Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Especialización en Estadística Aplicada

Técnicas de Muestreo I

Tarea 4

Adriana Haydeé Contreras Peruyero (haydeeperuyero@im.unam.mx)

Alejandro Jiménez Palestino (ajpalestino@gmail.com)

Andrés Peña Montalvo (andres.pena.montalvo12@gmail.com)

Arturo Sánchez González (arturo.sanchez@im.unam.mx)

Jesus Alberto Urrutia Camacho (urcajeal@gmail.com)

Ciudad de México

26 de enero de 2021

Índice

1. Alquiladora de autos

Una alquiladora de autos quiere estimar cuánto gastaría (costo) en arreglar todos los autos que tiene (300), para ello decide elegir una muestra de 50 autos. El encargado de seleccionar la muestra dispone de tres listas con el total de las unidades de la alquiladora:

- `autos.random.csv`: Se encuentran completamente revueltos (fue construida al azar).
- `autos.flota.csv`: Están organizados por la flota a la que están asignados. Se sabe que cada año se procura asignar la misma cantidad de autos por flota.
- `autos.modelos.csv`: Están organizados por el año del auto. Se sabe que los autos con mayor antigüedad tienen mayor deterioro.

Para los fines de este trabajo, guardamos llamamos la información de dichas bases de datos como sigue.

```
autos_random <- read.csv("bd/autos.random.csv")
autos_flota <- read.csv("bd/autos.flota.csv")
autos_modelo <- read.csv("bd/autos.modelo.csv")
```

1.1. Muestreo sistemático

Si se quieren elegir los autos en muestra usando muestreo sistemático ¿qué lista debe seleccionar? ¿por qué?

Ya que queremos realizar un muestreo sistemático, NO es conveniente utilizar la base de datos construida al azar pues, en ese caso, el muestreo aleatorio simple es equivalente al muestreo sistemático (ver [1, Capítulo 8]). Ahora, ya que deseamos que cada uno de los bloques en los cuales se dividirá a la muestra sea lo más heterogéneo posible para reducir la varianza (ver [2, Sección 4.3.3]), como la base `autos_modelo` está ordenada por año y en ella, por año, no hay muchas diferencias en la variable `costo` porque, como se indica en la descripción, en autos con mayor antigüedad hay mayor gasto, no tenemos la heterogeneidad deseada (en la variable `costo`). En virtud de lo anterior, elegimos la base de datos `autos_flota`, donde hay variedad de modelos y aproximadamente la misma cantidad de autos en cada flota.

Una vez elegida la lista adecuada:

- Con muestreo sistemático selecciona 50 autos.

```
N <- length(autos_flota$id)
n <- 50
set.seed(12345)
k <- floor(N/n)
r <- sample(1:k,1)
m <- seq(r,N,k)
muestra <- autos_flota[m,]
```

- Menciona el id de los primeros 5 autos seleccionados.

```
muestra$id[1:5]
## [1] "f6"  "f12" "f18" "f24" "f30"
```

- Estima cuánto deberá de invertir la alquiladora para reparar todos los autos. Estimación puntual y por intervalo.

En primer lugar calculamos el estimador puntual $\hat{Y}$. Para ello, calculamos la media muestral $\bar{y}$ y la multiplicamos por $N = 300$.

```
y_bar <- mean(muestra$costo)
y_gorro <- N * y_bar
y_gorro
## [1] 596172
```

Por lo tanto, el estimador puntual del total es $\hat{Y} = \$\ 596172$.

A continuación calculamos el intervalo del 95 % de confianza para el total Y. Primero calculamos la varianza del estimador del total

$$\hat{V}\left[\hat{Y}\right] = N^2\left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n}$$

donde $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Procedemos a los cálculos necesarios.

```
s_gorro_cuadrado <- sum( (muestra$costo - y_bar )^2 ) / (n-1)

est_var_total <- N^2 * ( 1 - n / N ) * (s_gorro_cuadrado / n)
est_var_total
## [1] 451985361
```

Finalmente, procedemos a la obtención del intervalo de confianza.

```
sd_est <- sqrt(est_var_total)
long_intervalo <- 1.96 * sd_est

lim_inf_intervalo <- y_gorro - long_intervalo
lim_sup_intervalo <- y_gorro + long_intervalo
lim_inf_intervalo
## [1] 554502.5
lim_sup_intervalo
## [1] 637841.5
```

Por lo tanto, el intervalo del 95 % de confianza para el total Y es (554502.5, 637841.5).

Es importante notar que el costo real de las reparaciones es $Y = 597922$, que está dentro del intervalo de confianza obtenido. Además, la diferencia entre el valor estimado puntualmente $\hat{Y}$ y Y es 1750, lo cual muestra explícitamente que es una buena estimación.

1.2. Muestreo por conglomerados

Usando los datos de cualquiera de las listas:

- Divide la lista en conglomerados.

Se podría dividir por `Flotas` o por `Modelo`, se sabe que cada año trataron de asignar la misma cantidad de autos por flota, esto es alrededor de 3.

Flota/Modelo	2009	2010	2011	2012	2013	2014	2015	2016	Total
1	3	3	4	3	4	3	2	3	25
2	3	3	3	4	3	3	3	3	25
3	3	3	4	3	4	3	2	3	25
4	3	3	3	4	3	3	3	3	25
5	3	3	4	3	4	3	2	3	25
6	3	3	3	4	3	3	3	3	25
7	3	3	4	3	4	3	2	3	25
8	3	3	3	4	3	3	3	3	25
9	3	3	4	3	4	3	2	3	25
10	3	3	3	4	3	3	3	3	25
11	3	3	4	3	4	3	2	3	25
12	3	3	3	4	3	3	3	3	25
Total	36	36	42	42	42	36	30	36	300

- Explica el criterio de división.

Como se vió en el apartado anterior, existen 2 agrupaciones naturales previamente establecidas: las `Flotas` a las que son asignados, y el `Modelo` del auto. Se podría construir los conglomerados también con la combinación `Flotas` y `Modelo`, pero en tal caso los conglomerados quedarían demasiado pequeños y la idea es buscar un balance entre tamaño y número de conglomerados, por tanto se empieza por descartar esta combinación.

Dado que existen 12 flotas, y sólamente 8 modelos diferentes de autos, en primera instancia se intuye que al construir conglomerados más pequeños se estaría controlando mejor la viariabilidad los estimadores. Sin embargo, la razón más fuerte de seleccionar como conglomerados a las `Flotas` es porque tienen el mismo tamaño (25), en tal sentido no se tendrá el problema de sobreestimación de varianzas que se da cuando el numero de elementos M_i en los conglomerados es distinto [1]. Entonces, los conglomerados son las flotillas, donde $M_i = 25$ y $N = 12$.

- Elige dos de los conglomerados y usando todos los datos de cada uno de los conglomerados elegidos, estima cuánto deberá invertir la alquiladora para reparar todos los autos. Estimación puntual y por intervalo.

Como se indica en el enunciado, se toman 2 conglomerados de los 12 existentes. La forma en que se seleccionan ambos conglomerados es mediante muestreo aleatorio simple. De esta manera han sido seleccionados los conglomerados 3 y 10.

```
N <- 12
M <- 300
n <- 2

#Se selecciona aleatorieamente 2 conglomerados de los 12, y obteniendo el 3ro y 10mo.
i <- sample(1:N, n)

(muestra <- conglomerados[i, ])
```

flota	costo
3	50589
10	49057

Se estima la media del costo de autos por flota:

$$\hat{\bar{Y}}_e = \frac{\hat{Y}}{M} = \frac{N\hat{\bar{Y}}}{M} = \frac{N}{Mn}\sum_{i=1}^{n} y_i$$

```
(promedio_por_unidad <- (N*sum(muestra$costo))/(M*n))
## [1] 1992.92
```

Y para estimar el costo que debe invertir la alquiladora para reparar todos los autos:

$$\hat{Y} = N\hat{\bar{Y}}$$

```
Y_hat_bar <- mean(muestra$costo)
(total_poblacion <- N*Y_hat_bar)
## [1] 597876
```

Es decir, el costo promedio por reparación de auto es de 1992.92. Mientras que el estimador del total de costo es igual a 597876.

Por otra parte, la varianza estimada es:

[1] Dada la generación de variabilidad entre los totales de los conglomerados.

$$\hat{V}(\hat{Y}) = N^2 \left(1 - \frac{n}{N}\right) \frac{\hat{S}_b^2}{n}$$

Con:

$$\hat{S}_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \hat{\bar{Y}}\right)^2 \text{ y } \hat{\bar{Y}} = \frac{\sum_{i=1}^{n} y_i}{n}$$

```
S_b2 <- sum((muestra$costo-Y_hat_bar)^2)/(n-1)

(var_estim_total <- (N**2)*(1-(n/N))*(S_b2/n))
## [1] 70410720

(desest_estim_total <- sqrt(var_estim_total))
## [1] 8391.11
```

Entonces la varianza estimada para el total del costo de reparación de los autos tiene el valor de 70410720. Por lo que su desviación estándar es igual a 8391.11.

Entonces, se podría arguir que las anteriores estimaciones cambiarían si es que se calculan mediante estimadores de razón, lo cual sería cierto sólo si el tamaño de los conglomerados son de diferente medida. Sin embargo, como ya se observó, los conglomerados son de la misma longitud ($M_i = 25$).

Finalmente, se procede a calcular el intervalo confianza del 95 % para los cálculos anteriores.

```
z <- qnorm(0.975)
IC <- total_poblacion+c(-1,1)*desest_estim_total*z
IC
## [1] 581429.7 614322.3
```

De esta manera, se concluye que el intervalo al 95 % de confianza para el total Y es (581429.7, 614322.3). Mientras que el costo real de las reparaciones $Y = 597922$ cae al interior del intervalo de confianza estimado.

1.3. Comparación de resultados y Conclusiones

Como se observa, los dos intervalos de confianza de las estimaciones del *costo total de reparación de los autos* contienen al verdadero parámetro poblacional. Sin embargo, la estimación del costo total con muestreo por comglomerados es superior a la que se obtuvo con muestreo sistemático y al mismo tiempo más cercana al verdadero valor del parámetro.

Se evidencia, que en este caso[2], la variabilidad (`Desv_Est`) de la estimación del muestreo por conglomerados es menor que en el sistemático, por tanto su intervalo de confianza más angosto, lo cual es un indicio de mayor precisión en la estimación al usar muestreo por conglomerados.

	Y_pob	Y_gorro	Diferencia	Desv_Est	LI	LS
Muestreo sistemático	597922	596172	1750	21259.95	554502.5	637841.5
Muestreo por conglomerados	597922	597876	46	8391.11	581429.7	614322.3

2. Baches en Ciudad de México

En la Ciudad de México existen 16 alcaldías, con la finalidad de investigar la cantidad de baches en la ciudad se eligieron aleatoriamente 4 de las alcaldías y en cada alcaldía 4 colonias en las que se contó la cantidad total de baches. Los valores obtenidos se pueden consultar en el Cuadro 1.

#Análisis

[2]Con la semilla elegida 12345.

Cuadro 1: Baches muestreados

Alcaldía	Baches por colonia en muestra				Colonias por alcaldía
Azcapotzalco	101	69	87	98	111
Iztapalapa	24	136	111	114	293
GAM	85	187	23	74	232
Coyoacán	48	98	305	68	153

En este problema podemos observar que se realizo un muestreo por conglomerados, seleccionando como la Unidad Primaria de Muestreo (UPM) a las 16 alcaldías que existen en la CDMX, y como Unidad Secundaria de Muestreo (USM) a las colonias que conforman dichas alcaldías.

Es importante mencionar que el número de colonias por alcaldía es diferente.

Tomando en cuenta lo anterior a nivel poblacional podemos decir que:

N = 16 (# de Alcaldías en CDMX)

A nivel muestral:

n.bache = 4

m_i.bache = 4

```
N.bache <- 16
n.bache <- 4
m_i.bache <- 4
azcapo <- c(101, 69, 87, 98)
iztapa <- c(24,136,111, 114)
GAM <- c(85, 187, 23, 74)
coyo <- c(48, 98, 305, 68)
i = c("Azcapotzalco", "Iztapalapa", "GAM", "Coyoacán")
prom.baches <- c(88.75, 96.25, 92.25, 129.75)
M_i <- c(111, 293, 232, 153)
Y.tot <- c(prom.baches*M_i)

baches_cdmx.p <- data.frame( cbind(i, M_i,
                        baches_cdmx[,2],
                        baches_cdmx[,3],
                        baches_cdmx[,4],
                        baches_cdmx[,5],
                        prom.baches)
)

# formato de la tabla
kbl(baches_cdmx.p, booktabs = T, align = "c",
    caption = "Baches muestreados", col.names = c(" ", " ", " ", " ", " ", " ",""") )%>%
#column_spec(1, bold=T) %>%
#collapse_rows(columns = 1:2, latex_hline = "major", valign = "middle") %>%
kable_styling(position = "center", latex_options = c("hold_position")) %>%
add_header_above(c("Alcaldía" = 1,  "Colonias por\nalcaldía" = 1,
                   "Baches por colonia\nen muestra" = 4, "prom.y_i"=1))
```

Cuadro 2: Baches muestreados

Alcaldía	Colonias por alcaldía	Baches por colonia en muestra				prom.y_i
Azcapotzalco	111	101	69	87	98	88.75
Iztapalapa	293	24	136	111	114	96.25
GAM	232	85	187	23	74	92.25
Coyoacán	153	48	98	305	68	129.75

Una vez calculado el promedio para cada USM dentro de las UPM podemos calcular el total estimado para cada UPM_i, para eso usamos la expresion:

$$\hat{Y}_i = M_i \hat{\bar{Y}}_i$$

```
Y.tot.tab <-cbind(i,Y.tot)

kbl(Y.tot.tab , booktabs = T, align = "c", caption = "Total estimado para las UPM's",
    col.names = c(" ", " ") )%>%
#column_spec(1, bold=T) %>%
#collapse_rows(columns = 1:2, latex_hline = "major", valign = "middle") %>%
kable_styling(position = "center", latex_options = c("hold_position")) %>%
add_header_above(c("Alcaldía" = 1,  "Total Estimado por UPM" = 1))
```

Cuadro 3: Total estimado para las UPM's

Alcaldía	Total Estimado por UPM
Azcapotzalco	9851.25
Iztapalapa	28201.25
GAM	21402
Coyoacán	19851.75

El Promedio de los totales estimados de UPM lo obtenemos de:

$$\bar{Y} = 1/n \sum_{i=1}^{n} \hat{Y}_i$$

Esta suma es igual a:

```
prom.est.upm <- sum(Y.tot)/4
prom.est.upm
## [1] 19826.56
```

Y el Estimador total poblacional:

$$\hat{Y} = N\hat{\bar{Y}}$$

```r
est.tot.pob <- (N.bache*prom.est.upm)
est.tot.pob
## [1] 317225
```

2.1. Estimación puntual del total de baches en CDMX

De acuerdo al muestreo bietápico realizo, el estimador total pblacional nos indica que en las 16 alcaldías de la CDMX hay un total de 317225 baches.

2.2. Estimación del intervalo del 95 % del total de baches en CDMX

Para Estimar el intervalo al 95 % tenemos que estimar la varianza entre USM dentro de las UPM (`S_wi`) y la varianza entre UPM (`S_b`) Usando las siguientes expresiones:

$$\hat{S}_b^2 = 1/(n-1)\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2$$

$$\hat{S}_{wi}^2 = 1/m_i - 1\sum_{j=1}^{m_i}(y_{ij} - \bar{y_i})^2$$

Sustituimos los valores que ya tenemos:

```r
S_b <- (1/3)*(sum((Y.tot-prom.est.upm)^2))

S_w1 <- sum(((azcapo-prom.baches[c(1)])^2)/3)
S_w2 <- sum(((iztapa-prom.baches[c(2)])^2)/3)
S_w3 <- sum(((GAM-prom.baches[c(3)])^2)/3)
S_w4 <- sum(((coyo-prom.baches[c(4)])^2)/3)
S_wi <- c(S_w1, S_w2, S_w3, S_w4)
cbind(i,S_wi)
##      i              S_wi
## [1,] "Azcapotzalco" "209.583333333333"
## [2,] "Iztapalapa"   "2444.25"
## [3,] "GAM"          "4719.58333333333"
## [4,] "Coyoacán"     "14072.25"
```

```r
betw <- (M_i^2)*((1/4)-(1/M_i))*S_wi
```

La varianza del estimador del total:

$$V(\hat{Y}) = N^2(\frac{1}{n} - \frac{1}{N})S_b^2 + \frac{N}{n}\sum_{i=1}^{N} M_i^2(\frac{1}{m_i} - \frac{1}{M_i})S_{wi}^2$$

```r
var.upm <- ((N.bache^2)*((1/n.bache)-(1/N.bache))*S_b)
var.usm <- ((N.bache/n.bache)*(sum(betw)))
var.bache <- var.upm+var.usm

lim.inf.bache <- est.tot.pob - ((1.96)*(sqrt(var.bache)))
lim.sup.bache <- est.tot.pob + ((1.96)*(sqrt(var.bache)))
round(lim.inf.bache, 2)
## [1] 200709.5
round(lim.sup.bache, 2)
## [1] 433740.5
```

En virtud de lo anterior, el intervalo del 95 % de confianza para el total de baches es (200709.45 , 433740.55).

2.3. Porcentaje de la varianza correspondiente a las UPM

```
perc.var <- var.upm/var.bache
perc.var
## [1] 0.779306
```

Las UPM componen el 77.93 % de la varianza estimada para el total de baches.

Referencias

[1] Cochran, William G., *Sampling techniques*, John Wiley & sons, 3rd. ed., "Willey series in probability and mathematical statistics", USA, 1977.

[2] Som, Ranjan Kumar, *Practical sampling techniques*, Marcel Dekker, Inc., 2nd. ed. revised and expanded, "Statistics, textbooks and monographs", vol. 148, USA, 1996.