Muestreo Estratificado

Curso de Muestreo Probabilístico en Encuestas de Hogares

Andrés Gutiérrez, Ph.D.

CEPAL - Unidad de Estadísticas Sociales

1. Estratificación
2. Muestreo aleatorio estratificado
3. Muestreo estratificado proporcional al tamaño
4. Estratificación implícita y sistemática

Motivación

La estratificación es una de las técnicas más difundidas y usadas en muestreo puesto que tiene funcionalidades estadísticas y administrativas que la hacen atractiva: permite tratar con subpoblaciones, aumenta la eficiencia de las estimaciones y contribuye a la administración eficiente de grandes encuestas..

Richard Valliant (2000)

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*.

Estratificación

Características

- En algunas ocasiones, la característica de interés tiende a tomar distintos valores promedio con respecto a subgrupos poblacionales.
- Si la población tiene un comportamiento diferente en estos subgrupos, es posible mejorar la precisión de las estimaciones tomando muestras independientes en cada uno de estos
- Cuando existe mucha variabilidad entre los subgrupos, pero dentro de ellos la variabilidad es constante, se incrementa la eficiencia.

Características

- Cuando existe en el marco de muestreo información auxiliar que permite la división de la población en H subgrupos, se dice que la estrategia de muestreo utiliza un **diseño de muestreo estratificado**
- El nombre de los subgrupos, formados antes de la recolección de la información, se denomina **estratos**.
- Nótese la diferencia con los subgrupos poblacionales llamados **dominios**, en donde la partición de la población se realiza después de la recolección de la información.

Características

No es solamente la disponibilidad de esta información auxiliar la que nos lleva a utilizar un diseño de muestreo estratificado, además de esto:

1. La variable de interés asume distintos valores promedio en diferentes sub-poblaciones.
2. Debido al proceso logístico y/o de recolección de datos es mejor estratificar y dividir la población en particiones.

Características

Algunas variables típicas de estratificación son de tipo:

- regional (municipio, estado o provincia),
- demográfico (género o grupo de edad) y
- socio-económico (grupo de ingresos).

Características

La necesidad de estratificar la población surge por una o más de las siguientes razones:

1. Por razones administrativas: existen marcos de muestreo que ya tienen dividida la población en subgrupos formados naturalmente.
2. Porque se desea garantizar que la muestra seleccionada sea representativa con respecto al comportamiento de cada subgrupo: al seleccionar una muestra aleatoria simple de una población de personas, podría suceder que la muestra seleccionada no incluyera a ningún hombre.

Características

3. Porque se requieren estimativos con alta precisión discriminados para cada sub-población: aumentar el tamaño de muestra en los estratos menos representados.
4. Porque hay un menor Coste: distintos esquemas operativos para diversos estratos. Encuestas por correo para empresas grandes. Menor tamaño de muestras en zonas de tolerancia o zonas de difícil manejo del orden público.
5. Porque ser quiere tener una reducción de la varianza en la estimación: se reduce la varianza pues los estratos son homogéneos por dentro, pero heterogéneos entre sí.

Fundamentos teóricos

Suponga que la población U se particiona en H sub-grupos poblacionales separados U_h $(h = 1, 2, \ldots, H)$ llamados estratos, de tal forma que

1. $\bigcup_{h=1}^{H} U_h = U$
2. $U_h \bigcap U_i = \emptyset \qquad h \neq i$

Cada estrato U_h es de tamaño N_h, por tanto

$$\sum_{h=1}^{H} N_h = N \tag{1}$$

Parámetros de interés: total poblacional

El total poblacional se define como

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} t_{yh} \tag{2}$$

donde $t_{yh} = \sum_{k \in U_h} y_k$

Parámetros de interés: media poblacional

La media poblacional se define como:

$$\bar{y} = \frac{\sum_{k \in U} y_k}{N} = \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h \tag{3}$$

donde $\bar{y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k$

Independencia en el muestreo

Dependiendo de la naturaleza de los estratos, diferentes estrategias de muestreo pueden ser utilizadas en diferentes estratos.

- En ausencia de información auxiliar, se puede usar una estrategia aleatoria simple en algunos estratos.
- Para los sub-grupos en donde el marco de muestreo permita el conocimiento de información auxiliar continua, es posible aplicar una estrategia de muestreo proporcional al tamaño.
- Pueden existir estratos en los que, por obligación (logística o técnica), se deba aplicar un censo.

Independencia en el muestreo

- La selección de las H muestras es realizada de manera independiente en cada estrato.
- Aunque se conozcan qué unidades serán incluidas en la muestra de algún estrato, este conocimiento no afecta la inclusión de cualquier otra unidad en los restantes estratos.
- La muestra aleatoria S queda definida por

$$S = \bigcup_{h=1}^{H} S_h. \tag{4}$$

Tamaño de muestra

Nótese que si el tamaño de muestra en cada estrato es igual a n_h, entonces el tamaño de la muestra seleccionada mediante un diseño de muestreo estratificado es

$$n = \sum_{h=1}^{H} n_h. \tag{5}$$

Estimación: insesgamiento

Si $\hat{t}_{yh}$ estima insesgadamente el total de la característica de interés t_{yh} del subgrupo poblacional h, entonces un estimador insesgado para el total poblacional t_y está dado por

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} \tag{6}$$

Estimación: varianza

Si $\hat{t}_{yh}$ tiene varianza igual a $Var(\hat{t}_{yh})$, entonces la varianza de $\hat{t}_y$ es

$$Var(\hat{t}_y) = \sum_{h=1}^{H} Var(\hat{t}_{yh}) \tag{7}$$

Muestreo aleatorio estratificado

Características

- El diseño de muestreo aleatorio estratificado es el más sencillo de los diseños estratificados.
- Se selecciona una muestra aleatoria simple en cada estrato, de tal forma que las selecciones sean independientes.
- Es utilizado cuando la variabilidad de la característica de interés dentro de los estratos es similar.

Características

- En cada estrato h una muestra aleatoria simple sin reemplazo de tamaño n_h es seleccionada, de manera independiente, de la población del estrato de tamaño N_h.
- El diseño estratificado puede resultar mucho más eficiente que utilizar un diseño de muestreo aleatorio simple sin dividir la población.

Diseño de muestreo aleatorio estratificado

Para tamaños de muestra fijos en cada estrato, denotados como $n_1, \ldots, n_H$, la probabilidad de seleccionar una muestra de tamaño n está dada por

$$p(s) = \begin{cases} \prod_{h=1}^{H} \frac{1}{\binom{N_h}{n_h}}, & \text{si } \sum_{h=1}^{H} n_h = n \\ 0, & \text{en otro caso} \end{cases} \tag{8}$$

Algoritmos de selección

- Separar la población en H subgrupos o estratos mediante la caracterización poblacional de información auxiliar.
- En cada estrato seleccionar una muestra aleatoria simple sin reemplazo.
- Cada una de las H selecciones es realizada de manera independiente

Suponga que nuestra población de ejemplo U está particionada de acuerdo a la sección anterior. Es necesario definir los dos estratos en R, de manera tal que ningún elemento tenga una doble pertenencia a algún estrato.

Ejemplo

Considere la siguiente estratificación de la población.

```
U1  <-  c("Erik", "Sharon")
N1  <-  length(U1)
U2  <-  c("Yves", "Ken", "Leslie")
N2  <-  length(U2)
```

Ejemplo

```
U <- union(U1,U2)
N <-  N1+N2

U
```

```
## [1] "Erik"   "Sharon" "Yves"   "Ken"    "Leslie"
```

```
N
```

```
## [1] 5
```

Ejemplo

- Se seleccionará una muestra aleatoria simple sin reemplazo de tamaño $n_1 = 1$ para U_1 y,
- Se seleccionará una muestra aleatoria simple sin reemplazo de tamaño $n_2 = 2$ para U_2.
- De tal forma que la muestra general será de tamaño $n = n_1 + n_2 = 3$.

Ejemplo

```
library(TeachingSampling)

sam1 <- sample(N1, 1, replace=FALSE)
U1[sam1]
```

```
## [1] "Sharon"
```

```
sam2 <- S.SI(N2,2)
U2[sam2]
```

```
## [1] "Ken"    "Leslie"
```

```
sam <- union(U1[sam1],U2[sam2])
sam
```

```
## [1] "Sharon" "Ken"    "Leslie"
```

Estimador del total en cada estrato

Bajo un diseño de muestreo aleatorio simple sin reemplazo en el estrato h, un estimador insesgado del total t_{yh} y su varianza estimada están dados por

$$\hat{t}_{yh,\pi} = \frac{N_h}{n_h} \sum_{k \in S_h} y_k \tag{9}$$

$$\widehat{Var}_{MAS}(\hat{t}_{yh,\pi}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{yS_h}^2 \tag{10}$$

respectivamente. En donde la **varianza muestral** de los valores de la característica de interés en la muestra aleatoria del estrato S_h es

$$S_{yS_h}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_{S_h})^2, \qquad h = 1, \ldots, H. \tag{11}$$

Estimador del total poblacional

El estimador de Horvitz-Thompson del total poblacional t_y y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \sum_{h=1}^{H} \hat{t}_{yh,\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} y_k \tag{12}$$

$$\widehat{Var}_{MAE}(\hat{t}_{y,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ys_h}^2 \tag{13}$$

respectivamente.

Asignación del tamaño de muestra

- El tamaño de la muestra general es n y existen H estratos fijos.
- Se quiere determinar los tamaños de muestra n_h para cada estrato h de tal manera que se garantice la ganancia de precisión del estimador.
- Cuando se incluyen varias características de interés, es imposible lograr ganancias en la eficiencia de manera global.

Asignación proporcional

- Se decide utilizar este tipo de asignación cuando la muestra debe ser representativa de la población de acuerdo al comportamiento de la información auxiliar
- Al utilizar la asignación proporcional, la muestra se puede ver como una versión miniatura de la población.

Asignación proporcional

- Se define la **fracción de muestreo** como $f_h = n_h/N_h$ en el estrato h.
- Al utilizar la asignación proporcional la fracción de muestreo será la misma para todos los estratos, tal que $f_h = f$.
- La probabilidad de inclusión de cualquier elemento en la población $\pi_k = f_h = f$ es constante y fija.

Asignación proporcional

Un diseño de muestreo aleatorio estratificado tiene asignación proporcional si

$$\frac{n_h}{N_h} = \frac{n}{N} \qquad h = 1, \ldots, H \tag{14}$$

Cada unidad en la muestra representará el mismo número de elementos en la población, independientemente del estrato al que pertenezca.

Asignación de Neyman

- Neyman trató con el problema de minimizar la varianza $Var_{MAE}(\hat{t}_{y,\pi})$ del estimador de Horvitz-Thompson fijando el tamaño de muestra general n.
- Bajo este método se producen las menores varianzas para la media muestral comparado con otras técnicas de asignación de tamaño de muestra.

Asignación de Neyman

Bajo la asignación de Neyman, el tamaño de muestra que minimiza la varianza del muestreo aleatorio estratificado está dado por

$$n_h = n\frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}} \tag{15}$$

donde $S_{yU_h} = \sqrt{S^2_{yU_h}}$.

Asignación de Neyman

- Al estimar proporciones se tienen buenos resultados si hay grandes diferencias entre los estratos.
- Este método funciona bien bajo el supuesto de que sólo existe una característica de interés.
- Cuando se trabaja en encuesta multi-propósito no se tiene una reducción de varianza para todas las características de interés incluidas en la investigación.

Estimación en dominios

- La estimación por dominios se caracteriza por el desconocimiento de la pertenencia de las unidades poblacionales al dominio.
- Para conocer cuáles unidades de la población pertenecen al dominio, es necesario realizar el proceso de medición.

Estimación en dominios

- Tanto estratos como dominios dividen la población en subgrupos poblacionales.
- El conocimiento a priori de la pertenencia de los elementos poblacionales a los estratos ayuda a mejorar la eficiencia de la estimación.
- El precio que se debe pagar por el desconocimiento de la pertenencia de los elementos poblacionales a los dominios resulta alto.

Estimación del total en un dominio

El estimador de Horvitz-Thompson para el total del dominio t_{yhd} en el estrato h y su varianza estimada están dados por

$$\hat{t}_{yhd,\pi} = \frac{N_h}{n_h} \sum_{S_h} y_{hdk} \tag{16}$$

$$\widehat{Var}(\hat{t}_{yhd,\pi}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{y_{ds_h}} \tag{17}$$

respectivamente. En donde y_{hdk} es el valor de la nueva característica y_{dk} en el h-ésimo estrato.

Estimación del total en un dominio

El estimador de Horvitz-Thompson para el total del dominio t_{yd} en la población y su varianza estimada están dados por

$$\hat{t}_{yd,\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{S_h} y_{hdk} \tag{18}$$

$$\widehat{Var}(\hat{t}_{yd,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{y_{ds_h}} \tag{19}$$

Estimación del tamaño de un dominio

El estimador de Horvitz-Thompson para el tamaño absoluto de un dominio N_{hd} en el estrato h y su varianza estimada están dados por

$$\hat{N}_{hd,\pi} = \frac{N_h}{n_h} \sum_{S_h} z_{dk} \tag{20}$$

$$\widehat{Var}(\hat{N}_{hd,\pi}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{ds_h}} \tag{21}$$

respectivamente, con $S^2_{z_{ds_h}}$ la varianza de los valores de la característica de interés z_{dk} en la muestra s_h.

Estimación del tamaño de un dominio

El estimador de Horvitz-Thompson para el tamaño absoluto de un dominio N_d en la población y su varianza estimada están dados por

$$\hat{N}_{d,\pi} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{S_h} z_{dk} \tag{22}$$

$$\widehat{Var}(\hat{N}_{d,\pi}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{z_{ds_h}} \tag{23}$$

respectivamente.

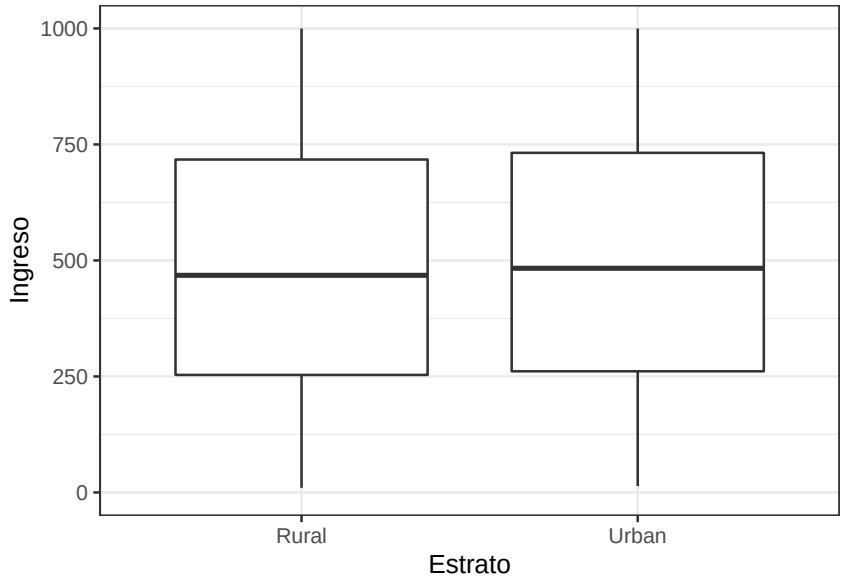
Práctica en R

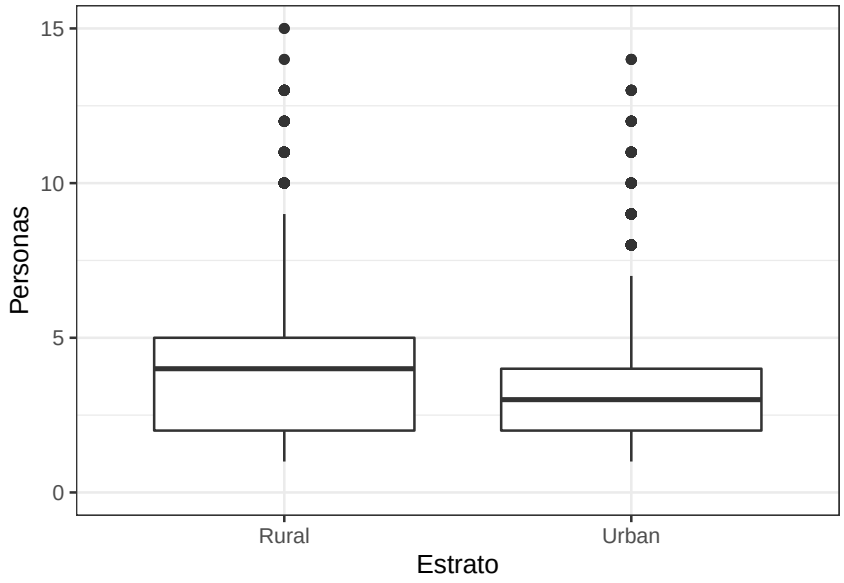
```r
library(TeachingSampling)
library(dplyr)
data("BigCity")

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

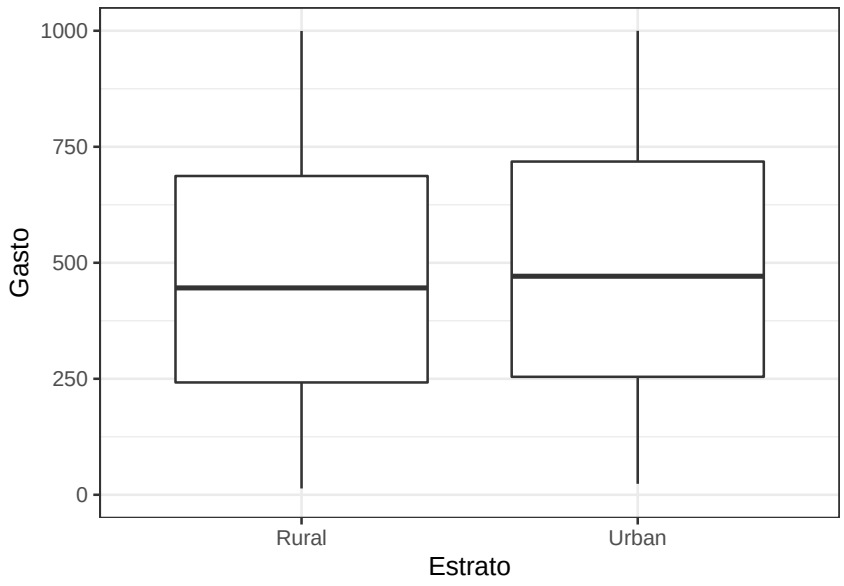
Práctica en R

```
library(ggplot2)
attach(Hogares)

q1 = qplot(Estrato, Ingreso , data =Hogares, geom=c("boxplot"),
           ylim = c(0,1000))
q2 = qplot(Estrato, Personas, data=Hogares, geom=c("boxplot"),
           ylim = c(0,15))
q3 = qplot(Estrato, Gasto, data=Hogares, geom=c("boxplot"),
           ylim = c(0,1000))
```

Práctica en R







Práctica en R

```
tamano.muestra = Hogares %>% group_by(Estrato) %>%
  summarise(Nh = n(),
  nh = round(2000*Nh/dim(Hogares)[1]))

tamano.muestra
```

Estrato	Nh	nh
Rural	19174	929
Urban	22116	1071

```
sum(tamano.muestra$Nh)
```

```
## [1] 41290
```

```
sum(tamano.muestra$nh)
```

```
## [1] 2000
```

Práctica en R

```
Nh <- tamano.muestra$Nh
nh <- tamano.muestra$nh

sam <- S.STSI(Estrato, Nh, nh)
muestra <- Hogares[sam,]
rownames(muestra) <- NULL
attach(muestra)
```

Práctica en R

```
head(muestra)
```

HHID	Estrato	Personas	Ingreso	Gasto	Pobreza
idHH00023	Rural	3	1115	630	NotPoor
idHH00039	Rural	5	829	849	Relative
idHH00074	Rural	4	1031	421	Relative
idHH00076	Rural	1	236	126	NotPoor
idHH00085	Rural	2	705	330	NotPoor
idHH00086	Rural	4	1260	650	NotPoor

Práctica en R

A través del siguiente codigo es posible estimar el tamaño poblacional y el total para cada variable.

```
estima <- data.frame(Ingreso, Gasto, Personas)
E.STSI(Estrato, Nh, nh, estima)
```

```
## , , N
##
##                Rural Urban Population
## Estimation     19174 22116      41290
## Standard Error     0     0          0
## CVE                0     0          0
## DEFF             NaN   NaN        NaN
##
## , , Ingreso
##
##                     Rural      Urban  Population
## Estimation     32236820.8 56241194.7 88478015.53
## Standard Error  1321062.0  2220359.4  2583641.01
## CVE                   4.1        3.9        2.92
## DEFF                  1.0        1.0        0.98
##
## , , Gasto
##
##                     Rural      Urban  Population
## Estimation     21476136.1 35494483.8 56970619.93
## Standard Error   721711.3  1085292.7  1303352.39
## CVE                   3.4        3.1        2.29
## DEFF                  1.0        1.0        0.97
##
## , , Personas
##
```

Práctica en R:

Estimación del tamaño poblacional

```
E.STSI(Estrato, Nh, nh, estima)[, , "N"]
```

	Rural	Urban	Population
Estimation	19174	22116	41290
Standard Error	0	0	0
CVE	0	0	0
DEFF	NaN	NaN	NaN

Práctica en R:

Estimación del total para el ingreso

```
E.STSI(Estrato, Nh, nh, estima)[, , "Ingreso"]
```

	Rural	Urban	Population
Estimation	32236820.8	56241195	88478015.53
Standard Error	1321062.0	2220359	2583641.01
CVE	4.1	4	2.92
DEFF	1.0	1	0.98

Práctica en R:

Estimación del total para el gasto

```
E.STSI(Estrato, Nh, nh, estima)[, , "Gasto"]
```

	Rural	Urban	Population
Estimation	21476136.1	35494483.8	56970619.93
Standard Error	721711.3	1085292.7	1303352.39
CVE	3.4	3.1	2.29
DEFF	1.0	1.0	0.97

Práctica en R:

Estimación de los totales poblacionales

```
E.STSI(Estrato, Nh, nh, estima)[, 3, ]
```

	N	Ingreso	Gasto	Personas
Estimation	41290	88478015.53	56970619.93	148973.0
Standard Error	0	2583641.01	1303352.39	1614.0
CVE	0	2.92	2.29	1.1
DEFF	NaN	0.98	0.97	1.0

Práctica en R:

En R es posible encontrar el estimador del total para cada nivel según el dominio.

```
Dominios <- Domains(Pobreza)
dom.NP <- Dominios[,3]*estima
dom.PR <- Dominios[,2]*estima
dom.PE <- Dominios[,1]*estima

E.STSI(Estrato, Nh, nh, Dominios)
```

Práctica en R:

Estimación del tamaño para `No pobre`

```
E.STSI(Estrato, Nh, nh, Dominios)[, , "NotPoor"]
```

	Rural	Urban	Population
Estimation	13167.9	15879.7	29047.7
Standard Error	284.8	296.8	411.3
CVE	2.2	1.9	1.4
DEFF	1.0	1.0	1.0

Práctica en R:

Estimación del tamaño para `Pobreza relativa`

```
E.STSI(Estrato, Nh, nh, Dominios)[, , "Relative"]
```

	Rural	Urban	Population
Estimation	4747.1	4997.3	9744.3
Standard Error	265.0	275.8	382.5
CVE	5.6	5.5	3.9
DEFF	1.0	1.0	1.0

Práctica en R:

Estimación del tamaño para Pobreza extrema

```
E.STSI(Estrato, Nh, nh, Dominios)[, , "Extreme"]
```

	Rural	Urban	Population
Estimation	1259	1239	2498.0
Standard Error	152	152	214.8
CVE	12	12	8.6
DEFF	1	1	1.0

Práctica en R:

Estimación del total para ingreso en Pobreza relativa

```
E.STSI(Estrato, Nh, nh, dom.PR)[, , "Ingreso"]
```

	Rural	Urban	Population
Estimation	749138	1437082	2186220
Standard Error	117108	267064	291611
CVE	16	19	13
DEFF	1	1	1

Práctica en R:

Estimación del total para `gasto` en `Pobreza relativa`

```
E.STSI(Estrato, Nh, nh, dom.PR)[, , "Gasto"]
```

	Rural	Urban	Population
Estimation	1142338	1721873	2864212
Standard Error	175872	291465	340416
CVE	15	17	12
DEFF	1	1	1

Práctica en R:

Estimación del total para `ingreso` en `No pobre`

```
E.STSI(Estrato, Nh, nh, dom.NP)[, , "Ingreso"]
```

	Rural	Urban	Population
Estimation	4847436.9	7326176.9	12173613.9
Standard Error	396130.8	597453.8	716847.7
CVE	8.2	8.2	5.9
DEFF	1.0	1.0	1.0

Práctica en R:

Estimación del total para `gasto` en `No pobre`

```
E.STSI(Estrato, Nh, nh, dom.NP)[, , "Gasto"]
```

	Rural	Urban	Population
Estimation	4670194.3	6570661.3	11240855.6
Standard Error	381620.0	510010.9	636981.1
CVE	8.2	7.8	5.7
DEFF	1.0	1.0	1.0

Práctica en R:

Estimación del total para ingreso en Pobreza extrema

```
E.STSI(Estrato, Nh, nh, dom.PE)[, , "Ingreso"]
```

	Rural	Urban	Population
Estimation	26640246.0	47477935.4	74118181.38
Standard Error	1373962.6	2293572.0	2673620.36
CVE	5.2	4.8	3.61
DEFF	1.0	1.0	0.98

Práctica en R:

Estimación del total para `gasto` en `Pobreza extrema`

```
E.STSI(Estrato, Nh, nh, dom.PE)[, , "Gasto"]
```

	Rural	Urban	Population
Estimation	15663603.4	27201949.2	42865552.64
Standard Error	736175.4	1119881.4	1340182.32
CVE	4.7	4.1	3.13
DEFF	1.0	1.0	0.98

Muestreo estratificado proporcional al tamaño

Características

- La ganancia de precisión al utilizar un diseño de muestreo estratificado es importante.
- Los resultados pueden mejorarse al utilizar una característica continua auxiliar x_k bien relacionada con la característica de interés y_k en cada estrato.
- Es posible estimar el parámetro de interés mediante el estimador de Hansen-Hurwitz con una varianza pequeña.

Características

- Entre mejor correlación exista entre y y x, asumiendo que el comportamiento promedio de la variable de interés es diferente en cada estrato, menor varianza tendrá el estimador de Hansen-Hurwitz.
- El marco de muestreo debe tener dos características auxiliares: una *variable de estratificación* y la *información auxiliar continua*, ambas disponibles para cada elemento en todos los estratos.

Algoritmo de selección

En la selección de las muestras PPT con reemplazo en cada estrato es posible utilizar los algoritmos de muestreo vistos anteriormente:

- Separar la población en H estratos mediante la variable de estratificación.
- En cada estrato U_h, seleccionar una muestra PPT con reemplazo. Los algoritmos utilizados en la selección de la muestra dentro de cada estrato pueden ser los métodos acumulativo total.
- Cada una de las H selecciones es realizada de manera independiente.

Estimador del total en cada dominio

El estimador de Hansen-Hurwitz del total poblacional t_{yh} y su varianza estimada están dados por:

$$\hat{t}_{yh,p} = \frac{t_{xh}}{m_h} \sum_{\substack{i=1 \\ k \in S_h}}^{m_h} \frac{y_{ki}}{x_{ki}} \tag{24}$$

$$\widehat{Var}_{PPT}(\hat{t}_{yh,p}) = \frac{1}{m_h(m_h-1)} \sum_{\substack{i=1 \\ k \in S_h}}^{m_h} \left(\frac{y_{ki}}{p_{ki}} - \hat{t}_{yh,p} \right)^2 \tag{25}$$

respectivamente.

Estimador del total poblacional

El estimador de Hansen-Hurwitz del total poblacional t_y y su varianza estimada están dados por:

$$\hat{t}_{yh,p} = \sum_{h=1}^{H} \frac{t_{xh}}{m_h} \sum_{\substack{i=1 \\ k \in S_h}}^{m_h} \frac{y_{ki}}{x_{ki}} \tag{26}$$

$$\widehat{Var}_{EPPT}(\hat{t}_{yh,p}) = \sum_{h=1}^{H} \frac{1}{m_h(m_h-1)} \sum_{\substack{i=1 \\ k \in S_h}}^{m_h} \left(\frac{y_{ki}}{p_{ki}} - \hat{t}_{yh,p} \right)^2 \tag{27}$$

respectivamente.

Práctica en R

```
attach(Hogares)

tamano.muestra = Hogares %>%
  group_by(Estrato) %>%
  summarise(Nh = n(),
            mh = round(2000*Nh/dim(Hogares)[1]))

sum(tamano.muestra$Nh)
```

```
## [1] 41290
```

```
sum(tamano.muestra$mh)
```

```
## [1] 2000
```

Práctica en R

```
Nh <- tamano.muestra$Nh
mh <- tamano.muestra$mh

res <- S.STPPS(Estrato, Personas, mh)
sam <- res[,1]
pk <- res[,2]
muestra <- Hogares[sam,]
```

Práctica en R

```
rownames(muestra) <- NULL
attach(muestra)
head(muestra)
```

HHID	Estrato	Personas	Ingreso	Gasto	Pobreza
idHH01828	Rural	4	1182	1226	NotPoor
idHH33336	Rural	5	5514	2409	NotPoor
idHH00924	Rural	5	1101	1291	Relative
idHH40682	Rural	3	521	445	Relative
idHH39290	Rural	7	6329	2028	NotPoor
idHH21551	Rural	5	2058	916	NotPoor

Práctica en R

Estimación del tamaño poblacional y total de las variables.

```
estima <- data.frame(Ingreso, Gasto)
E.STPPS(estima, pk, mh, Estrato)
```

```
## , , N
##
##                    Rural   Urban Population
## Estimation       19588.2 22547.5    42135.7
## Standard Error     419.2   423.0      595.5
## CVE                  2.1     1.9        1.4
## DEFF                 Inf     Inf        Inf
##
## , , Ingreso
##
##                      Rural       Urban  Population
## Estimation     31449583.08 54499364.40 85948947.48
## Standard Error   831322.20  1543995.80  1753573.39
```

Práctica en R:

Estimación del tamaño poblacional

```
E.STPPS(estima, pk, mh, Estrato)[, , "N"]
```

	Rural	Urban	Population
Estimation	19588.2	22547.5	42135.7
Standard Error	419.2	423.0	595.5
CVE	2.1	1.9	1.4
DEFF	Inf	Inf	Inf

Práctica en R:

Estimación del total para Ingreso

```
E.STPPS(estima, pk, mh, Estrato)[, , "Ingreso"]
```

	Rural	Urban	Population
Estimation	31449583.08	54499364.40	85948947.48
Standard Error	831322.20	1543995.80	1753573.39
CVE	2.64	2.83	2.04
DEFF	0.24	0.22	0.22

Práctica en R:

Estimación del total para `Gasto`

```
E.STPPS(estima, pk, mh, Estrato)[, , "Gasto"]
```

	Rural	Urban	Population
Estimation	20889312.28	34363019.64	55252331.92
Standard Error	428907.20	701696.74	822398.75
CVE	2.05	2.04	1.49
DEFF	0.23	0.21	0.21

Práctica en R:

Estimación para el tamaño poblacional y el total de las variables por dominio.

```
Dominios <- Domains(Pobreza)
dom.NP <- Dominios[,3] * estima
dom.PR <- Dominios[,2] * estima
dom.PE <- Dominios[,1] * estima
```

Práctica en R:

Estimación de los tamaños en los dominios

```
E.STPPS(Dominios, pk, mh, Estrato)[, 3 , ]
```

	N	NotPoor	Extreme	Relative
Estimation	42135.7	28862.7	2918.94	10354.1
Standard Error	595.5	722.3	233.08	438.2
CVE	1.4	2.5	7.99	4.2
DEFF	Inf	2.7	0.82	1.1

Práctica en R:

Estimación del tamaño para `No pobre`

```
E.STPPS(Dominios, pk, mh, Estrato)[, , "NotPoor"]
```

	Rural	Urban	Population
Estimation	13138.9	15723.8	28862.7
Standard Error	504.4	517.1	722.3
CVE	3.8	3.3	2.5
DEFF	2.7	2.6	2.7

Práctica en R:

Estimación del tamaño `Pobreza relativa`

```
E.STPPS(Dominios, pk, mh, Estrato)[, , "Relative"]
```

	Rural	Urban	Population
Estimation	4903.6	5450.5	10354.1
Standard Error	303.9	315.7	438.2
CVE	6.2	5.8	4.2
DEFF	1.1	1.1	1.1

Práctica en R:

```
E.STPPS(dom.PR, pk, mh, Estrato)
```

```
## , , N
##
##                    Rural   Urban Population
## Estimation      19588.2 22547.5    42135.7
## Standard Error    419.2   423.0      595.5
## CVE                 2.1     1.9        1.4
## DEFF                Inf     Inf        Inf
##
## , , Ingreso
##
##                      Rural      Urban Population
## Estimation      964502.01 1075845.35 2040347.37
## Standard Error 106521.55  131871.55  169519.75
## CVE                11.04      12.26       8.31
## DEFF                0.24       0.24       0.24
##
```

Práctica en R:

Estimación del total para Ingreso en Pobreza relativa

```
E.STPPS(dom.PR, pk, mh, Estrato)[, , "Ingreso"]
```

	Rural	Urban	Population
Estimation	964502.01	1075845.35	2040347.37
Standard Error	106521.55	131871.55	169519.75
CVE	11.04	12.26	8.31
DEFF	0.24	0.24	0.24

Práctica en R:

Estimación del total para `Gasto` en `Pobreza relativa`

```
E.STPPS(dom.PR, pk, mh, Estrato)[, , "Gasto"]
```

	Rural	Urban	Population
Estimation	1248274.22	1397779.59	2646053.81
Standard Error	142075.12	177648.85	227474.07
CVE	11.38	12.71	8.60
DEFF	0.25	0.24	0.24

Práctica en R:

Estimación del total para Ingreso en No pobre

```
E.STPPS(dom.NP, pk, mh, Estrato)[, , "Ingreso"]
```

	Rural	Urban	Population
Estimation	5184728.31	8277014.39	13461742.71
Standard Error	303793.90	475101.12	563925.36
CVE	5.86	5.74	4.19
DEFF	0.24	0.14	0.16

Práctica en R:

Estimación del total para `Gasto` en `Pobreza extrema`

```
E.STPPS(dom.PE, pk, mh, Estrato)[, , "Gasto"]
```

	Rural	Urban	Population
Estimation	14875725.29	25879973.41	40755698.71
Standard Error	537951.00	849511.14	1005515.01
CVE	3.62	3.28	2.47
DEFF	0.33	0.32	0.32

Estratificación implícita y sistemática

Características

- Las subpoblaciones de interés o estratos son divisiones de la población objetivo en grupos que comparten características comunes.
- La estratificación explícita consiste en agrupar las unidades de interés (hogares) en estratos que serán tratados de forma independiente (pues se tienen marcos de muestreo separados para cada subpoblación).
- La estratificación implícita consiste en clasificar las unidades de interés (UPMs) dentro de cada estrato explícito haciendo uso de un conjunto de variables definidas.

Estratificación explícita

Los estratos explícitos son útiles para reducir la varianza de muestreo y asegurar la representatividad de las unidades de interés en cada uno de los grupos. Algunas variables que se consideran en el proceso de estratificación explícita son:

- Estados o regiones de un país.
- Zona en la que están ubicadas la UPMs (urbana o rural).

Estratificación implícita

- Este tipo de estratificación es una forma de garantizar una asignación estrictamente proporcional de las unidades de interés en algunos grupos de interés.
- También puede conducir a una mayor confiabilidad de las estimaciones del estudio, siempre que las variables de estratificación implícita que se consideran estén correlacionadas con la variable de interés.

Estratificación implícita

Por ejemplo, algunas variables de estratificación implícita son:

- Municipalidades: denotadas como la segund división administrativa de una país.
- Presencia de grupos minoritarios: regiones con grupos minoritarios (étnias o tribus indígenas).
- Nivel socio-económico: clasificación de cada UPM en un subgrupo de bienestar (útil cuando sólo se considera la ubicación cartográfica como estratifcación explícita).
- Distancia a la ciudad más importante: de esta forma se garantiza una mejor dispersión geográfica en cada estrato explícito.

Ejemplo

Un país puede estar interesado en particionar las UPMs del marco con respecto al número de habitantes de los municipios, de la siguiente manera:

- *Municipios capitales*: que son las ciudades capitales de departamentos o estados.
- *Municipios grandes*: referentes a municipios con más de 100.000 habitantes
- *Municipios intermedios*: que son municipios entre 50.000 y 100.000 habitantes.
- *Municipios medianos*: que representan a los municipios entre 20.000 y 50.000 habitantes.
- *Municipios pequeños*: con menos de 20.000 habitantes.

Pasos para la selección

1. Las UPMs se clasifican según las variables de estratificación implícita dentro de cada estrato explícito. Primero se usa la primera variable de estratificación implícita, luego la segunda (dentro de los niveles generados por la primera) y así sucesivamente hasta aplicar todas las variables. La última variable de ordenamiento siempre será la medida de tamaño.
2. Luego se ejecuta un muestreo (sistemático) proporcional para seleccionar la muestra de UPMs.

Reemplazos

- El escenario ideal en la aplicación de cualquier encuesta es que todas las unidades de interés muestreadas originalmente accedan a ser parte del estudio.
- No es usual tener una tasa de participación del 100%.
- Algunas veces, para evitar una reducción en el tamaño de la muestra y garantizar un nivel de precisión y confiabilidad adecuados para el análisis estadístico, cada unidad seleccionada cuenta con dos reemplazos en caso de que se rehúse a participar.

Reemplazos

- Estos corresponden, por lo general, a las UPMs u hogares que se encuentran inmediatamente antes y después de la originalmente muestreada en el marco muestral (que debe estar estrictamente ordenado por las variables de estratificación explícitas, implícitas y por la medida de tamaño descendentemente).
- Estos reemplazos siempre se encontrarán en el mismo estrato explícito (aunque es posible que no haga parte del mismo estrato implícito) y tendrá un tamaño similar a la muestreada originalmente.

Reemplazos

- Sin embargo, aun cuando las características son similares realizar estos cambios en la muestra puede producir sesgo.
- Se enfatiza que solo se usen los reemplazos en caso de que sea estrictamente necesario.

Reemplazos

Estratificación		Código UPM	Hogares	Estado - Diseño
Explícita	Implícita			
Urbana	SE Alto	PSU0001	160	No seleccionada
		PSU0002	155	Reemplazo
		PSU0003	153	Seleccionada
		PSU0004	152	Reemplazo
		PSU0005	155	No seleccionada
		.	.	.
		.	.	.
		.	.	.
	SE Medio	PSU0101	100	No seleccionada
		PSU0102	97	Reemplazo
		PSU0103	93	Seleccionada
		PSU0104	91	Reemplazo
		PSU0105	90	No seleccionada
		.	.	.
		.	.	.
		.	.	.
	SE Bajo	PSU0201	40	No seleccionada
		PSU0202	37	Reemplazo
		PSU0203	35	Seleccionada
		PSU0204	33	Reemplazo
		PSU0205	31	No seleccionada
		.	.	.
		.	.	.
		.	.	.

Estratificación		Código UPM	Hogares	Estado - Diseño
Explícita	Implícita			
Rural	SE Alto	PSU1101	100	No seleccionada
		PSU1102	98	Reemplazo
		PSU1103	96	Seleccionada
		PSU1104	93	Reemplazo
		PSU1105	91	No seleccionada
		.	.	.
		.	.	.
		.	.	.
	SE Medio	PSU1201	70	No seleccionada
		PSU1202	67	Reemplazo
		PSU1203	65	Seleccionada
		PSU1204	62	Reemplazo
		PSU1205	61	No seleccionada
		.	.	.
		.	.	.
		.	.	.
	SE Bajo	PSU1301	40	No seleccionada
		PSU1302	37	Reemplazo
		PSU1303	35	Seleccionada
		PSU1304	33	Reemplazo
		PSU1305	31	No seleccionada
		.	.	.
		.	.	.
		.	.	.

Figure 1: *Reemplazos en los estratos urbano y rural*

¡Gracias!

Andrés Gutiérrez

Experto Regional en Estadísticas Sociales

Division de Estadísticas

Email: andres.GUTIERREZ@cepal.org