Trabajo 1 - Muestreo y Planificación de Encuestas I

Matias Bajac - Lucas Pescetto - Andres Vidal

2023-04-10

Introducción

Partimos de la base de datos del censo de hogares de 2011 en Rio Branco para estudiar los diseños **Simple sin reposición (SI)**, **Simple con reposición (SIR)** y **Bernoulli (BER)**. Nos interesa estimar el total poblacional de dos variables calculadas a partir de la base de datos en cuestión:

- `nbi`: vale 0 si el hogar tiene 3 o menos necesidades básicas insatisfechas (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.

Esto es, estaremos estimando la cantidad de hogares con 4 o más NBI y la cantidad de hogares sin computadoras XO. Además del cálculo de las variables `nbi` y `xo` a partir de la base de datos, esto requirió remover observaciones duplicadas, puesto que la base está a nivel de personas e interesa calcular los totales a nivel de hogares.

Distribución empírica del estimador

En esta parte presentamos el marco de trabajo del análisis. Definimos funciones auxiliares para automatizar el análisis y estandarizar y simplificar la presentación de los resultados. El objetivo es abstraer los procedimientos a ser realizaos de los tamaños de muestra y de los diseños de muestreo. Para esto, utilizamos el paquete `survey`.

Funciones Auxiliares

Definimos funciones auxiliares para facilitar el resto del análisis:

- `estimate_total` recibe un nombre de variable y un diseño de muestreo para estimar un total poblacional. Envuelve la función `svytotal` del paquete `survey` para facilitar su uso.
- `estimate_totals` recibe un nombre de variable y una lista de diseños de muestreo y estima el total poblacional para cada diseño. Se utiliza especialmente para automatizar la aplicación sobre varios tamaños de muestra.
- `show_results` recibe un nombre de variable y una lista de resultados para mostrarlos de forma estándar.
- `confint_norm` recibe un estimador (resultado de `svytotal`) y calcula el intervalo de confianza al 95% asumiendo distribución normal.

```
estimate_total <- function(var, design) {
  svytotal(as.formula(paste0("~", var)), design, deff = TRUE)
}
```

```r
estimate_totals <- function(var, design_list) {
  lapply(design_list, function(design) estimate_total(var, design))
}
show_results <- function(var, named_result_list) {
  t(as.data.frame(t(named_result_list), row.names = var))
}
confint_norm <- function(t_estimate) {
  t <- coef(t_estimate)
  se <- SE(t_estimate)
  ci <- cbind(
    t - qnorm(0.975) * se,
    t + qnorm(0.975) * se
  )
  colnames(ci) <- c("2.5%", "97.5%")
  ci
}
```

Muestreo del estimador

Obtenemos 1000 muestras y calculamos el estimador del total poblacional para cada una de ellas. La función `sample_t_estimate` implementa este procedimiento genéricamente, recibiendo como parámetros:

- `var` La variable que se desea estimar
- `sample_size` el tamaño de las muestras que se deben tomar
- `get_sample` una función que dado valor `n` devuelve una muestra de tamaño `n`
- `get_design` una función que dada una muestra devuelve un objeto de diseño de muestreo generado con `svydesign`

```r
sample_t_estimate <- function(var, sample_size, get_sample, get_design) {
  n_simulations <- 1000
  replicate(n_simulations, {
    sample <- get_sample(sample_size)
    design <- get_design(sample, sample_size)
    coef(estimate_total(var, design))
  })
}
```

Utiliza `estimate_total` para estimar el total poblacional para la variable indicada. Además, definimos la función `empirical_distribution` (cuyo código no está expuesto en el informe) que toma los mismos parámetros, utiliza `sample_t_estimate`, grafica la distribución empírica y retorna resúmenes de la distribución empírica del estimador, como la media, la varianza y los intervalos de confianza empíricos al 95% de confianza, para cada tamaño de muestra.

Diseño de muestreo SIR

En esta parte analizamos el diseño **Simple con Reposición (SIR)**. En primera instancia, definimos los artefactos necesarios para utilizar nuestro marco de trabajo. Luego, estimamos la distribución empírica del estimador del total poblacional para `nbi` y para `xo` y comparamos sus características con el parámetro poblacional y con el estimador teórico de la varianza del estimador poblacional.

Definición del diseño

Definimos la función `get_sir_design` que recibe una muestra y genera un objeto de diseño de muestreo con `svydesign` del paquete `survey` de la siguiente manera:

```
get_sir_design <- function(sample, expected_sample_size = nrow(sample)) {
  svydesign(
    ids = ~1,
    data = sample,
    probs = nrow(sample) / N
  )
}
```

En este caso, la estrategia de muestreo es diseño SIR con estimador t_{pwr}. Entonces, la probabilidad de inclusión de cada unidad es la misma para todas las unidades y está definida como n N, dónde n es el tamaño de la muestra y N es el tamaño de la población.

Algoritmo de selección

Definimos la función `get_sir_sample` que recibe el tamaño de la muestra y devuelve una muestra de tamaño `n` utilizando el método SIR. Específicamente, esto se implementa utilizando la función `srswr` del paquete `sampling`.

```
get_sir_sample <- function(sample_size) {
  index <- srswr(sample_size, N)
  getdata(data, index)
}
```

Obtener muestras finales

Utilizamos la función `get_sir_sample` para obtener muestras de tamaño 150, 600 y 1000 y almacenarlas en la lista `sir_samples`.

```
sample_sizes <- c(150, 600, 1000)
sir_samples <- lapply(sample_sizes, get_sir_sample)
names(sir_samples) <- sample_sizes
```

Obtener objetos de diseño finales

Utilizamos la función `get_sir_design` para obtener los objetos de diseño de muestreo finales para cada muestra. Almacenamos estos objetos en la lista `sir_designs`.

```
sir_designs <- lapply(sir_samples, get_sir_design)
```

Análisis de NBI

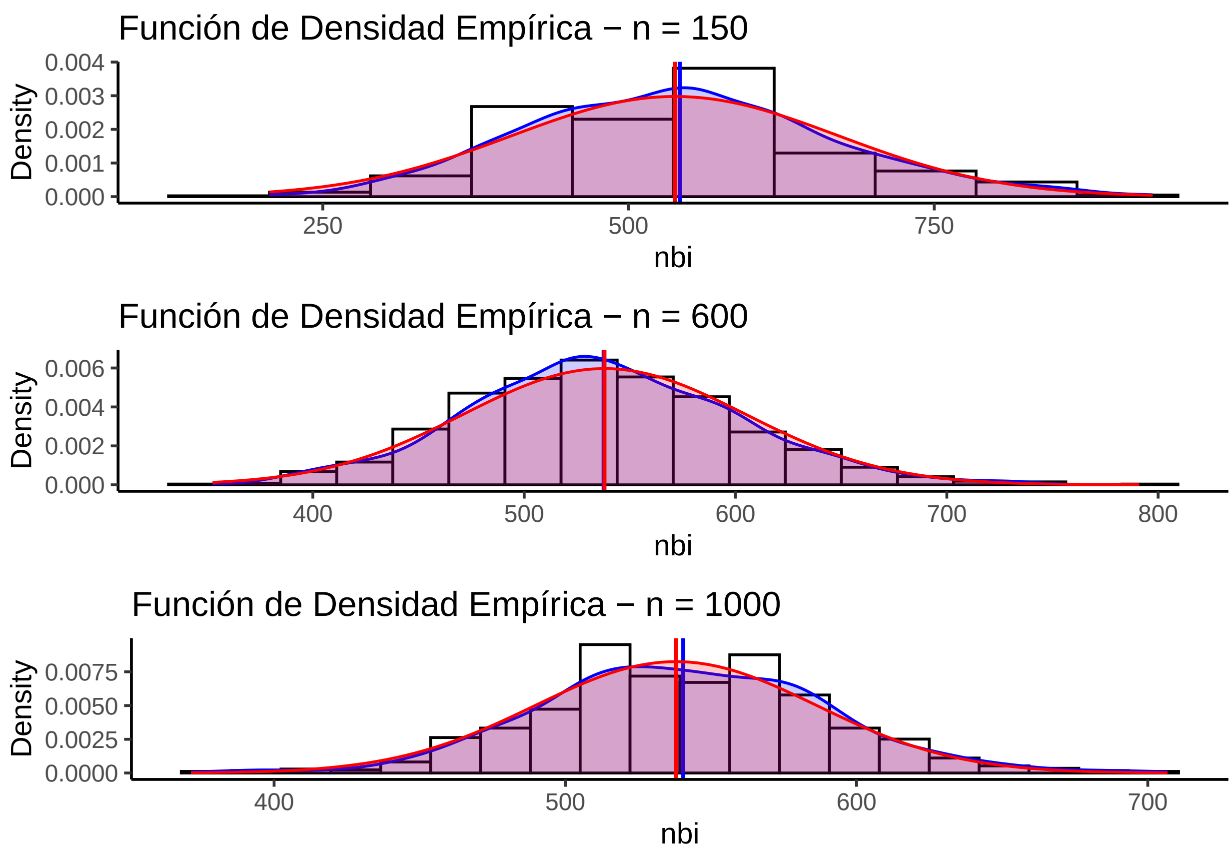
En esta parte analizamos los resultados obtenidos para la variable `nbi`. Interesa estimar el total de hogares con 4 o más NBI.

Distribución empírica del estimador para el total de NBI

Utilizamos la función `empirical_distribution` para obtener y visualizar la distribución empírica del estimador. Esta función utiliza internamente a la función `sample_t_estimate` presentada anteriormente. Para esto es necesario pasar las funciones `get_sir_sample` y `get_sir_design` definidas recién como parámetros.

Como resultado, obtenemos resúmenes de la distibución empírica y los almacenamos en la variable `t_nbi_dist`. Además visualizamos la distribución empírica para cada tamaño de muestra mediante histogramas y la función de densidad empírica. Mostramos el valor real del total poblacional de `nbi` como una línea vertical roja y el promedio muestral como una línea vertical azul. La función de densidad empírica se muestra como un área azul y la función de densidad teórica (Normal con media en el parámetro poblacional y la varianza teórica del estimador) y como un área roja.

```
sir_t_nbi_dist <- empirical_distribution(
  "nbi",
  sample_sizes,
  get_sir_sample,
  get_sir_design
)
```



A seguir podemos observar las estimaciones puntuales para el total poblacional, los promedios empíricos y el sesgo de cada uno (la diferencia con el parámetro real) para cada tamaño de muestra. Observamos bajo sesgo de la distribución empírica, lo cual puede visualizarse en los gráficos arriba. El estimado, sin embargo, presenta un sesgo de aproximadamente 20 unidades, con una leve reducción a medida que aumenta el tamaño de la muestra.

```
##      t_estimate      bias t_dist_mean      bias difference
```

```
## 150       515.9     22.1    541.9014  3.90136   26.00136
## 600    558.8917 20.89167    537.5936 0.406405   21.29807
## 1000    557.172   19.172     540.493 2.492953   16.67905
```

Abajo reportamos las estimaciones del desvío estándar teórico del estimador y su varianza empírica, respectivamente, para cada tamaño de muestra. Observamos reducción en ambas estimaciones del desvío estándar al aumentar el tamaño de muestra. Por otro lado, al observar la diferencia entre ambos estimadores, se observa una reducción del 71% al aumentar el tamaño de la muestra de 150 a 600 y un aumento del 37% al aumentar el tamaño de muestra de 600 a 1000. Esto quiere decir que la varianza resultó mejor estimada por la fórmula teórica para el tamaño de muestra de 600. Sin embargo, la diferencia para ambos tamaños de muestra (600 y 1000) es pequeña.

El bajo error al estimar la varianza del estimador puede apreciarse también en los gráficos, al observar la semejanza entre la curva roja (densidad teórica asintótica) y la curva azul (densidad empírica).

```
##      Var(t_nbi) Var(t_dist) difference
## 150    126.7925    125.3152   1.477358
## 600    65.51405    63.94801   1.566038
## 1000   50.66139     48.9199   1.741492
```

Efecto diseño del estimador para el total de NBI

A seguir presentamos los valores del efecto diseño para el estimador del total poblacional en cada tamaño de muestra. El efecto diseño resultó mayor que 1 para todos los casos, por la estrategia de muestreo (SIR con estimador t_{pwr}) causa pérdida de eficiencia en varianza respecto al diseño SI con estimador HT.

Observamos también que a mayores tamaños de muestra aumenta el efecto diseño, por lo que concluimos que esta estrategia pierde eficiencia al aumentar tamaño de muestra.

```
##      deff
## 150  1.029946
## 600  1.131608
## 1000 1.240442
```

Intervalos de confianza para el total de NBI

Abajo reportamos los intervalos de confianza **empíricos** al 95% de confianza para el total de la variable `nbi` en cada tamaño de muestra. En la tercera columna incluímos el rango de cada intervalo, como medida de su precisión.

Observamos que, además de que todos los intervalos incluyen al total real de la variable `nbi` (538 hogares), la precisión de los intervalos aumentan con el tamaño de muestra. Al aumentar el tamaño de muestra de 150 a 600 observamos una reducción del 48% en el rango y al aumentar el tamaño de muestra de 600 a 1000 observamos una reducción del 23%.

```
##          2.5%   97.5%   range
## X150  309.540 825.440 515.900
## X600  412.720 670.670 257.950
## X1000 448.833 639.716 190.883
```

Abajo reportamos los intervalos de confianza al 95% asumiendo que el estimador se distribuye con normalidad. En este caso, tambien observamos una reducción del 48% en el rango al aumentar el tamaño de muestra de 150 a 600 y una reducción del 23% al aumentar de 600 a 1000 elementos. Por lo tanto, concluimos que en el caso normal el intervalo de confianza es más preciso a medida que aumenta el tamaño de muestra.

```
##          2.5%    97.5%    range
## 150  267.3912 764.4088 497.0176
## 600  430.4865 687.2968 256.8103
## 1000 457.8775 656.4665 198.5890
```

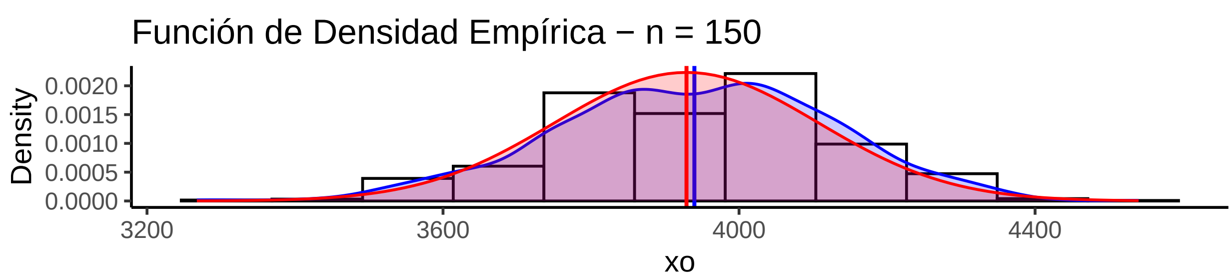
Al comparar los intervalos de confianza empíricos y los intervalos de confianza asumiendo normalidad, observamos su similaridad en términos de rango y posición, lo que implica que también coincidan respecto a su variación con el tamaño de muestra. Destacamos, sin embargo, que los intervalos al considerar tamaño de muestra de 600 o 1000 son más cercanos que cuando el tamaño de muestras es 150. Esto es razonable, puesto que la normalidad del estimador del total poblacional es asintótica, por lo que será más evidente a medida que crece el tamaño de muestra.

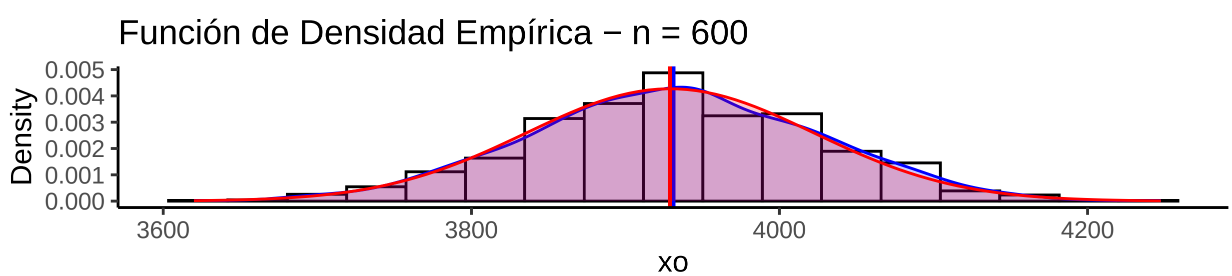
Análisis de XO

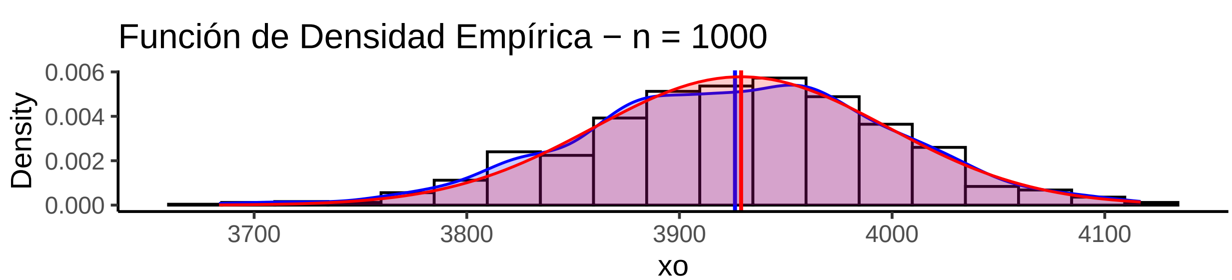
En esta parte se presenta el mismo procedimiento realizado para la variable `nbi`, pero considerando a la variable `xo` como característica de interés. Interesa estimar el total de hogares con al menos un dispositivo XO.

Distribución empírica del estimador para el total de XO

A seguir presentamos los gráficos de la distribución empírica del estimador del total de `xo`. Destacamos la similiridad entre la distribución empírica (área azul) y la distribución teórica (área roja).







Respecto a las estimaciones puntuales para el total de `xo`, ov

```
##      t_estimate     bias t_dist_mean     bias difference
## 150    3886.447 42.55333    3939.653 10.65315   53.20649
## 600    3912.242 16.75833    3931.201 2.200992   18.95933
## 1000   3925.999    3.001    3926.123 2.877184   0.123816
```

```
##      Var(t_xo) Var(t_dist) difference
## 150   182.1886    185.9691   3.780531
## 600   90.23823      92.873   2.634774
## 1000  69.61039    71.37506   1.764675
```

```
##      deff
## 150  1.029946
## 600  1.131608
## 1000 1.240442
```

```
##          2.5%    97.5%    range
## 150  3529.364 4243.530 714.1662
## 600  3735.378 4089.105 353.7274
## 1000 3789.565 4062.433 272.8677
```

Diseño de muestreo SI

En este caso estudiaremos el diseño Simple con la misma estrategia que los anteriores Diseños. A difrencia del Bernoulli, el diseño simple tiene la particularidad de ser de tamaño fijo. La probabiliad de inclusion de primer orden la definimos como n/N, siendo n el tamaño esperado de la muestra. Definiendo la probabilidad de esta forma logramos que el tamaño de las muestras sea lo más cercano posible al buscado. Nos basaremos en el estmiador Horvitz thompson para estimar el total poblacional de la variable de interes.

Definición del diseño

Definimos la función `get_si_design` que recibe una muestra y genera un objeto de diseño de muestreo con `svydesign` del paquete `survey` de la siguiente manera:

```
get_si_design <- function(sample, expected_sample_size = nrow(sample)) {
  svydesign(
    ids = ~1,
    data = sample,
    probs = rep(nrow(sample)/N,nrow(sample)),
     fpc =rep(N,nrow(sample))
  )
}
```

En este caso, la estrategia de muestreo es diseño SI con estimador t_π. La probabilidad de inclusión de primer orden la definimos como n N, siendo n el tamaño esperado de la muestra.

#Algoritmo de seleccion

Definimos la función `get_si_sample` que recibe el tamaño de la muestra y devuelve una muestra de tamaño `n` utilizando el método SI. Específicamente, esto se implementa utilizando la función `srswor` del paquete `sampling`.

```r
get_si_sample <- function(sample_size) {
  index <- srswor(sample_size, N)
  getdata(data, index)
}
```

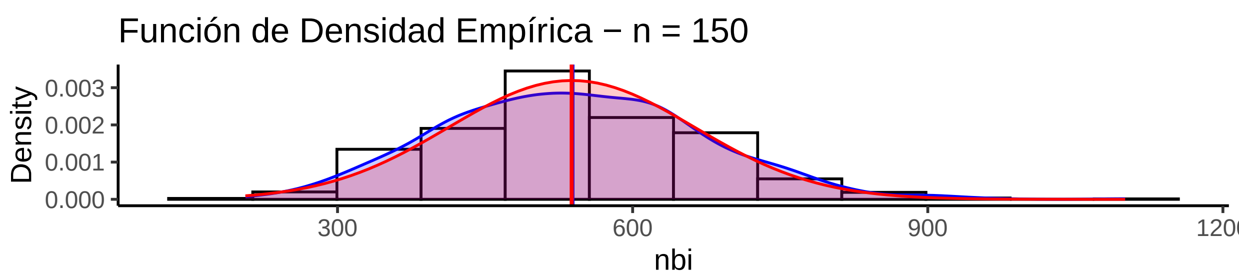
```r
sample_sizes <- c(150, 600, 1000)
si_samples <- lapply(sample_sizes, get_si_sample)
names(si_samples) <- sample_sizes
```

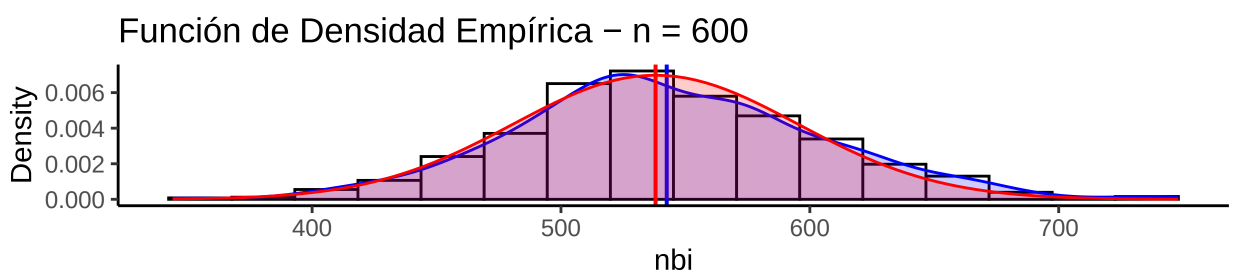
Obtener objetos de diseño finales

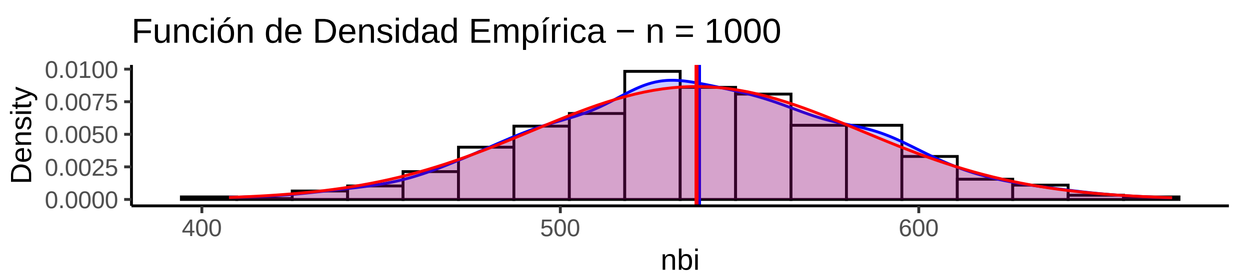
Utilizamos la función `get_si_design` para obtener los objetos de diseño de muestreo finales para cada muestra. Almacenamos estos objetos en la lista `si_designs`.

```r
si_designs <- lapply(si_samples,get_si_design)
```

```r
si_t_nbi_dist <- empirical_distribution(
  "nbi",
  sample_sizes,
  get_si_sample,
  get_si_design
)
```







```
##      t_estimate   bias t_dist_mean      bias difference
## 150       515.9   22.1    538.8404 0.8403533   22.94035
## 600     593.285 55.285    542.4946  4.494645   50.79036
## 1000     464.31  73.69    538.6925  0.692462   74.38246
```

A seguir podemos observar las estimaciones puntuales para el total poblacional, los promedios empíricos y el sesgo de cada uno (la diferencia con el parámetro real) para cada tamaño de muestra. Observamos un bajo sesgo de la distribución empírica, lo cual puede visualizarse en los gráficos de arriba. El parametro estimado cumple con el principio de insesgades respecto a la media. A medida que aumenta el tamaño de muestra, se aproxima mas al parametro real

```
##      Var(t_nbi) Var(t_dist) difference
## 150    124.9357    129.8519   4.916215
## 600    63.21566    61.11042   2.105243
## 1000   41.94077    44.17701   2.236236
```

Intervalos de confianza para el 95%

Abajo reportamos los intervalos de confianza **empíricos** al 95% de confianza para el total de la variable `nbi` en cada tamaño de muestra.

```
##           2.5%    97.5%    range
## X150  309.5400 791.0467 481.5067
## X600  429.9167 670.6700 240.7533
## X1000 453.8630 624.3680 170.5049
```

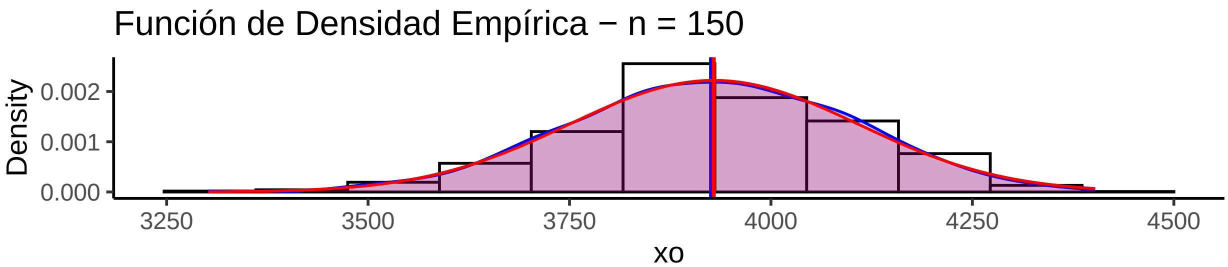
Se puede apreciar que a medida que aumenta la el tamaño de muestra, disminuyo la amplitud del intervalo. En todos los casos, el intervalo contiene al parametro estimado.

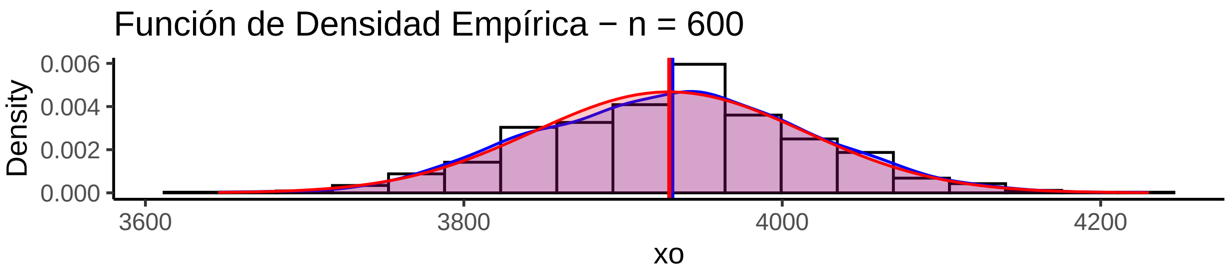
Abajo reportamos los intervalos de confianza al 95% asumiendo que el estimador se distribuye con normalidad. Podemos ver que hay diferencia entre el empirico y asumiento normalidad, siendo que a medida que aumenta el tamaño de muestra, se aproxima mas a la empirica

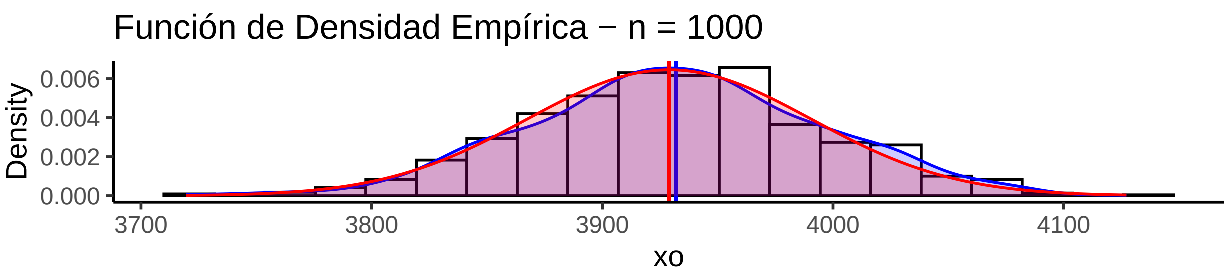
```
##          2.5%    97.5%    range
## 150  271.0306 760.7694 489.7388
## 600  469.3846 717.1854 247.8008
## 1000 382.1076 546.5124 164.4048
```

Distribución empírica del estimador para el total de XO

A seguir presentamos los gráficos de la distribución empírica del estimador del total de `xo`. Destacamos la similiridad entre la distribución empírica (área azul) y la distribución teórica (área roja).







Respecto a las estimaciones puntuales para el total de `xo`, ov

```
##      t_estimate     bias t_dist_mean     bias difference
## 150    3852.053 76.94667    3925.689  3.31054   73.63613
## 600    3877.848 51.15167    3930.943 1.943042   53.09471
## 1000   3879.568   49.432    3931.947 2.947327   52.37933
```

```
##      Var(t_xo) Var(t_dist) difference
## 150   182.1886    185.9691   3.780531
## 600   90.23823      92.873   2.634774
## 1000  69.61039    71.37506   1.764675
```

Intervalos de confianza para el 95%

Abajo reportamos los intervalos de confianza **empíricos** al 95% de confianza para el total de la variable `xo` en cada tamaño de muestra.

```
##          2.5%    97.5%    range
## 150  3497.058 4207.049 677.6852
## 600  3710.052 4045.645 310.2667
## 1000 3755.524 4003.612 214.0473
```

```
##           2.5%    97.5%    range
## X150  3576.907 4264.773 687.8667
```

```
## X600  3774.668 4092.807 318.1383
## X1000 3812.501 4054.974 242.4730
```

Podemos obserar en este caso que los intervalos de confianza empirico y asumiendo normalidad son muy similares para cualquier tamaño de muestra