Entrega Final - Muestreo I

Lucca Frachelle, Cecilia Waksman

2024-09-07

Introducción

El objetivo de este trabajo es seleccionar una muestra aleatoria de hogares de Montevideo bajo un diseño estratificado, por conglomerados y en dos etapas de selección. Los estratos son 5 y son definidos a nivel socioeconómico (1 = Bajo, 2 = Medio Bajo, 3 = Medio, 4 = Medio Alto, 5 = Alto). LA UPM es la manzana y la USM es el hogar.

Introducción

Las UPM son seleccionadas bajo un diseño PPS sin reemplazo utilizando como medida de tamaño la cantidad de personas por UPM. Luego, dentro de cada UPM seleccionada en la primera etapa, se deben seleccionar 5 viviendas dentro de cada UPM con igual probabilidad de selección.

Una vez seleccionada la muestra, se computarán estimaciones puntuales para distintos parámetros (junto con medidas de calidad de las mismas). Las estimaciones para dichos parámetros pueden ser calculadas, ya sea, a nivel de toda la población, como para distintos dominios/áreas de estimación.

Parte 1

Calcule el tamaño de muestra para obtener un margen de error de $\pm 3\%$ a un 95% de confianza para estimar cualquier proporción poblacional. Asuma un efecto de diseño de 1.5.

Paso 1) $n_0 = (\frac{z^* \sigma}{moe})^2 = (\frac{1.96 \times 0.5}{0.03})^2$

Paso 2) $n_1 = \frac{n_0}{1+(n_0/N)}$

Paso 3) $n_{deff} = n_1 \times deff$

```
  neff
1 1597
```

Parte 2

Asignar por estrato de forma óptima el tamaño de muestra calculado en la parte anterior, utilizando como variable auxiliar el ingreso del hogar (x).

Tamaño de muestra por estrato según asignación óptima:
$n_h = n \times \frac{N_h sd_{U_h}[x]}{\sum_{h=1}^{H} N_h sd_{U_h}[x]}$, donde, $x \propto y$ aproximadamente.

Parte 2

```
estratos = df %>% group_by(estrato) %>%
  summarise(N=n(), sd_ing_hog=sd(ingreso_hog))
estratos = estratos %>% mutate(n_opt=
  round(neff*N*sd_ing_hog/sum(N*sd_ing_hog)))
estratos %>% kable()
```

estrato	N	sd_ing_hog	n_opt
1	66632	901.0174	137
2	92165	1063.8405	223
3	124197	1328.8254	375
4	113171	1797.5697	463
5	56556	3108.8159	400

Parte 3

En la primera etapa del muestreo por conglomerados, se selecciona una muestra de manzanas en cada estrato. El diseño utilizado es πps *sistemático* y las probabilidades de inclusión de cada manzana se calculan en función de su cantidad de personas que viven en la misma.

La cantidad de manzanas por estrato a seleccionar en la muestra se calcula como el tamaño de muestra por asigniación óptima respectivo (calculado en punto 2) dividido la cantidad de individuos a seleccionar en cada manzana en la segunda etapa, en este caso 5.

Primera etapa

```r
set.seed(5)
s_upm=sampling::strata(data=U_upm,
                       stratanames = "estrato",
                       size=round(estratos$n_opt/5),
                       method='systematic',
                       pik=U_upm$Mi,
                       description=T)
```

```
Stratum 1

Population total and number of selected units: 361 27
Stratum 2

Population total and number of selected units: 545 45
Stratum 3

Population total and number of selected units: 768 75
Stratum 4
```

Segunda etapa

En la segunda etapa, se seleccionan de cada manzana 5 hogares mediante un diseño aleatorio simple sin reposición con probabilidades de inclusión $\frac{n_h}{N_h}$, según el estrato.

```
U_usm = df %>% left_join(s_upm %>%
select(manzana, prob_upm), by="manzana") %>%
filter(is.na(prob_upm)==FALSE)

U_usm= U_usm %>% arrange(manzana)
set.seed(5)
s= sampling::strata(data=U_usm,
                    stratanames = 'manzana',
                    size=rep(5,nrow(U_usm)),
                    method='srswor')
s = getdata(U_usm,s) %>%
  rename(prob_usm=Prob)
```

Parte 4

Calcular la estimación puntual del ingreso promedio, proporción de hogares pobres y total de personas, a nivel de toda la población. Para cada estimación se debe computar: error estándar (SE), coeficiente de variación, efecto de diseño y márgenes de error al 95%.

Estimador Horvitz-Thompson en un diseño estratificado:

- Total: $\hat{Y}_{HT} = \sum_{h=1}^{H} \sum_{i \in s} w_{hi} y_{hi}$
- Promedio: $\hat{\bar{Y}}_{HT} = \frac{1}{N} \hat{Y}_{HT} = \frac{1}{N} \sum_{h=1}^{H} \sum_{i \in s_h} w_{hi} y_{hi}$
- Proporsión: $\hat{\bar{Y}}_{HT}$ con y una variable booleana.

Ingreso Promedio

```
res <- svymean(~ingreso_hog, ps1, deff = TRUE)
coef_res <- coef(res)
conf_int <- confint(res)
cv_res <- cv(res)
deff_res <- deff(res)
```

Estadística	Valor
Ingreso promedio por hogar	4646.796
Límite inferior del IC	4534.264
Límite superior del IC	4759.327
Desvío	57.415
Coeficiente de variación	0.012
Efecto de diseño	1.435

Hogares Pobres

```
res <- svymean(~pobre, ps1 , deff = TRUE)
```

Estadística	Valor
Proporción de hogares pobres	0.0739
Límite inferior del IC	0.0543
Límite superior del IC	0.0935
Desvío	0.0100
Coeficiente de variación	0.1352
Efecto de diseño	2.3423

Total de Personas

```
res <- svytotal(~cant_personas, ps1, deff = TRUE)
```

Estadística	Valor
Total de personas	1233250.294
Límite inferior del IC	1204811.779
Límite superior del IC	1261688.809
Desvío	14509.713
Coeficiente de variación	0.012
Factor de diseño	0.783

Parte 5

La varianza a partir de la cual se calcula el desvío estándar ($SE(\hat{\theta}) = \sqrt{V(\hat{\theta})}$) de los parámetros a estimar en la parte anterior se obtiene por el **método del último conglomerado** de la siguiente manera:

$$\hat{V}_{UC}(\hat{\theta}) = \sum_{h=1}^{H} \frac{1}{m_h(m_h-1)} \sum_{j \in s_h} (\hat{\theta}_j^* m_h - \hat{\theta}_h)^2$$

donde $\hat{\theta}_j^*$ es la estimación del parámetro en la j-ésima UPM (manzana), $\hat{\theta}_h$ la estimación del parámetro para el h-ésimo estrato y m_h la cantidad de UPMs del mismo estrato.

Parte 6

Calcular el ingreso per cápita en Montevideo (junto con su error estándar). Indique el tipo de parámetro y qué método fue utilizado por defecto por el paquete survey para la estimación del error estándar. Tenga en cuenta que el ingreso per cápita se calcula como: $\frac{ingrsos\ totales\ en\ Montevideo}{Cantidad\ de\ habitantes}$.

Esta estimación es un ratio, por tanto se calcula como la razón entre dos totales, es decir $R = \frac{Y}{Z}$, y se estima como $\hat{R} = \frac{\hat{Y}_{HT}}{\hat{Z}_{HT}}$.

Parte 6

```
res <- svyratio(~ingreso_hog, ~cant_personas,
                ps1, deff = TRUE)
```

Estadística	Valor
Ingreso per cápita	1705.0376
Límite inferior del IC	1647.0081
Límite superior del IC	1763.0671
Desvío	29.6074
Coeficiente de variación	0.0174

Parte 6

Para este tipo de estimador, la varianza se calcula mediante el método de **estimador de razón** que aproxima pa varianza por linealización de Taylor. Entonces la varianza aproximada se calcula como:

$$AV(\hat{R}) = V(\hat{R}) = \frac{1}{Z^2} \sum_{i \in U} \sum_{i \in U} \triangle_{ij} \frac{y_i - Rz_i}{\pi_i} \frac{y_i - Rz_j}{\pi_j}$$

Y su estimación es: $\hat{V}(\hat{R}) = \frac{1}{\hat{Z}^2_{HT}} \sum_{i \in s} \sum_{i \in s} \frac{\triangle_{ij}}{\pi_{ij}} \frac{y_i - \hat{R}z_i}{\pi_i} \frac{y_i - \hat{R}z_j}{\pi_j}$

Por último, el error estándar será $SE(\hat{\theta}) = \sqrt{V(\hat{\theta})}$.

Parte 7: Jackknife

Este método de remuestreo se basa en eliminar una observación (UPM para conlomerados) por réplica. Hay entonces tantas réplicas como observaciones.

- Cálculo de nuevos ponderadores: $w_{hj(i)} = \frac{n_h}{n_h - 1} w_{hj}$, donde *j* hace referencia a la UPM, *h* al estrato e *i* al índice de la réplica respectiva.
- Cálculo de un total: $\hat{Y}_{(i)} = \sum_{j \in s_{(i)}} w_{j(i)} y_j$.
- Cálculo de un ratio: $\hat{R}_{(i)} = \frac{\hat{Y}_{(i)}}{\hat{Z}_{(i)}}$.
- Cálculo de varianza: $V_J(\hat{Y}) = \frac{n_I}{n_I - 1} \sum_{i=1}^{n_I} (\hat{Y}_{(i)} - \hat{Y})^2$.

Parte 7: Jackknife

```
pps1 <- svydesign(
  strata = ~estrato,
  ids = ~manzana + ID,
  probs  = ~prob_upm + prob_usm ,
  data = s)
jkn <- as.svrepdesign(design = pps1, type = "JKn")

te=svyratio(~ingreso_hog, ~cant_personas,
            jkn , return.replicates=TRUE)
estimacion <- te$ratio
desvio <- sqrt(te$var)
```

Estimación	Desvío
1705.04	28.72

Parte 7: Bootstrap

Para este diseño, el cual presenta estratos, usaremos la variación del método Bootstrap, Rao-Wu.

Se extraen 1000 muestras aleatorias simples con reposición (réplicas) de tamaño m'_h de las UPM entre las m_h seleccionadas en la muestra original.

Sea m^b_{hj} la cantidad de veces que aparece la j-ésima UPM es seleccionada en la réplica b, el *factor de multiplicidad*. Entonces el ponderador a utilizar será: $w^b_{khj} = \frac{m_h}{m_h - 1} w_{khj}$.

A partir de esto se calculan para cada réplica los totales pertinentes por el método de Horvitz-Thompson y con ellos el ratio.

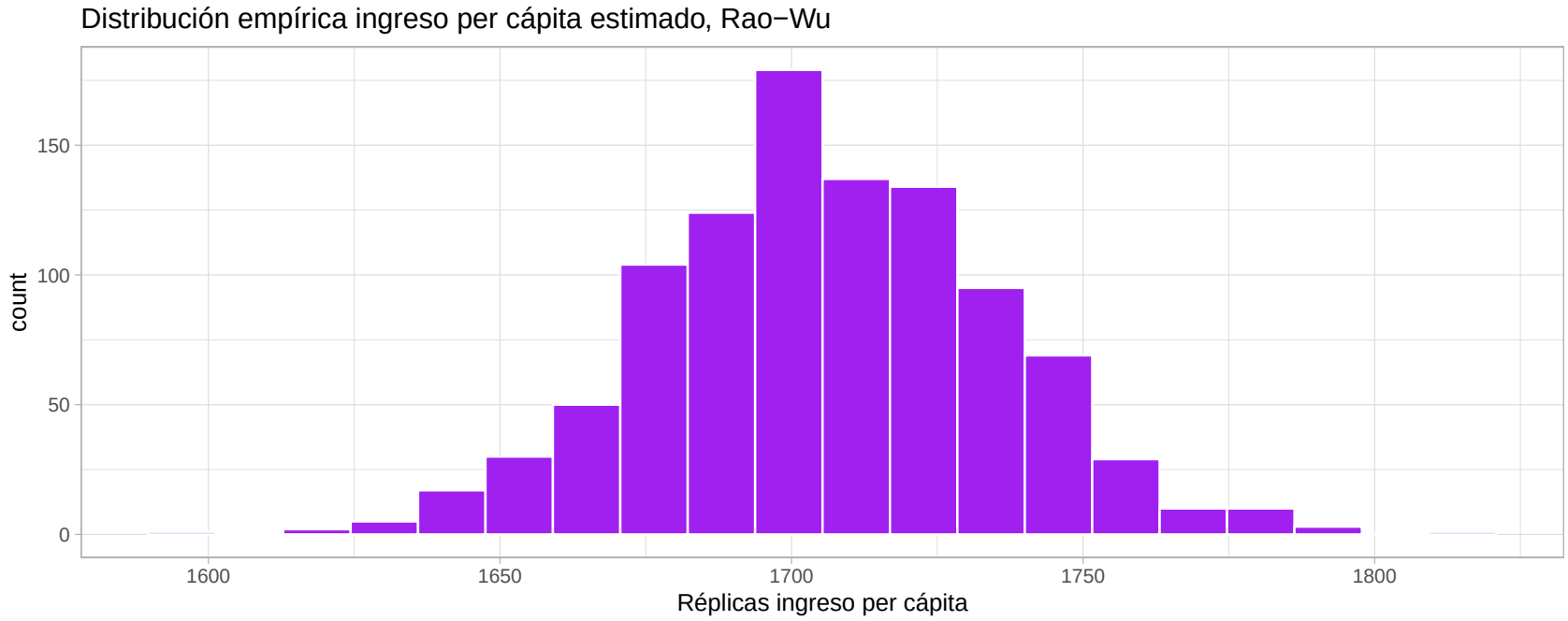
Parte 7: Bootstrap

```
#|output=FALSE
boot=as.svrepdesign(design=ps1, type='subbootstrap',
                    replicates=1000)

te=svyratio(~ingreso_hog, ~cant_personas,
            boot,return.replicates=TRUE)
estimacionb <- te$ratio
desviob <- sqrt(te$var)
```

Método	Estimación	Desvío
Estimador de razón	1705.04	29.61
Jackknife	1705.04	28.72
Bootstrap	1705.04	29.14

Parte 8



Según el método de Bootstrap, las estimaciónes en las réplicas de la variable, ingreso per cápita, tienden a concentrarse en valores en torno a 1700, con una distribución aproximadamente simétrica.

Parte 9

Estimación de la cantidad de personas pobres y no pobres, junto con sus márgenes de error, utilizando el Bootstrap realizado en los puntos anteriores.

```
res_no <- svytotal(~(pobre==0), boot)
coef_no <- coef(res_no)
conf_int_no <- confint(res_no)
se_no <- SE(res_no)
```

Estado	Estimación	LI.IC	LS.IC	SE
Personas pobres	33444	24911	41977	4354
Personas no pobres	419070	402946	435193	8227

Parte 9

Los dominios de estimación, en este caso son no planeados, ya que la distinción entre pobre y no pobre (dominios) no es tenida en cuenta en el diseño.

Si se quisieran mejorar dichas estimaciones se podría:

- estratificar también según si son o no pobres, o en el caso de hacer conglomerados por manzanas, si la proporción de pobres en dicha manzana pasa o no cierto umbral;
- asignar el tamaño de muestra por estratos según una variable que se encuentre más relacionada con la variable de pobreza;
- usar un tamaño de muestra mayor.