Muestreo en Varias Etapas

Curso de Muestreo Probabilístico en Encuestas de Hogares

Andrés Gutiérrez, Ph.D.

CEPAL - Unidad de Estadísticas Sociales

Motivación

En encuestas complejas, los grupos poblacionales de elementos que se forman naturalmente como barrios, municipios o escuelas pueden ser tratados como unidades de muestreo. Este tipo de esquemas de muestreo ayuda a aumentar el tamaño de muestra manteniendo el costo de la encuesta.

Risto Lehtonen (2004)

Bibliografía y referencias

- Kish, L. (1965) *Survey Sampling*. John Wiley and Sons.
- Cochran, W. G. (1977) *Sampling Techniques*. John Wiley and Sons.
- Särndal, et. al. (2003) *Model-assisted Survey Sampling*. Springer.
- Gutiérrez, H. A. (2016) *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Ediciones de la U.
- Gutiérrez, H. A. (2017) `TeachingSampling`. *R package*.

Aglomeración

Características

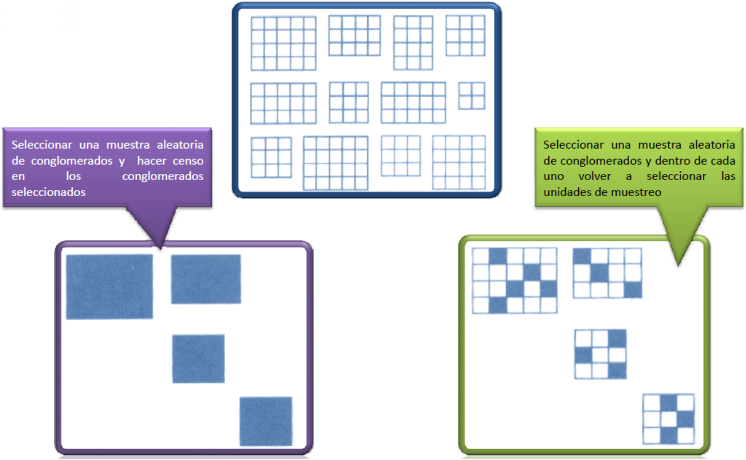


Figure 1: *Muestreo de conglomerados y en etapas*

Características

Utilizamos muestreo por conglomerados sí:

- La construcción de un marco de muestreo de elementos es muy difícil, muy costosa o imposible de conseguir. Enumerar abejas, enumerar clientes, enlistar árboles en un sector, enlistar hogares en los barrios conglomerados (dispersión geográfica, reducción de costos).
- La población objetivo se encuentra muy dispersa (geográficamente) o aparece en agrupaciones naturales: familias, escuelas, etc.

Características

- Los elementos individuales de una población sólo participan en la muestra si pertenecen a un conglomerado incluido en la muestra.
- El muestreo estratificado aumenta la precisión de las estimaciones, mientras que el muestreo por conglomerados tiende a disminuirla. Es un precio que se paga al no poseer un marco de muestreo definido para los elementos de la población objetivo.
- Al obtener una muestra de elementos que pertenecen a un conglomerado repetimos la información del conglomerado (dada la agrupación natural). Lo ideal es conseguir información nueva en cada individuo, por lo anterior se pierde precisión en las estimaciones.

Muestreo de conglomerados

Suponga que la población de elementos

$$U = \{1, ..., k, ..., N\}.$$

se divide en N_I sub-grupos poblacionales, llamados **conglomerados** y denotados como $U_I = \{U_1, \ldots, U_{N_I}\}$.

Características

La población de conglomerados estará dada, sin pérdida de generalidad, por

$$U_I = \{1, \ldots, N_I\}.$$

Estos definen una partición de la población en tal forma que

1. $U = \bigcup_{i=1}^{N_I} U_i$
2. $U_i \bigcap U_j = \emptyset$ para todo $i \neq j$

Características

El número de unidades N_i en el conglomerado i-ésimo se llama **tamaño del conglomerado** tal que

$$N = \sum_{i=1}^{N_I} N_i,$$

donde N es el tamaño de la población U. Con la población dividida en N_I conglomerados,

Parámetros poblacionales

El total poblacional puede escribirse como:

$$t_y = \sum_{k \in U} y_k = \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \sum_{i=1}^{N_I} t_{yi} \qquad (1)$$

donde $t_{yi} = \sum_{k \in U_i} y_k$ es el total del i-ésimo conglomerado.

Parámetros poblacionales

La media poblacional puede escribirse como:

$$\bar{y}_U = \frac{\sum_{k \in U} y_k}{N} = \frac{1}{N} \sum_{i=1}^{N_I} \sum_{k \in U_i} y_k = \frac{1}{N} \sum_{i=1}^{N_I} N_i \bar{y}_i \tag{2}$$

donde $\bar{y}_i = \dfrac{1}{N_i} \sum_{k \in U_i} y_k$ es la media del i-ésimo conglomerado.

Diseño de muestreo

El esquema general del diseño de muestreo por conglomerados está definido de la siguiente forma

1. Seleccionar una muestra probabilística s_I de conglomerados de la población U_I mediante un diseño de muestreo tal que

$$Pr(S_I = s_I) = p_I(s_I) \qquad \text{para todo } s_I \in Q_I. \tag{3}$$

   donde Q_I es el soporte de muestras de conglomerados.

2. Todos y cada uno de los elementos pertenecientes a los conglomerados seleccionados son observados y medidos.

Diseño de muestreo

En muestreo por conglomerados, se utiliza un diseño $p_I(s_I)$ que puede ser cualquiera de los vistos anteriormente:

- **Sin reemplazo:** si todas las posibles muestras en Q_I son sin reemplazo. Muestreo aleatorio simple, Bernoulli, Sistemático, Poisson, πPT o estratificado simple.
- **Con reemplazo:** si todas las posibles muestras en Q_I son con reemplazo. Muestreo aleatorio simple con reemplazo o muestreo PPT.
- **De tamaño fijo:** si todas las posibles muestras en Q tienen el mismo tamaño de muestra $n(S_I) = n_I$.

La muestra aleatoria de elementos viene caracterizada por

$$S = \bigcup_{i \in S_I} U_i \tag{4}$$

El tamaño de muestra

El tamaño de la muestra de elementos está dado por

$$n(S) = \sum_{i \in S_I} N_i \tag{5}$$

Ejemplo

Nuestra población ejemplo U dada por

$$U = \{\textbf{Yves, Ken, Erik, Sharon, Leslie}\}$$

se divide en tres conglomerados de la siguiente forma

$$U_1 = \{\textbf{Yves, Ken}\}$$

el segundo conformado por

$$U_2 = \{\textbf{Erik,Sharon}\}$$

y el último conglomerado dado por

$$U_3 = \{\textbf{Leslie}\}$$

Ejemplo

Se tienen $N_I = 3$ conglomerados de tamaño diferentes. De esta manera, la población de conglomerados queda definida por

$$U_I = \{U_1, U_2, U_3\}$$

Ejemplo

Suponga que se selecciona una muestra s_I de conglomerados de tamaño $n_I = 2$.

```r
U <- c("Yves", "Ken", "Erik", "Sharon", "Leslie")
U1 <- c("Yves", "Ken")
U2 <- c("Erik", "Sharon")
U3 <- c("Leslie")

UI <- c("U1","U2","U3")
```

Ejemplo

Las posibles muestras (sin reemplazo y de tamaño fijo) de conglomerados son:

```
library(TeachingSampling)

NI <- 3
nI <- 2

QI  <-  Support(NI, nI, UI)
QI
```

U1	U2
U1	U3
U2	U3

Estimador del total poblacional

El estimador de Horvitz-Thompson para el total t_y y su varianza estimada están dados por

$$\hat{t}_{y,\pi} = \sum_{i \in S_I} \frac{t_{yi}}{\pi_{Ii}} \tag{6}$$

$$\widehat{Var}_1(\hat{t}_{y,\pi}) = \sum\sum_{S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{t_{yi}}{\pi_{Ii}} \frac{t_{yj}}{\pi_{Ij}} \tag{7}$$

respectivamente, con $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$ y t_{yi} el total del i-ésimo conglomerado seleccionado.

Comentarios

- La eficiencia de la estrategia de muestreo toma su máximo valor cuando los valores $\frac{t_{yi}}{\pi_{Ii}}$ son constantes para todo $i = 1, \ldots, N_I$.
- Cuando el diseño por conglomerados asigna probabilidades de inclusión idénticas a cada conglomerado, la estrategia pierde eficiencia, a menos que el comportamiento de los totales de cada conglomerado sea similar.

Comentarios

- Es preferible escoger diseños de muestreo que asignen probabilidades de inclusión proporcionales al tamaño del conglomerado.
- Para esto se debería disponer de información auxiliar continua disponible para toda la población U_I que estuviera bien correlacionada con los totales de la característica de interés en cada conglomerado t_{yi}.
- Si t_{xi} es el total de la información auxiliar en el i-ésimo conglomerado, la correlación entre t_{xi} y t_{yi} debería ser bastante fuerte.

Comentarios

- Si la selección de los conglomerados se hace con reemplazo, haciendo uso de un diseño de muestreo PPT, es posible utilizar los principios del estimador de Hansen-Hurwitz para completar la estrategia de muestreo.
- En caso de conocer los tamaños N_i de cada cluster $i = 1, \ldots, N_I$, estos mismos pueden ser utilizados como medidas de tamaño para desarrollar un plan de muestreo con probabilidades proporcionales.

Estimador del total poblacional

El estimador de Hansen-Hurwitz para el total t_y y su varianza estimada están dados por

$$\hat{t}_{y,p} = \frac{1}{m_I} \sum_{v=1}^{m_I} \frac{t_{yi_v}}{p_{Ii_v}} \tag{8}$$

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m_I(m_I - 1)} \sum_{v=1}^{m_I} \left(\frac{t_{yi_v}}{p_{Ii_v}} - \hat{t}_{y,p} \right)^2 \tag{9}$$

respectivamente, en donde $p_{Ii} = \frac{t_{xi}}{t_x}$.

Muestreo aleatorio simple de conglomerados

- La muestra s_I de n_I conglomerados es seleccionada mediante un diseño de muestreo aleatorio simple sin reemplazo.
- Este diseño de muestreo asume que el comportamiento del total de la característica de interés es constante en cada uno de los conglomerados.

Muestreo aleatorio simple de conglomerados

- En la práctica esta situación se presenta en muy pocas ocasiones, es por esto que este diseño pierde precisión, en la mayoría de ocasiones, ante el muestreo aleatorio simple.
- Para que este diseño de muestreo sea más eficiente el valor promedio de la característica de interés en cada cluster $\bar{y}_{U_i}$ debería ser proporcional a $\frac{c}{N_i}$.

Tamaño de muestra

El tamaño de la muestra de elementos s es aleatorio y su esperanza está dada por

$$E(n(S)) = N\frac{n_I}{N_I} \tag{10}$$

Muestreo sistemático

- Nótese que el diseño de muestreo sistemático en un caso especial del muestreo aleatorio de conglomerados cuando se selecciona una muestra s_I de tamaño igual a $n_I = 1$.
- Al igual que en muestreo sistemático no se tiene un estimador de la varianza cuando se selecciona sólo un conglomerado.

Práctica en R

```r
library(TeachingSampling)
library(dplyr)
data("BigCity")

Hogares <- BigCity %>% group_by(HHID) %>%
  summarise(Estrato = unique(Zone),
            UPM = unique(PSU),
            Personas = n(),
            Ingreso = sum(Income),
            Gasto = sum(Expenditure),
            Pobreza = unique(Poverty))
```

Práctica en R

```r
attach(Hogares)

UI <- levels(as.factor(Hogares$UPM))
NI <- length(UI)
nI <- 100
```

Práctica en R

```
samI <- S.SI(NI, nI)
muestra <- UI[samI]
muestra
```

```
##   [1] "PSU0026" "PSU0033" "PSU0043" "PSU0047" "PSU0061" "PSU0068" "PSU0070"
##   [8] "PSU0082" "PSU0123" "PSU0142" "PSU0159" "PSU0173" "PSU0174" "PSU0207"
##  [15] "PSU0213" "PSU0222" "PSU0243" "PSU0251" "PSU0289" "PSU0298" "PSU0302"
##  [22] "PSU0303" "PSU0338" "PSU0385" "PSU0387" "PSU0402" "PSU0413" "PSU0415"
##  [29] "PSU0455" "PSU0463" "PSU0465" "PSU0488" "PSU0496" "PSU0502" "PSU0504"
##  [36] "PSU0529" "PSU0530" "PSU0544" "PSU0545" "PSU0562" "PSU0573" "PSU0618"
##  [43] "PSU0620" "PSU0629" "PSU0631" "PSU0639" "PSU0669" "PSU0676" "PSU0679"
##  [50] "PSU0728" "PSU0731" "PSU0762" "PSU0773" "PSU0798" "PSU0806" "PSU0821"
##  [57] "PSU0851" "PSU0892" "PSU0901" "PSU0946" "PSU0999" "PSU1001" "PSU1013"
##  [64] "PSU1046" "PSU1061" "PSU1070" "PSU1099" "PSU1147" "PSU1157" "PSU1168"
##  [71] "PSU1171" "PSU1173" "PSU1212" "PSU1214" "PSU1249" "PSU1254" "PSU1258"
##  [78] "PSU1268" "PSU1299" "PSU1328" "PSU1370" "PSU1379" "PSU1383" "PSU1404"
##  [85] "PSU1409" "PSU1411" "PSU1426" "PSU1441" "PSU1462" "PSU1477" "PSU1504"
##  [92] "PSU1525" "PSU1538" "PSU1574" "PSU1609" "PSU1639" "PSU1643" "PSU1645"
##  [99] "PSU1657" "PSU1658"
```

Práctica en R

```r
CityI <- Hogares[which(Hogares$UPM %in% muestra),]
attach(CityI)
head(CityI)
```

HHID	Estrato	UPM	Personas	Ingreso	Gasto	Pobreza
idHH00355	Rural	PSU0026	4	4283	1796	NotPoor
idHH00356	Rural	PSU0026	2	1782	390	NotPoor
idHH00357	Rural	PSU0026	4	895	593	Relative
idHH00358	Rural	PSU0026	4	4190	2094	NotPoor
idHH00359	Rural	PSU0026	6	1358	2055	Relative
idHH00360	Rural	PSU0026	4	2525	945	NotPoor

Práctica en R

```
Area <- as.factor(CityI$UPM)
estimaI <- data.frame(CityI$Personas, CityI$Ingreso, CityI$Gasto)
head(estimaI)
```

CityI.Personas	CityI.Ingreso	CityI.Gasto
4	4283	1796
2	1782	390
4	895	593
4	4190	2094
6	1358	2055
4	2525	945

Práctica en R

```
E.1SI(NI, nI, estimaI, Area)
```

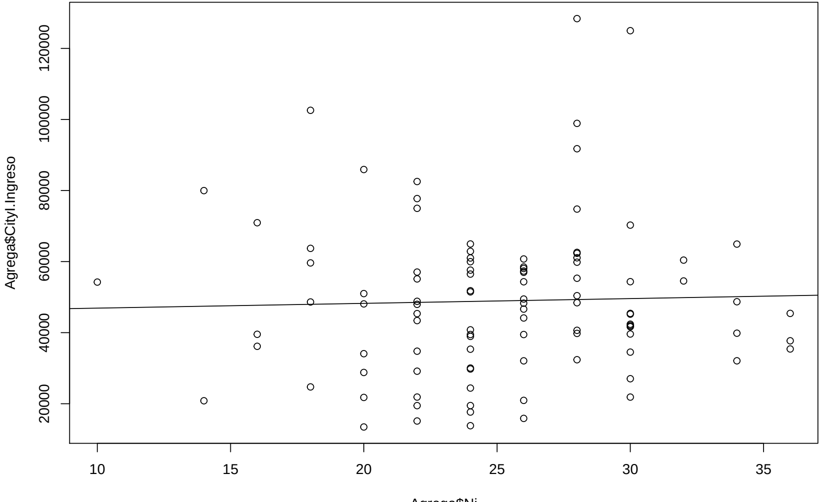
	N	CityI.Personas	CityI.Ingreso	CityI.Gasto
Estimation	41766.4	152389.1	81420759.7	53805486.3
Standard Error	814.4	3691.1	3492922.8	2362556.2
CVE	1.9	2.4	4.3	4.4
DEFF	Inf	5.6	2.7	3.6

Práctica en R

- Es claro que los resultados de esta estrategia de muestreo no son satisfactorios.
- La explicación de la deficiencia de esta estrategia es inmediata al analizar el comportamiento estructural de los totales en los conglomerados.

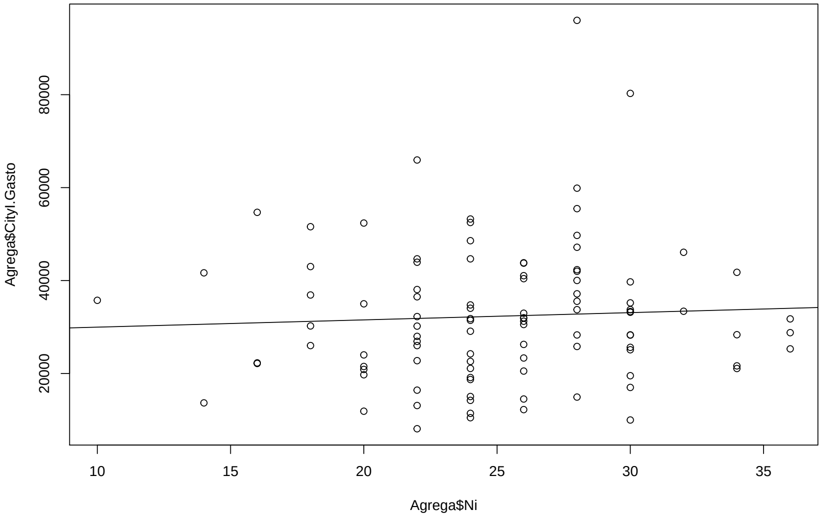
Práctica en R

```
Agrega <- as.data.frame(T.SIC(estimaI, Area))
plot(Agrega$Ni, Agrega$CityI.Ingreso)
abline(lm(Agrega$CityI.Ingreso ~ Agrega$Ni))
```



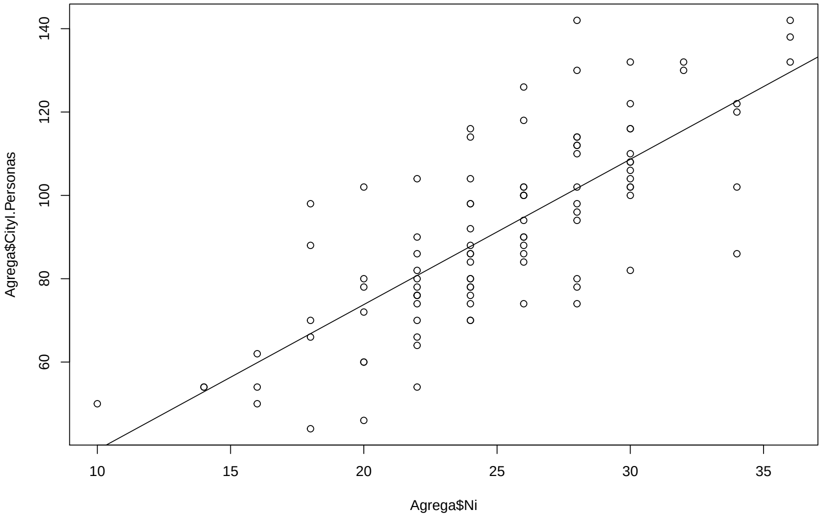
Práctica en R

```
plot(Agrega$Ni, Agrega$CityI.Gasto)
abline(lm(Agrega$CityI.Gasto ~ Agrega$Ni))
```



Práctica en R

```r
plot(Agrega$Ni, Agrega$CityI.Personas)
abline(lm(Agrega$CityI.Personas ~ Agrega$Ni))
```



Muestreo en varias etapas

Características

- Anteriormente se utilizó la agrupación natural de los elementos en la población para ahorrar costes financieros y logísticos al planear una estrategia de muestreo por conglomerados.
- El ahorro en términos operativos se ve reflejado en un alto precio por pagar con respecto a la eficiencia estadística de la estrategia.
- Una posible solución para disminuir la varianza es aumentar el tamaño de muestra de conglomerados, solución que aumentaría los costos operativos.

Características

- Para mantener un equilibrio entre los costos financieros y las bondades de la estrategia de muestreo es posible aprovechar la homogeneidad dentro de los conglomerados.
- No vamos a realizar un censo dentro de cada conglomerado seleccionado sino a seleccionar una sub-muestra dentro de cada uno.

Características

- Como el comportamiento estructural de la característica de interés al interior de los conglomerados es homogéneo, entonces una estimación del total del conglomerado tendría una varianza pequeña.
- Como no se tienen acceso a un marco de muestreo de elementos, se debe realizar un empadronamiento para levantar un marco de muestreo de elementos en cada uno de los conglomerados seleccionados.

Características

El principio básico del muestreo en varias etapas se puede definir como el proceso jerárquico que realiza I veces los siguientes pasos:

1. Construcción de I marcos de muestreo de unidades (conglomerados en las primeras $I-1$ etapas del diseño muestral y de elementos en la última etapa).
2. Aplicación de un diseño muestral y selección de la muestras (o sub-muestras) de cada marco de muestreo.

Unidades de muestreo

Es interesante observar cómo la población, en el estado de la naturaleza, se subdivide gracias al comportamiento “jerárquico” de las poblaciones.

$$\underbrace{\textbf{Ciudad}}_{\text{Población}U} \Rrightarrow \underbrace{\textbf{Escuelas}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Niveles}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Alumnos}}_{\text{UTM}}$$

Unidades de muestreo

1. **Unidad Primaria de Muestreo** o **UPM** es la primera subdivisión en conglomerados de la población original.
2. **Unidad Secundaria de Muestreo** o **USM** es la sub-subdivisión de las unidades primarias.
3. La **Unidad Terciaria de Muestreo** o **UTM** corresponde a los elementos de la población objetivo, que en el caso anterior corresponde a los alumnos de la ciudad.

Unidades de muestreo

- No siempre las unidades finales de muestreo son elementos.
- Es posible planear un diseño en dos etapas de conglomerados, refiriéndose a que la unidad secundaria de muestreo son conglomerados.
- También es posible aplicar un diseño en cuatro etapas de elementos, en donde las unidades finales de muestreo sean elementos

Unidades de muestreo

Considere el siguiente caso:

$$\underbrace{\textbf{Ciudad}}_{\text{Población}\,U} \Rrightarrow \underbrace{\textbf{Sección}}_{\text{UPM}} \Rrightarrow \underbrace{\textbf{Manzana}}_{\text{USM}} \Rrightarrow \underbrace{\textbf{Vivienda}}_{\text{UTM}} \Rrightarrow \underbrace{\textbf{Persona}}_{\text{UCM}}$$

Supuestos

El principio básico de una estrategia de muestreo en varias etapas es construir estimaciones desde abajo hasta arriba.

1. **Invariancia:** sugiere que la probabilidad de selección de una muestra de unidades de muestreo (conglomerados o elementos) no depende del diseño de muestreo de la anterior etapa.
2. **Independencia:** interpretado como que el sub-muestreo de cualquier unidad de muestreo se lleva a cabo de manera independiente con las otras unidades de muestreo, en la misma etapa o en etapas superiores o inferiores.

Muestreo en dos etapas

Características

1. Este diseño de muestreo estima el total de cada cluster t_i mediante una sub-muestra dentro de los conglomerados seleccionados de la población.
2. En la estimación de los parámetros de interés se encuentran dos fuentes de variabilidad en cada etapa.

- Existe variabilidad debido a la selección de las unidades primarias de muestreo.
- También existe variabilidad debido a la selección de una muestra de elementos, unidades secundarias de muestro en los conglomerados seleccionados.

Características

- Una muestra s_I de unidades primarias de muestreo es seleccionada de U_I de acuerdo a un diseño de muestreo $p_I(s_I)$.
- Nótese que S_I representa la muestra aleatoria de conglomerados tal que $Pr(S_I = s_I) = p_I(s_I)$.

Características

- Para cada conglomerado U_i $i = 1, \ldots, N_I$ seleccionado en la muestra s_I, se selecciona una muestra s_i de elementos de acuerdo a un diseño de muestreo $p_i(s_i)$.
- Nótese que S_i representa la muestra aleatoria de elementos tal que $Pr(S_i = s_i) = p_i(s_i)$.

Estimador del total poblacional

El estimador de Horvitz-Thompson es insesgado para el total poblacional y toma la forma

$$\hat{t}_{y,\pi} = \sum_{i \in S_I} \sum_{k \in S_i} \frac{y_k}{\pi_{Ii}\pi_{k|i}} = \sum_{i \in S_I} \frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \tag{11}$$

donde

$$\hat{t}_{yi,\pi} = \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}}$$

Estimador de la varianza total poblacional

La varianza estimada del estimador de Horvitz-Thompson es

$$\widehat{Var}_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum\sum_{S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{t}_{yi,\pi}}{\pi_{Ii}} \frac{\hat{t}_{yj,\pi}}{\pi_{Ij}}}_{\widehat{Var}(UPM)} + \underbrace{\sum_{i \in S_I} \frac{\widehat{Var}(\hat{t}_{yi,\pi})}{\pi_{Ii}}}_{\widehat{Var}(USM)} \tag{12}$$

donde

$$\widehat{Var}(\hat{t}_i) = \sum\sum_{S_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \tag{13}$$

Diseño de muestreo MAS-MAS

- En el muestreo aleatorio simple de conglomerados se medían todos y cada una de los elementos pertenecientes a los conglomerados seleccionados en la muestra s_I.
- En la mayoría de situaciones, los conglomerados tienden a ser muy similares en el comportamiento estructural de la característica de interés.
- Se consideraría un desperdicio de recursos económicos y logísticos la incorporación de elementos que no traen consigo nueva información.

Diseño de muestreo MAS-MAS

- Se supone que la población está divida en N_I unidades primarias de muestreo, de las cuales se selecciona una muestra s_I de n_I unidades mediante un diseño de muestreo aleatorio simple.
- El sub-muestreo dentro de cada unidad primaria seleccionada es también aleatorio simple.
- Para cada unidad primaria de muestreo seleccionada $i \in s_{Ih}$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i.

Probabilidades de inclusión

- Cuando el diseño de muestreo es aleatorio simple en las dos etapas, se tienen las siguientes probabilidades de inclusión de primer y segundo orden

$$\pi_{Ii} = \frac{n_I}{N_I} \tag{14}$$

- La probabilidad de inclusión de un elemento o unidad secundaria de muestreo perteneciente a la i-ésima unidad primaria de muestreo $i \in U_I$ está dado por

$$\pi_k = \frac{n_I}{N_I}\frac{n_i}{N_i} \tag{15}$$

Estimación de subtotales

Como no se miden todos los elementos de las unidades primarias seleccionadas, se deben estimar los totales t_{yi} mediante la siguiente expresión

$$\hat{t}_{yi,\pi} = \frac{N_i}{n_i} \sum_{k \in S_i} y_k = N_i \bar{y}_{U_i} \tag{16}$$

Estimador del total poblacional

El estimador de Horvitz-Thompson toma la forma

$$\hat{t}_{y,\pi} = \frac{N_I}{n_I} \sum_{i \in S_I} \frac{N_i}{n_i} \sum_{k \in S_i} y_k \tag{17}$$

con estimación de varianza dada por

$$\widehat{Var}_{MM}(\hat{t}_{y,\pi}) = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S^2_{\hat{t}_y S_I} + \frac{N_I}{n_I} \sum_{i \in S_I} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S^2_{y_{S_i}} \tag{18}$$

donde $S^2_{\hat{t}_y S_I}$ y $S^2_{y_{S_i}}$ son las varianzas muestrales de los totales y de los elementos, respectivamente.

Práctica en R

```
UI <- levels(as.factor(Hogares$UPM))
NI <- length(UI)
nI <- 200

samI <- S.SI(NI, nI)
muestraI <- UI[samI]
```

```
muestraI
```

```
##   [1] "PSU0010" "PSU0019" "PSU0039" "PSU0040" "PSU0041"
##   [8] "PSU0066" "PSU0072" "PSU0078" "PSU0079" "PSU0086"
##  [15] "PSU0111" "PSU0116" "PSU0119" "PSU0141" "PSU0155"
##  [22] "PSU0191" "PSU0205" "PSU0213" "PSU0218" "PSU0220"
##  [29] "PSU0230" "PSU0231" "PSU0233" "PSU0234" "PSU0240"
##  [36] "PSU0247" "PSU0255" "PSU0269" "PSU0278" "PSU0284"
##  [43] "PSU0291" "PSU0295" "PSU0299" "PSU0322" "PSU0350"
##  [50] "PSU0383" "PSU0387" "PSU0438" "PSU0440" "PSU0450"
##  [57] "PSU0475" "PSU0480" "PSU0488" "PSU0491" "PSU0492"
##  [64] "PSU0523" "PSU0527" "PSU0528" "PSU0530" "PSU0538"
##  [71] "PSU0550" "PSU0552" "PSU0573" "PSU0575" "PSU0584"
##  [78] "PSU0630" "PSU0636" "PSU0637" "PSU0638" "PSU0639"
##  [85] "PSU0659" "PSU0661" "PSU0668" "PSU0687" "PSU0688"
##  [92] "PSU0701" "PSU0705" "PSU0723" "PSU0732" "PSU0734"
##  [99] "PSU0747" "PSU0754" "PSU0761" "PSU0764" "PSU0767"
## [106] "PSU0786" "PSU0789" "PSU0791" "PSU0795" "PSU0796"
```

Práctica en R

```
CityI <- Hogares[which(Hogares$UPM %in% muestraI),]
head(CityI)
```

HHID	Estrato	UPM	Personas	Ingreso	Gasto	Pobreza
idHH00125	Urban	PSU0010	5	2102	1506	Relative
idHH00126	Urban	PSU0010	1	744	220	NotPoor
idHH00127	Urban	PSU0010	7	9663	4306	NotPoor
idHH00128	Urban	PSU0010	4	5107	2958	NotPoor
idHH00129	Urban	PSU0010	6	4440	4049	NotPoor
idHH00130	Urban	PSU0010	3	1615	1411	NotPoor

```r
Ni <- as.vector(table(CityI$UPM))
ni <- round(Ni * 0.4)
ni
```

```
##   [1]  9 13 10 10 12 13 12 10 14 10 10 10 10  9 14 10 10
##  [26] 10 11 10 10 10 12 14 10 14  9 10 11 11 11  9 10 10
##  [51] 14 10  3  6 13 12  9 12  7  9  6 11  9  7  6  6  7
##  [76] 11 10  8 12 10 10  9  4 10  7  5  8  6  8  6  7  7
## [101]  9  9 10 12 11  8  9 10 10 12  6  5  9  5  6  8  6
## [126]  9  8 14 10 11 11 10 14 11 11 12 10 10 12 10 12  8
## [151]  9 11  9 14 10 10 10 11 10  8 13 11  8 14 14 10  7
## [176] 10 12  8  5 10 10 11 10 10 10 12 10 12 14 10 12 10
```

```r
sum(ni)
```

```
## [1] 1965
```

Práctica en R

```
sam = S.SI(Ni[1], ni[1])
conglomerado = Hogares[which(Hogares$UPM == muestraI[1]),]
muestra = conglomerado[sam,]
for (i in 2:length(Ni)) {
  sam = S.SI(Ni[i], ni[i])
  conglomerado = Hogares[which(Hogares$UPM == muestraI[i]),
  muestra1 = conglomerado[sam,]
  muestra = rbind(muestra, muestra1)
}
```

Práctica en R

```r
rownames(muestra) <- NULL
attach(muestra)
dim(muestra)
```

```
## [1] 1965    7
```

```r
head(muestra)
```

HHID	Estrato	UPM	Personas	Ingreso	Gasto	Pobreza
idHH00129	Urban	PSU0010	6	4440	4049	NotPoor
idHH00131	Urban	PSU0010	2	3094	749	NotPoor
idHH00133	Urban	PSU0010	3	4950	2411	NotPoor
idHH00135	Urban	PSU0010	1	260	227	NotPoor
idHH20773	Urban	PSU0010	4	5107	2958	NotPoor
idHH20775	Urban	PSU0010	3	1615	1411	NotPoor

```r
estima <- data.frame(Personas, Ingreso, Gasto)
area <- as.factor(UPM)
E.2SI(NI, nI, Ni, ni, estima, area)
```

	N	Personas	Ingreso	Gasto
Estimation	40884	144114	80755918	52423458
Standard Error	NA	NA	NA	NA
CVE	NA	NA	NA	NA
DEFF	NA	NA	NA	NA

Otros diseños en varias etapas

Diseños en r etapas

- Los principios de independencia e invarianza se siguen manteniendo en todas las etapas del diseño muestral.
- El fundamento de este diseño de muestreo es la acumulación de las estimaciones desde la última etapa hasta la primera.
- A pesar de su complejidad, los diseños con tres o más etapas son ampliamente usados en las grandes encuestas.

Diseños en r etapas sin reemplazo

El estimador de Horvitz-Thompson es

$$\hat{t}_{y,\pi} = \sum_{i \in S_I} \frac{\hat{t}_{yi}}{\pi_{Ii}} \tag{19}$$

con varianza estimada

$$\widehat{Var}_{BI}(\hat{t}_{y,\pi}) = \underbrace{\sum\sum_{S_I} \frac{\Delta_{Iij}}{\pi_{Iij}} \frac{\hat{t}_{yi}}{\pi_{Ii}} \frac{\hat{t}_{yj}}{\pi_{Ij}}}_{\widehat{Var}(UPM)} + \underbrace{\sum_{i \in S_I} \frac{\hat{V}_i}{\pi_{Ii}}}_{\widehat{Var}(Resto)} \tag{20}$$

Donde $V_i = Var(\hat{t}_{yi}|S_I)$ y $\hat{V}_i$ es un estimador insesgado de V_i tal que $E(\hat{V}_i|S_I) = V_i$ para todo $i \in U_i$.

Diseños en r etapas con reemplazo

El estimador de Hansen-Hurwitz es

$$\hat{t}_{y,p} = \frac{1}{m_I} \sum_{v=1}^{m_I} \frac{\hat{t}_{yi_v}}{p_{Ii_v}} \tag{21}$$

con varianza estimada

$$\widehat{Var}(\hat{t}_{y,p}) = \frac{1}{m_I(m_I - 1)} \sum_{v=1}^{m_I} \left(\frac{\hat{t}_{yi_v}}{p_{Ii_v}} - \hat{t}_{y,p} \right)^2 \tag{22}$$

respectivamente.

¡Gracias!

Andrés Gutiérrez

Experto Regional en Estadísticas Sociales

Division de Estadísticas

Email: andres.GUTIERREZ@cepal.org