Datos","RB (1).RData"))
datos <-  rio_branco
rm(rio_branco)

# convertimos los 8 y 9 en 0
var_names <- names(datos)[grepl("^NBI_", names(datos))][-13]
for (var_name in var_names) {
 datos[[var_name]] <- gsub("[89]", "0", datos[[var_name]])
}
# pasamos las variables a numericas
for (var_name in var_names) {
 datos[[var_name]] <- as.numeric(datos[[var_name]])
}

datos_hogares <- datos %>%
  mutate(ID = paste(ID_VIVIENDA,HOGID))  %>%
  filter(!duplicated(ID))  %>%
  mutate(NBI = NBI_EDUCACIÓN + NBI_HAC + NBI_MAT + NBI_COC + NBI_VIV + NBI_AGUA + NBI_SANEA + NBI_ELECT
  mutate(NBI = if_else(NBI>3, 1, 0), XO = if_else(HOGCE09 == 0,1,0)) %>%
  select(ID, NBI, XO)
```

El total poblacional de NBI a nivel hogares es 340:

$$Nos\ basaremos\ en\ el\ estmiador\ Horvitz\ thompson\ para\ estimar\ el\ total\ poblacional\ de\ la\ variable\ NBI$$

$$en\ el\ Diseño\ Simple\ la\ probabilidad\ de\ inclusion\ de\ primer\ orden\ es\ \pi_k = n/N$$

$$del\ estimador\ H-T\ sabemos\ que\ t_\pi = \sum_s y_k/\pi_k$$

$$por\ lo\ tanto\ t_\pi = N * \bar{y_s}$$

Una vez obtenida la base, procedemos a crear funciones que permiten obtener las muestras y los respectivos estimadores para cada diseño. Elegimos trabajar con el Bernoulli y el SIR además del simple. Para el diseño Bernoulli, la muestra se hace simulando una U ~ (0,1) para cada observacion de la poblacion y luego seleccionando las filas en las cuales los valores sean menor a la probabiliad de inclusion de primer orden

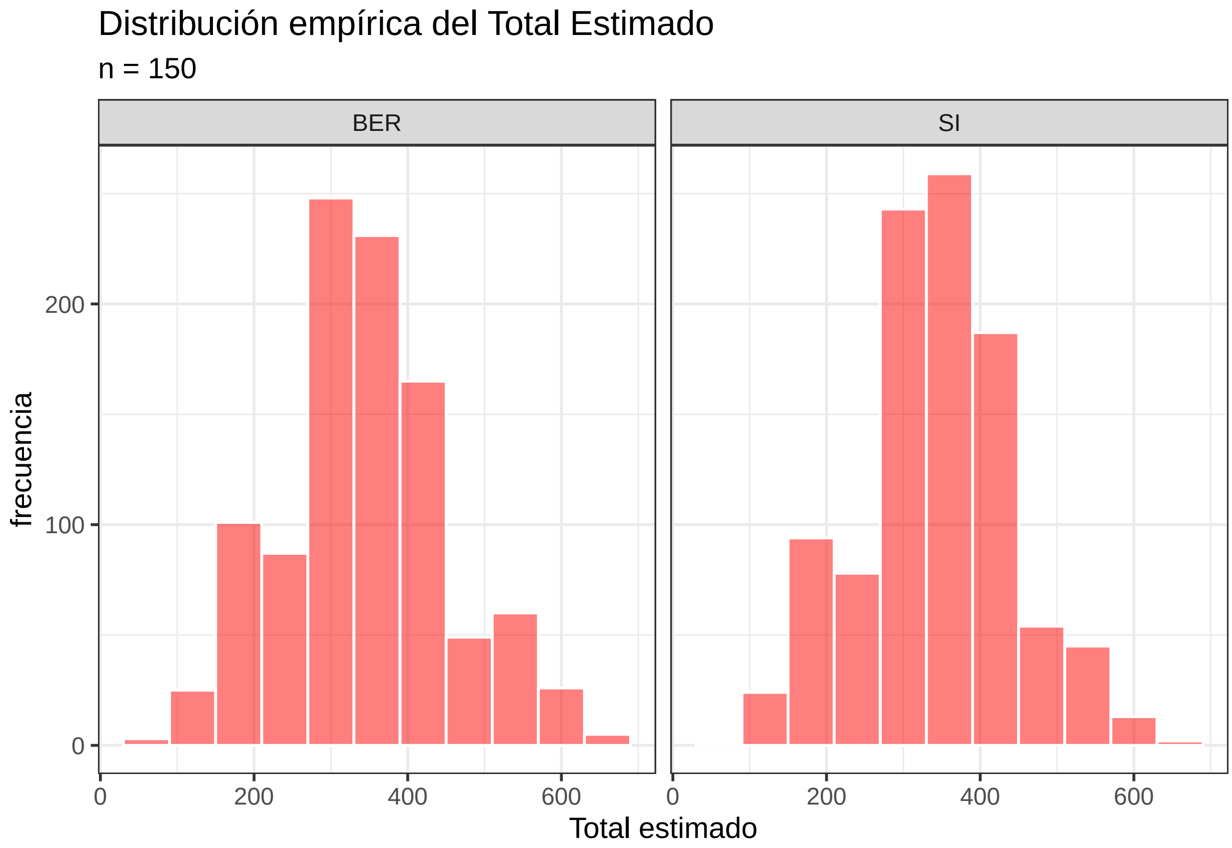
```r
# Obtenemos el tamaño de la población y establecemos la cantidad de simulaciones
set.seed(1234)

N <- nrow(datos_hogares)
R <- 1000

SI <-  function(n, var) {
  t_si <- numeric()
  for (i in 1:R){
    s <- srswor(n, N)
    m <- getdata(datos_hogares, s)
    t_si[i] <- N*mean(m[[var]])
  }
  return(t_si)
}

BER <- function(n, var) {
  t_ber <- numeric()
  # elegimos pi_k para que el tamaño de muestra esperado sea el requerido
  pi_k <- n/N
  for (i in 1:R) {
    datos_hogares$epsilon <- runif(nrow(datos_hogares))
    m <- datos_hogares %>% filter(epsilon < pi_k)
    t_ber[i] <- sum(m[[var]])/pi_k
  }
  return(t_ber)
}
```

Parte 1

Total de hogares con 4 o más NBI

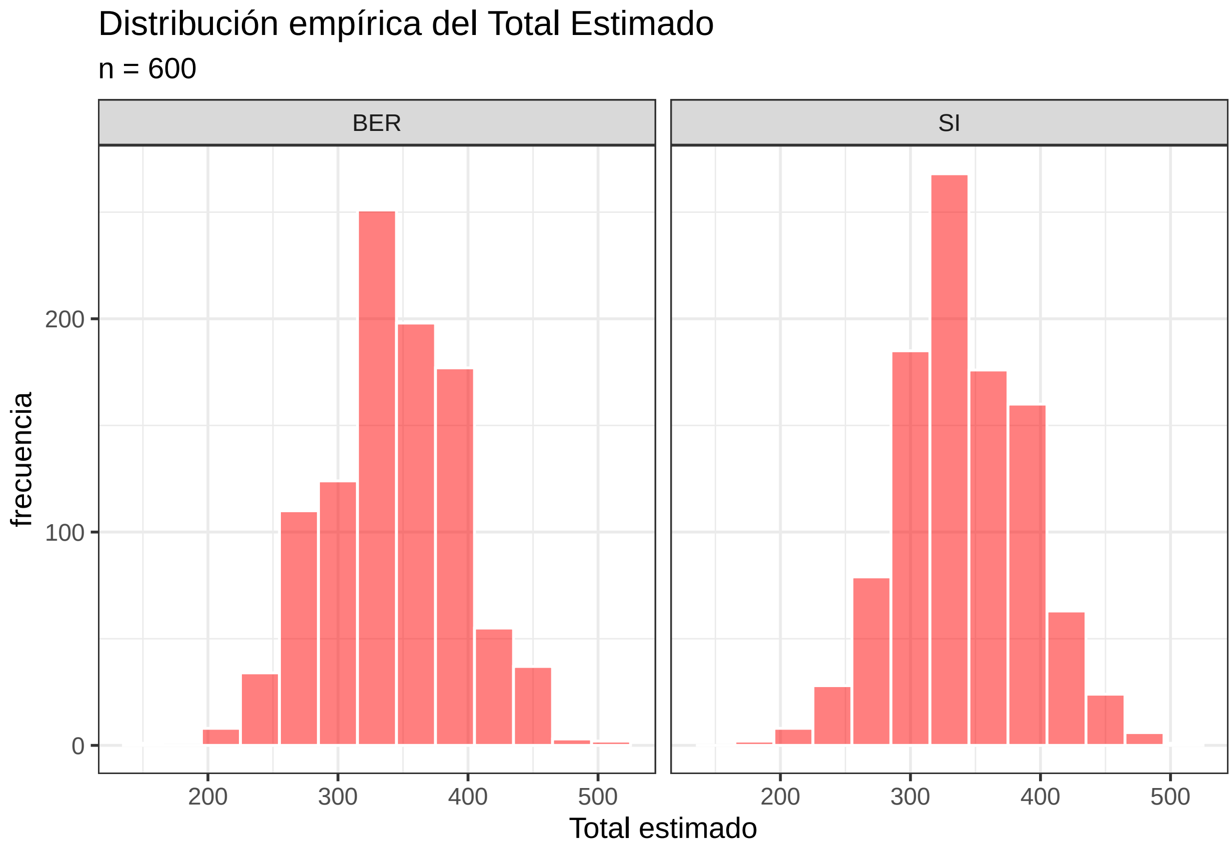
Tamaño de muestra n = 150

```r
set.seed(1234)
t1_SI <-  SI(150, "NBI")
t1_BER <- BER(150, "NBI")

ggplot(data.frame(t1_SI, t1_BER) %>% pivot_longer(cols = everything()), aes(x = value)) +   geom_histog
    theme_bw() +
  labs(x = "Total estimado", y = "frecuencia", title = "Distribución empírica del Total Estimado", subti
  facet_wrap(~name, labeller = labeller(name = function(var, name) {
    labels <- c("BER", "SI")
    return(labels[name])
  }))
```



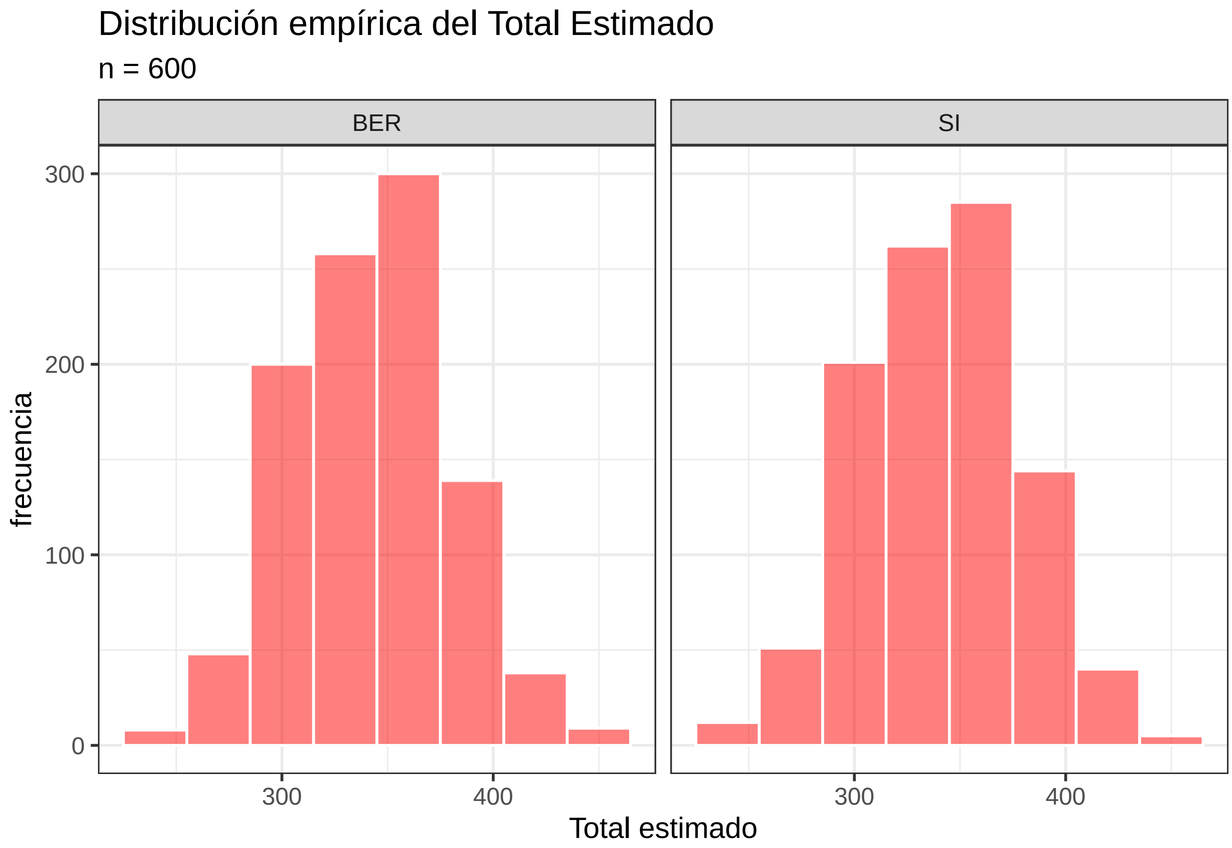
Tamaño de muestra n = 600

```r
set.seed(1234)

t2_SI <-  SI(600, "NBI")
t2_BER <- BER(600, "NBI")

ggplot(data.frame(t2_SI, t2_BER) %>% pivot_longer(cols = everything()), aes(x = value)) +   geom_histog
  theme_bw() +
  labs(x = "Total estimado", y = "frecuencia", title = "Distribución empírica del Total Estimado", subt
  facet_wrap(~name, labeller = labeller(name = function(var, name) {
    labels <- c("BER", "SI")
    return(labels[name])
  }))
```



Tamaño de muestra n = 1000

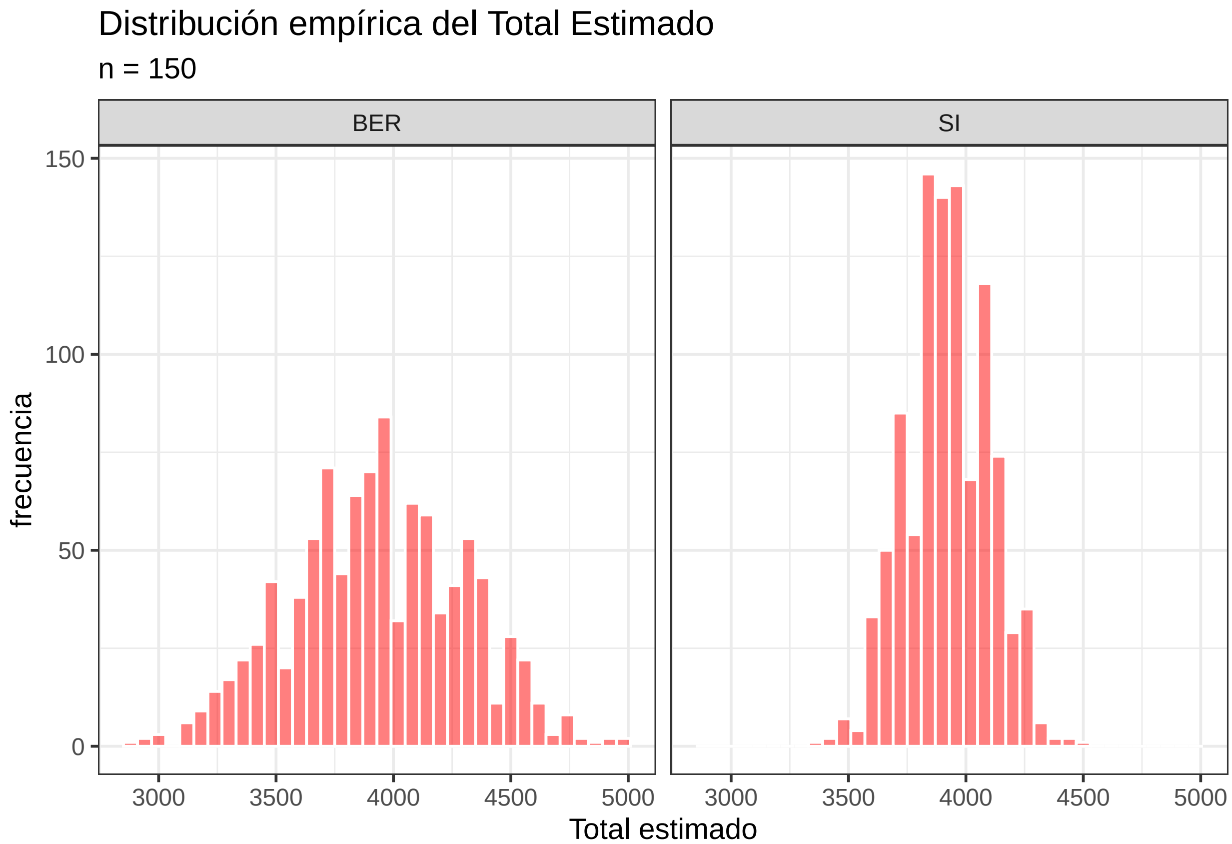
```
set.seed(1234)
t3_SI <-  SI(1000, "NBI")
t3_BER <- BER(1000, "NBI")

ggplot(data.frame(t3_SI, t3_BER) %>% pivot_longer(cols = everything()), aes(x = value)) +   geom_histog
  theme_bw() +
  labs(x = "Total estimado", y = "frecuencia", title = "Distribución empírica del Total Estimado", subt
  facet_wrap(~name, labeller = labeller(name = function(var, name) {
    labels <- c("BER", "SI")
    return(labels[name])
  }))
```



Total de hogares con acceso a XO

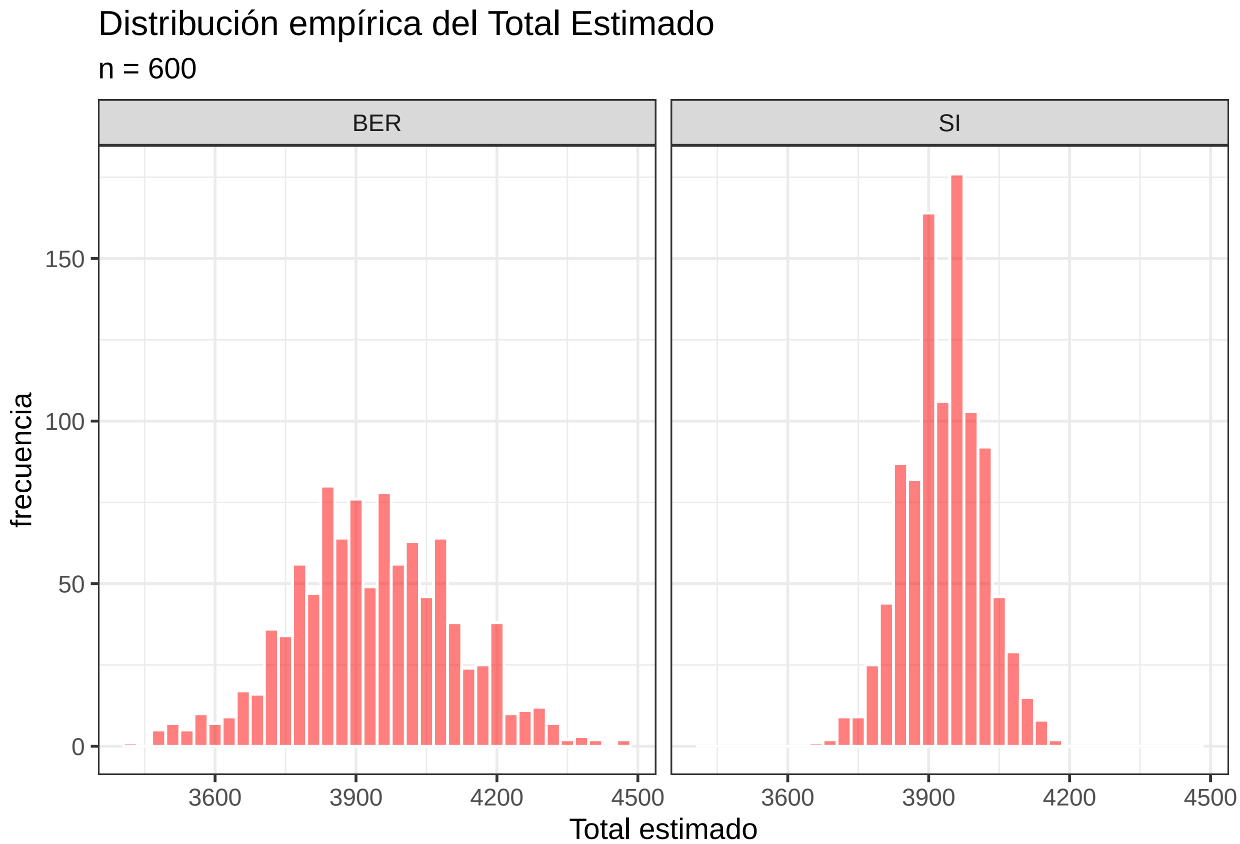
Tamaño de muestra n = 150

```
set.seed(1234)
t4_SI <-  SI(150, "XO")
t4_BER <- BER(150, "XO")


ggplot(data.frame(t4_SI, t4_BER) %>% pivot_longer(cols = everything()), aes(x = value)) +   geom_histog
    theme_bw() +
  labs(x = "Total estimado", y = "frecuencia", title = "Distribución empírica del Total Estimado", subt
  facet_wrap(~name, labeller = labeller(name = function(var, name) {
    labels <- c("BER", "SI")
    return(labels[name])
  }))
```



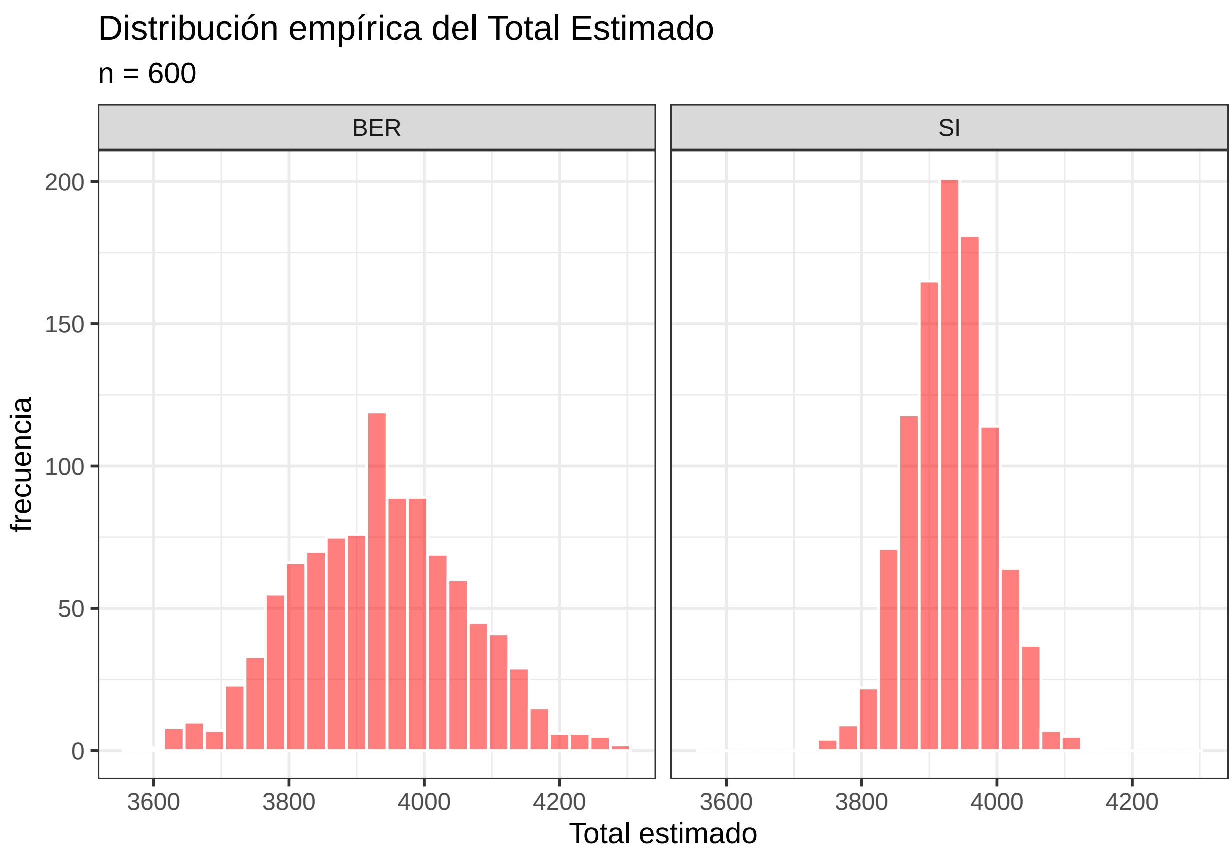
Tamaño de muestra n = 600

```
set.seed(1234)
t5_SI <-  SI(600, "XO")
t5_BER <- BER(600, "XO")

ggplot(data.frame(t5_SI, t5_BER) %>% pivot_longer(cols = everything()), aes(x = value)) +   geom_histog
  theme_bw() +
  labs(x = "Total estimado", y = "frecuencia", title = "Distribución empírica del Total Estimado", subti
  facet_wrap(~name, labeller = labeller(name = function(var, name) {
    labels <- c("BER", "SI")
    return(labels[name])
  }))
```



Tamaño de muestra n = 1000

```
set.seed(1234)
t6_SI <-  SI(1000, "XO")
t6_BER <- BER(1000, "XO")

ggplot(data.frame(t6_SI, t6_BER) %>% pivot_longer(cols = everything()), aes(x = value)) +   geom_histog
  theme_bw() +
  labs(x = "Total estimado", y = "frecuencia", title = "Distribución empírica del Total Estimado", subt
  facet_wrap(~name, labeller = labeller(name = function(var, name) {
    labels <- c("BER", "SI")
    return(labels[name])
  }))
```



Observamos que en los 3 casos se cumple que el estimador $\hat{t}$ es insesgado

Calculamos la varianza del estimador t para cada numero de muestra como paso previo para luego calcular el efecto diseño.

Calculamos con la varianza teorica y la comparamos con la simulada

cuadro de texto con max y min

```
annotate("text", x = 495, y = 290, hjust = 1, vjust = 1, label = paste("min:", round(min(t3_SI),2), "\n
  geom_rect(aes(xmin = 425, xmax = 500, ymin = 250, ymax = 300), fill = NA, color = "black")
```

#tabla comparativa

```
options(xtable.comment = FALSE)
n1= 150

v_si1 <- N^2*(1-n1/N)*var(datos_hogares$NBI)/n1
v_ber1 <- ((1-n1/N)/(n1/N))*sum(datos_hogares$NBI**2)


df = datos_hogares %>%  summarise( total_NBI = sum(NBI), total_estimado = mean(t1_SI), varianza_estimad

df2 = datos_hogares %>%  summarise( total_NBI = sum(NBI), total_estimado = mean(t1_BER), varianza_estim
```

```
df3 = rbind(df,df2)

 df3$diseño<- c("SI", "BER", rep(NA, nrow(df3) - 2))
 df3 <- df3[, c("diseño", names(df3)[-1])]

 df3= df3[,-5]

df3 = df3 %>%  mutate( deff= var(t1_BER)/var(t1_SI))  %>%  mutate( total_NBI = 340)


df_long <- df3 %>%
  pivot_longer(cols = -diseño, names_to = "variable", values_to = "value")

df_f = df_long %>%  pivot_wider(names_from = diseño, values_from = value)



df_f %>% xtable(caption = "Comparacion de estimacion entre diseño SI y BER de la variable NBI  para n =
```

Table 1: Comparacion de estimacion entre diseño SI y BER de la variable NBI para n = 150

variable	SI	BER
total_estimado	344.24	344.11
varianza_estimada	9947.41	11773.33
varianza_teorica	10607.53	11353.73
deff	1.18	1.18
total_NBI	340.00	340.00

```
options(xtable.comment = FALSE)

n2=600




v_si2 <- N^2*(1-n2/N)*var(datos_hogares$NBI)/n2
v_ber2 <- ((1-n2/N)/(n2/N))*sum(datos_hogares$NBI**2)

df.1 = datos_hogares %>%  summarise( total_NBI = sum(NBI), total_estimado = mean(t2_SI), varianza_estima

df.2 = datos_hogares %>%  summarise( total_NBI = sum(NBI), total_estimado = mean(t2_BER), varianza_estim


df.3 = rbind(df.1,df.2)

 df.3$diseño <- c("SI", "BER", rep(NA, nrow(df.3) - 2))
 df.3 <- df.3[, c("diseño", names(df.3)[-1])]

 df.3= df.3[,-5]

df.3 = df.3 %>%  mutate( deff= var(t2_BER)/var(t2_SI))  %>%  mutate( total_NBI = 340)
```

```
df_long2 <- df.3 %>%
  pivot_longer(cols = -diseño, names_to = "variable", values_to = "value")

df_f2 = df_long2 %>%  pivot_wider(names_from = diseño, values_from = value)

df_f2 %>% xtable(caption = "Comparacion de estimacion entre diseño SI y BER  de la variable NBI para n =
```

Table 2: Comparacion de estimacion entre diseño SI y BER de la variable NBI para n = 600

variable	SI	BER
total_estimado	341.72	343.52
varianza_estimada	2469.33	2641.97
varianza_teorica	2413.64	2583.43
deff	1.07	1.07
total_NBI	340.00	340.00

```
options(xtable.comment = FALSE)

n3= 1000
v_si3 =  N^2*(1-n3/N)*var(datos_hogares$NBI)/n3
v_ber3 = ((1-n3/N)/(n3/N))*sum(datos_hogares$NBI**2)


df.1 = datos_hogares %>%  summarise( total_NBI = sum(NBI), total_estimado = mean(t3_SI), varianza_estima

df.2 = datos_hogares %>%  summarise( total_NBI = sum(NBI), total_estimado = mean(t3_BER), varianza_esti


df.3 = rbind(df.1,df.2)

 df.3$diseño <- c("SI", "BER", rep(NA, nrow(df.3) - 2))
 df.3 <- df.3[, c("diseño", names(df.3)[-1])]

 df.3= df.3[,-5]

df.3 = df.3 %>%  mutate( deff= var(t3_BER)/var(t3_SI))  %>%  mutate( total_NBI = 340)


df_long2 <- df.3 %>%
  pivot_longer(cols = -diseño, names_to = "variable", values_to = "value")

df_f2 = df_long2 %>%  pivot_wider(names_from = diseño, values_from = value)

df_f2 %>% xtable(caption = "Comparacion de estimacion entre diseño SI y BER de la variable NBI  para n =
```

```
deff1 = var(t1_BER)/var(t1_SI)
```

Table 3: Comparacion de estimacion entre diseño SI y BER de la variable NBI para n = 1000

variable	SI	BER
total_estimado	340.90	342.03
varianza_estimada	1375.63	1330.42
varianza_teorica	1321.12	1414.06
deff	0.97	0.97
total_NBI	340.00	340.00