

Muestreo Aleatorio Simple

Rodrigo Zepeda (rodrigo.zepeda@itam.mx)

5 de julio 2020

Inicio

Siempre que inicies un nuevo trabajo en R ¡no olvides borrar el historial!

```
rm(list = ls()) #Clear all
```

Librerías

Para este análisis vamos a tener que llamar a las siguientes librerías previamente instaladas (por única vez) con `install.packages()`:

```
library(tidyverse)
library(dplyr)
library(imager)
library(rlist)
library(gridExtra)
```

Notación

Supongamos una matriz de datos de tamaño $N \times L$ dada por:

$$U = \left(z_1 \mid z_2 \mid \dots \mid z_\ell \right)$$

donde las z_i representan las columnas de la matriz. La U será conocida como la **matriz universo** (el *universo* ó *la población*) si contiene *toda* la información de la población. Intuitivamente, la matriz U son todos los datos de un censo: esta es una matriz *ideal* donde están todos los datos.

A cualquier permutación en las filas de una submatriz S (tamaño $n \times \ell$ con $0 < n \leq N$ y $0 < \ell \leq L$) de U se le conoce como una *muestra* de U . Si S es una variable aleatoria (por ejemplo, porque se construyó a partir de un mecanismo de aleatoriedad) decimos que S es una **muestra aleatoria** (denotamos \mathcal{S} a la variable aleatoria y S a un valor específico de la misma).¹

En particular en esta sección (y hasta establecer lo contrario) consideraremos que el universo U es de tamaño $N \times 1$ y la submatriz S (resp. \mathcal{S}) es de tamaño $n \times 1$. En notación $U = (x_1, x_2, \dots, x_N)^T$ y *técnicamente* $S = (x_{i_1}, x_{i_2}, \dots, x_{i_n})^T$ para un conjunto de índices i_1, i_2, \dots, i_n . Sin embargo para simplificar la notación consideraremos que en S están los primeros n de los x_i ; es decir:

$$S = (x_1, x_2, \dots, x_n)^T$$

Cuando estemos hablando de la muestra *como variable aleatoria* \mathcal{S} y no como *valores observados (fijos)* S , denotaremos:

$$\mathcal{S} = (X_1, X_2, \dots, X_n)^T$$

¹En la literatura muchas referencias establecen una muestra aleatoria como un conjunto de valores. Yo utilizo vectores para poder hablar de repeticiones (por ejemplo si extraes el mismo valor varias veces en la muestra).

donde \mathcal{S} representa la muestra posible y cada X_i es una variable aleatoria con el valor posible de la i -ésima entrada.

Un **esquema muestral** es una función \mathbb{P} de probabilidad definida en el conjunto de submatrices de U . Ésta es el punto medular de todas las estrategias de muestreo: distintos esquemas muestrales generan diferentes distribuciones y pueden llevar a distintas inferencias sobre un fenómeno.

Ejemplo

Considera la matriz universo con tres letras:

$$U = \begin{pmatrix} A \\ B \\ C \end{pmatrix}$$

Ésta es la matriz universo. Las submatrices² que pueden crearse a partir de dicho universo son:

1. De dimensión $n = 1$: $S^1 = (A)^T$, $S^2 = (B)^T$, $S^3 = (C)^T$.
2. De dimensión $n = 2$: $S^4 = (A, B)^T$, $S^5 = (A, C)^T$, $S^6 = (B, C)^T$, $S^7 = (B, A)^T$, $S^8 = (C, A)^T$, $S^9 = (C, B)^T$.
3. De dimensión $n = 3$: $S^{10} = (A, B, C)^T$, $S^{11} = (B, A, C)^T$, $S^{12} = (A, C, B)^T$, $S^{13} = (C, B, A)^T$, $S^{14} = (B, C, A)^T$, $S^{15} = (C, A, B)^T$.

Un esquema muestral sería la función de probabilidad:

$$\mathbb{P}(\mathcal{S} = S^k) = \begin{cases} 0.1 & \text{si } k = 1, \\ 0.2 & \text{si } k = 3, \\ 0.5 & \text{si } k = 11, \\ 0.2 & \text{si } k = 15, \\ 0 & \text{en otro caso.} \end{cases}$$

Otro esquema muestral posible sería:

$$\mathbb{P}(\mathcal{S} = S^k) = \begin{cases} \frac{1}{3} & \text{si } k = 1, \\ \frac{1}{3} & \text{si } k = 2, \\ \frac{1}{3} & \text{si } k = 3, \\ 0 & \text{en otro caso.} \end{cases}$$

Este último esquema, intuitivamente, corresponde a la selección aleatoria de un elemento de U con una probabilidad uniforme de que cada elemento salga.

A fin de simplificar el problema (y hasta que se diga lo contrario) agregaremos la hipótesis de **intercambiabilidad**; es decir, consideraremos es irrelevante el orden de las filas de las submatrices de datos. Por ejemplo, bajo intercambiabilidad, $S^4 = (A, B)^T$ es la misma matriz que $S^7 = (B, A)^T$.

Un ejemplo de muestra donde el orden sí importa (*i.e.* no son intercambiables) es cuando se realizan exámenes orales según una selección aleatoria de la lista. La tercera persona en presentar el examen estará informada por el *¿qué te preguntó el profe?*, *¿estuvo difícil?* que las primeras dos le cuenten.

Bajo intercambiabilidad, los esquemas muestrales estarán definidos únicamente sobre los siguientes conjuntos:

1. De dimensión $n = 1$: $S^1 = (A)^T$, $S^2 = (B)^T$, $S^3 = (C)^T$.
2. De dimensión $n = 2$: $S^4 = (A, B)^T$, $S^5 = (A, C)^T$, $S^6 = (B, C)^T$.

²Enumero las submatrices para luego poder hablar de ellas

3. De dimensión $n = 3$: $S^7 = (A, B, C)^T$.

En este caso un esquema muestral sería:

$$\mathbb{P}(\mathcal{S} = S^k) = \begin{cases} \frac{1}{16} & \text{si } k = 1, \\ \frac{3}{16} & \text{si } k = 2, \\ 0 & \text{si } k = 3, \\ \frac{7}{16} & \text{si } k = 4, \\ \frac{1}{16} & \text{si } k = 5, \\ \frac{4}{16} & \text{si } k = 6, \\ 0 & \text{en otro caso.} \end{cases}$$

Dado un elemento x_i del universo, podemos preguntarnos por la probabilidad de que dicho x_i esté en la muestra. Siguiendo el ejemplo anterior:

$$\mathbb{P}(A \in \mathcal{S}) = \mathbb{P}(\mathcal{S} = S^1) + \mathbb{P}(\mathcal{S} = S^4) + \mathbb{P}(\mathcal{S} = S^5) + \mathbb{P}(\mathcal{S} = S^7) = \frac{13}{16}.$$

Como notación, para una población $U = (x_1, x_2, \dots, x_N)^T$ y una muestra aleatoria \mathcal{S} denotamos la probabilidad de que x_k esté en la muestra como:

$$\pi_k = \mathbb{P}(x_k \in \mathcal{S})$$

Estas probabilidades (para $k = 1, 2, \dots, N$) se conocen como **probabilidades de inclusión de primer orden**. La probabilidad conjunta de que x_k y x_l (ambos) estén en la muestra (**probabilidad de inclusión de segundo orden**) está dada por:

$$\pi_{k,l} = \mathbb{P}(x_k \in \mathcal{S}, x_l \in \mathcal{S})$$

Notamos que por definición $\pi_{kk} = \pi_k$. Análogamente se pueden crear probabilidades de inclusión de cualquier orden deseado.

Finalmente, una población $U = (x_1, x_2, \dots, x_N)^T$ y una muestra aleatoria \mathcal{S} definimos la variable indicadora de que x_k esté en la muestra como:

$$\mathbb{I}_{\mathcal{S}}(x_k) = \begin{cases} 1 & \text{si } x_k \in \mathcal{S} \\ 0 & \text{si } x_k \notin \mathcal{S} \end{cases}$$

Notamos que para una muestra aleatoria \mathcal{S} las indicadoras tienen una distribución conocida:

$$\mathbb{I}_{\mathcal{S}}(x_k) \sim \text{Bernoulli}(\pi_k)$$

pues

$$\mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) = 1) = \mathbb{P}(x_k \in \mathcal{S}) = \pi_k$$

Como las $\mathbb{I}_{\mathcal{S}}(x_k)$ son Bernoulli podemos [calcular su varianza](#):

$$\text{Var}(\mathbb{I}_{\mathcal{S}}(x_k)) = \pi_k(1 - \pi_k)$$

Finalmente, recordamos que la covarianza entre dos variables aleatorias X, Y se define como:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Por lo que calculamos la covarianza entre dos indicadoras (de x_k y x_l):

$$\begin{aligned} \text{Cov}(\mathbb{I}_{\mathcal{S}}(x_k), \mathbb{I}_{\mathcal{S}}(x_l)) &= \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_k) \cdot \mathbb{I}_{\mathcal{S}}(x_l)] - \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_k)] \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_l)] \\ &= 1 \cdot \mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) \cdot \mathbb{I}_{\mathcal{S}}(x_l) = 1) + 0 \cdot \mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) \cdot \mathbb{I}_{\mathcal{S}}(x_l) = 0) - \pi_k \pi_l \\ &= \mathbb{P}(\mathbb{I}_{\mathcal{S}}(x_k) = 1, \mathbb{I}_{\mathcal{S}}(x_l) = 1) - \pi_k \pi_l \\ &= \pi_{k,l} - \pi_k \pi_l \end{aligned} \tag{1}$$

La cantidad $\pi_{k,l} - \pi_k \pi_l$ usualmente se denota $\Delta_{k,l}$:

$$\Delta_{k,l} = \pi_{k,l} - \pi_k \pi_l$$

A continuación hablaremos de algunos esquemas de muestreo comunmente utilizados y, finalmente, llegaremos a una generalización de los mismos.

Ejercicio

Demuestra las siguientes propiedades de los π_k para un diseño muestral \mathbb{P} con tamaño fijo de la muestra $n \in \mathbb{N}$:

1. $\sum_{k=1}^N \pi_k = n$
2. $\sum_{k=1}^N \sum_{\substack{l=1 \\ k \neq l}}^N \pi_{k,l} = n(n-1)$
3. $\sum_{\substack{l=1 \\ l \neq k}}^N \pi_{k,l} = (n-1)\pi_k$

Muestreo Aleatorio Simple sin Reemplazo (MAS/sR)

Vamos a considerar una de las formas más sencillas de muestreo: el aleatorio simple *sin reemplazo*. Para ello seleccionamos de $U = (x_1, x_2, \dots, x_N)^T$ a $n \in \mathbb{N}$ (fijo) observaciones asignándole la probabilidad de ser seleccionada a cada una de $\frac{1}{N}$. Una vez se selecciona la primera, se selecciona una de las que restan de U con probabilidad $\frac{1}{N-1}$. El proceso se repite hasta extraer n elementos.

Comencemos por un ejemplo, supongamos tenemos una población de cinco personas:

$$U = (\text{Ana}, \text{Beto}, \text{Carlos}, \text{Diana}, \text{Enriqueta})^T$$

Si queremos tomar una muestra de 3 personas sin reemplazo, las muestras posibles son:

1. $(\text{Ana}, \text{Beto}, \text{Carlos})^T$
2. $(\text{Ana}, \text{Carlos}, \text{Diana})^T$
3. $(\text{Ana}, \text{Beto}, \text{Diana})^T$
4. $(\text{Ana}, \text{Beto}, \text{Enriqueta})^T$
5. $(\text{Ana}, \text{Carlos}, \text{Enriqueta})^T$
6. $(\text{Ana}, \text{Diana}, \text{Enriqueta})^T$
7. $(\text{Beto}, \text{Carlos}, \text{Diana})^T$
8. $(\text{Beto}, \text{Diana}, \text{Enriqueta})^T$
9. $(\text{Beto}, \text{Carlos}, \text{Enriqueta})^T$

10. $\left(\text{Carlos, Diana, Enriqueta}\right)^T$

Obtener una muestra aleatoria se puede hacer en R con un vector mediante `sample`:

```
#Vector de nombres
nombres <- c("Ana", "Beto", "Carlos", "Diana", "Enriqueta")

#Muestra
sample(nombres, 3, replace = FALSE)
```

```
## [1] "Enriqueta" "Ana"          "Beto"
```

Formalmente, un esquema de muestreo es **aleatorio simple sin reemplazo** si dada una constante $n \in \mathbb{N}$ (con $0 < n \leq N$) se tiene:

$$\mathbb{P}(\mathcal{S} = S) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{si } \#S = n \\ 0 & \text{en otro caso.} \end{cases}$$

En el caso de muestreo aleatorio simple sin reemplazo podemos calcular las probabilidades de inclusión como siguen:

$$\pi_k = \mathbb{P}(x_k \in \mathcal{S}) = \sum_{i=1}^{M_1} \frac{1}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} = f$$

donde la tercera igualdad se sigue de que hay $M_1 = \binom{N-1}{n-1}$ muestras que contienen al x_k . (La lógica es, fijo el x_k y entonces me quedan $N - 1$ valores de x a acomodar en $n - 1$ espacios). Por otro lado:

$$\pi_{k,j} = \mathbb{P}(x_k \in \mathcal{S}, x_j \in \mathcal{S}) = \sum_{i=1}^{M_2} \frac{1}{\binom{N}{n}} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

pues hay $M_2 = \binom{N-2}{n-2}$ muestras conteniendo a x_k y x_j a la vez.

Para estimar el total poblacional dado por:

$$t = \sum_{i=1}^N x_i$$

bajo MAS/sR podemos tomar:

$$\hat{t} = N \cdot \bar{x}_{\mathcal{S}} = N \frac{1}{n} \sum_{i=1}^n x_k = \sum_{i=1}^n \frac{x_k}{n/N} = \sum_{k=1}^N \frac{x_k}{\pi_k} \cdot \mathbb{I}_{\mathcal{S}}(x_k)$$

Notamos entonces que el estimador \hat{t} es una variable aleatoria pues depende de las indicadoras de la muestra. En particular:

$$\mathbb{E}[\hat{t}] = \mathbb{E}\left[\sum_{k=1}^N \frac{x_k}{\pi_k} \cdot \mathbb{I}_{\mathcal{S}}(x_k)\right] = \sum_{k=1}^N \frac{x_k}{\pi_k} \underbrace{\mathbb{E}\left[\mathbb{I}_{\mathcal{S}}(x_k)\right]}_{\pi_k} = t$$

de donde se sigue que en promedio el estimador \hat{t} vale el total.

Definición

Un estimador $\hat{\theta}$ es un estimador insesgado de θ si:

$$\mathbb{E}[\hat{\theta} - \theta] = 0$$

En nuestro caso \hat{t} es *insesgado*.

De manera numérica, podemos simular la estimación del total en 1000 simulaciones como sigue:

```

nsim <- 1000
N     <- 1000
n     <- 100
base.completa <- data.frame(x = rnorm(N))
total        <- sum(base.completa$x)
total.muestra <- rep(NA, nsim)
for (i in 1:nsim){
  muestra      <- sample(base.completa$x, n)
  total.muestra[i] <- N*mean(muestra)
}
mean(total.muestra)

```

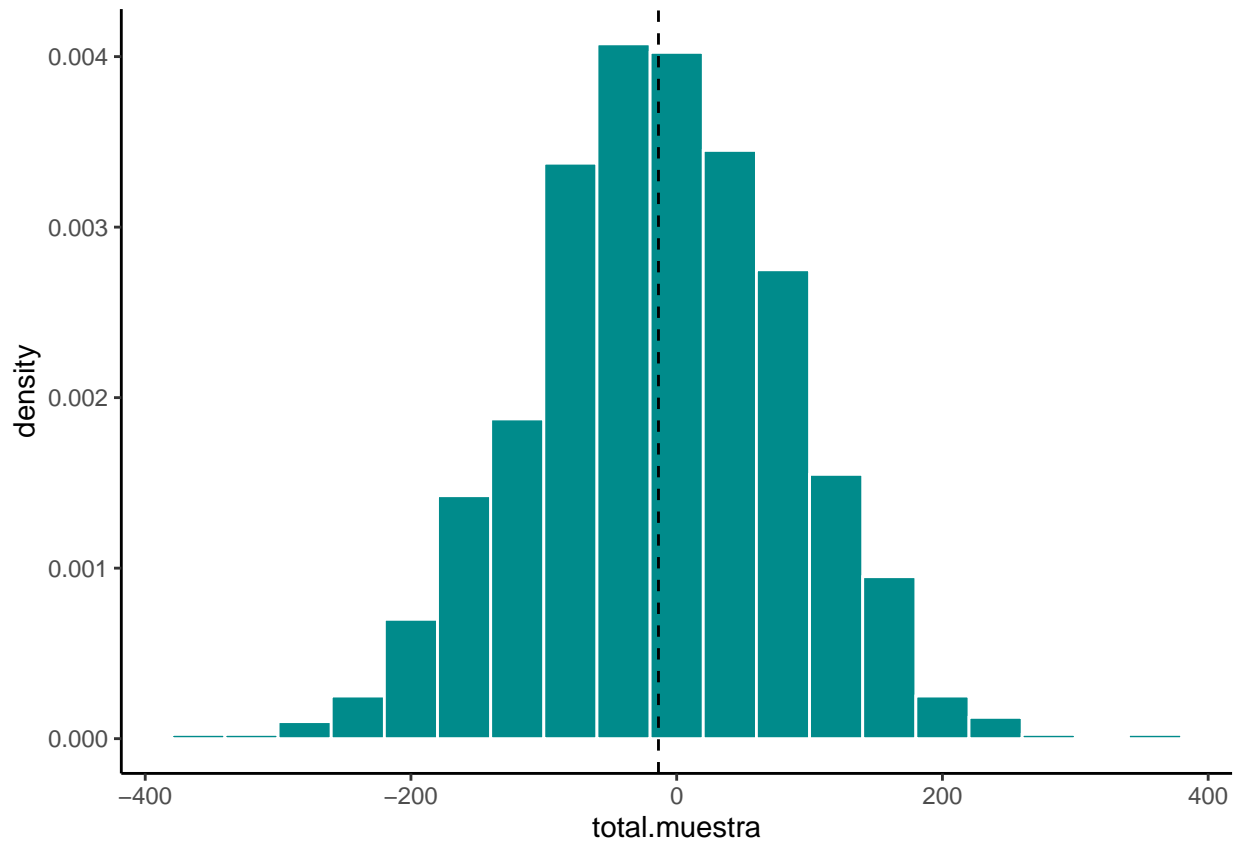
```
## [1] -13.42996
```

Podemos ver las simulaciones como sigue:

```

ggplot() +
  geom_histogram(aes(x = total.muestra, y = ..density..), fill = "#008B8B",
                 color = "white", binwidth = 40) +
  geom_vline(aes(xintercept = total), linetype = "dashed") +
  theme_classic()

```



Como podrás notar la \hat{t} es una variable aleatoria y por tanto tiene varianza. De hecho:

$$\text{Var}(\hat{t}) = \sum_{k=1}^N \sum_{l=1}^k \Delta_{k,l} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l}$$

Para demostrarlo seguimos las igualdades:

$$\begin{aligned}
\text{Var}(\hat{t}) &= \text{Var}\left(\sum_{k=1}^N \frac{x_k}{\pi_k} \cdot \mathbb{I}_{\mathcal{S}}(x_k)\right) \\
&= \sum_{k=1}^N \frac{x_k^2}{\pi_k^2} \cdot \text{Var}\left(\mathbb{I}_{\mathcal{S}}(x_k)\right) + \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \cdot \text{Cov}\left(\mathbb{I}_{\mathcal{S}}(x_k), \mathbb{I}_{\mathcal{S}}(x_l)\right) \\
&= \sum_{k=1}^N \frac{x_k^2}{\pi_k^2} \cdot \underbrace{\pi_k(1 - \pi_k)}_{\Delta_{k,k}} + \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \cdot \underbrace{\text{Cov}\left(\mathbb{I}_{\mathcal{S}}(x_k), \mathbb{I}_{\mathcal{S}}(x_l)\right)}_{\Delta_{k,l}} \\
&= \sum_{k=1}^N \sum_{\substack{l=1 \\ l \neq k}}^N \Delta_{k,l} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l}
\end{aligned}$$

Numéricamente, en el ejemplo anterior la varianza (simulada) de \hat{t} es:

```
var(total.muestra)
```

```
## [1] 9239.745
```

mientras que la *real* está dada por (ver ejercicio más adelante):

```
f      <- n/N
varianza <- N^2*(1 - f)/n*var(base.completa$x)
print(varianza)
```

```
## [1] 9309.829
```

Nota que tenemos un problema: para estimar $\text{Var}(\hat{t})$ necesitamos conocer todas las x_k de la población ¡lo cual es imposible! Entonces necesitamos un estimador de la varianza de \hat{t} para lo cual proponemos:

$$\widehat{\text{Var}}(\hat{t}) = \sum_{k=1}^n \sum_{l=1}^n \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l}$$

Para demostrar que el estimador es insesgado tomamos el valor esperado y agregamos las variables indicadoras correspondientes:

$$\widehat{\text{Var}}(\hat{t}) = \sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_l)$$

Se sigue la demostración:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\text{Var}}(\hat{t})\right] &= \mathbb{E}\left[\sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_l)\right] \\
&= \sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \underbrace{\mathbb{E}\left[\mathbb{I}_{\mathcal{S}}(x_k) \mathbb{I}_{\mathcal{S}}(x_l)\right]}_{*}
\end{aligned}$$

donde notamos que:

$$\begin{aligned}
* &= \mathbb{E} \left[\mathbb{I}_S(x_k) \mathbb{I}_S(x_l) \right] = \text{Cov} \left(\mathbb{I}_S(x_k), \mathbb{I}_S(x_l) \right) + \mathbb{E} \left[\mathbb{I}_S(x_k) \right] \mathbb{E} \left[\mathbb{I}_S(x_l) \right] \\
&= \pi_{k,l} - \pi_k \pi_l + \pi_k \pi_l \\
&= \pi_{k,l}
\end{aligned}$$

de donde se sigue:

$$\begin{aligned}
\mathbb{E} \left[\widehat{\text{Var}}(\hat{t}) \right] &= \sum_{k=1}^N \sum_{l=1}^N \frac{\Delta_{k,l}}{\pi_{k,l}} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} \underbrace{\pi_{k,l}}_* \\
&= \sum_{k=1}^N \sum_{l=1}^N \Delta_{k,l} \frac{x_k}{\pi_k} \frac{x_l}{\pi_l} = \text{Var}(\hat{t})
\end{aligned}$$

Podemos calcular la varianza estimada para una muestra aleatoria simple sin reemplazo como sigue (ver ejercicio):

```
f      <- n/N
varianza <- N^2*(1 - f)/n*var(muestra)
print(varianza)

## [1] 5919.348
```

Observaciones

1. La media muestral $\bar{x}_S = \frac{1}{n} \sum_{i=1}^n x_i$ es un estimador insesgado de la media poblacional $\bar{x}_U = \frac{1}{N} \sum_{i=1}^N x_i$. Se sigue de una factorización de n del total (t y \hat{t} respectivamente).
2. Se puede obtener $\text{Var}(\bar{x}_S)$ y $\widehat{\text{Var}}(\bar{x}_S)$ factorizando las n de manera cuadrática del \hat{t} .

Ejercicio

Definimos:

$$s_{x,U}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{x}_U)^2$$

como la **varianza poblacional ajustada** y

$$s_{x,S}^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}_S)^2$$

como la **varianza muestral ajustada**. Sea $f = \frac{n}{N}$ la **fracción muestral**. Demuestra que en el caso de muestreo aleatorio simple sin reemplazo:

$$\text{Var}(\hat{t}) = N^2 \frac{1-f}{n} s_{x,U}^2$$

mientras que el estimador insesgado se transforma en:

$$\widehat{\text{Var}}(\hat{t}) = N^2 \frac{1-f}{n} s_{x,S}^2$$

Teorema del Límite Central (Aplicación)

En esta sección hablaremos del teorema central del límite correspondiente a muestreo aleatorio simple con poblaciones finitas. Éste no es el mismo que el de Proba 2 (en términos de hipótesis) aunque las conclusiones sean las mismas. El teorema de Proba 2 establece que si se tiene una colección $\{X_i\}$ de variables aleatorias independientes idénticamente distribuidas (todas con distribución acumulada F_X) con media μ y varianza $\sigma^2 < \infty$, entonces, si definimos Z como:

$$Z = \lim_{n \rightarrow \infty} \sqrt{\frac{n}{\sigma^2}} \cdot \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right)$$

se tiene que $Z \sim \text{Normal}(0, 1)$.

En este teorema central podemos observar que hay algo muy parecido a la media muestral embebido en el teorema (la $\frac{1}{n} \sum_{i=1}^n X_i$) pero no es exactamente la media muestral (aquí se supone que todas las X_i son independientes con distribución F_X y en el caso de muestreo aleatorio sin reemplazo se sabe que las indicadoras **NO** son independientes y que de hecho tampoco son idénticamente distribuidas cuando analizamos $\sum_{i=1}^n x_i \mathbb{I}_S(x_i)$). Entonces *técnicamente* no podemos aplicar el teorema central del límite así como está a nuestra muestra. Sin embargo, Hájek (y más tarde [Rosen](#)) encontraron condiciones *sin tener que pedir independencia ni distribución idéntica* que permiten sustituir las X_i por las de la media muestral ($x_i \mathbb{I}_S(x_i)$) y que, cuando N y n tienden a infinito “de buena manera”, se tiene algo similar a esta expresión (**OJO** no es una expresión *correcta* pero es la idea):

$$Z = \lim_{N, n \rightarrow \infty} \sqrt{\frac{1}{\text{Var}(\bar{x}_S)}} \cdot \left(\frac{1}{n} \sum_{i=1}^N x_i \mathbb{I}_S(x_i) - \bar{x}_U \right)$$

donde $\mu = \sum_{k=1}^N x_k$ es la media poblacional y $\sigma^2 = \frac{1}{N} \sum_{k=1}^N (x_k - \mu)^2$ la varianza poblacional no ajustada. La demostración propia de este teorema la posponemos para una sección posterior. Por ahora, ejemplificaremos el teorema del límite central en R, utilizaremos la expresión anterior para deducir y explicar el concepto de intervalo de confianza y, finalmente, haremos un ejemplo de estimación de intervalo.

Programación en R del teorema del límite central con variables aleatorias independientes idénticamente distribuidas

Lo que programaremos (por facilidad) en esta sección corresponde a ejemplos del teorema de proba 2: dadas variables aleatorias independientes idénticamente $\{X_i\}$ distribuidas con media μ y varianza finita σ^2 tenemos que:

$$Z = \lim_{n \rightarrow \infty} \sqrt{\frac{n}{\sigma^2}} \cdot \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right) \sim \text{Normal}(0, 1)$$

donde el símbolo \sim se lee “se distribuye”. En este caso la interpretación va a ser que para n muy grande tendremos que

$$\sqrt{\frac{n}{\sigma^2}} \cdot \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right) \sim \text{Normal}(0, 1)$$

donde \sim se lee como “se distribuye aproximadamente”. Programaremos una función en R que para n grande muestre eso:

```
TeoremaCentrallimite <- function(numero_simulaciones = 1000,
  n = c(10, 100, 1000, 10000),
  distribucion = rpois, mu = 1, sigma = 1,
  bins = 50,
  ncol = 2, distname = "Poisson(1)",
```

```

                                rcolor = sample(rainbow(100),1), ...){

#Creamos
plot_list <- list()

for (k in n){

  #Guardamos las Zi en un vector
  Z <- rep(NA, numero_simulaciones)

  #Simulamos todas las simulaciones
  for (i in 1:numero_simulaciones){
    simulaciones_X <- distribucion(n = k, ...)
    Z[i] <- sqrt(k)*(sum(simulaciones_X/k) - mu)
  }

  #Graficación
  x <- seq(min(Z)-1, max(Z) + 1, length.out = 1000)
  y <- dnorm(x, sd = sigma)
  plot_list <- list.append(
    plot_list,
    ggplot() +
    geom_histogram(aes(x = Z, y = ..density..), bins = bins, fill = rcolor,
                    data = data.frame(Z = Z)) +
    geom_line(aes_string(x = x, y = y), color = "black", data = data.frame(x = x, y = y)) +
    ggtitle(paste0("Simulaciones con ", distname, "\nn = ", k)) +
    xlab("Z") + ylab("Densidad de Z") +
    theme_bw()
  )
}
do.call("grid.arrange", c(plot_list, ncol = ncol))
}

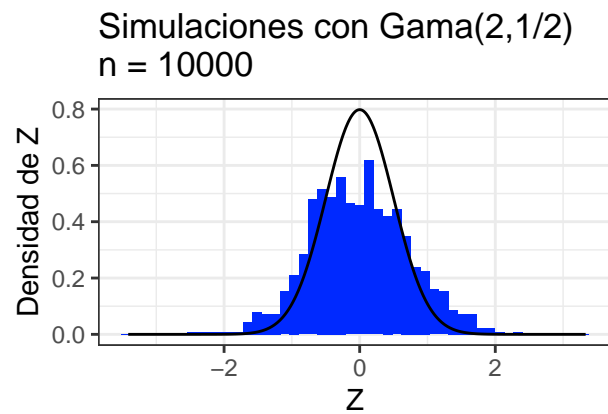
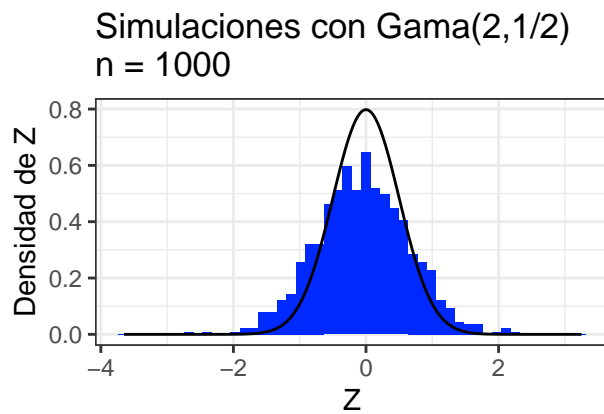
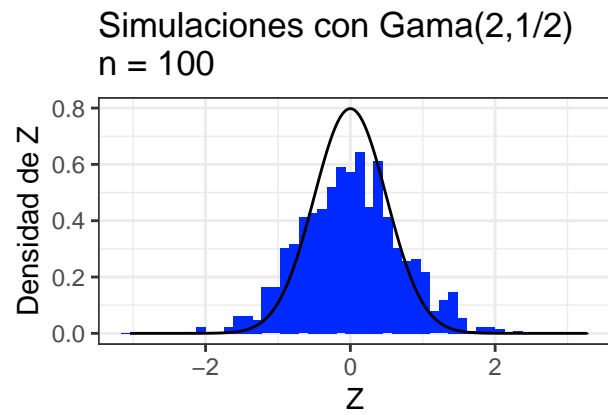
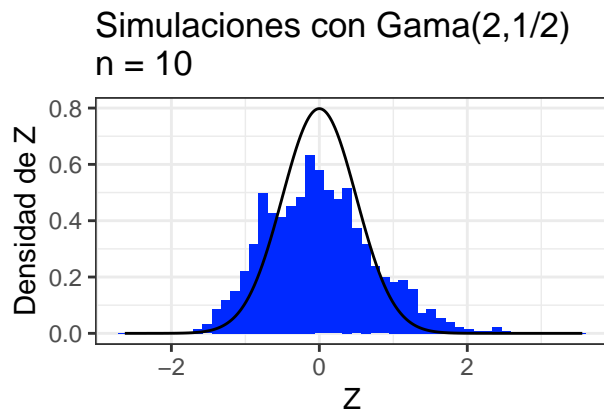
```

donde podemos ver la aproximación normal si tomamos, por ejemplo, las X_i siguen una distribución Gamma:

```

TeoremaCentralLmite(distribucion = rgamma, mu = 1, sigma = 0.5, shape = 2,
                     scale = 0.5, distname = "Gama(2,1/2)")

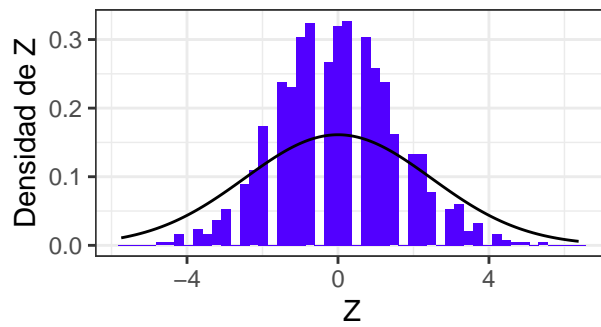
```



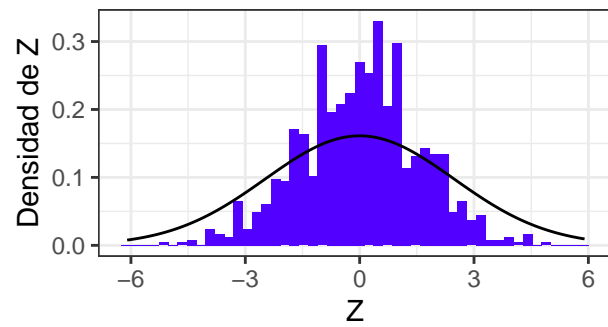
La binomial se ve así:

```
TeoremaCentralLmite(distribucion = rbinom, mu = 4.5, sigma = 2.475, size = 10,
  prob = 0.45, distname = "Binomial(10,0.45)")
```

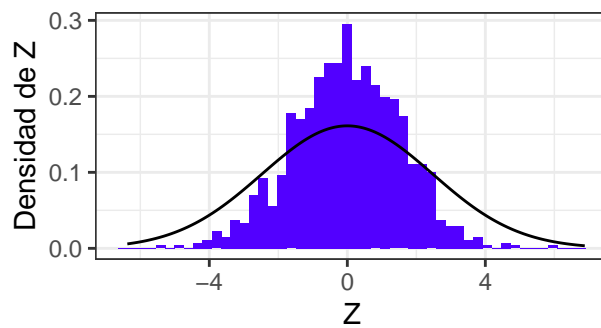
Simulaciones con Binomial(10,0.4)
n = 10



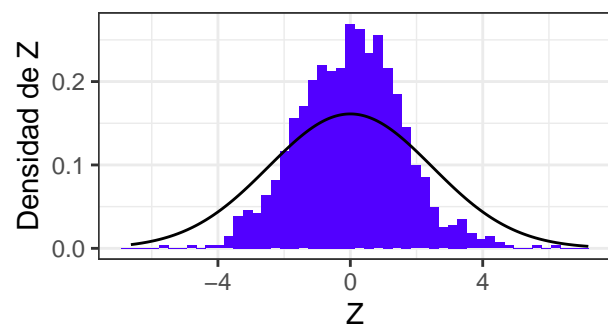
Simulaciones con Binomial(10,0.4)
n = 100



Simulaciones con Binomial(10,0.4)
n = 1000

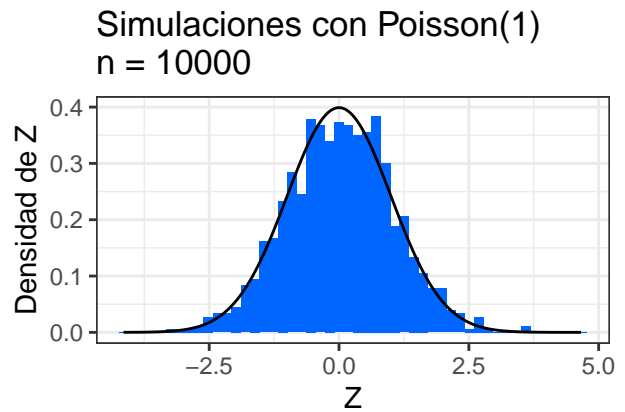
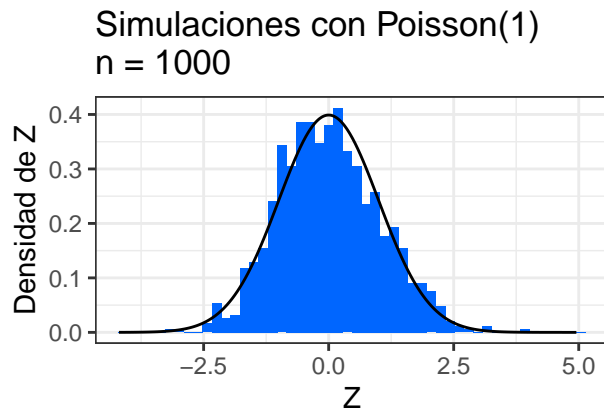
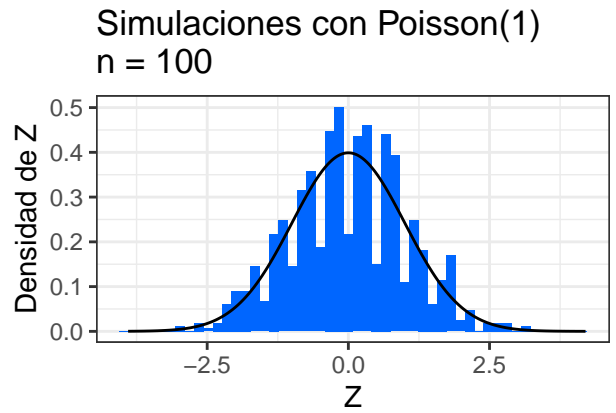
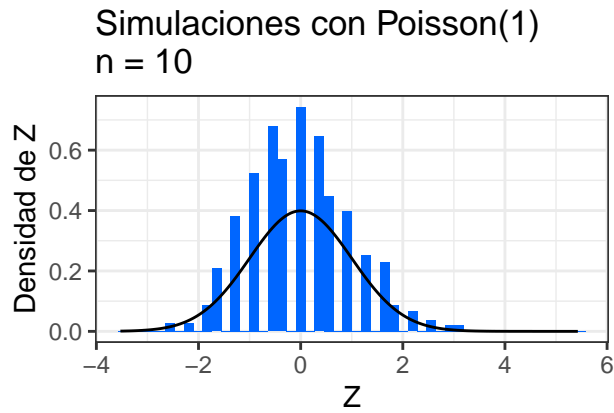


Simulaciones con Binomial(10,0.4)
n = 10000



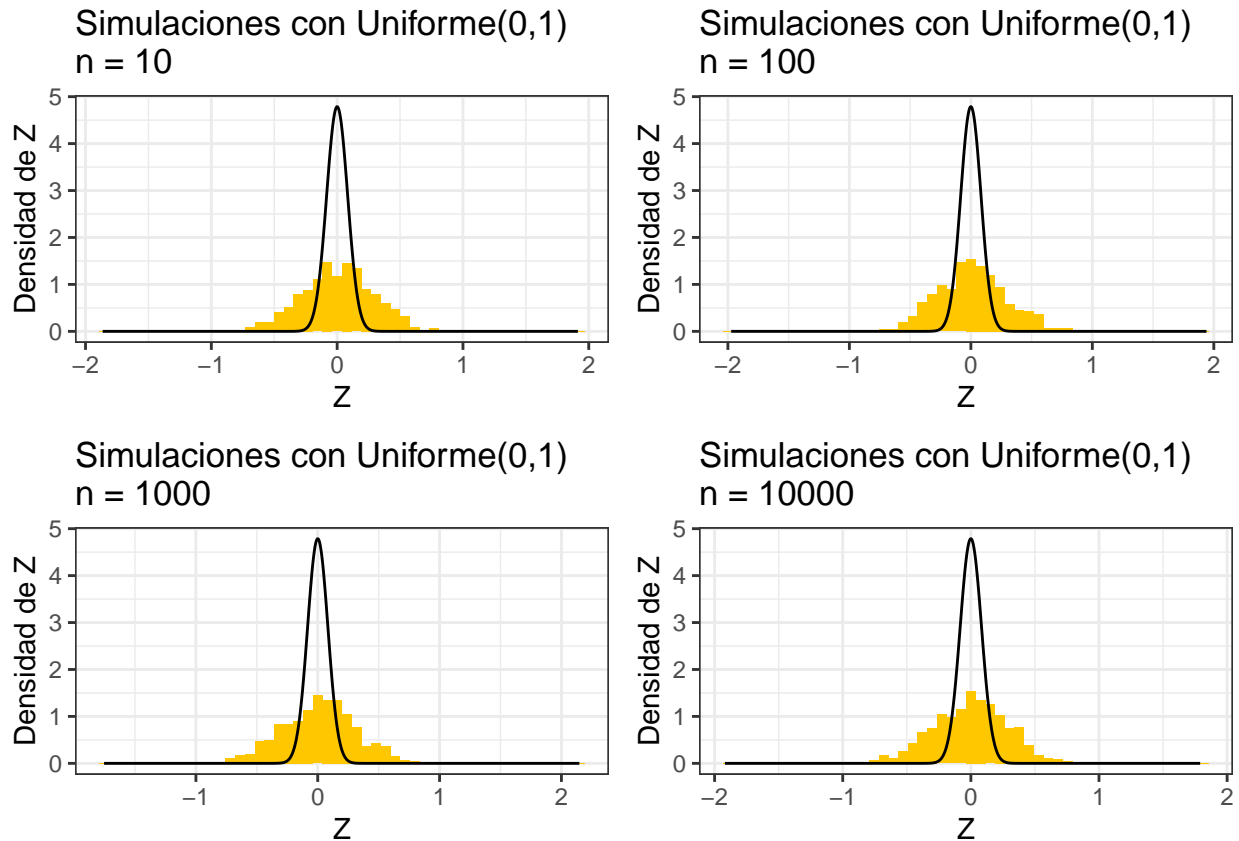
Poisson:

```
TeoremaCentralLimite(lambda = 1, distname = "Poisson(1)")
```



E inclusive uniformes:

```
TeoremaCentralLmite(distribucion = runif, mu = 1/2, sigma = 1/12,
                     distname = "Uniforme(0,1)")
```



Experimenta con otras distribuciones ¿puedes encontrar alguna para la que no funcione?

Ejercicio

Repite la programación del teorema del límite central pero ahora tomando las X_k con distintas distribuciones siempre y cuando X_k tenga media μ_k finita y las variables aleatorias satisfagan la [condición de Lindberg](#) (una forma de hacerlo es teniendo varianzas finitas que no incrementan con la k).

Estimación de intervalos de confianza para el total

Un intervalo de confianza de $(1 - \alpha) \times 100\%$ de un estimador poblacional desconocido $\theta = \theta(x_1, x_2, \dots, x_N)$ (constante) es un intervalo aleatorio de la forma $[L(\mathcal{S}), U(\mathcal{S})]$ (donde L, U son variables aleatorias que dependen de la muestra) tal que

$$\mathbb{P}(\theta \in [L(\mathcal{S}), U(\mathcal{S})]) = 1 - \alpha$$

Notamos que lo aleatorio del intervalo son las cotas del mismo y que, dadas distintas muestras \mathcal{S} el valor de interés θ no siempre va a caer ahí. La idea de un intervalo es poder dar una cota de más o menos dónde anda un valor. Veamos un ejemplo con el total.

Recordamos que el estimador del total es insesgado $\mathbb{E}[\hat{t}] = t$ y que por definición:

$$\hat{t} = N \frac{1}{n} \sum_{i=1}^N x_i \cdot \mathbb{I}_{\mathcal{S}}(x_i)$$

luego usando la versión de muestreo finito del teorema central del límite (factorizando N) tenemos que:

$$\sqrt{\frac{1}{\text{Var}(\bar{x}_S)}} \cdot \left(\frac{1}{n} \sum_{i=1}^N x_i \mathbb{I}_S(x_i) - \bar{x}_U \right) = \cdot N \frac{\left(\frac{1}{n} \sum_{i=1}^N x_i \mathbb{I}_S(x_i) - \bar{x}_U \right)}{N \sqrt{\text{Var}(\bar{x}_S)}} = \frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \sim \text{Normal}(0, 1)$$

De donde se sigue que si se desea tener un intervalo de tamali $(1 - \alpha) \times 100\%$ lo que hay que hacer es buscar $L(S)$ y $U(S)$ tales que:

$$\mathbb{P} \left(L(S) \leq \frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \leq U(S) \right) = 1 - \alpha$$

En este caso las probabilidades (por aproximación asintótica) se modelan bajo la hipótesis de normalidad. Y tomamos ventaja de que la normal es simétrica respecto a la media para proponer que $L(S) = -U(S)$ y ambas correspondan a $\pm \Phi^{-1}(\alpha/2)$ (la función de distribución acumulada inversa de la normal). Es decir, ambos deben corresponder a los cuantiles con probabilidad $\alpha/2$ y $1 - \alpha/2$, denotados $z_{\alpha/2}$ y $z_{1-\alpha/2}$. Por simetría de la normal tenemos que: $z_{\alpha/2} = -z_{1-\alpha/2}$ y por tanto:

$$\mathbb{P} \left(z_{\alpha/2} \leq \frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \leq z_{1-\alpha/2} \right) = 1 - \alpha$$

de donde despejamos:

$$\mathbb{P} \left(z_{\alpha/2} \sqrt{\text{Var}(\hat{t})} \leq \hat{t} - t \leq z_{1-\alpha/2} \sqrt{\text{Var}(\hat{t})} \right) = \mathbb{P} \left(\hat{t} - z_{1-\alpha/2} \sqrt{\text{Var}(\hat{t})} \leq t \leq \hat{t} + z_{\alpha/2} \sqrt{\text{Var}(\hat{t})} \right) = 1 - \alpha$$

Notamos que como no conocemos $\text{Var}(\hat{t})$ la podemos aproximar mediante $\widehat{\text{Var}}(\hat{t})$ (hay mejores aproximaciones mediante una t de Student asintótica pero no lo usaremos ahora) y tener intervalos aproximados de la forma:

$$\begin{aligned} L(S) &= \hat{t} - z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{t})} \\ U(S) &= \hat{t} + z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{t})} \end{aligned} \tag{2}$$

de manera concisa muchas veces los escribimos como:

$$\hat{t} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{t})}$$

Ejemplo con simulación:

Veamos cómo se ven múltiples intervalos simulados con confianza del 90% y suponiendo la varianza es conocida

```
nsim <- 100
n     <- 100

total.muestra <- rep(NA, nsim)
confianza.bajo <- rep(NA, nsim)
confianza.alto <- rep(NA, nsim)
f <- n/N
z <- qnorm(1 - 0.1/2)

var.total      <- N^2*(1 - f)/n*var(base.completa$x)

for (i in 1:nsim){
```

```

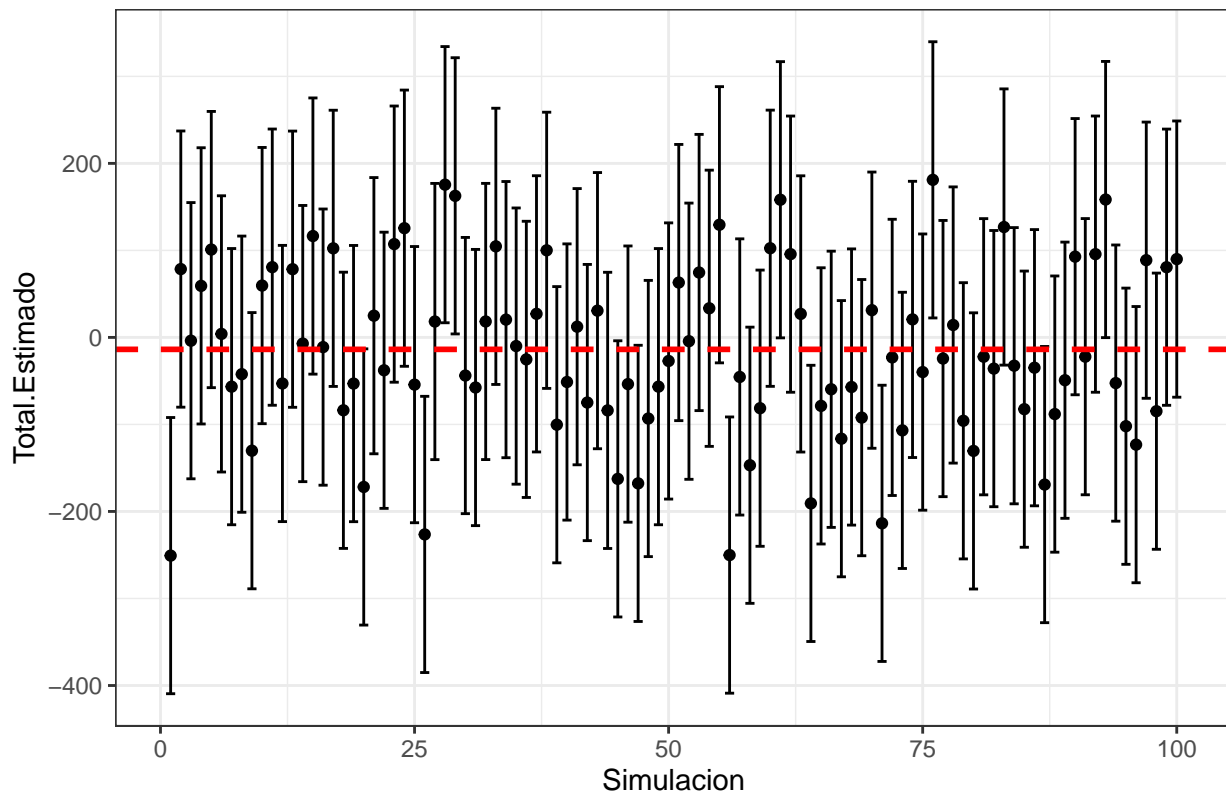
muestra          <- sample(base.completa$x, n, replace = FALSE)
total.muestra[i] <- N*mean(muestra)
#var.total[i]    <- N^2*(1 - f)/n*var(muestra)
confianza.bajo[i] <- total.muestra[i] - z*sqrt(var.total)
confianza.alto[i] <- total.muestra[i] + z*sqrt(var.total)
}

intervalos.simulados <- data.frame(
  Simulacion = 1:nsim,
  Intervalo.Bajo = confianza.bajo,
  Total.Estimado = total.muestra,
  Intervalo.Alto = confianza.alto
)

ggplot(intervalos.simulados) +
  geom_point(aes(x = Simulacion, y = Total.Estimado)) +
  geom_errorbar(aes(x = Simulacion, ymin = Intervalo.Bajo,
                    ymax = Intervalo.Alto)) +
  geom_hline(aes(yintercept = sum(base.completa$x)),
             linetype = "dashed",
             size = 1, color = "red") +
  theme_bw() +
  ggtitle("Simulación de intervalos de confianza")

```

Simulación de intervalos de confianza



Nota que estos intervalos son aproximados y no siempre van a funcionar. (¿Puedes hallar un ejemplo donde no sirvan a pesar de que n y N sean grandes?) Luego veremos correcciones a esto; por ahora, supondremos que la aproximación es buena.

Ejemplo Resumen: Estimación de una proporción bajo muestreo aleatorio simple sin reemplazo

Se realiza una encuesta mediante muestreo aleatorio simple sin reemplazo a la población del ITAM $N = 5000$ donde interesa conocer la proporción de gente que apoya al gobierno en turno p . Implícitamente, se supone que alguien apoya (proporción p de toda la población) o no lo apoya (proporción $1 - p$), que dichos conjuntos son disjuntos y que no hay una tercera opción (como NO RESPONDE / DESCONOCE QUIÉN GOBIERNA). La pregunta es: ¿a cuántas personas hay que encuestar si interesa estimar p con un error máximo de tamaño $\epsilon = 0.05$ al 99% de confianza (es decir, que el estimador \hat{p} de la proporción esté, a lo más, a ± 0.05 de distancia del valor verdadero p con un intervalo de confianza al 99%)?

Supongamos tomamos una muestra de tamaño n dada por $\mathcal{S} = (x_1, x_2, \dots, x_n)^T$ de una población $\mathcal{U} = (x_1, x_2, \dots, x_N)^T$ de tamaño N . Pensemos, además, existen N_1 personas que aprueban al gobierno actual y $N - N_1$ que desaprueban del mismo y por tanto la proporción que nos interesa estimar es:

$$p = \frac{N_1}{N}$$

Por otro lado, la proporción muestral de personas que aprueban está dada por:

$$\hat{p} = \frac{\sum_{i=1}^n \mathbb{I}_{\text{Aprueba}}(x_i)}{n}$$

donde si definimos $H = \frac{\sum_{i=1}^n \mathbb{I}_{\text{Aprueba}}(x_i)}{n}$ notamos que la distribución de H está dada por una variable [Hipergeométrica](#) (pues de una población de N se seleccionan n donde N_1 cumplen la categoría deseada). Su media y varianza están dadas respectivamente por:

$$\mathbb{E}[H] = n \frac{N_1}{N} = np$$

así como por:

$$\text{Var}[H] = n \frac{N_1}{N} \left(1 - \frac{N_1}{N}\right) \left(\frac{N-n}{N-1}\right) = np(1-p) \left(\frac{N-n}{N-1}\right)$$

Se sigue entonces que $\mathbb{E}[\hat{p}] = p$ y por tanto \hat{p} es un estimador insesgado. La varianza por otro lado es:

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n} \left(\frac{N-n}{N-1}\right)$$

Finalmente, el estimador de la varianza es:

$$\widehat{\text{Var}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)$$

el cual también cumple que es insesgado (demuéstralo).

Podemos aplicar el Teorema Central del Límite para la proporción³ notando que la definición de \hat{p} coincide con una media (de las indicadoras):

$$\frac{\hat{p} - p}{\underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)}}_{\widehat{\text{Var}}(\hat{p})}} \sim \text{Normal}(0, 1)$$

De donde se tiene que:

$$\begin{aligned} \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)}} \leq z_{\alpha/2}\right) &\approx 1 - \alpha \\ \Rightarrow \mathbb{P}\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1}\right)}\right) &\approx 1 - \alpha \end{aligned} \tag{3}$$

³Una mejor distribución sería una t de Student; empero eso lo verás en Estadística Matemática.

Nota Es común encontrar en Internet que para los intervalos de confianza la gente supone una población muy grande N respecto a la muestra n y entonces eliminan el término $\frac{N-n}{N-1}$ argumentando que $\frac{N-n}{N-1} \approx 1$ y obtienen la siguiente fórmula:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

esto simplifica algunos cálculos (a mano) pero nosotros tenemos R y podemos hacer cálculos más exactos sin tener que suponer semejantes atrocidades.

Como el error deseado es de tamaño ϵ queremos $|p - \hat{p}| \leq \epsilon$ esto se traduce en:

$$|p - \hat{p}| \leq \underbrace{z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right)}}_{\epsilon}$$

de donde igualamos para despejar la n :

$$\begin{aligned} \epsilon &= z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right)} \\ &= \frac{\epsilon^2}{z_{\alpha/2}^2} = \frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right) \\ &= \frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\ &= \frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1 = \frac{N}{n} \\ \Rightarrow n &= \frac{N}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} = \frac{\frac{z_{\alpha/2}^2}{\epsilon^2} \hat{p}(1-\hat{p})}{\frac{N-1}{N} + \frac{1}{N} \frac{z_{\alpha/2}^2}{\epsilon^2} \hat{p}(1-\hat{p})} = \frac{m}{1 + \frac{m-1}{N}} \end{aligned}$$

donde

$$m = \frac{z_{\alpha/2}^2}{\epsilon^2} \hat{p}(1-\hat{p})$$

Ahora el problema es que el tamaño de muestra n depende de la muestra a través de \hat{p} ¡y no hemos tomado la muestra! Para ello entonces analizamos el peor caso que puede ocurrir de \hat{p} de tal forma que obtengamos la n que puede salir con la peor proporción \hat{p} posible. Para ello maximizamos con derivadas:

$$\begin{aligned} \frac{\partial n}{\partial \hat{p}} &= \frac{\partial}{\partial \hat{p}} \left(\frac{N}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right) \\ &= N \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \cdot \frac{\partial}{\partial \hat{p}} \left(\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1 \right) \\ &= \underbrace{N(N-1) \frac{\epsilon^2}{z_{\alpha/2}^2}}_C \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \cdot \frac{\partial}{\partial \hat{p}} \left(\frac{1}{\hat{p}(1-\hat{p})} \right) \\ &= C \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \left(\frac{1}{\hat{p}(1-\hat{p})} \right)^2 \frac{\partial}{\partial \hat{p}} \hat{p}(1-\hat{p}) \\ &= C \left(\frac{1}{\frac{N-1}{\hat{p}(1-\hat{p})} \frac{\epsilon^2}{z_{\alpha/2}^2} + 1} \right)^2 \left(\frac{1}{\hat{p}(1-\hat{p})} \right)^2 (1-2\hat{p}) = 0 \end{aligned}$$

de donde se sigue que $\hat{p} = \frac{1}{2}$ es un punto crítico. De hecho puede verificarse que es el máximo (por ejemplo a través de la segunda derivada). Luego, podemos estimar la n de la muestra mediante:

$$n = \left\lceil \frac{m}{1 + \frac{m-1}{N}} \right\rceil$$

donde $m = \frac{1}{4} \frac{z_{\alpha/2}^2}{\epsilon^2}$. En el caso particular de este ejercicio, $N = 5000$, $\epsilon = 0.05$, $\alpha = 0.01$ y $z_{\alpha/2}^2 \approx \text{qnorm}(0.9)$. Luego podemos calcular:

```
alpha <- 0.01
z <- qnorm(1 - alpha/2)
epsilon <- 0.05
m <- (1/4)*(z/epsilon)^2
N <- 5000
n <- ceiling(m/(1 + (m-1)/N))

print(paste0("El tamaño de muestra es ", n))

## [1] "El tamaño de muestra es 586"
```

Ejemplo Resumen: Estimación del total de individuos en una fotografía

En este ejercicio vamos a determinar cuánta gente aparece en la siguiente foto:

```
knitr::include_graphics("concierto.jpg")
```



Figure 1: Imagen de un concierto extraída de <https://www.youtube.com/watch?v=pJ1YKwyH5bk>

Hay varias opciones para determinar la cantidad de gente que está en dicha foto. Una sería contar todas las cabecitas que aparecen; otra, diseñar un modelo de redes neuronales (o de [convolución](#) porque a la gente le encanta eso) que identifique una cabeza y la cuente. Nosotros lo que haremos (por ser un curso de estadística) será muestrear. Como investigador me interesa responder la siguiente pregunta:

¿Cuánta gente está en la fotografía con un intervalo de error de ± 50 casos al 95%?

Para ello dividiremos la fotografía en N pedazos (a determinar), muestrearemos n de ellos y contaremos la cantidad de personas que aparecen en cada pedazo. Finalmente, generamos intervalos de confianza y de muestreo. Para ello repetimos el ejercicio anterior de despejar la n del intervalo de confianza; por el teorema del límite central tenemos:

$$\frac{\hat{t} - t}{\sqrt{\text{Var}(\hat{t})}} \sim \text{Normal}(0, 1)$$

de donde obtenemos intervalos (¡verifícalo!) de la forma:

$$\hat{t} \pm z_{1-\alpha/2} \cdot \sqrt{\text{Var}(\hat{t})}$$

Donde podemos aproximar la varianza mediante $\widehat{\text{Var}}(\hat{t}) = N^2 \frac{1-f}{n} s_{x,S}^2$ donde recordamos que $f = n/N$ y $s_{x,S}^2$ es la varianza muestral. Tomamos $\epsilon = 50$ y despejamos:

$$\begin{aligned} \epsilon &= z_{1-\alpha/2} \cdot \sqrt{\text{Var}(\hat{t})} \\ \Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2} &= N^2 \frac{1-f}{n} s_{x,S}^2 \\ \Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N^2} &= \frac{1 - \frac{n}{N}}{n} \\ \Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N^2} &= \frac{1}{n} - \frac{1}{N} \\ \Rightarrow \frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N^2} + \frac{1}{N} &= \frac{1}{n} \\ \Rightarrow \frac{1}{N} \left(\frac{\epsilon^2}{z_{1-\alpha/2}^2 s_{x,S}^2 N} + 1 \right) &= \frac{1}{n} \\ \Rightarrow \frac{1}{N} \left(\frac{\epsilon^2 + z_{1-\alpha/2}^2 s_{x,S}^2 N}{z_{1-\alpha/2}^2 s_{x,S}^2 N} \right) &= \frac{1}{n} \\ \Rightarrow \left(\frac{(z_{1-\alpha/2} s_{x,S} N)^2}{\epsilon^2 + z_{1-\alpha/2}^2 s_{x,S}^2 N} \right) &= n \end{aligned}$$

El problema aquí es que la n depende de la varianza muestral $s_{x,S}^2$ (actualmente desconocida) así como de la cantidad de cuadritos originales N en los que dividimos la foto. Hay en la literatura varias técnicas que se pueden utilizar para estimar el $s_{x,S}^2$:

1. Realizar un estudio piloto (es decir un pequeño ejemplo de lo que vas a hacer en una población chica y de ahí tener la varianza). Esta es la mejor opción.
2. Buscar otros estudios similares donde se analicen objetos similares de estudio y ver sus varianzas; suponer que la de este estudio es similar. Esta es la segunda mejor opción.
3. Inventártela (sí, es una opción pero no la mejor). Vamos, ¿cuál es la probabilidad de que nadie en todo el mundo haya hecho un análisis similar al tuyo? Si realmente estás haciendo algo completamente nuevo *sin estudio piloto* pues... podrías inventarla. ¿Lo recomiendo? No; pero pasa.

En nuestro caso utilizaremos la varianza estimada [de este artículo](#) reportada en 1.02; luego $s_{x,S}^2 \approx 1.02$ para nuestro análisis.

Finalmente, como éste es sólo un ejercicio de clase tomaremos $N = 100$ (dividir la foto en 100 cuadritos). De manera profesional, de nuevo habría que ver diferencias en los resultados de las estimaciones en función de los cuadritos, o bien asignar un costo a la cantidad de cuadros. Concluimos entonces que para nuestro estudio:

$$n = \left\lceil \frac{(z_{1-\alpha/2} s_{x,S} N)^2}{\epsilon^2 + z_{1-\alpha/2}^2 s_{x,S}^2 N} \right\rceil = \left\lceil \frac{(1.95 \cdot \sqrt{1.02} \cdot 100)^2}{50^2 + 1.95^2 \cdot 1.02 \cdot 100} \right\rceil$$

Podemos calcular en R:

```
n <- ceiling((qnorm(0.975)*sqrt(1.02)*100)^2/(50^2 + (qnorm(0.975)^2*1.02*100)))
print(paste0("El tamaño de muestra es ", n))
```

```
## [1] "El tamaño de muestra es 14"
```

Podemos proceder a dividir la foto en los $N = 100$ pedazos:

```
#División con base en el siguiente link:
#https://rpubs.com/issacttoast/cutimage
library(imager)

#Cargamos la imagen
img <- load.image("concierto.jpg")

#Función auxiliar del link superior
make.vr <- function( x, name ){
  assign( name, x, envir = .GlobalEnv)
}

#División en N
N <- 100
par(mfrow=c(sqrt(N),sqrt(N)), mar = c(0.1,0.1,0.1,0.1))
k <- 1
for (j in 1:sqrt(N)){
  for (i in 1:sqrt(N)){
    vr.name <- paste0("sub", k)
    k <- k + 1
    imsub(img, (width/sqrt(N))*(i-1) < x & x < i * (width/sqrt(N)),
          (height/sqrt(N))*(j-1) < y & y < j * (height/sqrt(N))) %>%
      make.vr(name = vr.name) %>%
      # save.image( file = paste0(vr.name, ".jpg")) %>%
      plot(axes = FALSE,
           xaxt="n", yaxt="n",
           xlab = "", ylab = "", ann = FALSE )
  }
}
```




Podemos acceder a cada una de las imágenes que se tienen a través de su nombre (sub seguido de un número entre 0 y 100). Muestreamos entonces los nombres de las 15 imágenes:

```
#Obtenemos los dígitos a muestrear
imagenes.muestreadas <- sample(1:100, n, replace = FALSE)

#Agregamos el prefijo sub
imagenes.muestreadas <- paste0("sub", imagenes.muestreadas)
```

Y graficamos cada una de ellas:

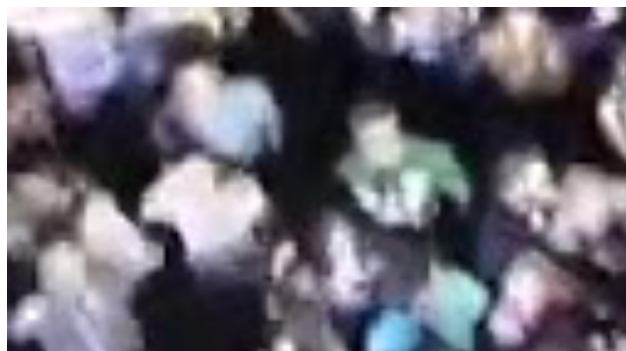
```
par(mfrow = c(1,1))

for (imagen in imagenes.muestreadas){
  plot(get(imagen), main = imagen, axes = FALSE)
}
```

sub8



sub82



sub67



sub52



sub96



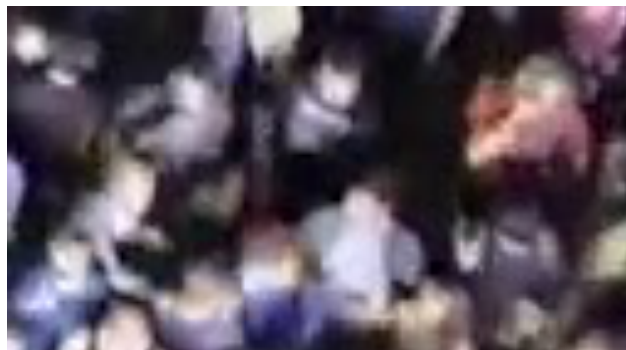
sub49



sub25



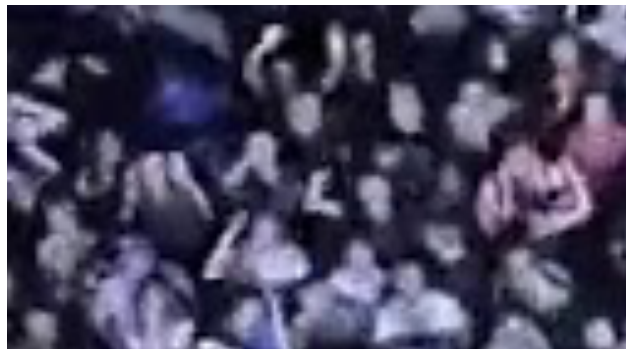
sub70



sub27



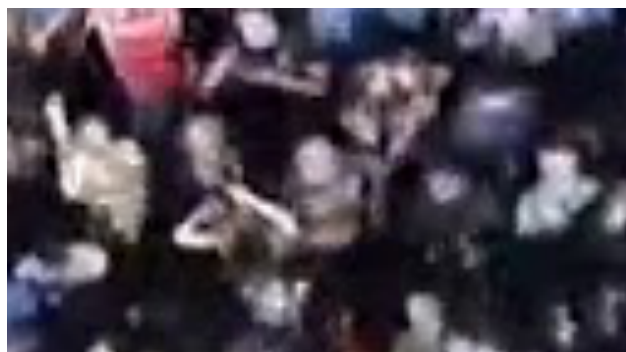
sub13



sub89



sub42



sub32



sub30



Para cada una de las imágenes contamos las cabecitas que aparecen:

```
datos <- data.frame(
  Imagen = imagenes.muestreadas,
  Conteo = c(13, 11, 9, 14, 9, 15, 14, 10, 1, 22, 8, 9, 17, 16)
)
kable(datos) %>% kable_styling(latex_options = "striped")
```

Imagen	Conteo
sub8	13
sub82	11
sub67	9
sub52	14
sub96	9
sub49	15
sub25	14
sub70	10
sub27	1
sub13	22
sub89	8
sub42	9
sub32	17
sub30	16

Tenemos entonces que la estimación del total \hat{t} es: 1200, por otro lado la varianza muestral es $s_{x,S}$ está dada por: 25.2307692. Podemos entonces establecer un intervalo de confianza para el total:

```
x <- c(13, 11, 9, 14, 9, 15, 14, 10, 1, 22, 8, 9, 17, 16, 10)
s2 <- var(x)
N <- 100
n <- 15
total.muestra <- N*mean(x)
ci <- qnorm(0.975)*sqrt(N^2*(1 - n/N)/n*s2)
ci_low <- round(total.muestra - ci,2)
ci_up <- round(total.muestra + ci,2)
```

```
print(paste0("Se estiman ", round(total.muestra,2), " personas con intervalo de ",
"confianza al 95% de [", ci_low, " ,", ci_up,"]"))
```

```
## [1] "Se estiman 1186.67 personas con intervalo de confianza al 95% de [959.55 ,1413.78]"
```

Ejercicio:



Cuando se registra un paquete de R en [CRAN](#) estos se registran junto con sus autores como muestra la imagen:

```
knitr::include_graphics("CRAN.png")
```

ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics

A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Version: 3.3.2
 Depends: R (≥ 3.2)
 Imports: digest, glue, grDevices, grid, gtable (≥ 0.1.1), isoband, MASS, mgcx, rlang (≥ 0.3.0), scales (≥ 0.5.0), stats, tibble, withr (≥ 2.0.0)
 Suggests: covr, dplyr, ggplot2movies, hexbin, Hmisc, knitr, lattice, maptools, maps, mapproj, maps, maptools, multcomp, munsell, nlme, profvis, quantreg, RColorBrewer, rgeos, rmarkdown, rpart, sf (≥ 0.7-3), syglight (≥ 1.2.0.9001), testthat (≥ 2.1.0), ydiff (≥ 0.3.0)
 Enhances: sp
 Published: 2020-06-19

Author: Hadley Wickham  [aut], Winston Chang  [aut], Lionel Henry [aut], Thomas Lin Pedersen  [aut, cre], Kohske Takahashi [aut], Claus Wilke  [aut], Kara Woo  [aut], Hiroaki Yutani  [aut], Dewey Dunnington  [aut], RStudio [cph, fnd]
 Maintainer: Thomas Lin Pedersen <thomas.pedersen@rstudio.com>
 BugReports: <https://github.com/tidyverse/ggplot2/issues>
 License: [GPL-2 | file LICENSE](#)
 URL: <http://ggplot2.tidyverse.org>, <https://github.com/tidyverse/ggplot2>
 NeedsCompilation: no
 Citation: [ggplot2 citation info](#)
 Materials: [README NEWS](#)
 In views: [Graphics](#), [Phylogenetics](#), [TeachingStatistics](#)
 CRAN checks: [ggplot2 results](#)

Downloads:

Reference manual: [ggplot2.pdf](#)
 Vignettes: [Extending ggplot2](#), [Using ggplot2 in packages](#), [Aesthetic specifications](#)
 Package source: [ggplot2_3.3.2.tar.gz](#)
 Windows binaries: r-devel: [ggplot2_3.3.2.zip](#), r-release: [ggplot2_3.3.2.zip](#), r-oldrel: [ggplot2_3.3.2.zip](#)
 macOS binaries: r-release: [ggplot2_3.3.2.tgz](#), r-oldrel: [ggplot2_3.3.2.tgz](#)
 Old sources: [ggplot2 archive](#)

Reverse dependencies:

La información de un paquete puede encontrarse en la página de [CRAN](#) dando clic en **Packages** y luego en **Table of available packages, sorted by name** y buscando el paquete deseado.

Se desea conocer el número promedio de autores por paquete registrado en [CRAN](#) con un intervalo de confianza al 80% y un error de ± 1 . Obtén la n necesaria para muestrear, calcula un estimador de la media y obtén intervalos de confianza. Justifica tu elección de la varianza para la n mediante un estudio piloto (muestreando de manera inicial 10 y calculando la varianza de ellos).

Hint Para obtener una lista (censo) de todos los paquetes de R puedes utilizar la función `available.packages()` la cual devuelve una matriz con todos los paquetes e incluye la `url` de donde se encuentra.

Ejemplo Resumen: Estimación de una región crítica

En una elección existen dos candidatas A y B . Se realiza una encuesta de opinión mediante muestreo aleatorio simple sin reemplazo donde se les pregunta a una cantidad suficiente de votantes por quién votarían de las dos. En este análisis no hay NO SABE / NO RESPONDE sino que todos los individuos indican su preferencia. Se desea determinar la cantidad de puntos porcentuales que debe haber de diferencia entre la proporción de

individuos que reportan apoyan al candidato A y los que reportan que apoyan al B de tal forma que el 95% de las veces podamos declarar de manera adecuada al ganador.

Nota Si A no es el ganador entonces $p_A < 50\%$ (la proporción de votantes que van a elegir a A es menor a la mitad) ¿cierto?

Para ello el análisis es como sigue: sea \hat{p}_A un estimador de la proporción de individuos que van a elegir a A y p_A la verdadera proporción. Sin pérdida de generalidad supondremos que B es el ganador; es decir que $p_A < 0.5$. El problema puede traducirse en determinar una c tal que:

$$\mathbb{P}(\hat{p}_A > c | p_A < 0.5) \leq 0.05$$

Notamos que el evento $\{p_A < 50\%\}$ es por definición conocido (con probabilidad 0 ó 1) pues está dado por la población (constante). Notamos que por el teorema del límite central podemos escribir:

$$\frac{\hat{p}_A - p_A}{\sqrt{\text{Var}(\hat{p}_A)}} \sim \text{Normal}(0, 1)$$

donde $\hat{p}_A = \frac{1}{N} \sum_{i=1}^N x_i \mathbb{I}_S(x_i)$ como anteriormente hicimos para proporciones y su varianza está dada por:

$$\text{Var}(\hat{p}_A) = \frac{p_A(1-p_A)}{n} \left(\frac{N-1}{N-n} \right)$$

donde el cálculo se hizo en el primer ejemplo de esta sección. Podemos transformar el problema entonces en hallar c tal que:

$$\mathbb{P} \left(\underbrace{\frac{\hat{p}_A - p_A}{\sqrt{\text{Var}(\hat{p}_A)}}}_{Z \sim \text{Normal}(0,1)} > \frac{c - p_A}{\sqrt{\text{Var}(\hat{p}_A)}} \middle| p_A < 0.5 \right) \leq 0.05$$

Notamos que el lado izquierdo tiene una aproximación normal y entonces podemos reescribir el problema como hallar c tal que:

$$\mathbb{P} \left(Z > \frac{c - p_A}{\sqrt{\text{Var}(\hat{p}_A)}} \middle| p_A < 0.5 \right) \leq 0.05 \quad \text{donde } Z \sim \text{Normal}(0, 1).$$

Recordando la expresión para la varianza sustituyo:

$$\mathbb{P} \left(Z > \frac{c - p_A}{\sqrt{\frac{p_A(1-p_A)}{n} \left(\frac{N-1}{N-n} \right)}} \middle| p_A < 0.5 \right) \leq 0.05 \quad \text{donde } Z \sim \text{Normal}(0, 1).$$

En función del análisis pasado, observamos que $\frac{c - p_A}{\sqrt{\frac{p_A(1-p_A)}{n} \left(\frac{N-1}{N-n} \right)}}$ es una función decreciente en términos de

p_A (¡compruéballo!) y que el mínimo valor se alcanza en el máximo de la p_A en el intervalo; es decir cuando $p_A = \frac{1}{2}$. Luego el problema se transforma en hallar c tal que:

$$\mathbb{P} \left(Z > \frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}} \right) \leq 0.05 \quad \text{donde } Z \sim \text{Normal}(0, 1).$$

donde eliminamos el evento $p_A < 0.5$ por ser un evento seguro. Reescribimos el evento:

$$\underbrace{\mathbb{P} \left(Z < \frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}} \right)}_{\Phi(x)} \geq 0.95 \quad \text{donde } x = \frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}}$$

de tal forma que descubrimos la acumulada de la normal; terminamos de escribir todo:

$$\Phi(x) \geq 0.95$$

donde aplicamos la función inversa de la acumulada de la normal para descubrir:

$$\frac{c - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}(1 - \frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}} \geq \phi^{-1}(0.95) \Rightarrow c = \frac{1}{2} + \phi^{-1}(0.95) \sqrt{\frac{\frac{1}{2}(1 - \frac{1}{2})}{n} \left(\frac{N-1}{N-n} \right)}$$

de donde se sigue que:

$$\hat{p}_A > \frac{1}{2} \left(1 + \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}} \right) \Rightarrow 2\hat{p}_A = 1 + \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}}$$

Notando que los puntos porcentuales de B estimados mediante \hat{p}_B tienen la forma:

$$\hat{p}_B = 1 - \hat{p}_A$$

se tiene entonces que la diferencia entre puntos para determinar quien gana es:

$$\hat{p}_A - \hat{p}_B = 2\hat{p}_A - 1 \geq \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}}$$

El mismo análisis se seguiría bajo la hipótesis de que el perdedor es B ; por tanto se tiene que cumplir que:

$$|\hat{p}_A - \hat{p}_B| \geq \phi^{-1}(0.95) \sqrt{\frac{N-1}{n(N-n)}}$$

para poder declarar como ganador a aquél con más puntos porcentuales de manera correcta con una confianza del 95%.

Ejemplo Resumen: Estimación del total de una población

Consideremos una población de tiburones donde se desconoce el tamaño total de la población N . Algunas veces para determinar el tamaño poblacional se utiliza un modelo de *captura y recaptura*. En él se capturan ℓ individuos los cuales se identifican ([mediante etiquetas](#), por ejemplo) y se devuelven a convivir entre la población de N para mezclarse de vuelta. Una vez mezclados, seleccionamos n nuevos individuos por muestreo aleatorio simple sin reemplazo donde descubrimos que K están marcados. Suponiendo que $K \neq 0$, determinaremos un estimador \hat{N} del total poblacional (en el caso $K = 0$ tuvimos muy mala suerte y seguimos recapturando tiburones hasta encontrar alguno).

En primer lugar notamos que los K marcados que surgen en la segunda muestra siguen una distribución hipergeométrica:

$$\mathbb{P}(K = x) = \frac{\binom{\ell}{x} \binom{N-\ell}{n-x}}{\binom{N}{n}}$$

donde $x \in [\max\{0, \ell - N + n\}, \min\{n, \ell\}] \cap \mathbb{N}$. Para construir el estimador notamos que:

$$\mathbb{E}(K) = n \frac{\ell}{N}$$

de donde podemos despejar N :

$$N = n \frac{\ell}{\mathbb{E}(K)}$$

Ahora bien, dada una muestra donde se obtuvieron K (de n) marcados se propone un estimador de N dado por:

$$\hat{N} = \ell \cdot \frac{n}{K}$$

donde $K = \sum_{i=1}^n x_i$ donde las $x_i = 1$ si estaba marcado y $x_i = 0$ si no lo estaba. La K de hecho depende de la muestra y se puede escribir como:

$$K = \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)$$

Para estimar si \hat{N} es insesgado, habría que calcular su valor esperado condicional en que $K > 0$. Para ello notamos que:

$$\mathbb{E}[\hat{N}|K > 0] = (\ell n) \cdot \mathbb{E}\left[\frac{1}{K}|K > 0\right]$$

Sabemos (por la desigualdad de Jensen) que $\mathbb{E}\left[\frac{1}{K}\right] \neq \frac{1}{\mathbb{E}[K]}$ por lo cual aproximamos el valor esperado mediante una expansión de Taylor; es decir para una función $f \in \mathcal{C}^2$:

$$\mathbb{E}[f(X)] \approx \mathbb{E}[f(\mu) + (X - \mu)f'(\mu) + (X - \mu)^2 f''(\mu)] = f(\mu) + \text{Var}[X] f''(\mu)$$

donde $\mu = \mathbb{E}[X]$. En nuestro caso $f(k) = \frac{1}{k}$ y por tanto:

$$\mathbb{E}\left[\frac{1}{K}|K > 0\right] \approx \frac{1}{\mathbb{E}[K|K > 0]} + 2 \cdot \frac{\text{Var}[K|K > 0]}{(\mathbb{E}[K|K > 0])^3} = \frac{1}{\mu} + 2 \frac{\sigma^2}{\mu^3}$$

Calculamos los valores esperados:

$$\mathbb{E}[K] = \underbrace{\mathbb{E}[K|K=0]\mathbb{P}(K=0)}_{=0} + \mathbb{E}[K|K>0]\mathbb{P}(K>0) \Rightarrow \mathbb{E}[K|K>0] = \frac{\ell n}{N} \frac{1}{\mathbb{P}(K>0)}$$

de donde se sigue que:

$$\mathbb{E}[K|K>0] = \frac{\ell n}{N} \frac{1}{1 - \mathbb{P}(K=0)} = \frac{\ell n}{N} \frac{1}{1 - \frac{\binom{N-\ell}{n}}{\binom{N}{n}}} = \frac{\ell n}{N} \cdot \frac{\binom{N}{n}}{\binom{N}{n} - \binom{N-\ell}{n}} = \mu$$

Por otro lado el cálculo de la varianza:

$$\begin{aligned} \text{Var}[K|K>0] &= \mathbb{E}[K^2|K>0] - \mathbb{E}[K|K>0]^2 \\ &= \frac{\mathbb{E}[K^2]}{\mathbb{P}(K>0)} - \mu^2 \\ &= \frac{\text{Var}[K] + \mathbb{E}[K]^2}{1 - \mathbb{P}(K=0)} - \mu^2 \\ &= \frac{\text{Var}[K] + \left(n \frac{M}{N}\right)^2}{1 - \mathbb{P}(K=0)} - \mu^2 \\ &= \frac{\frac{n\ell}{N} \cdot \frac{(N-\ell)}{N} \cdot \left(\frac{N-n}{N-1}\right) + \left(n \frac{M}{N}\right)^2}{1 - \mathbb{P}(K=0)} - \mu^2 \\ &= \frac{\frac{n\ell}{N} \cdot \frac{(N-\ell)}{N} \cdot \left(\frac{N-n}{N-1}\right) + \left(n \frac{M}{N}\right)^2}{1 - \frac{\binom{N-\ell}{n}}{\binom{N}{n}}} - \mu^2 \\ &= \binom{N}{n} \frac{\frac{n\ell}{N} \cdot \frac{(N-\ell)}{N} \cdot \left(\frac{N-n}{N-1}\right) + \left(n \frac{M}{N}\right)^2}{\binom{N}{n} - \binom{N-\ell}{n}} - \mu^2 = \sigma^2 \end{aligned}$$

Donde se tiene entonces que:

$$\mathbb{E}[\hat{N}|K > 0] \approx (\ell n) \left[\frac{1}{\mathbb{E}[K|K > 0]} + 2 \cdot \frac{\text{Var}[K|K > 0]}{(\mathbb{E}[K|K > 0])^3} \right]$$

con los valores estimados en los renglones anteriores. En particular, \hat{N} no es insesgado pero puede demostrarse que en el límite $\lim_{N \rightarrow \infty} \frac{n \rightarrow \infty}{N \rightarrow \infty}$ lo es.

De manera similar puede obtenerse (ver Lohr capítulo 13):

$$\text{Var}[\hat{N}|K > 0] \approx \left(\frac{n\ell}{K} \right)^2 \frac{\ell - K}{K(\ell - 1)}$$

Misma que puede utilizarse para los intervalos de confianza.

Demostración del Teorema del Límite Central para Muestras Finitas

OJO ESTA SECCIÓN TIENE UN ERROR: CUANDO TERMINE DE CORREGIRLO LES AVISO CONTINUÁ HASTA LA PARTE DE BERNOULLI Éste no es el teorema más general pero es una adaptación suficiente para nuestros propósitos. Para ello el esquema de demostraciones es como sigue:

1. Primero probamos el Teorema del Límite Central bajo [la condición de Lindberg](#)
2. Usamos dicha versión del TLC (más general que el de Proba 2) para demostrar el teorema de límite central para poblaciones finitas la cual es una adaptación del de Hajek basada en la que presenta [Lehmann](#)

Teorema de Límite Central bajo condición de Lindberg

Eventualmente, si encuentro una forma de ponerla en términos de Proba 2 aparecerá por ahora te recomiendo checar [este artículo](#)

Teorema de Límite Central para poblaciones finitas

Introducimos la notación: $\{\Pi_N\}_{N \in \mathbb{N}}$ es una sucesión de poblaciones de tamaño creciente donde $\Pi_1 = \{\nu_{1,1}\}$, $\Pi_2 = \{\nu_{2,1}, \nu_{2,2}\}$ y en general $\Pi_N = \{\nu_{1,N}, \nu_{2,N}, \dots, \nu_{N,N}\}$. Supongamos además para cada Π_N se toma una muestra aleatoria de tamaño $n_N = n(N)$ que depende, de alguna forma, de N . Los n valores que se incluyen en la suma son $x_{N,1}, x_{N,2}, \dots, x_{N,n}$ y su suma es:

$$S_N = \sum_{i=1}^n x_{N,i}$$

Por otro lado el promedio de los N valores de Π_N es:

$$\mu_N = \frac{1}{N} \sum_{j=1}^N \nu_{n,j}$$

Bajo esta notación se tiene el siguiente teorema:

TEOREMA DEL LÍMITE CENTRAL PARA POBLACIONES FINITAS Sea

$$S_N^* = \frac{S_N - \mathbb{E}[S_N]}{\sqrt{\text{Var}(S_N)}}$$

entonces S_N^* converge en distribución a una variable aleatoria Normal(0, 1) si $n, (N - n) \rightarrow \infty$ y se cumple la siguiente condición de Lindberg:

$$\lim_{N \rightarrow \infty} \frac{1}{s_N^2} \sum_{k=1}^N \mathbb{E} \left[(X_k - \mu_k)^2 \cdot \mathbf{1}_{\{|X_k - \mu_k| > \varepsilon s_N\}} \right] = 0$$

con

$$s_N^2 = \sum_{i=1}^N (\nu_{N,i} - \nu_N)^2 \frac{n}{N} \left(1 - \frac{n}{N}\right)$$

y los μ_k dados por:

$$\mu_k = (\nu_{N,k} - \nu_N) \frac{n}{N}$$

Demostración Las S_N no son independientes entre sí pero construiremos una nueva sucesión de varianles aleatorias independientes que tenga el mismo límite y a la que podamos apliar el Teorema del Límite Central. Para ello tomamos una colección $\{U_i\}_{i=1}^N$ de N variables aleatorias independientes con distribución Uniforme(0, 1). Asignamos a cada una de estas U_i su rango; es decir:

$$R(U_i) = j \Leftrightarrow U_{(j)} = U_i$$

(con palabras, U_i es la j -ésima en el momento en el que ordenamos todas las U s). Una muestra de tamaño n la podemos construir como sigue: un elemento $\nu_{i,N}$ se incluye en la muestra si y sólo si $R(U_i) \leq n$. Como las U_i son uniformes e independientes se puede demostrar usando estadísticos de orden que cada una de las $\binom{N}{n}$ muestras es igualmente probable. Podemos entonces escribir cada S_N como:

$$S_N = \sum_{i=1}^N \nu_{N,i} \mathbb{I}_{[0,n]}(R(U_i))$$

Esta suma, sin embargo, está construida con variables *dependientes* pues los rangos $R(U_i)$ son dependientes entre sí. Podemos construir otra suma distinta dada por:

$$T_N = \sum_{i=1}^N (\nu_{N,i} - \mu_N) \mathbb{I}_{[0, \frac{n}{N}]}(U_i) + n\mu_N$$

Donde demostraremos que en el límite T_N toma los mismos valores que S_N . Por ahora pensemos que eso pasa; entonces basta con probar que $T_N \rightarrow \text{Normal}(0, 1)$ cuando $N \rightarrow \infty$. Podemos aplicar el teorema central del límite a las variables aleatorias dadas por:

$$X_{N,i} = (\nu_{N,i} - \nu_N) \mathbb{I}_{[0, \frac{n}{N}]}(U_i)$$

Éstas sí son independientes pues las U_i son independientes y por tanto que las U_i caigan o no en el intervalo $[0, \frac{n}{N}]$ es un evento independiente para cada i . Podemos calcular la media y varianza de las $X_{N,i}$:

$$\mu_{N,i} = (\nu_{N,i} - \nu_N) \frac{n}{N} \quad \text{y} \quad \sigma_{N,i}^2 = (\nu_{N,i} - \nu_N)^2 \frac{n}{N} \left(1 - \frac{n}{N}\right)$$

Pues

$$\mathbb{E}[X_{N,i}] = (\nu_{N,i} - \nu_N) \mathbb{E}[\mathbb{I}_{[0, \frac{n}{N}]}(U_i)] = (\nu_{N,i} - \nu_N) \mathbb{P}(U_i \leq \frac{n}{N}) = (\nu_{N,i} - \nu_N) \frac{n}{N}$$

y por otro lado:

$$\text{Var}[X_{N,i}] = (\nu_{N,i} - \nu_N)^2 \text{Var}[\mathbb{I}_{[0, \frac{n}{N}]}(U_i)] = (\nu_{N,i} - \nu_N)^2 p(1-p) = (\nu_{N,i} - \nu_N)^2 \frac{n}{N} \left(1 - \frac{n}{N}\right)$$

En este caso podemos traducir la condición de Lindberg dada por:

$$\lim_{N \rightarrow \infty} \frac{1}{s_N^2} \sum_{k=1}^N \mathbb{E} \left[(X_k - \mu_k)^2 \cdot \mathbf{1}_{\{|X_k - \mu_k| > \varepsilon s_N\}} \right] = 0$$

$\forall \varepsilon > 0$ con

$$s_N^2 = \sum_{i=1}^N (\nu_{N,i} - \nu_N)^2 \frac{n}{N} \left(1 - \frac{n}{N}\right)$$

y los μ_k de arriba. Lo único que falta por ver es que en el límite T_N toma los mismos valores que S_N . Para ello demostraremos que:

$$\lim_{N \rightarrow \infty} \frac{\mathbb{E}[(T_N - S_N)^2]}{\text{Var}(T_N)} = 0$$

Para ello introducimos la notación:

$$J_i = \mathbb{I}_{[0, \frac{n}{N}]} \left(\frac{R_i}{N} \right) \quad K_i = \mathbb{I}_{[0, \frac{n}{N}]}(U_i) = \mathbb{I}_{[0, \frac{n}{N}]}(U_{(R_i)})$$

donde $R_i = R(U_i)$. Tenemos entonces (demuestra) que:

$$T_N - S_N = \sum_{i=1}^N (\nu_{N,i} - \mu_N) [K_i - J_i]$$

Calcularemos la esperanza condicional de $\mathbb{E}[(T_N - S_N)^2 | U_{(1)}, U_{(2)}, \dots, U_{(N)}]$ y después usaremos la propiedad de torre para obtener $\mathbb{E}[(T_N - S_N)^2]$. Para ello notamos que:

$$\mathbb{E}[K_i - J_i | U_{(1)}, U_{(2)}, \dots, U_{(N)}] = 0$$

pues es independiente de i ; por otro lado, $\sum_{i=1}^N (\nu_{N,i} - \mu_N) = 0$ Luego $\mathbb{E}[(T_N - S_N) | U_{(1)}, U_{(2)}, \dots, U_{(N)}] = 0$ y por tanto:

$$\mathbb{E}[(T_N - S_N)^2 | U_{(1)}, U_{(2)}, \dots, U_{(N)}] = \text{Var}[(T_N - S_N) | U_{(1)}, U_{(2)}, \dots, U_{(N)}]$$

Una cota para la varianza está dada por [Lehmann](#) (ecuación A.49) donde:

$$\text{Var}[(T_N - S_N)^2 | U_{(1)}, U_{(2)}, \dots, U_{(N)}] \leq \frac{1}{N-1} \sum_{i=1}^N (\nu_{N,i} - \mu_N)^2 \sum_{i=1}^N (K_i - J_i)$$

donde $\sum_{i=1}^N (K_i - J_i) = |D - n|$ donde D es la cantidad de U_i de valor menor o igual a n/N . Si usamos la propiedad de torre:

$$\mathbb{E}[(T_N - S_N)^2] \leq \frac{1}{N-1} \sum_{i=1}^N (\nu_{N,i} - \mu_N)^2 \mathbb{E}[|D - n|]$$

Recordando la desigualdad de Jensen:

$$\mathbb{E}[|D - n|]^2 = \mathbb{E}[Y^2] = \text{Var}(Y)$$

Aplicamos dicha desigualdad a $Y = D - n$ donde $D \sim \text{Binomial}(n/N, N)$ (demuestra). Tenemos entonces:

$$\mathbb{E}[|D - n|] \leq \sqrt{N \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

y por tanto:

$$\mathbb{E}[(T_N - S_N)^2] \leq \frac{1}{N-1} \sum_{i=1}^N (\nu_{N,i} - \mu_N)^2 \sqrt{N \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

Finalmente, notamos que:

$$\text{Var}(T_N) = \sum_{i=1}^N (\nu_{N,i} - \mu_n)^2 \frac{n}{N} \left(1 - \frac{n}{N}\right)$$

de donde se sigue que:

$$\frac{\mathbb{E}[(T_N - S_N)^2]}{\text{Var}(T_N)} \leq \frac{N}{N-1} \sqrt{\frac{N}{n(N-n)}}$$

Este último límite es una cuestión de cálculo comprobar que tiende a 0 cuando $N - n$ tiende a infinito pues si tomamos $\alpha \in (0, 1)$ y suponemos sin pérdida de generalidad que $n \geq \alpha N$ entonces:

$$\sqrt{\frac{N}{n(N-n)}} \leq \sqrt{\frac{1/\alpha}{N-n}} \rightarrow 0 \quad \text{cuando} \quad (N-n) \rightarrow \infty$$

Muestreo Aleatorio Simple Bernoulli (BE)

En un esquema de muestreo Bernoulli (BE) se tiene una población de tamaño $N \in \mathbb{N}$ (constante) la cual se enlista de manera ordenada $U = (x_1, x_2, \dots, x_N)^T$. Se recorre la lista de 1 hasta N . Cada elemento de la población, se selecciona y se mide con probabilidad $\pi \in (0, 1)$ para generar una muestra $\mathcal{S} = (x_1, x_2, \dots, x_n)^T$ de tamaño $n = n(\mathcal{S})$ aleatorio (con $0 \leq n(\mathcal{S}) \leq N$).

Un ejemplo de muestreo Bernoulli ocurre en las aduanas del Sistema de Administración Tributaria (SAT) donde con probabilidad π se revisa la mercancía de un viajero (de un total predefinido de N viajeros) para verificar no haya contrabando y con probabilidad $1 - \pi$ se le deja entrar al país sin revisar su mercancía.

Un muestreo Bernoulli no necesariamente tiene muestras del mismo tamaño: como el que cada elemento esté en la muestra depende de π entonces $n(\mathcal{S})$ es una variable aleatoria con distribución Binomial:

$$n(\mathcal{S}) \sim \text{Binomial}(N, \pi)$$

con media y varianza dadas por:

$$\mathbb{E}[n(\mathcal{S})] = N\pi \quad \text{y} \quad \text{Var}[n(\mathcal{S})] = N\pi(1 - \pi)$$

Una forma de muestrear de un muestreo Bernoulli es recorrer uno a uno los elementos de la muestra y generar una variable aleatoria $B_i \sim \text{Bernoulli}(\pi)$ de tal forma que si $B_i = 1$ se incluye el elemento en la muestra. Este esquema está programado en R como sigue:

```
datos <- data.frame(Edad = c(10, 12, 5, 4, 1, 3, 14),
                    Raza = c("Labrador", "Pomeranio", "Labrador",
                             "Pastor Alemán", "Bulldog", "Bulldog", "Chihuahua"))

datos$en_muestra <- 0
proba <- 3/4
for (i in 1:nrow(datos)){
  Bi <- sample(c(0,1), 1, prob = c(1 - proba, proba))
  datos$en_muestra[i] <- Bi
}
muestra <- datos %>% filter(en_muestra == 1)
```

Bajo este esquema se tiene que:

$$\pi_k = \mathbb{P}(x_k \in \mathcal{S}) = \pi \quad \forall k$$

Además en este caso las $\{\mathbb{I}_{\mathcal{S}}(x_k)\}_k$ son independientes y por tanto:

$$\pi_{k,l} = \pi^2$$

En caso de muestreo aleatorio Bernoulli tenemos que un estimador del total es de la misma forma que en el caso de muestreo aleatorio simple:

$$\hat{t}_\pi = \frac{1}{\pi} \sum_{i=1}^{n(\mathcal{S})} x_i$$

El cual es insesgado pues usando indicadoras reescribimos $\hat{t}_\pi = \frac{1}{\pi} \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)$ y tomamos valor esperado:

$$\mathbb{E}[\hat{t}_\pi] = \frac{1}{\pi} \sum_{i=1}^N x_i \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_i)] = \frac{1}{\pi} \sum_{i=1}^N x_i \pi = \sum_{i=1}^N x_i = t$$

por otro lado su varianza está dada por:

$$\text{Var}_{\text{BE}}(\hat{t}_\pi) = \left(\frac{1}{\pi} - 1\right) \sum_{i=1}^N x_i^2$$

la cual puede estimarse de manera insesgada mediante:

$$\widehat{\text{Var}}_{\text{BE}}(\hat{t}_\pi) = \frac{1}{\pi} \left(\frac{1}{\pi} - 1\right) \sum_{i=1}^{n(\mathcal{S})} x_i^2$$

Ejercicio

1. Demuestra la expresión para $\text{Var}_{\text{BE}}(\hat{t}_\pi)$
2. Demuestra que $\widehat{\text{Var}}_{\text{BE}}(\hat{t}_\pi)$ es un estimador insesgado de $\text{Var}_{\text{BE}}(\hat{t}_\pi)$.

Ejemplo

Consideraremos un ejemplo presentado por *Särndal et al.* Un profesor corrige 600 exámenes. Quiere tener un estimado de la calificación de sus alumnos y para ello cada que aparece un examen tira un dado justo de 6 caras y si sale un 6 corrige dicho examen; en caso contrario lo deja pasar. Al final del análisis el profe obtiene una muestra de 90 estudiantes de los cuales 60 pasaron. Asignamos $x_i = 0$ si un alumno no pasó y $x_i = 1$ si pasó; de esta forma la estimación de la cantidad de alumnos que pasaron es un total dado por:

$$\hat{t} = \frac{1}{\pi} \sum_{i=1}^{90} x_k = \frac{1}{\frac{1}{6}} 60 = 360$$

El profe, después de pensarlo un rato se le ocurre otra manera de estimar la proporción de los alumnos que pasaron. Si pasaron 60/90 se tiene entonces que 2/3 de los alumnos pasan; aplicando el 2/3 a los 600 alumnos que tiene un estimador alternativo del total sería:

$$\hat{t}_{\text{Alt}} = \frac{2}{3} \cdot 600 = 400$$

El cual escrito en términos de las variables utilizadas es:

$$\hat{t}_{\text{Alt}} = \begin{cases} \frac{N}{n(\mathcal{S})} \cdot \sum_{i=1}^{n(\mathcal{S})} x_i & \text{si } n(\mathcal{S}) > 0 \\ 0 & \text{si } n(\mathcal{S}) = 0 \end{cases}$$

La pregunta obligada es ¿cuál es un mejor estimador si \hat{t} o bien \hat{t}_{Alt} ?

Un mejor estimador: el proporcional al tamaño

Para decidir si \hat{t}_{Alt} es un mejor estimador que \hat{t} calculemos su valor esperado y su varianza. En ambos casos tenemos dos cosas aleatorias: los elementos que sí quedaron en la muestra (las x_i) y el tamaño de muestra (la n). Para ello utilizamos [la propiedad de torre de la esperanza condicional](#):

$$\begin{aligned}
\mathbb{E}[\hat{t}_{\text{Alt}}] &= \mathbb{E}\left[\mathbb{E}[\hat{t}_{\text{Alt}} | n(\mathcal{S}) = k]\right] \\
&= \sum_{k=0}^N \mathbb{E}\left[\hat{t}_{\text{Alt}} | n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) \\
&= \sum_{k=1}^N \mathbb{E}\left[\frac{N}{n(\mathcal{S})} \cdot \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) \\
&= \sum_{k=1}^N \mathbb{E}\left[\frac{N}{k} \cdot \sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) \\
&= \sum_{k=1}^N \frac{N}{k} \mathbb{E}\left[\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k\right] \binom{N}{k} \pi^k (1-\pi)^{N-k} \\
&= \sum_{k=1}^N \left(\frac{N}{k} \sum_{i=1}^N x_i \mathbb{E}[\mathbb{I}_{\mathcal{S}}(x_i) | n(\mathcal{S}) = k]\right) \binom{N}{k} \pi^k (1-\pi)^{N-k} \\
&= \sum_{k=1}^N \left(\frac{N}{k} \sum_{i=1}^N x_i \frac{k}{N}\right) \binom{N}{k} \pi^k (1-\pi)^{N-k} \\
&= \left(\sum_{i=1}^N x_i\right) \cdot \sum_{k=1}^N \left(\binom{N}{k} \pi^k (1-\pi)^{N-k}\right) \\
&= t \cdot (1 - (1-\pi)^N)
\end{aligned}$$

en este caso el estimador *no* es insesgado y su sesgo es $(1-\pi)^N$. Este sesgo es prácticamente ignorable pues para aplicaciones con N grande $(1-\pi)^N \approx 0$ y no habrá mucha variación en el resultado.

Definición Dado $\hat{\theta}$ estimador de θ definimos el **sesgo** de $\hat{\theta}$ como:

$$\text{Sesgo}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \theta]$$

Podemos calcular la varianza de nuestro estimador; para ello denotamos

$$H(\pi, N) = \sum_{k=1}^N \frac{1}{k} \binom{N}{k} \pi^k (1-\pi)^{N-k} - \frac{(1 - (1-\pi)^N)}{N}$$

Luego:

$$\begin{aligned}
\text{Var}[\hat{t}_{\text{Alt}}] &= \mathbb{E}[\hat{t}_{\text{Alt}}^2] - \mathbb{E}[\hat{t}_{\text{Alt}}]^2 \\
&= \mathbb{E}[\hat{t}_{\text{Alt}}^2] - \left(t \cdot (1 - (1 - \pi)^N)\right)^2 \\
&= \mathbb{E}\left[\mathbb{E}[\hat{t}_{\text{Alt}}^2 | n(\mathcal{S}) = k]\right] - \left(t \cdot (1 - (1 - \pi)^N)\right)^2 \\
&= \sum_{k=1}^N \mathbb{E}[\hat{t}_{\text{Alt}}^2 | n(\mathcal{S}) = k] \cdot \mathbb{P}(n(\mathcal{S}) = k) - \left(t \cdot (1 - (1 - \pi)^N)\right)^2 \\
&= \sum_{k=1}^N \frac{N^2}{k^2} \mathbb{E}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right)^2 \middle| n(\mathcal{S}) = k\right] \cdot \mathbb{P}(n(\mathcal{S}) = k) - \left(t \cdot (1 - (1 - \pi)^N)\right)^2
\end{aligned}$$

Notamos que:

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right)^2 \middle| n(\mathcal{S}) = k\right] \\
&= \text{Var}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right) \middle| n(\mathcal{S}) = k\right] + \mathbb{E}\left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i)\right) \middle| n(\mathcal{S}) = k\right]^2 \\
&= \sum_{i=1}^N x_i^2 \text{Var}\left[\mathbb{I}_{\mathcal{S}}(x_i) \middle| n(\mathcal{S}) = k\right] + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \text{Cov}\left[\mathbb{I}_{\mathcal{S}}(x_i), \mathbb{I}_{\mathcal{S}}(x_j) \middle| n(\mathcal{S}) = k\right] \\
&\quad + \left(\sum_{i=1}^N x_i \mathbb{E}\left[\mathbb{I}_{\mathcal{S}}(x_i) \middle| n(\mathcal{S}) = k\right]\right)^2 \\
&= \sum_{i=1}^N x_i^2 \frac{k}{N} \left(1 - \frac{k}{N}\right) + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \left(\frac{k(k-1)}{N(N-1)} - \frac{k^2}{N^2}\right) + \left(\frac{k}{N} \sum_{i=1}^N x_i\right)^2 \\
&= \frac{k}{N} \left[\sum_{i=1}^N x_i^2 \left(1 - \frac{k}{N}\right) + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \left(\frac{k-1}{N-1} - \frac{k}{N}\right)\right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N} \left[\sum_{i=1}^N x_i^2 \left(\frac{N-k}{N}\right) - \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j \left(\frac{N-k}{N(N-1)}\right)\right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N} (N-k) \left[\frac{1}{N} \sum_{i=1}^N x_i^2 - \frac{1}{N} \frac{1}{N-1} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N x_i x_j\right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N} (N-k) \frac{1}{N-1} \sum_{i=1}^N \left[\frac{N-1}{N} x_i^2 - \frac{1}{N} x_i \sum_{\substack{j=1 \\ j \neq i}}^N x_j\right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N} (N-k) \frac{1}{N-1} \sum_{i=1}^N \left[x_i^2 - \frac{1}{N} x_i \sum_{j=1}^N x_j\right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N} (N-k) \frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i\right) \left(\sum_{j=1}^N x_j\right)\right] + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= \frac{k}{N} (N-k) \frac{1}{N-1} \left[\sum_{i=1}^N x_i^2 - \frac{1}{N} \left(\sum_{i=1}^N x_i\right)^2\right] + k^2 \bar{x}_{\mathcal{U}}^2
\end{aligned}$$

$$\begin{aligned}
&= \frac{k}{N}(N-k) \frac{1}{N-1} \sum_{i=1}^N \left(x_i - \frac{1}{N} \sum_{j=1}^N x_j \right)^2 + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= k \frac{(N-k)}{N} \frac{1}{N-1} \sum_{i=1}^N \left(x_i - \bar{x}_{\mathcal{U}} \right)^2 + k^2 \bar{x}_{\mathcal{U}}^2 \\
&= k \frac{(N-k)}{N} s_{\mathcal{U}}^2 + k^2 \bar{x}_{\mathcal{U}}^2
\end{aligned}$$

por lo cual si sustituimos en la ecuación anterior: %

$$\begin{aligned}
\text{Var}[\hat{t}_{\text{Alt}}] &= \sum_{k=1}^N \frac{N^2}{k^2} \mathbb{E} \left[\left(\sum_{i=1}^N x_i \mathbb{I}_{\mathcal{S}}(x_i) \right)^2 \middle| n(\mathcal{S}) = k \right] \cdot \mathbb{P}(n(\mathcal{S}) = k) - \left(t \cdot (1 - (1 - \pi)^N) \right)^2 \\
&= \sum_{k=1}^N \frac{N^2}{k^2} \left[k \frac{(N-k)}{N} s_{\mathcal{U}}^2 + k^2 \bar{x}_{\mathcal{U}}^2 \right] \cdot \binom{N}{k} \pi^k (1 - \pi)^{N-k} - \left(t \cdot (1 - (1 - \pi)^N) \right)^2 \\
&= N^2 \sum_{k=1}^N \left[\frac{(N-k)}{kN} s_{\mathcal{U}}^2 + \bar{x}_{\mathcal{U}}^2 \right] \cdot \binom{N}{k} \pi^k (1 - \pi)^{N-k} - t^2 \cdot (1 - (1 - \pi)^N)^2 \\
&= N^2 s_{\mathcal{U}}^2 \sum_{k=1}^N \left(\frac{1}{k} - \frac{1}{N} \right) \binom{N}{k} \pi^k (1 - \pi)^{N-k} + N^2 \bar{x}_{\mathcal{U}}^2 \sum_{k=1}^N \binom{N}{k} \pi^k (1 - \pi)^{N-k} \\
&\quad - N^2 \bar{x}_{\mathcal{U}}^2 (1 - (1 - \pi)^N) \\
&= N^2 s_{\mathcal{U}}^2 \sum_{k=1}^N \frac{1}{k} \binom{N}{k} \pi^k (1 - \pi)^{N-k} - \frac{1}{N} (1 - (1 - p)^N) + \\
&\quad (1 - (1 - p)^N) N^2 \bar{x}_{\mathcal{U}}^2 (1 - (1 - (1 - p)^N)) \\
&= N^2 [H(N, \pi) s_{\mathcal{U}}^2 + (1 - p)^N (1 - (1 - p)^N) \bar{x}_{\mathcal{U}}^2]
\end{aligned}$$

En nuestro caso para elegir el mejor estimador entre \hat{t} y \hat{t}_{alt} se calculan las varianzas de ambos. Una posible elección es aquél que tiene menos varianza (podría estar más cercano al valor dado que el sesgo de \hat{t}_{alt} es pequeñísimo⁴). Se puede demostrar (ver Särndal) que en general

$$\text{Var}[\hat{t}_{\text{Alt}}] \ll \text{Var}[\hat{t}]$$

Y usualmente se prefiere el estimador \hat{t}_{Alt} .

Ejemplo Resumen: Aduana

Se sabe que de manera diaria fluyen por un punto de la aduana 1000 cargamentos. Cada cargamento que entra debe ser analizado para buscar contrabando con probabilidad p (y con probabilidad $(1 - p)$ se deja pasar sin mayor análisis). Determina la probabilidad p si se desea estimar el total de cargamentos con contrabando que pasan por la aduana y, a la vez, se busca que el 75% de las ocasiones no se analicen más de 200 cargamentos.

Para encontrar la probabilidad p (correspondiente al π) recordamos que el tamaño de la muestra n tiene una distribución Binomial:

$$n(\mathcal{S}) \sim \text{Binomial}(1000, p)$$

Buscamos entonces una p tal que

$$\mathbb{P}(B \leq 200) = 0.75 \quad \text{donde } B \sim \text{Binomial}(1000, p)$$

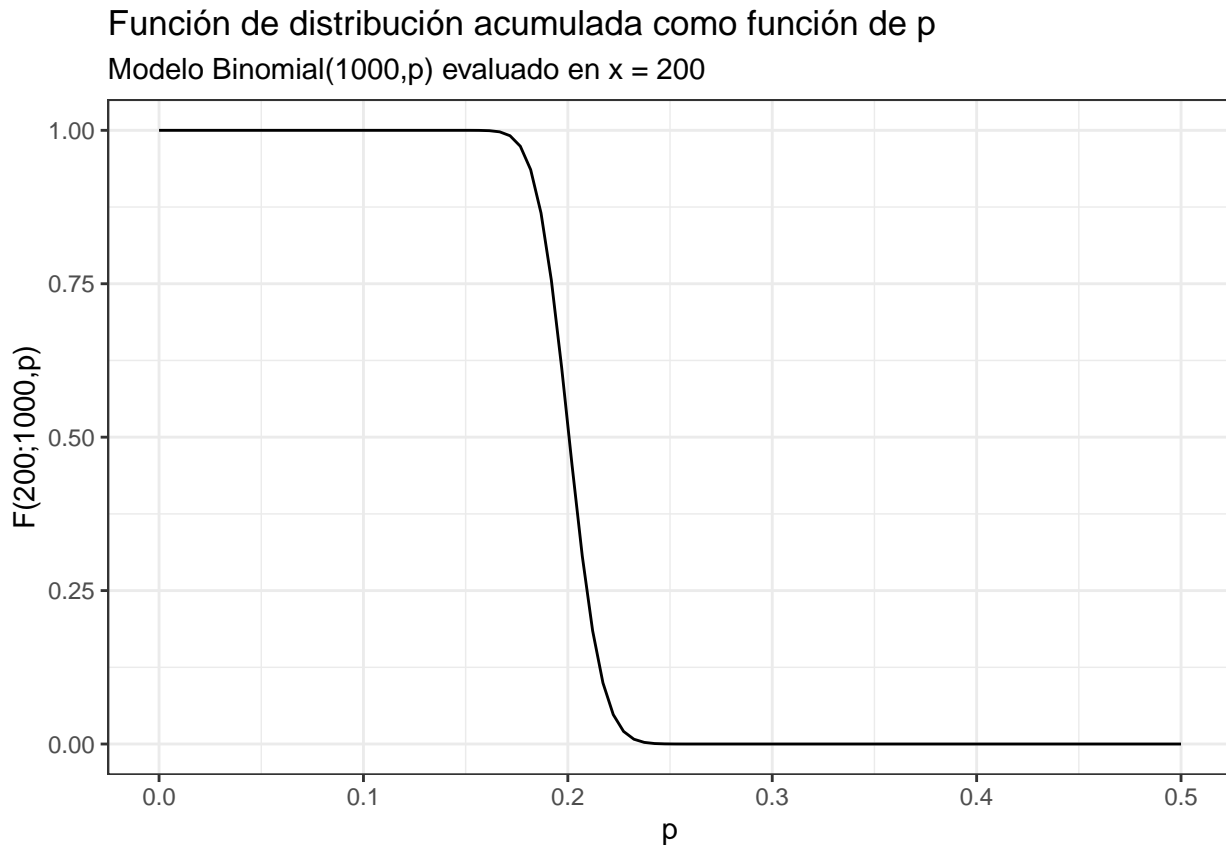
⁴Calcúlalo.

En particular, notamos que del lado izquierdo tenemos a la función de distribución acumulada $F_B(200) = \mathbb{P}(B \leq 200)$ la cual depende (de manera implícita) de p . ¡Hagamos explícita la dependencia de los parámetros p y N :

$$F_B(200; 1000, p) = 0.75 \quad \text{donde } B \sim \text{Binomial}(1000, p)$$

Podemos graficar la función de distribución acumulada como función de p :

```
p.val <- seq(0, 0.5, length.out = 100)
ggplot() +
  geom_line(aes(x = p.val, y = pbinom(200, 1000, p.val))) +
  theme_bw() +
  labs(
    x = "p",
    y = "F(200;1000,p)",
    title = "Función de distribución acumulada como función de p",
    subtitle = "Modelo Binomial(1000,p) evaluado en x = 200"
  )
```



Notamos entonces que lo que necesitamos es hallar la p donde la función de distribución acumulada (como función de p) toca al 0.75. Para ello, como no podemos despejar, utilizamos un método numérico a través de `uniroot` para encontrar el 0 de la función $g(p) = F_B(200; 1000, p) - 0.75$ (pues la p^* tal que $g(p^*) = 0$ es la respuesta):

```
g.fun <- function(p){pbinom(200, 1000, p) - 0.75}
raiz <- uniroot(g.fun, lower = 0, upper = 0.5)
print(paste0("El valor de p es ", raiz$root))
```

```
## [1] "El valor de p es 0.192159774829166"
```

De donde obtenemos el p necesario.

Muestreo Aleatorio Simple con Reemplazo (MAS/cR)

El muestreo aleatorio simple con reemplazo es idéntico al muestreo aleatorio sin reemplazo *pero* en este caso no se extrae un elemento de la muestra sino que se permite que se seleccione múltiples veces. En cada selección hay una probabilidad $1/N$ de que un individuo de la población sea seleccionado. Cada selección es independiente de la pasada. Aquí consideraremos un universo de tamaño constante $N \in \mathbb{R}$ dado por $U = (x_1, x_2, \dots, x_N)^T$ y las variables aleatorias N_k que denotan la cantidad de veces que x_k fue seleccionado para incluirse en la muestra⁵. El orden en el que fueron seleccionados los elementos no importa.

En el caso de muestreo aleatorio simple con reemplazo se fija un tamaño de muestra m y hay por tanto N^m muestras posibles. Cada una de las muestras sigue la siguiente función de probabilidad uniforme:

$$\mathbb{P}(S = S) = \begin{cases} \frac{1}{N^m} & \text{si } \#S = m \\ 0 & \text{en otro caso} \end{cases}$$

Dado un elemento x_k la probabilidad de que dicho x_k aparezca r veces en la muestra de tamaño m está dada por:

$$\binom{m}{r} \left(\frac{1}{N}\right)^r \left(1 - \frac{1}{N}\right)^{m-r}$$

En particular se tiene que la probabilidad de que x_k no esté en la muestra es:

$$\left(1 - \frac{1}{N}\right)^m$$

o bien de que esté en la muestra:

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m$$

lo cual se calcula por el complemento. Por otro lado, la probabilidad conjunta $\pi_{k,l}$ de que x_k y x_l estén en la muestra se puede computar usando inclusión exclusión:

$$\pi_{k,l} = 1 - \underbrace{\left(1 - \frac{1}{N}\right)^m}_{\text{No está } x_k} - \underbrace{\left(1 - \frac{1}{N}\right)^m}_{\text{No está } x_l} + \underbrace{\left(1 - \frac{2}{N}\right)^m}_{\text{No está ni } x_k \text{ ni } x_l}$$

En R puedes obtener un muestreo aleatorio simple con reemplazo cambiando en `sample` el `replace`:

```
sample(c("A", "B", "C"), 10, replace = TRUE)
```

```
## [1] "B" "C" "B" "C" "B" "C" "A" "C" "B" "A"
```

Una observación es bastante relevante aquí:

$$\begin{aligned} \pi_k &= 1 - \left(1 - \frac{1}{N}\right)^m = 1 - \sum_{j=0}^m \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\ &= 1 - \left[\sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} - \frac{m}{N} + 1 \right] \\ &= \frac{m}{N} - \sum_{j=0}^{m-2} \binom{m}{j} \left(-\frac{1}{N}\right)^{m-j} \\ &= \frac{m}{N} + \mathcal{O}\left(\frac{m^2}{N^2}\right) \end{aligned}$$

⁵Observa que las variables aleatorias N_k generalizan a las variables indicadoras.

donde $\mathcal{O}\left(\frac{m^2}{N^2}\right)$ es notación que implica que una función $f(n)$ es de orden $g(n)$ (denotado $f(n) = \mathcal{O}(g(n))$) si y sólo si existe M tal que para cualquier $n \in \mathbb{N}$ tal que $|f(n)|/g(n) \leq M$. Escrito con palabras en este caso esto significa que si m/N es pequeño entonces $\frac{m^2}{N^2}$ es caso 0 y entonces muestrear con reemplazo es casi lo mismo que muestrear sin reemplazo (lo cual tiene sentido: si tu población es muy grande $N \gg 0$ entonces está bien difícil que vuelvas a capturar a uno en tu encuesta y por tanto es casi lo mismo muestrear **con** que **sin** reemplazo en términos prácticos).

Ejemplo Resumen: Captura-Recaptura con reemplazo

Se realiza un estudio para determinar la cantidad de ratas en la CDMX. Para ello se pone una trampa en algún lugar aleatorio de la ciudad. Si se atrapa una rata se le marca y se le deja ir. Si para 50 ratas capturadas, contamos 42 marcadas determina el número de ratas en la isla suponiendo que las 50 fueron con reemplazo.

Para ello denotamos $p_N(r)$ a la probabilidad de tener r ratas distintas en m intentos con reemplazos ($m = 50$ es determinado por nosotros y no es aleatorio) en una población de tamaño desconocido N . Una vez que se fijan las r ratas que van a salir hay $\binom{N}{r}$ formas de elegir las. Entonces:

$$p_N(r) = \frac{N!}{r!(N-r)!} q_N(r)$$

donde $q_N(r)$ es la probabilidad de obtener r distintas ratas en m intentos con reemplazo. Fijado el número de ratas, el universo Ω de posibilidades se forma por el grupo de mapeos de $\{1, 2, \dots, m\} \rightarrow \{1, 2, \dots, N\}$ (todas las formas de haber acomodado las ratas). Tenemos $m \geq r$ y de hecho:

$$q_N(r) = \sum_{\omega \in \text{Fav}} p(\omega)$$

dpmde $p(\omega)$ es la probabilidad de obtener un mapeo favorable ($w \in \text{Fav}$). Tenemos que $p(\omega) = N^{-m}$ para toda ω . La cantidad de casos favorables es lo mismo que preguntarse por la cantidad de mapeos suprayectivos del conjunto $\{1, 2, \dots, m\}$ en $\{1, 2, \dots, r\}$ lo cual está dado por $r!$ multiplicado por el número de Stirling de segundo tipo $\mathfrak{s}_m^{(r)}$:

$$\mathfrak{s}_m^{(r)} = \frac{1}{r!} \sum_{i=1}^r \binom{r}{i} i^m (-1)^{r-i}$$

donde $\mathfrak{s}_m^{(r)}$ es la forma de encontrar de un grupo de m elementos r partes no vacías. Tenemos entonces:

$$p_N(r) = \frac{N!}{(N-r)!N^m} \mathfrak{s}_m^{(r)} \quad \text{para } r = 1, 2, \dots, \min\{m, N\}.$$

Lo que vamos a hacer es pensar que la N que generó los datos es la N máxima (*criterio de máxima verosimilitud*) y entonces lo que hay que maximizar es:

$$\frac{N!}{(N-r)!N^m} = \frac{\prod_{i=0}^{r-1} (N-i)}{N^m}$$

Hay dos formas de hacer esta maximización: enlistando todas las N y r para mi población o bien derivando el logaritmo:

$$\frac{d}{dN} \ln \left(\frac{\prod_{i=0}^{r-1} (N-i)}{N^m} \right) = \frac{d}{dN} \left[\sum_{i=0}^{r-1} \ln(N-i) - m \ln(N) \right] = \sum_{i=0}^{r-1} \frac{1}{N-i} - \frac{m}{N} = 0$$

de donde se sigue que:

$$\sum_{i=0}^{r-1} \frac{N}{N-i} = m$$

La cual es una ecuación no lineal que se puede resolver mediante `uniroot` como la pasada. ¡Inténtalo! Llegarás a que $m = 50$ y $r = 42$.

Muestreo Aleatorio Simple Ponderado (MAS/P)

En el caso más general posible cada uno de los elementos x_k tiene una probabilidad π_k de aparecer en la muestra. Análogamente se tienen probabilidades conjuntas de la forma $\pi_{k,l}$ donde no necesariamente hay independencia. El estimador del total está dado por el estimador Horvitz Thompson:

$$\hat{t} = \sum_{k=1}^n \frac{x_k}{\pi_k}$$

donde su varianza y sus estimadores fueron ya calculados desde muestreo aleatorio simple. En el siguiente capítulo comenzaremos a variar mucho más las π_k cuando entremos a muestreo estratificado. ¡Nos vemos pronto!

Ejercicios

1. Bajo muestreo aleatorio Bernoulli se propone el siguiente estimador para el total:

$$\hat{t}_{\text{BE}} = \frac{N}{\mathbb{E}[n(\mathcal{S})]} \sum_{i=1}^{n(\mathcal{S})} x_i$$

- a. Demuestra que es insesgado
- b. Obtén su varianza
- c. Para el ejemplo del profesor con los exámenes ¿es \hat{t}_{BE} una mejor opción que \hat{t}_{Alt} ? Justifica calculando en R todos los posibles casos (desde que 0 pasan hasta que los 600 pasan el examen) y analizando cuántas de esas veces la varianza de \hat{t}_{Alt} es menor que la de \hat{t}_{BE} .