Cálculo del Tamaño de Muestra

Curso de Muestreo Probabilístico en Encuestas de Hogares

Andrés Gutiérrez, Ph.D.

CEPAL - Unidad de Estadísticas Sociales

Confiabilidad y precisión

Tamaño de muestra en encuestas simples

El efecto de diseño

Tamaño de muestra en encuestas complejas

Ejemplos de cálculo del tamaño de muestra en Encuestas de Hogares

Cálculo del tamaño de muestra para la proporción de desocupados

Cálculo del tamaño de muestra para la media de los ingresos del hogar

Motivación

> *Una de las preguntas más básicas que debe enfrentar el investigador es* ***¿cuántas unidades debo seleccionar?****. No es un pregunta fácil de responder en encuestas con múltiples propósitos y estimadores.*

Robert M. Grooves (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) `samplesize4surveys`. *R package*.

Confiabilidad y precisión

Intervalo de confianza

Se define un estimador insesgado para un parámetro de interés θ con distribución normal

$$\hat{\theta} \sim \mathit{Normal}\ (\theta, \sqrt{\mathit{Var}(\hat{\theta})})$$

Luego, un intervalo de confianza para el parámetro θ está dado por:

$$\mathit{IC}(1-\alpha) = \left[\hat{\theta} - z_{1-\alpha/2}\sqrt{\mathit{Var}(\hat{\theta})} \ \ , \ \ \hat{\theta} + z_{1-\alpha/2}\sqrt{\mathit{Var}(\hat{\theta})}\right]$$

Cobertura

La cobertura del intervalo de confianza se define en términos del soporte:

$$1 - \alpha = \sum_{Q_0 \supset s} p(s),$$

donde Q_0 es el conjunto de todas las posible muestras cuyo intervalo de confianza contiene al parámetro de interés θ.

Medidas de precisión

Antes de introducir las metodologías básicas para el cálculo del tamaño de muestra mínimo, es necesario definir los diferentes tipos de error muestral que se definen en una encuesta.

Margen de error

Desde la expresión del intervalo de confianza, se define el *margen de error*, como aquella cantidad que se suma y se resta al estimador insesgado. En este caso, se define como:

$$ME = z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}$$

Error estándar

También es posible definir el *error estándar*, dado por

$$EE = \sqrt{Var(\hat{\theta})} \tag{1}$$

Margen de error relativo

Una medida que tiene en cuenta la precisión y el sesgo del estimador es el *margen de error relativo*, que se define como:

$$MER = z_{1-\alpha/2} \frac{\sqrt{Var(\hat{\theta})}}{E(\hat{\theta})} \tag{2}$$

Coeficiente de variación

De la misma manera, también se define el *coeficiente de variación* o *error estándar relativo* definido por:

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{E(\hat{\theta})} \tag{3}$$

Comentarios

- El tamaño de muestra dependerá del tipo de error que se quiera minimizar.
- Por ejemplo, para una población particular, el tamaño de muestra requerido para minimizar el margen de error, no será el mismo que el que se necesitaría para minimizar el coeficiente de variación.

Reportando los errores de muestreo

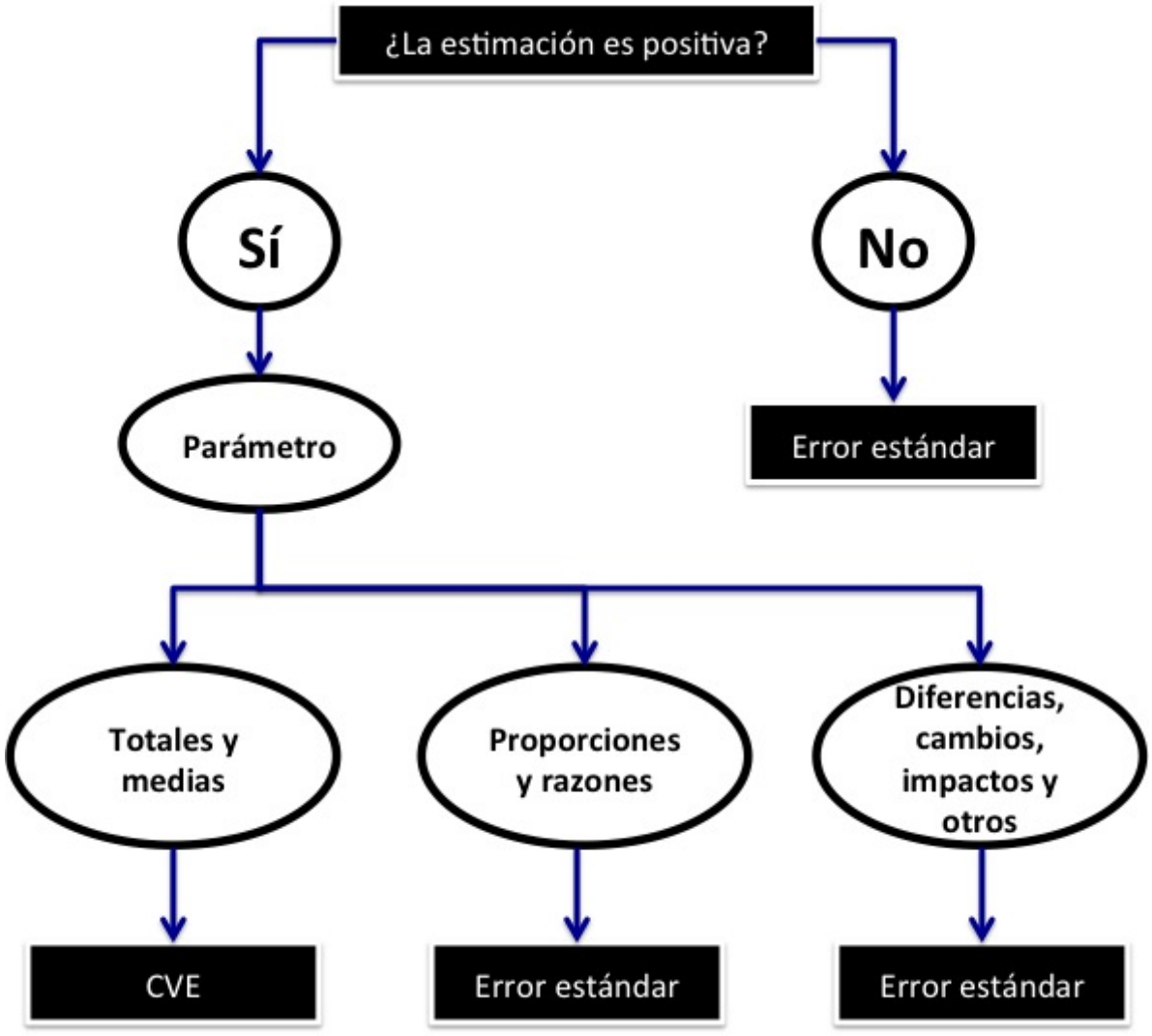


Figura 1: *Diferentes esquemas de reporte.*

Tamaño de muestra en encuestas simples

Estimador de una media

Bajo muestreo aleatorio simple, el estimador de Horvitz-Thompson para una media es:

$$\hat{\bar{y}}_\pi = \frac{1}{n} \sum_{k=1}^{n} y_k = \bar{y}_s \tag{4}$$

Varianza de una media

Bajo muestreo aleatorio simple, la varianza estimada del estimador de Horvitz-Thompson para una media es:

$$Var(\hat{\bar{y}}_\pi) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{yU} \tag{5}$$

Intervalo de confianza

Bajo muestreo aleatorio simple sin reemplazo, un intervalo de confianza de $100(1-\alpha)\,\%$ para la media de la población es:

$$\left(\bar{y}_s \pm z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S_{yU}}{\sqrt{n}}\right) \tag{6}$$

Tamaño de muestra para medias (ME)

Si ME es el margen de error que se quiere minimizar (que se encuentra en términos de las mismas unidades de la característica de interés), entonces despejando n, se tiene que:

$$n \geq \frac{S_{yU}^2}{\frac{ME^2}{z_\alpha^2} + \frac{S_{yU}^2}{N}} \tag{7}$$

Tamaño de muestra para medias (MER)

Si se quiere lograr una precisión relativa dada por el margen de error relativo (definido usualmente como 3 % o 5 %), entonces despejando n, se tiene que:

$$n \geq \frac{S_{yU}^2}{\frac{RME^2}{z_\alpha^2 \bar{y}_U^2} + \frac{S_{yU}^2}{N}} \tag{8}$$

Práctica en R: RME

Tamaño de muestra para estimar la media de la variable ingreso con un margen de error relativo menor al 5 %.

```
data("BigCity")

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

attach(Hogares)
```

Práctica en R: RME

Tamaño de muestra para estimar la media de la variable ingreso con un margen de error relativo menor al 5 %.

```
N <- nrow(Hogares)
mu <- mean(Ingreso)
sigma <- sd(Ingreso)
rme <- 0.05

ss4m(N = N, mu = mu, sigma = sigma,
     error = "rme", delta = rme)
```

```
## [1] 2678
```

Práctica en R: ME

Recuerde que el *margen de error relativo* se definió como:

$$MER = \frac{ME}{\theta} \tag{9}$$

Por tanto, para obtener un tamaño de muestra para estimar la media de la variable `ingreso` con un error relativo menor a 10 dólares.

```
me <- 10
rme <- me/mu

rme
```

```
## [1] 0.0047
```

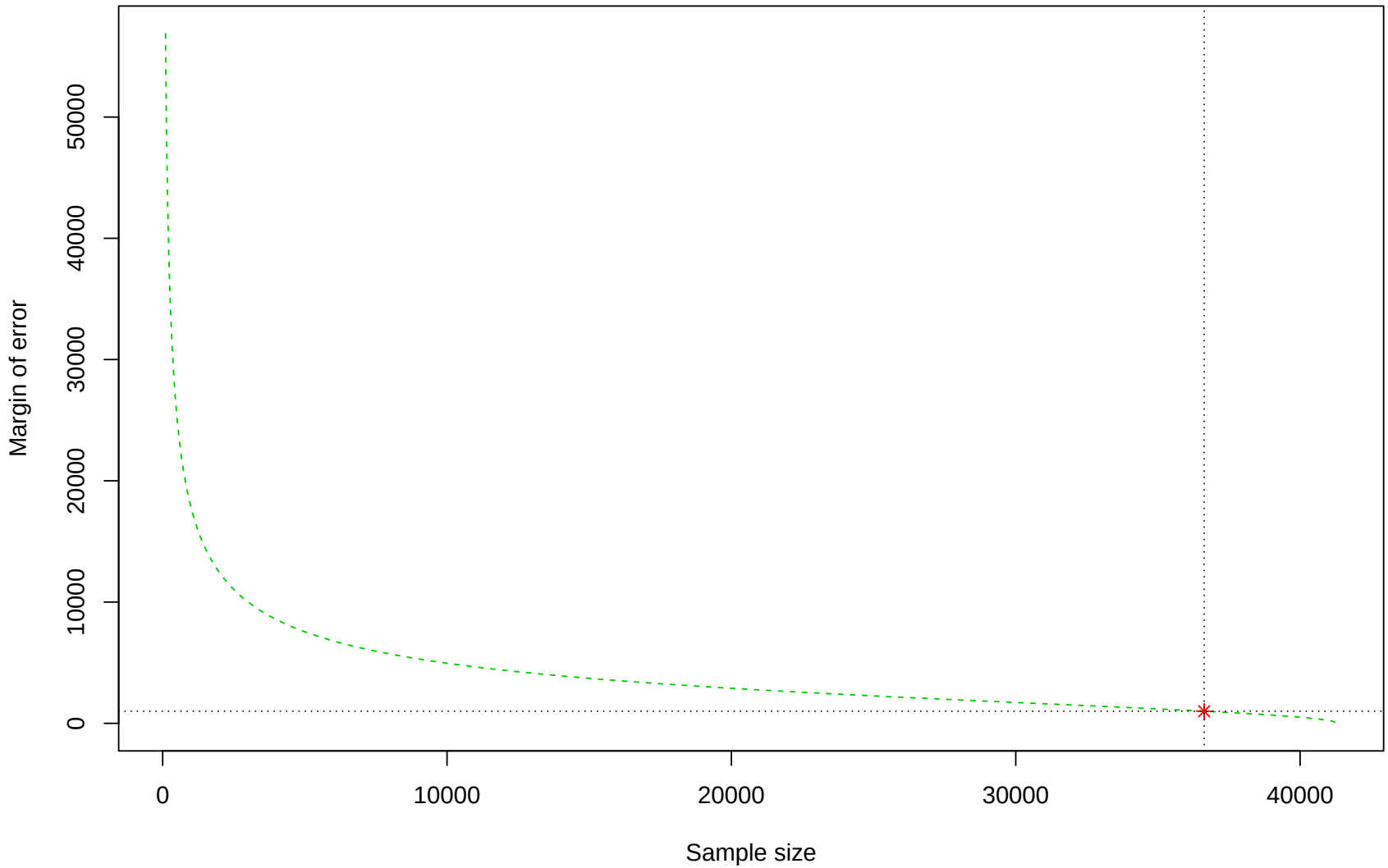
```
ss4m(N = N, mu = mu, sigma = sigma, error = "me", delta = 1
```

```
## [1] 36627
```

```
ss4m(N = N, mu = mu, sigma = sigma, error = "rme", delta =
```

Práctica en R: plot

```
ss4m(N = N, mu = mu, sigma = sigma,
     error = "me", delta = 10, plot = TRUE)
```




```
## [1] 36627
```

Estimador de una proporción

Bajo muestreo aleatorio simple, el estimador de Horvitz-Thompson para una proporción es:

$$\hat{P}_{d,\pi} = \frac{1}{n} \sum_{k=1}^{n} z_{dk} \tag{10}$$

Varianza de una proporción

Bajo muestreo aleatorio simple, la varianza estimada del estimador de Horvitz-Thompson para una proporción es:

$$Var(\hat{P}_{d,\pi}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{z_d} = \frac{1}{n}\left(1 - \frac{n}{N}\right) P_d Q_d$$

Intervalo de confianza

Bajo muestreo aleatorio simple sin reemplazo, un intervalo de confianza de $100(1-\alpha)\,\%$ para la proporción P_d es:

$$\left(\hat{P}_{d,\pi} \pm z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{P_d Q_d}{n}}\right) \tag{11}$$

Tamaño de muestra para proporciones (ME)

El tamaño de muestra mínimo necesario para lograr una estimación confiable de P_d, con menos de $ME\,\%$ de error, es:

$$n \geq \frac{P_d Q_d}{\frac{ME^2}{z_\alpha^2} + \frac{P_d Q_d}{N}} \tag{12}$$

Tamaño de muestra para proporciones (¡¡¡MER!!!)

Si se quiere lograr una precisión relativa dada por el margen de error relativo, entonces despejando n, se tiene que:

$$n \geq \frac{P_d Q_d}{\frac{RME^2}{z_\alpha^2 P_d^2} + \frac{P_d Q_d}{N}} \tag{13}$$

Práctica en R: ME

Tamaño de muestra para estimar la proporción de la variable `pobreza == "pobreza extrema"` con un margen de error menor al 2 %.

```
prop.table(table(Pobreza))
```

NotPoor	Extreme	Relative
0.69	0.07	0.24

```
N <- nrow(Hogares)
P <- prop.table(table(Pobreza))[2]
me <- 0.02

ss4p(N = N, P = P, error = "me", delta = me)
```

```
## Extreme
##     616
```

Práctica en R: RME

Note que el margen de error se escribe como:

$$ME = MER * \theta \tag{14}$$

Por tanto, para obtener un tamaño de muestra para estimar la proporción de la variable `pobreza == "pobreza extrema"` con un margen de error relativo menor al 2 %.

```
rme <- 0.02
me <- rme * P

me
```

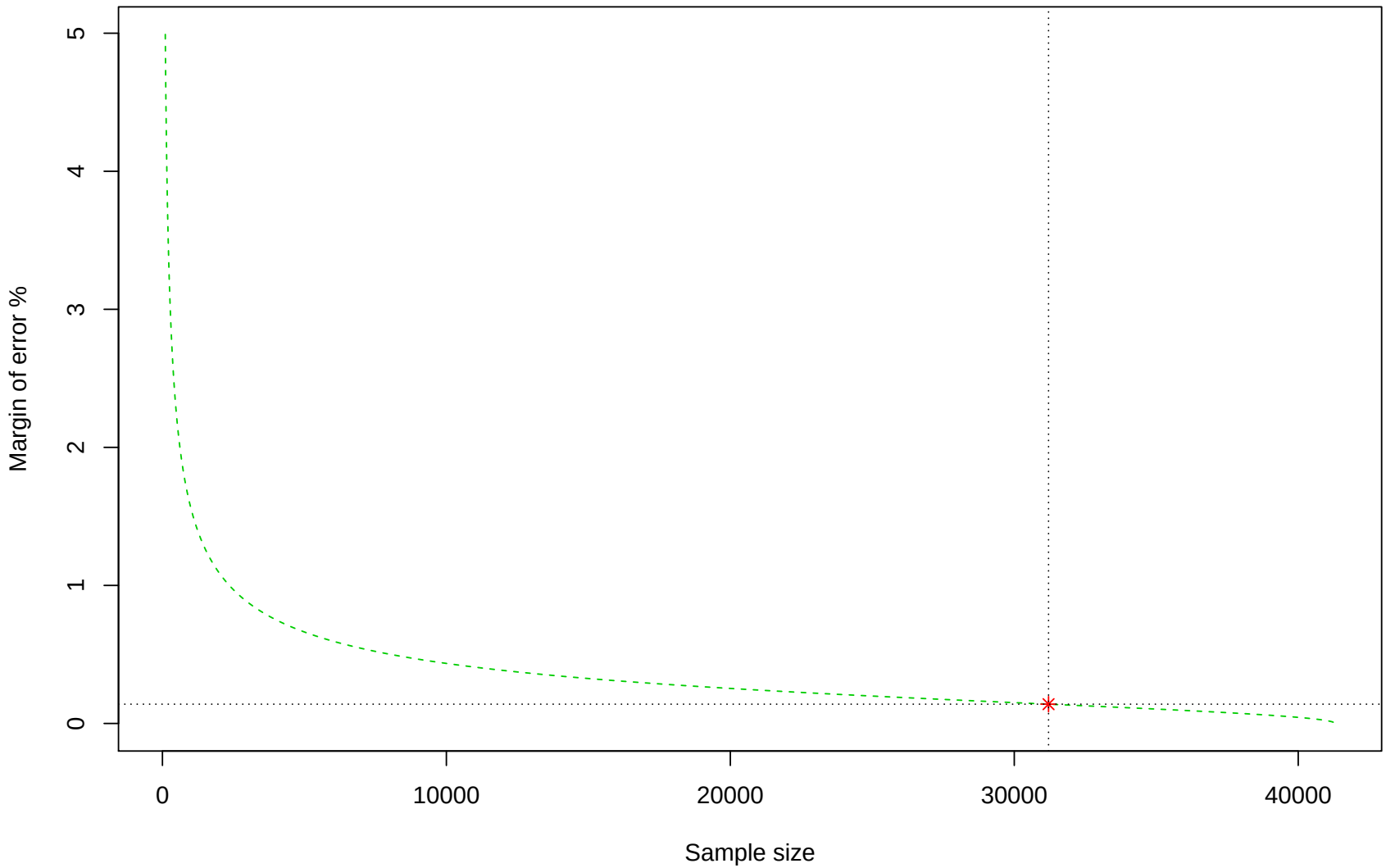
```
## Extreme
##  0.0014
```

```
ss4p(N = N, P = P, error =  "me", delta = me)
```

```
## Extreme
##   31208
```

Práctica en R: plot

```
ss4p(N = N, P = P, error =  "me",
     delta = me, plot = TRUE)
```




```
## Extreme
```

El efecto de diseño

Muestras complejas

- Cuando se selecciona una muestra utilizando un diseño de muestreo de conglomerados o en varias etapas, entonces es imposible afirmar que existe independencia entre las observaciones.
- No es adecuado usar las fórmulas clásicas para la determinación de un tamaño de muestra.
- Una forma sencilla de incorporar este efecto de aglomeración en las expresiones del muestreo aleatorio simple la da el efecto de diseño.

El efecto de diseño

El efecto de diseño se define como la relación entre la varianza del diseño complejo y la varianza del muestreo aleatorio simple

$$DEFF(\hat{\theta}) = \frac{Var_p(\hat{\theta})}{Var_{MAS}(\hat{\theta})} \tag{15}$$

Comentarios

1. Esta cifra da cuenta del efecto de aglomeración causado por la utilización de un diseño de muestreo complejo (p).
2. Siempre se toma como referencia al muestreo aleatorio simple (MAS).
3. El DEFF depende del parámetro de la población finita θ que se quiera estimar.
4. El DEFF no es el mismo para un total, una proporción o una razón.

Comentarios

- Cuando el efecto de diseño es más grande que la unidad, la varianza de la estrategia del numerador es más grande que la denominador, por tanto, se ha perdido precisión al utilizar una estrategia de muestreo más compleja.
- Si el cociente es menor que uno, se ha ganado precisión.

DEFF para un total

En particular, el efecto de diseño, restringido a la estimación de un total poblacional y al usar el estimador de Horvitz-Thompson en ambas estrategias, toma la siguiente forma

$$Deff = \frac{Var_p(\hat{t}_{y,\pi})}{\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yU}} \tag{16}$$

Diseño Bernoulli

Asumiendo que $\pi = n/N$, entonces:

$$n = E_{MAS}(n(S)) = E_{BER}(n(S)) = N\pi$$

De esta manera podemos introducir la medida de eficiencia del diseño de muestreo Bernoulli con respecto al MAS, así

$$deff = \frac{Var_{BER}(\hat{t}_{y,\pi})}{Var_{MAS}(\hat{t}_{y,\pi})} = 1 - \frac{1}{N} + \frac{1}{CV_y^2} \cong 1 + \frac{1}{CV_y^2} \qquad (17)$$

Muestreo aleatorio estatificado

El efecto de diseño en el muestreo aleatorio estratificado con asignación proporcional está dado por

$$Deff \cong \frac{\sum_{h=1}^{H} W_h S_{yU_h}^2}{\sum_{h=1}^{H} W_h \left[S_{yU_h}^2 + (\bar{y}_{U_h} - \bar{y}_U)^2 \right]} \tag{18}$$

$$\tag{19}$$

$$\cong \frac{\text{Varianza dentro de los estratos}}{\text{Varianza Total}} \tag{20}$$

Muestreo aleatorio estatificado

Ahora, intuitivamente tenemos que

$$\textbf{Varianza Total} = \textbf{Varianza dentro} + \textbf{Varianza entre}$$

Por tanto se concluye que, casi siempre, esta estrategia de muestreo arrojará mejores resultados que una estrategia aleatoria simple.

Diseño Sistemático

El efecto de diseño de la estrategia de muestreo que utiliza un diseño sistemático y el estimador de Horvitz-Thompson está dado por

$$Deff = \frac{Var_{SIS}\hat{t}_{\pi}}{Var_{MAS}\hat{t}_{\pi}} = \frac{N-1}{N-n}[1+(n-1)\rho] \tag{21}$$

Diseño Sistemático

Se define el coeficiente de correlación intra-clase como

$$\rho = 1 - \frac{n}{n-1}\frac{SCD}{SCT} \tag{22}$$

Diseño Sistemático

Dado el efecto de diseño, se concluye que esta estrategia de muestreo es

- Igual de eficiente al muestreo aleatorio simple sí $\rho = \frac{1}{1-N}$.
- Menos eficiente que el muestreo aleatorio simple sí $\rho > \frac{1}{1-N}$.
- Más eficiente que el muestreo aleatorio simple sí $\rho < \frac{1}{1-N}$.

Muestreo de conglomerados

El efecto de diseño en muestreo aleatorio de conglomerados está dado por

$$Deff = \frac{Var_{MAC}(\hat{t}_\pi)}{Var_{MAS}(\hat{t}_\pi)} \cong 1 + (M-1)\rho$$

En donde M es el número de hogares que en promedio componen las UPM y *rho* es el coeficiente de correlación intraclase entre la variable de interés y las UPM.

Muestreo en varias etapas

El efecto de diseño para un muestreo en varias etapas está dado por

$$Deff = \frac{Var_{Etapas}(\hat{t}_\pi)}{Var_{MAS}(\hat{t}_\pi)} \cong 1 + (\bar{m} - 1)\rho$$

En donde $\bar{m}$ es el número de hogares seleccionados en cada UPM y *rho* es el coeficiente de correlación intraclase entre la variable de interés y las UPM.

Comentarios

- Los conglomerados se forman física y geográficamente como agrupaciones contiguas de elementos que comparten un ambiente natural.
- Se espera que el comportamiento de los elementos internamente sea similar.
- ρ es generalmente positivo.
- El muestreo por conglomerados tendrá una mayor varianza que el muestreo aleatorio simple.

Tamaño de muestra en encuestas complejas

Varianza del diseño complejo

El efecto de diseño se definió así

$$DEFF(\hat{\theta}) = \frac{Var_p(\hat{\theta})}{Var_{MAS}(\hat{\theta})} \tag{23}$$

Varianza del diseño complejo

Por tanto, es posible escribir la varianza del estimador $\hat{\theta}$ bajo el diseño de muestreo complejo como

$$Var_p(\hat{\theta}) = DEFF(\hat{\theta})\ Var_{MAS}(\hat{\theta}) \tag{24}$$

Varianza del diseño complejo

Por ejemplo, si $\hat{\theta} = \hat{\bar{y}}_s$, entonces la varianza del diseño complejo es

$$Var_p(\hat{\bar{y}}_s) = DEFF(\hat{\bar{y}}_s)\ Var_{MAS}(\hat{\bar{y}}_s) \tag{25}$$

$$= \frac{DEFF(\hat{\bar{y}}_s)}{n}\left(1 - \frac{n}{N}\right) S^2_{yU} \tag{26}$$

Tamaño de muestra con el DEFF

Si se implementara un muestreo aleatorio simple con un tamaño de muestra n_0, para una precisión deseada, entonces el valor del tamaño de muestra que tendrá en cuenta el efecto de aglomeración para un diseño complejo estará cercano a

$$n \approx n_0 * DEFF$$

Intervalo de confianza

Bajo condiciones de regularidad, un intervalo de confianza de $100(1-\alpha)\,\%$ para la media de la población en un muestreo complejo es:

$$\left(\hat{\bar{y}}_s \pm z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{Deff \;\; S^2_{yU}}{n}} \right)$$

Tamaño de muestra para medias (ME)

Si ME es el margen de error que se quiere minimizar (que se encuentra en términos de las mismas unidades de la característica de interés), entonces despejando n, se tiene que:

$$n \geq \frac{Deff \ S^2_{yU}}{\frac{ME^2}{z^2_\alpha} + \frac{Deff \ S^2_{yU}}{N}}$$

Tamaño de muestra para medias (MER)

Si se quiere lograr una precisión relativa dada por el margen de error relativo (definido usualmente como 3 % o 5 %), entonces despejando n, se tiene que:

$$n \geq \frac{Deff \ S_{yU}^2}{\frac{RME^2}{z_\alpha^2 \bar{y}_U^2} + \frac{Deff \ S_{yU}^2}{N}}$$

Práctica en R: RME

Tamaño de muestra para estimar la media de la variable ingreso con un margen de error relativo menor al 5 %.

```
attach(Hogares)
N <- nrow(Hogares)
mu <- mean(Ingreso)
sigma <- sd(Ingreso)
deff <- 2.5
rme <- 0.05
```

Práctica en R: RME

Tamaño de muestra para estimar la media de la variable ingreso con un margen de error relativo menor al 5 %.

```
ss4m(N = N, mu = mu, sigma = sigma,
     DEFF = deff, delta = rme, error = "rme")
```

```
## [1] 6100
```

Práctica en R: RME

Recuerde que el *margen de error relativo* se definió como:

$$MER = \frac{ME}{\mu} \tag{27}$$

Por tanto, para obtener un tamaño de muestra para estimar la media de la variable `ingreso` con un error relativo menor a 10 dólares.

```
me <- 10
rme <- me/mu

rme
```

```
## [1] 0.0047
```

Práctica en R: ME

Tamaño de muestra para estimar la media de la variable ingreso con un error relativo menor a 10 dólares.

```
ss4m(N = N, mu = mu, sigma = sigma,
     DEFF = deff, delta = rme, error = "rme")
```

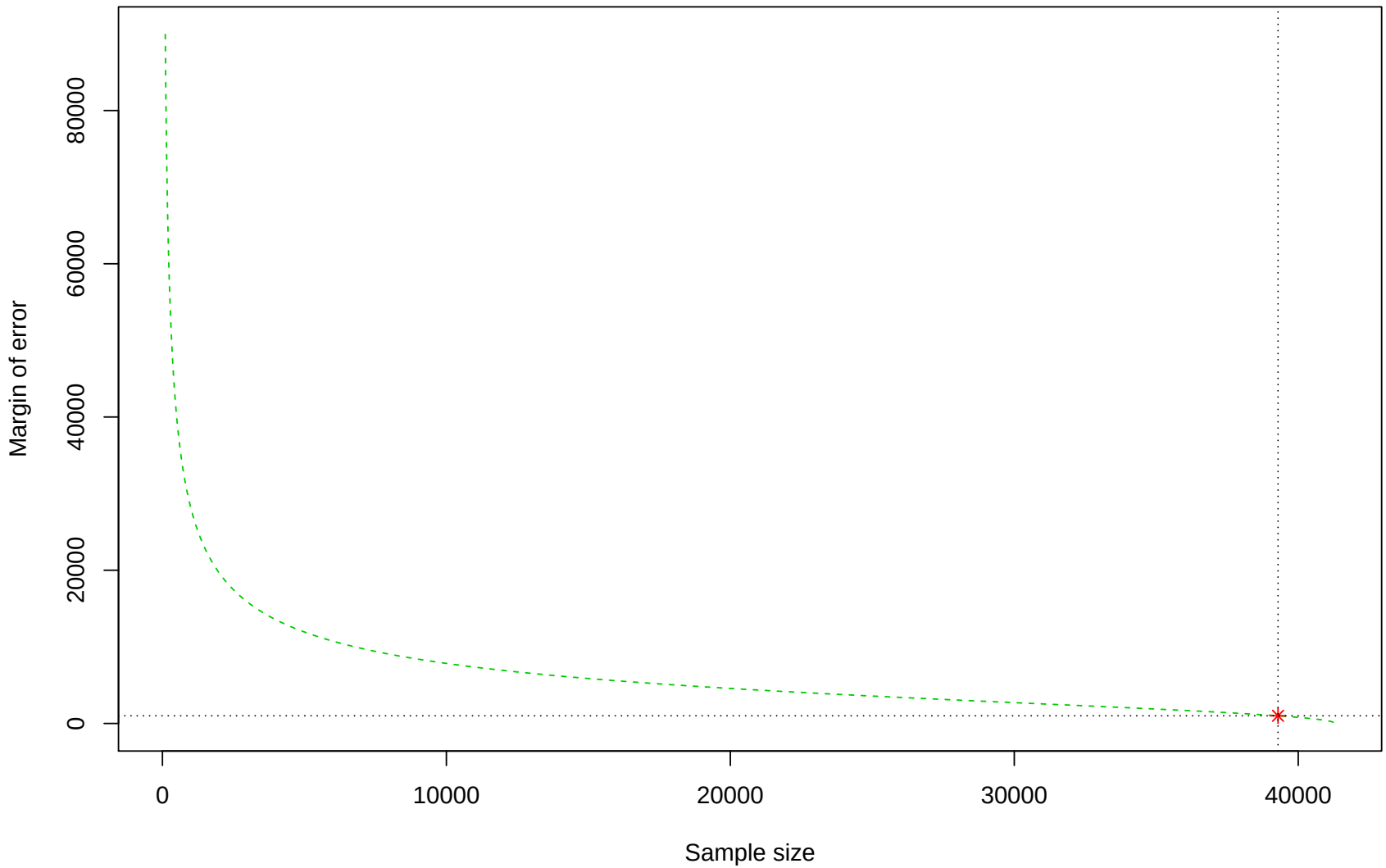
```
## [1] 39290
```

```
ss4m(N = N, mu = mu, sigma = sigma,
     DEFF = deff, delta = me, error = "me")
```

```
## [1] 39290
```

Práctica en R: plot

```
ss4m(N = N, mu = mu, sigma = sigma,
     DEFF = deff, delta = me, error = "me", plot = TRUE)
```




```
## [1] 39290
```

Varianza de una proporción

Bajo muestreo aleatorio simple, la varianza estimada del estimador de Horvitz-Thompson para una proporción es

$$Var(\hat{P}_{d,\pi}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) Deff \; S^2_{z_d} = \frac{1}{n}\left(1 - \frac{n}{N}\right) Deff \; P_d Q_d$$

Intervalo de confianza

Bajo condiciones de regularidad, un intervalo de confianza de $100(1-\alpha)\,\%$ para la proporción poblacional en un muestreo complejo es:

$$\left(\hat{P}_{d,\pi} \pm z_{1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{Deff\ P_d Q_d}{n}}\right) \tag{28}$$

Tamaño de muestra para proporciones (ME)

El tamaño de muestra mínimo necesario para lograr una estimación confiable de P_d, con menos de $ME\,\%$ de error, es:

$$n \geq \frac{Deff\ \ P_d Q_d}{\frac{ME^2}{z_\alpha^2} + \frac{Deff\ \ P_d Q_d}{N}} \tag{29}$$

Práctica en R: ME

Tamaño de muestra para estimar la proporción de la variable `pobreza == "pobreza extrema"` con un margen de error menor al 2 %.

```
prop.table(table(Pobreza))
```

NotPoor	Extreme	Relative
0.69	0.07	0.24

```
N <- nrow(Hogares)
P <- prop.table(table(Pobreza))[2]
deff <- 2.5
me <- 0.02
```

Práctica en R: ME

Tamaño de muestra para estimar la proporción de la variable `pobreza == "pobreza extrema"` con un margen de error menor al 2 %.

```
ss4p(N = N, P = P, DEFF = deff, delta = me, error = "me")
```

```
## Extreme
##    1504
```

Práctica en R: RME

Note que el margen de error se escribe como:

$$ME = MER * Pd$$

Por tanto, para obtener un tamaño de muestra para estimar la proporción de la variable `pobreza == "pobreza extrema"` con un margen de error relativo menor al 2 %.

```
rme <- 0.02
me <- rme * P

me
```

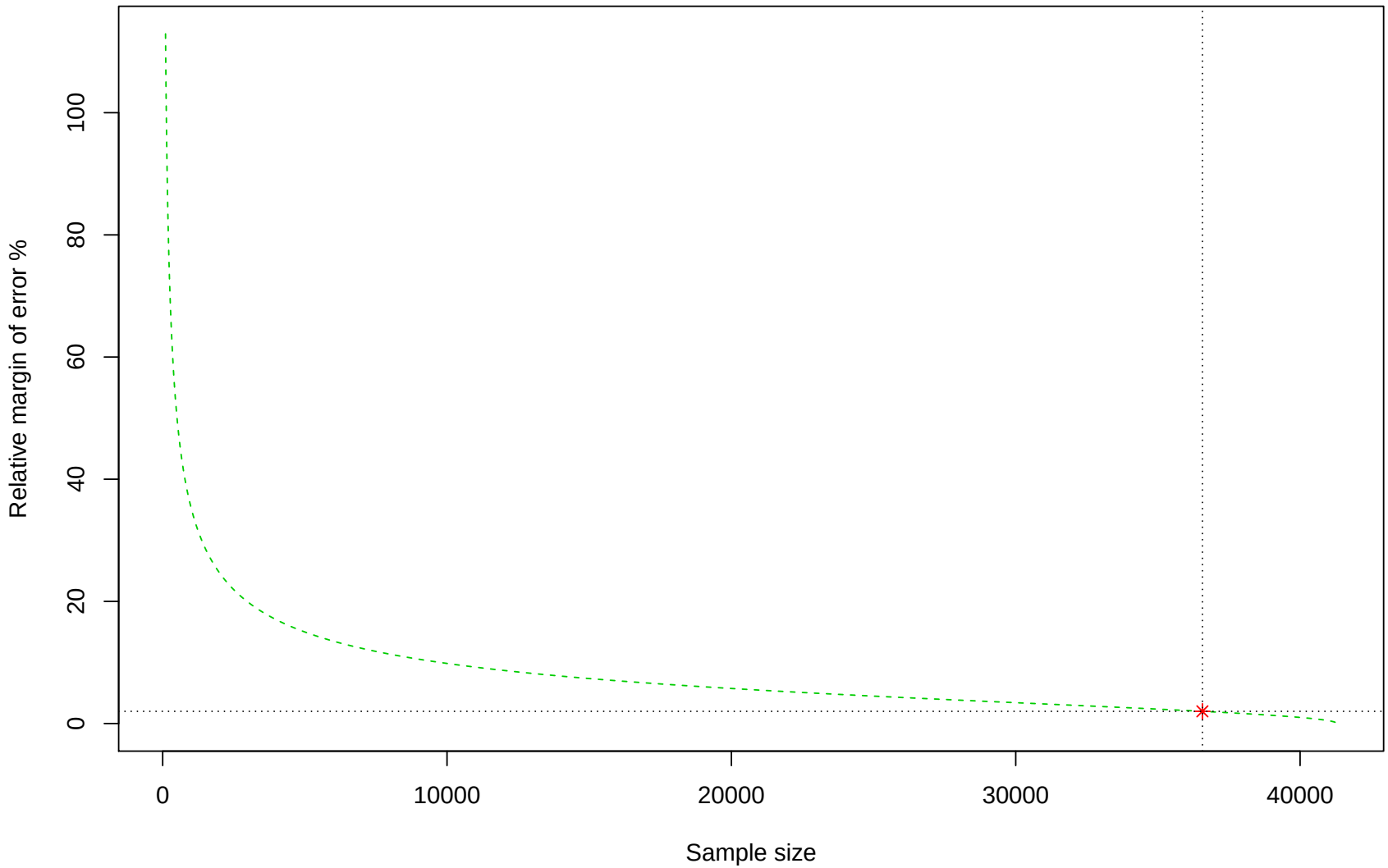
```
## Extreme
##  0.0014
```

```
ss4p(N = N, P = P, DEFF = deff, delta = me, error = "me")
```

```
## Extreme
##   36565
```

Práctica en R: plot

```
ss4p(N = N, P = P, DEFF = deff, delta = rme, error = "rme",
```




```
## Extreme
##   36565
```

Ejemplos de cálculo del tamaño de muestra en Encuestas de Hogares

Dos posible escenarios

1. Un caso común en las ONE trata con la asignación del tamaño de muestra en problemas de inferencia que tienen que ver con la estimación de parámetros de personas.
2. Cuando la variable de diseño y en general, las variables más importantes de la encuestas están presentes a nivel de hogar, entonces no es necesario realizar un submuestreo de personas.

Enfoque general

En general, se seguirá el enfoque tradicional que afirma que en un muestreo en varias etapas, el efecto de diseño se puede aproximar mediante la siguiente expresión

$$Deff \approx 1 + (\bar{n} - 1)\rho$$

En donde, para muestreos con conglomerados de tamaño desigual, $\bar{n} = n/n_I$ representa el tamaño promedio de la submuestra en cada conglomerado.

Enfoque general

El algoritmos de cálculo de tamaño de muestra se puede resumir en los siguientes pasos:

1. Escoger el parámetro que se desea estimar.
2. Fijar un margen de error máximo para la estimación del parámetro de interés.
3. Estimar el efecto de diseño para un valor predefinido $\bar{n}$ con base en alguna encuesta anterior o en el último censo disponible.
4. Calcular el tamaño de muestra apropiado para un muestreo aleatorio simple sin reemplazo.
5. Calcular el tamaño de muestra apropiado para el muestreo complejo ajustando el paso anterior por el efecto de diseño.
6. Determinar los tamaños de muestra de la primera y segunda etapa.

Algunas definiciones importantes

- N es el tamaño de la población U.
- n es el tamaño de la muestra s.
- N_I es el número de UPM en el marco de muestreo.
- n_I es el número de UPM que se selecciona en la muestra de la primera etapa s_I.
- N_{II} es el número de hogares existentes en el país.
- n_{II} es el número de hogares seleccionados en la muestra de la segunda etapa s_{II}.
- $\bar{n}$ es el número de personas promedio que se van a seleccionar en cada UPM.
- $\bar{n}_{II}$ es el número de hogares promedio que se van a seleccionar en cada UPM.
- ρ es el coeficiente de correlación intraclase, calculado para la variable de interés sobre las UPM.
- b es el número promedio de personas por hogar.
- r es el porcentaje de personas en la subpoblación de interés.

Cálculo del tamaño de muestra para la proporción de desocupados

Definir la población de interés de manera explícita.

En particular es necesario aclarar si la unidad de análisis son las personas o los hogares. De esta forma, se debe fijar los valores para r y b.

Si la unidad de análisis son todas las personas del hogar, entonces el porcentaje de personas con la característica de interés será $r = 1$, de otra forma $r < 1$. Por otro lado, el número promedio de personas por hogar b dependerá de la región o estrato en la que se requiera el cálculo.

Definir el número promedio de hogares

El número promedio de hogares que se desea encuestar en cada una de las UPM está dado por $\bar{n}_{II}$.

Este proceso debería ser repetido de forma iterativa en los pasos subsiguientes para poder evaluar la calidad del diseño. De las varias escogencias de $\bar{n}_{II}$ será necesario escoger solo una.

Calcular el número promedio de personas que serán encuestadas

Al igual que en el paso anterior es necesario probar varios escenarios que redundarán en la escogencia de un número óptimo de personas por UPM.

Los valores de $\bar{n}$ dependen directamente del paso anterior al escoger $\bar{n}_{II}$. Debido a que la selección de las personas está supeditada a la selección de los hogares, entonces $\bar{n}$ se puede descomponer manteniendo la relación con r y b, de la siguiente manera:

$$\bar{n} = \bar{n}_{II} * r * b$$

Calcular el efecto de diseño

Es necesario definir (o calcular con encuestas o censos anteriores) la correlación intraclase de la variable de interés con el agrupamiento por UPM ρ. Luego de esto se debe calcular *Deff* como función de ρ y de $\bar{n}$.

Ahora, el efecto de diseño *Deff*, definido como una función de la correlación existente entre la variable de interés y la conformación de las UPM, está dado por la siguiente expresión

$$Deff \approx 1 + (\bar{n} - 1)\rho$$

Tamaño de muestra de personas

A partir de las expresiones de tamaño de muestra para muestreos complejos, calcular el tamaño de muestra necesario para lograr una precisión adecuada en la inferencia.

En primer lugar, el tamaño de muestra necesario para alcanzar un error relativo máximo de $\delta\,\%$ es de

$$n \geq \frac{P\,(1-P)\;Deff}{\frac{\delta^2 P^2}{z_\alpha^2} + \frac{P\,(1-P)\;Deff}{N}}$$

Tamaño de muestra de hogares

Es necesario calcular el número total de hogares de deben ser seleccionados para lograr entrevistar a todas las personas que serán observadas en el punto anterior. El número de hogares que deben se seleccionados estará determinado por las cantidades n, b y r, de la siguiente forma

$$n_{II} = \frac{n}{r \times b}$$

Cálculo del número de UPMS

Los hogares y las personas se observan a partir de las UPM. En este paso final es necesario calcular el número de UPM que deben ser seleccionadas en el muestreo a partir de la relación

$$n_I = \frac{n}{\bar{n}} = \frac{n_{II}}{\bar{n}_{II}}$$

Práctica en R

```
library(TeachingSampling)
data(BigCity)

BigCity1 <- BigCity[!is.na(BigCity$Employment), ]
summary(BigCity1$Employment)
```

```
## Unemployed   Inactive   Employed
##       4630      44104      62188
```

Práctica en R

```
BigCity1$Unemp <- Domains(BigCity1$Employment)[, 1]
BigCity1$Active <- Domains(BigCity1$Employment)[, 1] +
                   Domains(BigCity1$Employment)[, 3]
```

Práctica en R

```
N <- nrow(BigCity)
M <- length(unique(BigCity$PSU))
r <- sum(BigCity1$Active)/N
b <- N/length(unique(BigCity$HHID))
rho <- ICC(BigCity1$Unemp, BigCity1$PSU)$ICC
P <- sum(BigCity1$Unemp)/sum(BigCity1$Active)

delta <- 0.07
conf <- 0.95
m <- c(5:15)
```

Práctica en R

```
N
```

```
## [1] 150266
```

```
M
```

```
## [1] 1664
```

```
r
```

```
## [1] 0.44
```

```
b
```

```
## [1] 3.6
```

```
rho
```

```
## [1] 0.03
```

```
P
```

```
## [1] 0.069
```

Práctica en R

```
margen.error <- P * delta
margen.error
```

```
## [1] 0.0049
```

```
P - margen.error
```

```
## [1] 0.064
```

```
P + margen.error
```

```
## [1] 0.074
```

```
P
```

```
## [1] 0.069
```

Práctica en R

```
ss4HHSp(N, M, r, b, rho, P, delta, conf, m)
```

HouseholdsPerPSU	PersonsPerPSU	DEFF	PSUinSample	HouseholdsInSample	PersonsInSample
5	8	1.2	1454	7269	11763
6	10	1.3	1256	7535	12194
7	11	1.3	1114	7800	12623
8	13	1.4	1008	8064	13049
9	15	1.4	925	8325	13472
10	16	1.4	859	8585	13893
11	18	1.5	804	8843	14311
12	19	1.6	758	9100	14726
13	21	1.6	720	9355	15139
14	23	1.6	686	9609	15550
15	24	1.7	657	9861	15958

Cálculo del tamaño de muestra para la media de los ingresos del hogar

Definir la población de interés de manera explícita.

En algunas ocasiones la variable de interés como la unidad de observación están a nivel de hogar y es posible modificar el algoritmo anterior para seleccionar los hogares en la muestra.

En este caso algunas cantidades desaparecen porque no son objeto de la población de hogares, por ejemplo r y b; algunas otras expresiones deben ser redefinidas al contexto de los hogares, como por ejemplo, el coeficiente de correlación intraclase ρ, el efecto de diseño y todas las expresiones de tamaños de muestra.

Definir el número promedio de hogares

El número promedio de hogares que se desea encuestar en cada una de las UPM está dado por $\bar{n}_{//}$.

Esta cifra sigue siendo el insumo del algoritmo y se propone crear escenarios de muestreo a partir de su modificación y evaluación del tamaño de muestra final.

Calcular el efecto de diseño

Es necesario definir (o calcular con encuestas o censos anteriores) la correlación intraclase ρ_{II} de la variable de interés *a nivel del hogar* con el agrupamiento por UPM.

Además de los anterior, el efecto de diseño *Deff* será función $\bar{n}_{II}$, como se muestra a continuación:

$$Deff \approx 1 + (\bar{n}_{II} - 1)\rho_{II}$$

Tamaño de muestra de hogares

Partiendo de las expresiones de tamaño de muestra generales para muestreos complejos y teniendo en cuenta que la población de interés son lo hogares y que la variable de interés y_{II} está a nivel de hogar, entonces el tamaño de muestra necesario para alcanzar un error relativo máximo de $\delta\,\%$ es de

$$n_{II} \geq \frac{S^2_{y_{II}} Deff}{\frac{\delta^2 \bar{y}^2_{II}}{z^2_\alpha} + \frac{S^2_{y_{II}} Deff}{N_{II}}}$$

Cálculo del número de UPMS

Los hogares se observan a partir de las UPM. En este paso final es necesario calcular el número de UPM que deben ser seleccionadas en el muestreo a partir de la relación

$$n_I = \frac{n_{II}}{\bar{n}_{II}}$$

Práctica en R

```
data(BigCity)

BigCity1 <- BigCity %>%
            group_by(HHID) %>%
            summarise(IncomeHH = sum(Income),
                      PSU = unique(PSU))
```

Práctica en R

```
head(BigCity1, 10)
```

HHID	IncomeHH	PSU
idHH00001	2775	PSU0001
idHH00002	1492	PSU0001
idHH00003	4280	PSU0001
idHH00004	2200	PSU0001
idHH00005	3119	PSU0001
idHH00006	675	PSU0001
idHH00007	8038	PSU0001
idHH00008	4905	PSU0001
idHH00009	607	PSU0001
idHH00010	1396	PSU0001

Práctica en R

```
summary(BigCity1$IncomeHH)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10	597	1303	2129	2646	131680

```
mean(BigCity1$IncomeHH)
```

```
## [1] 2129
```

```
sd(BigCity1$IncomeHH)
```

```
## [1] 2906
```

Práctica en R

```
N <- nrow(BigCity1)
M <- length(unique(BigCity1$PSU))
rho <- ICC(BigCity1$IncomeHH, BigCity1$PSU)$ICC
mu <- mean(BigCity1$IncomeHH)
sigma <- sd(BigCity1$IncomeHH)
```

Práctica en R

```
N
```

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

```
M
```

```
## [1] 1664
```

```
rho
```

```
## [1] 0.11
```

```
mu
```

```
## [1] 2129
```

```
sigma
```

```
## [1] 2906
```

Práctica en R

```r
delta <- 0.05
conf <- 0.95
m <- c(5:15)

margen.error <- mu * delta
margen.error
```

```
## [1] 106
```

```r
mu - margen.error
```

```
## [1] 2022
```

```r
mu + margen.error
```

```
## [1] 2235
```

```r
mu
```

```
## [1] 2129
```

Práctica en R

```
ss4HHSm(N, M, rho, mu, sigma, delta, conf, m)
```

HouseholdsPerPSU	DEFF	PSUinSample	HouseholdsInSample
5	1.4	748	3738
6	1.5	666	3993
7	1.6	607	4246
8	1.8	562	4495
9	1.9	527	4741
10	2.0	498	4983
11	2.1	475	5222
12	2.2	455	5459
13	2.3	438	5692
14	2.4	423	5922
15	2.5	410	6149

¡Gracias!

Andrés Gutiérrez

Experto Regional en Estadísticas Sociales

Division de Estadísticas

Email: andres.GUTIERREZ@cepal.org