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.

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

El tamaño de muestra será entonces:

```
  neff
1 1597
```

Parte 2

Con el tamaño de muestra calculado en la parte anterior, asigne el mismo por estrato de forma óptima, utilizando como variable auxiliar el ingreso del hogar.

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, x es la variable auxiliar, y la variable de interés.

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.

```
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

Population total and number of selected units: 683 93
Stratum 5

Population total and number of selected units: 410 80
Number of strata  5
Total number of selected units 320
```

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.

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

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

La media del ingreso de los hogares es de 4646.8, con desvío estándar de 57.4 un intervalo de confianza al 95% de $\pm$122.5. Considerando que esta variable toma valores entre 1287 y 60345 en la población, esta estimación es relativamente precisa.

El coeficiente de variación estimado, 0.012, explica que los valores que toma la variable, ingreso de los hogares, presentan una dispersión significativamente pequeña respecto a la media estimada, por lo que la misma resume bien a la variable.

Por otro lado, el efecto de diseño es mayor que 1; este resultado implica que el diseño utilizado es 1.4 veces menos eficiente que si se hubiera usado un diseño aleatorio simple (MAS).

Hogares Pobres

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

Se estima que un 7.39% de la población son pobres, con un desvío de un 1% y, por consiguiente, con un $IC_{95\%} = \pm 1.96\%$, lo cual permite pensar que esta es una estimación precisa.

El coeficiente de variación estimado es de 0.135, este es menor que 0.3, por lo que todavía se puede considerar que los datos se encuentran concentrados en torno a la proporción estimada.

En este caso, el efecto de diseño estimado permite observar que el diseño utlilizado es 2.3 veces más ineficiente que si se usar un *MAS* para estimar esta proporción.

Total de Personas

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

El total de personas estimado es de 1.233 millones, con un desvío de 14509, con $IC_{95\%} = \pm 28439$, lo cual parece ser una estimación suficientemente precisa.

El coeficiente de variación estimado es de 0.012, por lo que la variación de los datos despecto de su media es relativamente pequeña.

Por último, el efecto de diseño estimado es de 0.783, es decir que el diseño elegido es más eficiente que si se hubiera realizado un *MAS*.

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 $\hat{R} = \frac{\hat{Y}_{HT}}{\hat{Z}_{HT}}$.

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

El ingreso per cápita estimado en Montevideo es de 1705, con un desvío estándar estimado de 29.6 y un $IC_{95\%} = \pm 58$. El coeficiente de variación estimado es de 0.017, por lo que la dispersión en los datos es relativamente pequeña.

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

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

Parte 7 y 8

Calcular la estimación del error estándar del punto anterior, utilizando dos métodos de remuestreo: Jackknife y Bootstrap (con 1000 réplicas). Compare los resultados obtenidos.

Realizar una visualización de la distribución empírica del estimador utilizando Bootstrap. Interprete los resuultados.

Jackknife

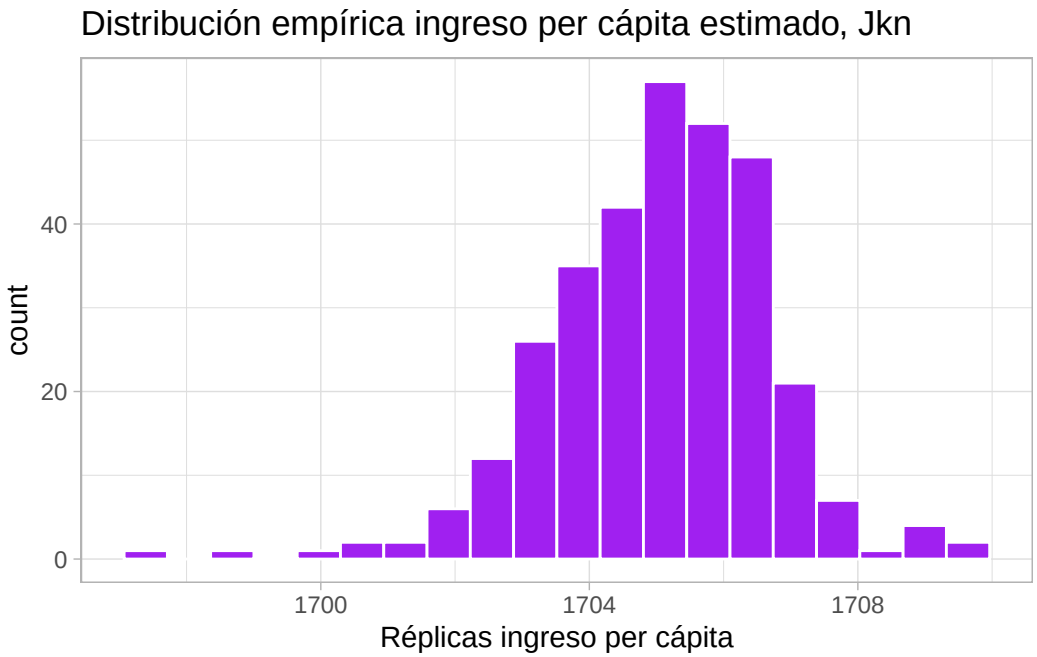
Este método de remuestreo se basa en eliminar una observación por réplica, es decir que hay tantas réplicas como observaciones. En el caso del diseño por conglomerados se elimina una UPM, por lo que los ponderadores se reescriben como: $w_{hj(i)} = \frac{n_h}{n_h - 1} w_{hj}$, donde j hace referencia a la UPM, h al estrato e i al índice de la réplica respectiva.

Un total en cada réplica se calcula entonces como $\hat{Y}_{(i)} = \sum_{j \in s_{(i)}} w_{j(i)} y_j$. Y el ratio respectivo es $\hat{R}_{(i)} = \frac{\hat{Y}_{(i)}}{\hat{Z}_{(i)}}$.

La varianza de un estimador en este método se calcula como: $V_J(\hat{Y}) = \frac{n_I}{n_I - 1} \sum_{i=1}^{n_I} (\hat{Y}_{(i)} - \hat{Y})^2$. Aplicando esta fórmula a las enstimaciones del ratio de interés y calculando su raíz cuadrada, obtenemos el error estándar.

Estimación	Desvío
1705.04	28.72

La estimaciones del ingreso per cápita y su error estándar por el método de *Jackknife* es aproximadamente igual a las etimaciones por el método de *estimador de razón*.



Según el método Jackkife, las estimaciónes en las réplicas del ingreso per cápita parecen estar más concentradas en valores en torno a 1705, con una distribución asimétirca hacia valores bajos.

Bootstrap

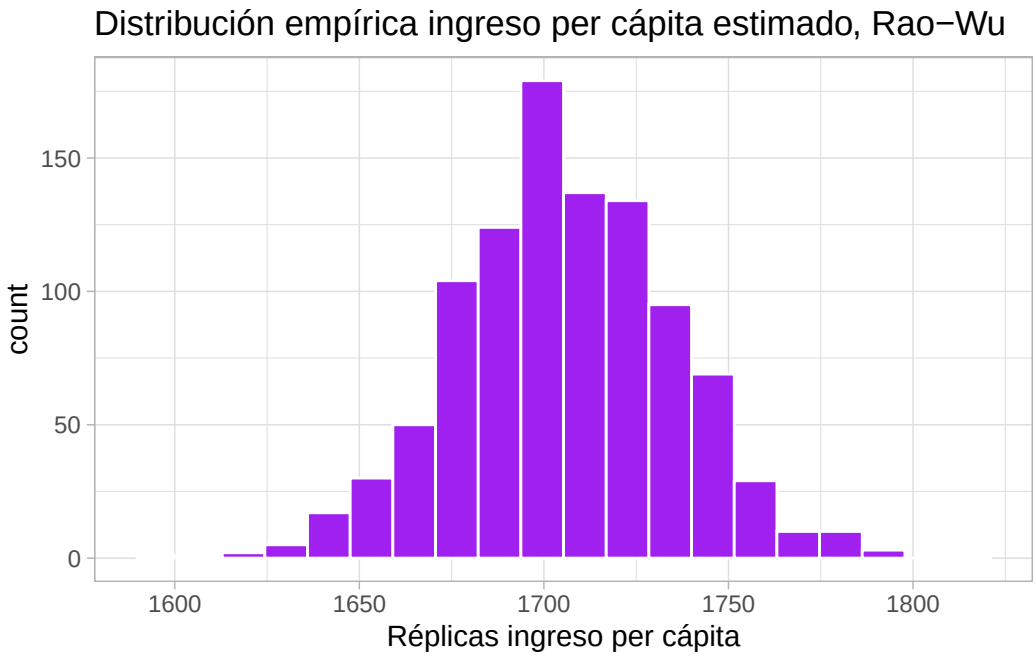
Para este diseño, el cual presenta estratos, usaremos la variación del método Bootstrap, Rao-Wu. La misma consiste en extraer, en este caso, 1000 muestras aleatorias simples con reposición (réplicas) de tamaño m'_h de las UPM entre las m_h seleccionadas en la muestra original.

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.

Estimación	Desvío
1705.04	29.14

La estimaciones del ingreso per cápita y su error estándar por el método de *Bootstrap* es aproximadamente igual a las etimaciones por los métodos de *estimador de razón* y *Kackknife*.



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

Estimar la cantidad de personas pobres y no pobres (junto con sus márgenes de error). Indicar si los dominios de estimación son planeados o no planeados. Para este caso, utilices el Bootstrap realizado en los puntos anteriores. Comente los resultados obtenidos y proponga estrategias para mejorar la precisión de las estimación obtenida.

Estado	Estimación	Límite.inferior.IC	Límite.superior.IC	Error.estándar
Personas pobres	33444	24911	41977	4354
Personas no pobres	419070	402946	435193	8227

Parte 9

Se estima que en la población hay 33444 ± 8533 personas pobres contra 419070 ± 16123 no pobres (siendo el intervalo de confianza al 95%), con errores estándar estimados de 4354 y 8227, respectivamente. La precisión de la estimación del total de no pobres es relativamente alta, pero la de pobres presenta una dispersión grande en comparación con su valor estimado.

Los dominios de estimación, es decir los subconjuntos de la población que distinguen entre pobre y no pobre, en este caso son no planeados, ya que los musmos no son tenidos en cuenta en el diseño muestral.

Si se quisieran mejorar dichas estimaciones se podría estratificar también según si son o no pobres (dominios serían entonces planeados), 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; como alguos ejemplos.