Muestreo Aleatorio Estratificado

Rodrigo Zepeda (rodrigo.zepeda@itam.mx)

13 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)
```

Muestreo Aleatorio Estratificado

Vamos a considerar una población $\mathcal{U}$ la cual suponemos podemos particionar en una cantidad finita (no vacía) de subpoblaciones: $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_H$ de tamaños $N_1, N_2, \ldots, N_H$ con $\bigcup_{h=1}^{H} \mathcal{U}_h = \mathcal{U}$ (lo que se traduce en $\sum_{h=1}^{H} N_h = N$). Lo que busca el muestreo aleatorio estratificado es *estimar* en cada uno de los estratos así como de manera global. Por ejemplo, puede interesarnos conocer la estatura en hombres y mujeres, las ganancias en empresas agrícolas, ganaderas, de servicios y de transformación, la cantidad de enfermos que hay en cada entidad federativa, etc. En cada uno de estos casos estamos hablando de estratos de en los cuales interesa realizar la estimación. Un punto importante aquí es que *los estratos son conocidos y decididos por la investigadora*. La extracción en cada uno de los estratos se realiza de manera independiente obteniéndose muestras $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_H$ de tamaños $n_1, n_2, \ldots, n_H$ para cada uno de ellos.

> **Notación** Si $x_i \in \mathcal{U}_h$ lo denotaremos como $x_{i,h}$ para de esta manera distinguir el x_i que está en $\mathcal{U}_h$ del que está en $\mathcal{U}_k$

La muestra total es:

$$\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_H)^T$$

un vector de tamaño $n = \sum_{h=1}^{H} n_h$. Por independencia, se tiene:

$$\mathbb{P}(\mathcal{S} = S) = \mathbb{P}_1(\mathcal{S}_1 = S_1) \cdot \mathbb{P}_2(\mathcal{S}_2 = S_2) \cdots \mathbb{P}_H(\mathcal{S}_H = S_H)$$

donde cada $\mathbb{P}_h$ es un esquema muestral para el estrato h. En el caso del muestreo aleatorio estratificado tenemos un estimador de la media dada por la media ponderada de las medias:

$$\bar{x}_{\mathcal{S}} = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \bar{x}_{\mathcal{S}_h}$$

donde por comodidad denotaremos

$$\bar{x}_h = \bar{x}_{\mathcal{S}_h}$$

En particular, por la independencia se tiene:

$$\mathrm{Var}\big(\bar{x}_{\mathcal{S}}\big) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \cdot \mathrm{Var}\big(\bar{x}_h\big)$$

donde un estimador de la varianza es:

$$\widehat{\mathrm{Var}}\big(\bar{x}_{\mathcal{S}}\big) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \cdot \widehat{\mathrm{Var}}\big(\bar{x}_h\big)$$

En particular tenemos que para muestreo aleatorio simple tenemos un estimador insesgado:

$$\mathbb{E}\Big[\bar{x}_{\mathcal{S}}\Big] = \sum_{h=1}^{H} \frac{N_h}{N} \mathbb{E}\big[\bar{x}_h\big] = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \frac{1}{N_h} \sum_{i=1}^{N_h} x_{i,h} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{N} x_{i,h} = \bar{x}_{\mathcal{U}}$$

Su varianza es:

$$\mathrm{Var}\big(\bar{x}_{\mathcal{S}}\big) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s^2_{\mathcal{U}_h}$$

donde $f_h = n_h/N_h$ es la fracción de muestreo del estrato h y:

$$s^2_{\mathcal{U}_h} = \frac{1}{N_h - 1} \sum_{\mathcal{U}_h} (x_k - \bar{x}_{\mathcal{U}_h})^2$$

es la varianza del estrato con

$$\bar{x}_{\mathcal{U}_h} = \sum_{\mathcal{U}_\langle} x_k$$

siendo la media del mismo. El estimador insesgado de la varianza en este caso es:

$$\widehat{\mathrm{Var}}\big(\bar{x}_{\mathcal{S}}\big) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s^2_{\mathcal{S}_h}$$

donde

$$s^2_{\mathcal{S}_h} = \frac{1}{n_h - 1} \sum_{\mathcal{S}_h} (x_k - \bar{x}_{\mathcal{S}_h})^2$$

es la varianza muestral ajustada.

Ejemplo

Dada la población $\mathcal{U} = \{x_1, x_2, x_3, x_4\}$ con $x_1 = x_2 = 0$, $x_3 = 1$, $x_4 = -1$:

1. Calcula la varianza del estimador de la media de un muestreo aleatorio simple sin reemplazo de tamaño 2
2. Calcula la varianza del estimador de la media de un muestreo estratificado de donde se selecciona una unidad por cada estrato y los estratos están dados por $U_1 = \{x_1, x_2\}$ y $U_2 = \{x_3, x_4\}$.

Solución Tenemos que:

$$\bar{x} = 0$$

mientras que por otro lado,

$$s^2_{\mathcal{S}} = \frac{1}{4-1}\big(1^2 + (-1)^2\big) = \frac{2}{3}$$

Finalmente:

$$\mathrm{Var}\big(\bar{x}\big) = \frac{N-n}{N}\frac{s^2_{\mathcal{S}}}{n} = \frac{1}{6}$$

Por otro lado para resolver **2**:

$$\bar{x}_1 = \bar{x}_2 = 0$$

Además de que:

$$s^2_{\mathcal{S}_1} = 0$$

y:

$$s^2_{\mathcal{S}_2} = 2$$

Tenemos entonces que:

$$\mathrm{Var}(\bar{x}) = \frac{N-n}{nN}\sum_{h=1}^{2}\frac{N_h}{N}s^2_{\mathcal{S}_h} = \frac{1}{4}$$

Notamos que la varianza del muestreo estratificado es mayor a la varianza del muestreo simple. Por lo que concluimos que **estratificar no necesariamente reduce la varianza**.

Ejercicio de clase

De entre 7500 empleados de una compañía deseamos conocer la proporción P que tiene un vehículo por lo menos. Se construyeron 3 estratos para la población según el ingreso (bajo, medio, alto). Se tiene N_h el total del estrato, n_h el total muestreado y p_h el estimador del total de vehículos para el estrato h según la muestra. Determina un estimador $\hat{p}$ y su intervalo de confianza.

	Bajo	**Medio**	**Alto**
N_h	3500	2000	2000
n_h	500	300	200
p_h	0.13	0.45	0.50

Table 1: Tabla de datos de trabajadores

Alocación

Una pregunta importante para el caso de muestreo estratificado es el cálculo de la(s) n. En este caso ¿cómo determinar cuánto muestrear de cada población? Veamos un ejemplo:

> Supongamos que se desean muestrear hombres y mujeres en México. En este país el 48% de los habitantes son hombres y el 52% son mujeres.

Una opción en este caso podría ser tomar una muestra que refleje exactamente esas proporciones. Ésta se conoce como *proporcional al tamaño*.

Alocación proporcional al tamaño

Dada una población de tamaño N con H estratos de tamaños $N_1, N_2, \ldots, N_H$ para $h = 1, 2, \ldots, H$ la alocación proporcional consiste en tomar n_h como:

$$n_h = n \cdot \frac{N_h}{N}$$

Ésta forma de asignar variables no necesariamente es la mejor (mucho menos para estudios con costo) por lo cual se tienen otras alocaciones.

Un alocación proporcional al tamaño representa usualmente una ganancia en la precisión (ver último ejemplo, el de los doctores)

Alocación óptima

Si consideramos muestreo aleatorio simple sin reemplazo y analizamos su varianza podemos reescribirla:

$$V = \mathrm{Var}(\bar{x}_{\mathcal{S}}) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} s_{\mathcal{U}_h}^2 = \sum_{h=1}^{H} \frac{A_h}{n_h} + B$$

donde

$$A_h = N_h^2 s_{\mathcal{U}_h}^2$$

y

$$B = -\sum_{h=1}^{H} N_h s_{\mathcal{U}_h}^2$$

(Ésta no es la única que se puede escribir de esta forma, también la de la Bernoulli, por ejemplo). Supongamos, además que asociado a muestrear cada estrato h hay un costo diferenciado c_h para cada elemento muestreado de h. El costo total sería:

$$C = c_0 + \sum_{h=1}^{H} n_h c_h$$

El problema de alocación de muestras es determinar las n_h que minimizan las varianzas V sujetas a los costos C (o puede verse de igual forma como hallar aquellas n_h que dadas varianzas predefinidas V minimizan los costos C).

Teorema Bajo un muestreo aleatorio estratificado donde V puede escribirse como:

$$V = \sum_{h=1}^{H} \frac{A_h}{n_h} + B$$

y con una función de costo lineal $C = c_0 + \sum_{h=1}^{H} n_h c_h$ la muestra óptima se alcanza tomando n_h proporcional a $(A_h/c_h)^{1/2}$.

Demostración Sea $V^* = V - B$ y $C^* = C - c_0$. El problema de optimización se puede reescribir como minimizar el producto:

$$V^* C^* = \Big(\sum_{h=1}^{H} \frac{A_h}{n_h} \Big) \cdot \Big(\sum_{h=1}^{H} n_h c_h \Big)$$

Utilizamos la desigualdad de Cauchy

$$\big(\sum_h a_h^2 \big) \big(\sum_h b_h^2 \big) \geq \big(\sum_h a_h b_h \big)^2$$

con $a_h = (A_h/n_h)^2$ y $b_h = (n_h c_h)^{1/2}$ tenemos:

$$V^* C^* \geq \Big[\sum_{h=1}^{H} (A_h c_h)^{1/2}\Big]^2$$

Recordamos que la igualdad en el caso de Cauchy se mantiene cuando b_h/a_h es constante para toda h:

$$\Big(\frac{n_h c_h}{A_h/n_h}\Big)^{1/2} = \text{Constante}$$

De donde se sigue el resultado que $n_h \propto (A_h/c_h)^{1/2}$

Nota Minimizar la varianza para un costo fijo C nos lleva a :

$$n_h = \frac{(C - c_0)(A_h/c_h)^{1/2}}{\sum_{h=1}^{H} (A_h c_h)^{1/2}}$$

en particular para muestreo aleatorio simple sin reemplazo puede demostrarse:

$$n_h = \frac{(C - c_0) N_h s_{\mathcal{S}_h}/c_h^{1/2}}{\sum_{h=1}^{H} N_h s_{\mathcal{S}_h} c_h^{1/2}}$$

Por otro lado, minimizar el costo para una varianza fija V nos lleva a:

$$n_h = \Big(\frac{A_h}{c_h}\Big)^{1/2} \Big[\sum_{h=1}^{H} (A_h c_h)^{1/2}\Big]/(V - B)$$

Nota 2 Cuando se toman todas las c_h idénticas y constantes se le conoce como alocación de Neymann o sólo alocación óptima.

Ejemplo

Se quiere estimar las ventas promedio de una población de empresas. Las empresas se enlistan según tres clases: según sus ventas en la siguiente tabla:

Ventas en millones	**Cantidad de negocios**
0 a 1	1000
1 a 10	100
Más de 10	10

Table 2: Tabla de datos de empresas

Se sabe que se quieren estimar 111 empresas. Se supone, además que dentro de cada clase la distribución de ventas es uniforme. Obtén las varianzas de los estimadores cuando se toma alocación proporcional y cuando se toma óptima con costos constantes (Neyman).

Solución Como la distribución intra-clase es uniforme podemos completar la tabla recordando que la varianza de una variable uniforme es:

$$\frac{(b-a)^2}{12}$$

de donde obtenemos la tabla actualizada:

de donde se sigue que (para convertir a $1/N - 1$ de $1/N$):

$$s^2_{h_1} = \frac{1}{12} \cdot \frac{1000}{999} \approx 0.0834168$$
$$s^2_{h_2} = \frac{81}{12} \cdot \frac{100}{99} \approx 0.81818$$
$$s^2_{h_3} = \frac{8100}{12} \cdot \frac{10}{9} \approx 750$$

Ventas en millones	Cantidad de negocios	Varianza
0 a 1	1000	1/12
1 a 10	100	81/12
Más de 10	10	8100/12

Table 3: ACTUALIZACIÓN Tabla de datos de empresas

Luego:

Estratificación proporcional al tamaño

$$\text{Var}(\bar{x}) = \frac{N-n}{nN}\sum_{h=1}^{3}\frac{N_h}{N}s_h^2 \approx 0.0604$$

Estratificación óptima Por un lado tenemos que:

$$N_1 s_1^2 = 288.82$$
$$N_2 s_2^2 = 261.116$$
$$N_3 s_3^2 = 273.861$$

Lo que nos da las alocaciones óptimas:

$$n_1 = \frac{nN_1 s_1}{\sum_{h=1}^{3} N_h s_h} = 38.9161$$
$$n_2 = \frac{nN_2 s_2}{\sum_{h=1}^{3} N_h s_h} = 35.18$$
$$n_3 = \frac{nN_3 s_3}{\sum_{h=1}^{3} N_h s_h} = 36.90$$

En el caso del tercer estrato $n_3 > N_3$ por lo que seleccionamos $n_3 = 10$. En este caso es necesario redistribuir de manera óptima los restantes 101 entre los estratos 1 y 2 por lo que recalculamos las n:

$$n_1 = \frac{101N_1 s_1}{N_1 s_1 + N_2 s_2} = 53.0439$$
$$n_2 = \frac{101N_2 s_2}{N_1 s_1 + N_2 s_2} = 47.9561$$

La distribución óptima entonces es $n_1 = 53, n_2 = 48, n_3 = 10$. Finalmente la varianza está dada por:

$$\text{Var}(\bar{x}_{\mathcal{S}}) = \sum_{h=1}^{3}\frac{N_h^2}{N^2}\frac{1-f_h}{n_h}s_h^2 = 0.0018$$

Ejercicio de clase:

En una ciudad grande se estudia el número promedio de pacientes que ven los médicos en su día laboral. Comenzamos con algunas hipótesis *a priori*: entre más experiencia tiene un médico más pacientes ve. Esto nos lleva a clasificar a la población de médicos en tres grupos: `recién graduados` (grupo 1), `intermedios` (grupo 2) y `experimentados` (grupo 3). Tenemos una lista de 500 doctores en el grupo 1, 1000 en el grupo 2 y 2500 en el grupo 3. Seleccionamos mediante muestreo aleatorio simple sin reemplazo 200 doctores por cada clase y calculamos el número de pacientes por día y por doctor: 10 para el grupo 1, 15 para el grupo 2 y 20 para el grupo 3. Finalmente calculamos las varianzas del número de pacientes por doctor en cada una de las siguientes muestras y encontramos respectivamente que son 4 (grupo 1), 7 (grupo 2) y 10 (grupo 3).

1. Estima la media del número de pacientes que ve un doctor por día y obtén un intervalo de confianza.

2. Si al año siguiente se volviera a repetir el mismo análisis con 600 médicos (de nuevo) una hipótesis usual es que las varianzas no cambian. Determina la alocación de Neyman y la proporcional en este caso.
3. Determina la ganancia en precisión de hacer alocación proporcional al tamaño por encima de hacer muestreo aleatorio simple

Solución 1. Consideramos el estimador de la media:

$$\bar{x}_{\mathcal{S}} = \sum_{h=1}^{3} \frac{N_h}{N} \bar{x}_h = 17.5$$

Por otro lado, su varianza está estimada por:

$$\widehat{\mathrm{Var}}(\bar{x}_{\mathcal{S}}) = \sum_{h=1}^{3} \Big(\frac{N_h}{N}\Big)^2 \Big(1 - \frac{n_h}{N_h}\Big) \frac{s_h^2}{n_h} \approx 0.0199$$

De donde tenemos el intervalo de confianza:

$$\bar{Y} \pm 1.95\sqrt{\widehat{\mathrm{Var}}} \Rightarrow \mathrm{IC}_{95\%} = [17.5 - 0.28, 17.5 + 0.28]$$

2. Una alocación proporcional al tamaño nos lleva a que $n_h = N_h/N$ en este caso, $n_1 = 75$, $n_2 = 150$, $n_3 = 375$. Por otro lado, si utilizamos para la de Neyman las varianzas del actual y suponemos serán similares el próximoa ño podemos estimar: $N_1 s_1 = 1000$, $N_2 s_2 = 2646$ y $N_3 s_3 = 7906$. En este caso, $n_1 = 52$, $n_2 = 137$ y $n_3 = 411$.

3. A partir de la fórmula de descomposición de la varianza (demuestra) podemos aproximar la varianza poblacional a partir de las de la muestra:

$$s_{\mathcal{U}}^2 \approx \sum_{h=1}^{3} \frac{N_h}{N} s_{\mathcal{U}_h}^2 + \sum_{h=1}^{3} \frac{N_h}{N} (\bar{x}_{\mathcal{U}_h} - \bar{x}_{\mathcal{U}})^2$$

Sabemos que $\mathbb{E}[s_{\mathcal{S}_h}^2] = s_{\mathcal{U}_h}^2$ (el estimador es insesgado en cada estrato). Nos interesa el valor esperado de:

$$A = \sum_{h=1}^{3} \frac{N_h}{N} (\bar{x}_{\mathcal{S}_h} - \bar{x}_{\mathcal{S}})^2 = \sum_{h=1}^{3} \frac{N_h}{N} \bar{x}_{\mathcal{S}_h}^2 - \bar{x}_{\mathcal{S}}^2$$

Luego:

$$\begin{aligned}
\mathbb{E}[A] &= \sum_{h=1}^{3} \frac{N_h}{N} \mathbb{E}[\bar{x}_{\mathcal{S}_h}^2] - \mathbb{E}[\bar{x}_{\mathcal{S}}^2] \\
&= \sum_{h=1}^{3} \frac{N_h}{N} \Big(\mathrm{Var}(\bar{x}_{\mathcal{S}_h}^2) + \bar{x}_{\mathcal{S}_h}^2\Big) - \Big(\mathrm{Var}[\bar{x}_{\mathcal{S}}^2] + \bar{x}_{\mathcal{S}}^2\Big) \\
&= \sum_{h=1}^{3} \frac{N_h}{N} \Big(\bar{x}_{\mathcal{U}_h} - \bar{x}_{\mathcal{U}}\Big)^2 + \sum_{h=1}^{3} \frac{N_h}{N} \mathrm{Var}(\bar{x}_{\mathcal{S}_h}) - \mathrm{Var}(\bar{x}_{\mathcal{S}})
\end{aligned}$$

En nuestro caso

$$\widehat{\mathrm{Var}}(\bar{x}_{\mathcal{S}_h}) = \Big(1 - \frac{n_h}{N_h}\Big) \frac{s_{\mathcal{S}_h}^2}{n_h}$$

es un estimador de $\mathrm{Var}(\bar{x}_{\mathcal{S}_h})$. Si juntamos todo tenemos un estimador insesgado de $s_{\mathcal{U}}^2$ dado por:

$$\hat{S}_{\mathcal{U}}^2 = \sum_{h=1}^{3} \frac{N_h}{N} s_{\mathcal{S}_h}^2 + \sum_{h=1}^{3} \frac{N_h}{N} (\bar{x}_{\mathcal{S}_h} - \bar{x}_{\mathcal{S}})^2 - \sum_{h=1}^{3} \frac{N_h}{N} \widehat{\mathrm{Var}}(\bar{x}_{\mathcal{S}_h}) + \widehat{\mathrm{Var}}(\bar{x}_{\mathcal{S}}) = 20.983$$

La varianza estimada entonces con muestreo aleatorio simple sin reemplazo es:

$$\widehat{V}_{\mathrm{MAS}} = \frac{1 - f}{n} \hat{S}_{\mathcal{U}}^2$$

Mientras que la estimada con alocación proporcional:

$$\widehat{V}_{\text{Prop}} = \frac{1-f}{n}\Big(\sum_{h=1}^{3}\frac{N_h}{N}s^2_{\mathcal{S}_h}\Big)$$

de donde la ganancia de la proporcional está dada por:

$$\frac{\widehat{V}_{\text{Prop}}}{\widehat{V}_{\text{MAS}}} = \frac{\sum_{h=1}^{3}\frac{N_h}{N}s^2_{\mathcal{S}_h}}{\hat{S}^2_{\mathcal{U}}} \approx 40.5\%$$

Lo cual nos muestra que, en este caso, estratificar sí resulta en estimaciones más precisas.

Ejercicio en R tipo control

La base de datos `Base_a_estratificar` (en este link) contiene una base con un millón de entradas correspondientes a los registros de un millón de clientes de una empresa. Se registró el grupo de edad, la entidad federativa y el género de la persona. Interesa realizar un muestreo aleatorio simple para estudiar el ingreso promedio de los clientes estratificando por grupo de edad, entidad y género. Se sabe además que los costos de muestreo por cada persona muestreada varían según el estado y puedes encontrarlos en la base `Costos_x_entidad` este link.

1. Determina las n óptimas para el muestreo suponiendo que la varianza sólo varía por edad de acuerdo a la siguiente tabla (varianza son las s^2) pensando, además que nos interesa un error de ± 50 al 95%.

Edad	Varianza
< 20	100
[20,60]	200
>60	500

2. Suponiendo un costo basal de $500,000$, ¿cuánto es el costo total de la encuesta?
3. La base de datos `Muestra` link contiene una muestra estratificada por muestreo aleatorio simple sin reemplazo de los datos. Obtén el estimador del ingreso promedio y su intervalo de confianza para el total y para cada uno de los estratos.

Solución

1. En primer lugar calculamos N_h de cada estrato así como el N:

```
library(readr)
base.datos <- read_rds("Base_a_estratificar.RDS")
base.nh    <- base.datos %>% group_by(Género, Entidad, Edad) %>% tally()
```

Por otro lado calculamos la varianza a partir del error usando que

$$\epsilon = Z_{1-\alpha/2}\sqrt{\text{Var}(\bar{x}_{\mathcal{S}})} \qquad \text{con} \qquad \alpha = 0.05$$

Entonces:

```
eps.error <- 50
zalpha    <- qnorm(0.975)
var.x     <- (eps.error/zalpha)^2
print(paste0("La varianza es ", var.x))
```

```
## [1] "La varianza es 650.794429067515"
```

De la ecuación calculamos el término $V^* = V - B$. Para ello recordamos que:

$$B = -\sum_{h=1}^{H} N_h s^2_{\mathcal{U}_h}$$

Entonces:

```
edad     <- c("< 20","[20,60]",">60")
dats     <- data.frame(Edad = edad, Varianza = c(100, 200, 500))
base.nh  <- base.nh %>% left_join(dats, by = "Edad")
base.nh  <- base.nh %>% mutate(BSumandos = Varianza*n)
B        <- -sum(base.nh$BSumandos)
print(paste0("B = ", B))
```

```
## [1] "B = -21616500"
```

Por otro lado, obtenemos los costos:

```
base.costos <- read_rds("Costos_x_entidad.RDS")
base.nh     <- base.nh %>% left_join(base.costos, by = "Entidad")
```

Y calculamos los A_h:

```
base.nh <- base.nh %>% mutate(Ah = Varianza*n^2)
```

Finalmente obtenemos los n_h:

```
sumaAh  <- sum(sqrt(base.nh$Ah*base.nh$Costo))
base.nh <- base.nh %>% mutate(nh = sqrt(Ah/Costo)*sumaAh/(var.x - B))
base.nh <- base.nh %>% mutate(nh = ceiling(nh))
```

Verificamos que no haya ningún $n_h > N_h$:

```
base.nh <- base.nh %>% mutate(nh = ifelse(nh > n, n, nh))
```

2. Para determinar el costo total de la encuesta,

```
base.nh <- base.nh %>% mutate(Costo_estrato = Costo*nh)
costo   <- 500000 + sum(base.nh$Costo_estrato)
print(paste0("El costo es de $", scales::comma(costo)))
```

```
## [1] "El costo es de $985,409"
```

3. Analizamos la base de datos muestra:

```
muestra <- read_rds("Muestra_estratificada.RDS")
```

Obtenemos los estimadores puntuales de cada uno

```
promedios.muestra <- muestra %>% group_by(Género, Entidad, Edad) %>%
  summarise(Media = mean(Ingreso), S_h = var(Ingreso), n = n())
```

```
## `summarise()` regrouping output by 'Género', 'Entidad' (override with `.groups` argument)
```

Agrego los N_h y los N:

```
promedios.muestra <- promedios.muestra %>%
  left_join(base.nh, by = c("Género", "Entidad", "Edad")) %>%
  rename(`N_mayusc_h` = n.y) %>% rename(`n_minusc_h` = n.x)

Ntotal <- sum(promedios.muestra$N_mayusc_h)
```

El estimador total es el promedio ponderado de los de cada grupo:

```
promedios.muestra <- promedios.muestra %>%
  mutate(factor_pop = N_mayusc_h/!!Ntotal)
promedios.muestra <- promedios.muestra %>%
  mutate(sumando_media = factor_pop*Media)
xbarra <- sum(promedios.muestra$sumando_media)
print(paste0("La media se estima con ", xbarra))
```

```
## [1] "La media se estima con 1193.65573405234"
```

Mientras que la varianza se estima mediante:

```
promedios.muestra  <- promedios.muestra %>%
  mutate(varianza_intra_clase = (1 - n_minusc_h/N_mayusc_h)/n_minusc_h*S_h)
promedios.muestra <- promedios.muestra %>%
  mutate(sumando_var = (factor_pop^2)*varianza_intra_clase)
varianza.est <- sum(promedios.muestra$sumando_var)
```

Luego el intervalo está dado por:

```
c(
  Lower = xbarra - zalpha*sqrt(varianza.est),
  Upper = xbarra + zalpha*sqrt(varianza.est)
)
```

```
##    Lower    Upper
## 1178.268 1209.043
```

Para los intervalos de cada estrato usamos la varianza específica de los mismos:

```
promedios.muestra <- promedios.muestra %>%
  mutate(ic_lower = Media - !!zalpha*sqrt(varianza_intra_clase)) %>%
  mutate(ic_upper = Media + !!zalpha*sqrt(varianza_intra_clase))
```