UPC

UNIVERSITAT POLITÈCNICA DE CATALUNYA

PROPUESTA DE CALCULO DE ERRORES DE MUESTREO: ENCUESTA DE MOVILIDAD DE MÉRIDA (VENEZUELA)

Lídia Montero i Mercadé
Dep. d'Estadística i Investigació Operativa. Edifici U
Universitat Politècnica de Catalunya.
Report Intern de Treball DR98/02 Febrer de 1.998

DOCUMENT DE RECERCA

DEPARTAMENT D'ESTADÍSTICA I
INVESTIGACIÓ OPERATIVA

1. PREMISAS Y NOTACIÓN

1.1 Tipo de muestreo

Por conglomerados estratificados en H estratos. En el presente estudio los estratos corresponden a zonas de transporte o similares (H ≈ 75/80).

La estratificación tiene por efecto la reducción de la varianza provocada por el muestreo en conglomerados.

Para cada estrato h ∈ {1,.....H} **se dispone**.

- N_h Número poblacional de unidades primarias (UP o conglomerados) (familias en el presente estudio)
- M_h Número poblacional de unidades secundarias (US) (habitantes) por cuotas por sexo y edad.

Se nota $\overline{M}_h = {M_h}/{N_h}$ Número medio de individuos por familia en el estrato h

f para

Tamaño muestral de familias: n = 3.000 (UP's) (fijado a priori).
Tamaño poblacional de familias: N = ? (conocido).

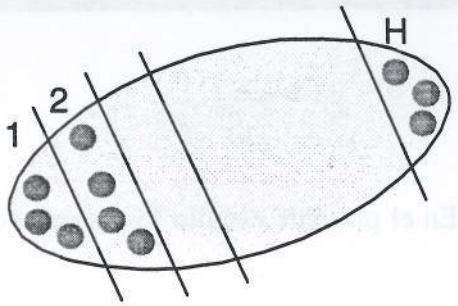
Se estima una muestra de individuos (US) de m = 12.000, sobre un total poblacional M=250.000.

Al hacer el trabajo de campo y presuponiendo una selección de UP a nivel de estrato no probabilista (no muestreo aleatorio simple sin reposición) se habría de implementar un método por cuotas A NIVEL DE ESTRATO (ZONA) según:

- Nivel socio-económico de la familia
- Tamaño familiar (crítico)

El resultado final de la muestra ha de ser representativo a nivel de individuo y MUY PROBABLEMENTE, si la muestra resultante no satisface a nivel de individuo las cuotas por SEXO y GRUPO DE EDAD (disponibles) se deberá aplicar una POST-ESTRATIFICACIÓN para conseguir cuotas individuales representativas: la validez de las estimaciones realizadas a partir de la muestra en gran medida de la representatividad de la muestra.

Los estimadores propuestos en el presente documento para el muestreo en conglomerados estratificados a llevar a cabo en la ciudad de Mérida son todos NO sesgados.



Estratificación: Proporcional al tamaño según las UP's por estrato

Sea $f = n/N$; $n_h = f \cdot N_h$

f: factor de muestreo (en familias)
n_h: tamaño muestral (UP o familias) en el estrato h

m_h desconocido a priori (tamaño muestral en individuos estrato h)

Universo: *M ~ 250.000*
N ~ 80.000

1.2 Notación

Sea Y una variable de estudio definida a nivel de individuo. A continuación se define la notación para diversos estimadores asociados a la variable de interés, tanto a nivel global (universo), como a nivel de estrato (zona).

Global	*Estadístico*	*vs. Estimador del estadístico*
Total	τ_y	$\hat{\tau}_y$ *o* T_y
Valor medio por familia (UP)	$\bar{\tau}_y$	$\hat{\bar{\tau}}_y$ *o* $\bar{t}_y$
Valor medio por individuo (US)	μ_y	$\bar{y}$
Por zona o estrato (h)		
Total	τ_y^h	$\hat{\tau}_y^h$ *o* T_y^h
Valor medio por familia (UP)	$\bar{\tau}_y^h$	$\hat{\bar{\tau}}_y^h$ *o* $\bar{t}_y^h$
Valor medio por individuo (US)	μ_y^h	$\bar{y}^h$

Se distinguen entre *los estimadores*, en caligrafía normal o con un símbolo ^ sobre la notación de los *valores verdaderos* o poblacionales de los estadísticos, que se notan con letras griegas o caligráficas.

La notación empleada para los distintos tipos de variancias es la siguiente:

σ_y^2: Varianza poblacional de la variable Y

$\sigma_y'^2$: Varianza poblacional corregida de Y (donde $\sigma'^2_y = \frac{M}{M-1}\sigma^2_y$)

S_y^2: Varianza muestral de Y

$S_y'^2$: Varianza muestral corregida de Y

$$S_y^2 = \frac{\sum_i (y_i - \bar{y})^2}{m} \qquad S'^2_y = \frac{\sum_i (y_i - \bar{y})^2}{m-1} \qquad \text{donde } \bar{y} = \sum_i y_i \Big/ m$$

Las propiedades de los estimadores anteriores son:

- $E\left[S'^2_y\right] = \sigma'^2_y$: para el muestreo sin reposición
- $V(\bar{y}) = (1 - \frac{m}{M})\frac{\sigma'^2_y}{m}$ y su estimador $\hat{V}(\bar{y}) = (1 - \frac{m}{M})\frac{S'^2_y}{m}$

El error estandard del estimador de la media se nota por $\sqrt{\hat{V}(\bar{y})}$ y un intervalo de confianza bilateral al 95% de μ_Y es:

$$\bar{y} \pm 2\sqrt{\hat{V}(\bar{y})}$$

1.2.1 Enfoque de los estimadores

Por el enfoque dado a la formulación, la variable Y se asocia a individuos (US), pero a nivel de familias (conglomerados o UP) interesa trabajar con totales: $\tau_y^{h,i}$ Suma de todos los valores de Y de las US de la UP *i* del estrato *h*

$$(W_h) \quad \tau_y^{h,i} = \sum_{\substack{j\ US\ de\ la\ UP\ i \\ del\ estrato\ h}} y_{ij}$$

Se podría definir W: *Total de Y en las UPs* y definir $\sigma^2_{w_h}, \sigma'^2_{w_h}, s^2_{w_h}, s'^2_{w_h}$, pero para no forzar en exceso la abstracción es mejor escribir más específicamente:

$\sigma^2_{\tau_y^h}$: Varianza poblacional del total de Y en el estrato *h*

$\sigma'^2_{\tau_y^h}$: Varianza corregida del total de Y en el estrato h

$s^2_{\tau_y^h}$: Varianza muestral del total de Y en el estrato h

$s'^2_{\tau_y^h}$: Varianza muestral corregida del total de Y en el estrato h

$$s^2_{\tau_y^h} = \frac{1}{n_h} \sum_{\substack{UP\ i\ del \\ estrato\ h}} (\tau_y^{h,i} - \bar{t}_y^h)^2$$

$$s'^2_{\tau_y^h} = \frac{1}{n_h - 1} \sum_{\substack{UP\ i\ del \\ estrato\ h}} (\tau_y^{h,i} - \bar{t}_y^h)^2$$

donde, $\bar{t}_y^h = \sum_i \tau_y^{h,i} / n_h$

1.2.2 Ejemplificación de la notación

1.2.2.1 Ejemplo 1.

Y: N° de viajes en autobús de un individuo.

$\overline{Y}$: N° medio de viajes en autobús por persona.

$\overline{Y}_h$: N° medio de viajes en autobús por persona en el estrato h.

$\hat{\tau}_y$ o T_y: Total de viajes en autobús.

$\bar{t}_y$: N° medio de viajes en autobús por familia (total medio por UP).

$\bar{t}_y^h$: N° medio de viajes en autobús en el estrato h (total medio por UP en el estrato h) .

1.2.2.2 Ejemplo 2.

Y: N° de viajes en autobús hacia el estrato (la zona) j.

$\hat{\tau}_y^h$ o T_y^h: Total de viajes en autobús de la zona h a la j.

$\bar{t}_y^h$: N° medio de viajes en autobús de h a j en familias de la zona h.

$\overline{y}^h$: N° medio de viajes en autobús (h, j) para individuos de la zona h.

La variable Y puede ser cualquier variable extraída de la muestra a nivel *individual* , incluso una variable binaria 0 ó 1, lo que da lugar a estimadores de proporciones individuales.

1.2.2.3 Ejemplo 3.

Y: Indicador de si un individuo es o no estudiante (1: lo es, 0: no lo es) (Y ~ Bernoulli p).

Si se emplea la idea propuesta en la notación de basar los estimadores en variables auxiliares relativas a totales por UP (familia), hecho fundamental en las fórmulas de los estimadores y los cálculos de errores estandard de los estimadores que se proponen en la Sección 2, el carácter binario de Y no supone ninguna dificultad ni especifidad en el tratamiento.

$\hat{\tau}_y$: Total de estudiantes en la población.
$\bar{t}_y$: Nº medio de estudiantes por familia
$\bar{y}$: Proporción de estudiantes en la población.
$\bar{y}^h$: Proporción de estudiantes en la zona h.

1.2.2.4 Ejemplo 4.

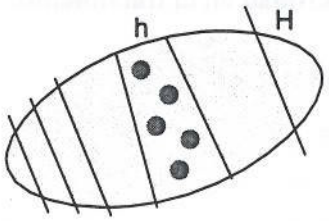
Si se quiere estimar un total familiar, será necesario usar los estimadores y fórmulas de error estandard del muestreo estratificado, pues no cumple la definición dada de Y, ya que la variable base debe estar definida a nivel de US (individuo) y no a nivel UP (familia).

Si se desea estimar el *total de cabezas de familia que son empresarios*, se debe considerar una variable X a nivel de UP (familia):

X: Indicador de cabeza de familia empresario (definido por UP o familia)

Las UP's o familias aparecen, respecto a la variable de interés X, como el último nivel de muestreo, lo que equivale al uso de estimadores y fórmulas de error relativos a un muestreo estratificado proporcional: los conglomerados no intervienen en absoluto (véase 3.4.1.2).

2. ESTIMADORES Y ERRORES ESTANDARD POR ESTRATO h



n_h: f N_h = n N_h/N
n: Tamaño muestral global de UP
n_h: Tamaño muestral estrato h de UP
(familias o conglomerados)

Para cálculo intervalo confianza de un estimador $\hat{E}$:

$$\frac{E-\mu_E}{\sqrt{\hat{V}(\hat{E})}} \sim t_\nu \text{ - Student} \qquad \textit{Condicion NO SESGO: } E\left[\hat{E}\right]=\mu_E$$

donde los grados de libertad de la distribución de t-Student son $\nu = n_h - 1$ y el nivel de confianza se indica (1-α)%. Para simplificar, se suele efectuar una aproximación

$$t_{\nu\ cualquiera}^{1-\alpha/2} = 2{,}0 \qquad \text{para } \alpha = 0{,}05 \quad (IC\ 95\%) \text{ ó en general}$$

en el cálculo de $IC(1-\alpha)\%$ *en lugar de* $\hat{E} \pm t_\nu^{1-\alpha/2}\sqrt{\hat{V}(\hat{E})}$, se emplea una aproximación normal que obvía el problema de los grados de libertad, $\hat{E} \pm z^{1-\alpha/2}\sqrt{\hat{V}(\hat{E})}$.

Siempre se calcula un estimador del error estandard ($\hat{V}(\hat{E})$) pues $V(\hat{E})$ es inasequible en la práctica.

2.1 Total por estrato (zona *h*), $\hat{\tau}_y^h$ *o* T_y^h

$$\hat{\tau}_y^h = \frac{N_h}{n_h}\sum_{i=1}^{n_h}\tau_y^{h,i} = N_h\,\bar{t}_y^h$$

$$\hat{V}(\hat{\tau}_y^h) = \hat{V}(N_h\,\bar{t}_y^h) = N_h^2\,\hat{V}(\bar{t}_y^h) = N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S'^2_{\tau_y^h}}{n_h}$$

donde $\quad S'^2_{\tau_y^h} = \frac{1}{n_h-1}\sum(\tau_y^{h,i}-\bar{t}_y^h)^2 \quad y \quad \bar{t}_y^h = \frac{1}{n_h}\sum_i \tau_y^{h,i}$

y $\tau_y^{h,i} = \sum_{\substack{j's\ de\ UP\ i \\ en\ el\ estrato\ h}} y_j$.

2.2 Valor medio por familia (UP), $\hat{\bar{\tau}}_y^h$ o $\bar{t}_y^h$

$$\bar{t}_y^h = \frac{1}{n_h} \sum_{i=1}^{n_h} \tau_y^{h,i}$$

$$\hat{V}(\bar{t}_y^h) = (1 - \frac{n_h}{N_h}) \frac{S'^2_{\tau_y^h}}{n_h}$$

$S'^2_{\tau_y^h}$: Varianza muestral corregida de los totales de las UP en el estrato.

2.3 Valor medio por individuo (US), $\bar{y}^h$

$$\bar{y}^h = \frac{N_h}{M_h} \frac{1}{n_h} \sum_{i=1}^{n_h} \tau_y^{h,i} = \frac{N_h}{M_h} \bar{t}_y^h = \bar{t}_y^h / \overline{M}_h$$

$$\hat{V}(\bar{y}^h) = \frac{N_h^2}{M_h^2} (1 - \frac{n_h}{M_h}) \frac{S'^2_{\tau_y^h}}{n_h} = \frac{1}{\overline{M}_h^2} (1 - \frac{n_h}{M_h}) \frac{S'^2_{\tau_y^h}}{n_h} = \frac{1}{\overline{M}_h^2} \hat{V}(\bar{t}_y^h)$$

Se suele definir $\overline{M}_h$: *Número poblacional medio de individuos por familia en el estrato h.*

$$\hat{V}(\hat{t}_x) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s'^2_{x^h}}{n_h} \equiv N^2 \frac{(1-f)}{n} s'^2_{\text{intra}}$$

donde $\sigma^2_{\text{intra}} \approx \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2$.

3.4.2.2 Valor medio global, $\bar{x}$

$$\bar{x} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{x}^h$$

$$\hat{V}(\bar{x}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{N_h}{N}\right) \frac{S'^2_{x^h}}{n_h} \approx (1-f) \frac{S'^2_{\text{intra}}}{n}$$

3.4.3 Caso particular: X ~ Bernoulli p

$\bar{x}^h$:	Proporción en familias del estrato *h*
$\bar{x}$:	Proporción en familias, global
$\bar{x}^h \equiv \hat{p}_h$:	Proporción por estrato $\hat{p}_h$
$\bar{x} \equiv \hat{p}$:	Proporción global $\hat{p}$

Estimadores:

$$\bar{x}^h = \sum_{i=1}^{n_h} x_i^h / n_h \qquad \bar{x} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{x}^h$$

Errores estandard:

$$\hat{V}(\bar{x}^h) = \left(1 - \frac{n_h}{N_h}\right) \frac{S'^2_{x^h}}{n_h} = \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

$S'^2_{x^h}$: varianza corregida de una Bernoulli (muestral)

$$S'^2_{x^h} = \frac{n_h}{n_h - 1} \hat{p}_h (1 - \hat{p}_h)$$

$$\hat{V}(\bar{x}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

4. JUSTIFICACIÓN: UN MUESTREO ASSR ES MEJOR QUE UN MUESTREO EN CONGLOMERADOS

Los muestreos en conglomerados suelen facilitar una mayor cobertura de la población a un coste de realización de la encuesta más barato. Para un mismo tamaño muestral final de individuos m, un muestreo en conglomerados representa una pérdida de precisión de los estimadores respecto a un muestreo aleatorio simple sin reposición (ASSR) de igual tamaño; en general, debido a la similaridad entre los individuos de una misma UP.

Los conglomerados han de ser:

- Los más hetereogeneos posibles, para mejorar la representatividad de la población total.
- El tamaño de los conglomerados ha de ser pequeño y similar entre ellos. Las familias europeas tienen pocos miembros y el tamaño es muy regular, entre 2 y 5 miembros en el 95% de los casos.
- Cuantos más conglomerados se muestreen, mejor, la calidad de los estimadores depende más de n (número de conglomerados o UP) que de m (número de individuos o US).

El efecto del conglomerado se puede medir por un coeficiente ρ llamado *coeficiente de correlación intraconglomerado:*

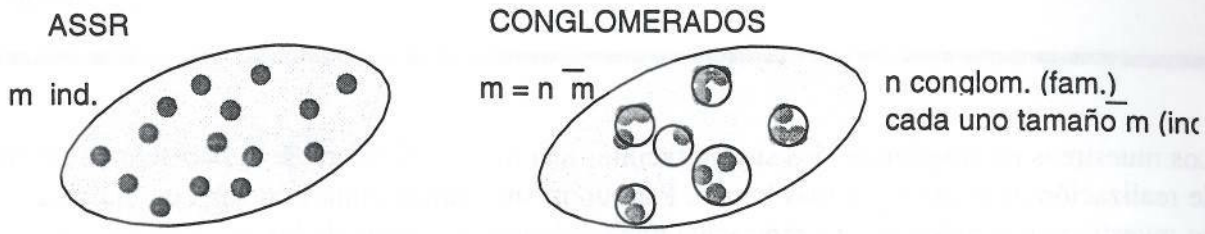
$$f = \frac{\sum_{i=1}^{n_i} \sum_{j=1}^{M_i} \sum_{k=1 \text{ y } k \neq j}^{M_i} (y_{ij} - \overline{y})(y_{ik} - \overline{y})}{\sum_{i=1}^{n_i} \sum_{j=1}^{M_i} (y_{ij} - \overline{y})^2} \cdot \frac{1}{\overline{M} - 1}$$

donde M_i: número de individuos del conglomerado i y

$$\overline{M} = \frac{M}{N}$$

Si $\rho >> 0$ Existe mucha similitud en el interior del conglomerado (desfavorable).
Si $\rho << 0$ Conglomerados hetereogéneos (favorable).

Ejemplo. Y: Viajes en autobús por individuo.



Estimador del total de viajes en autobús:

- Muestreo ASSR:

$$V_1(\hat{\tau}) = M^2 \frac{\sigma'^2}{n\,\overline{m}}\left(1-\frac{m}{M}\right) \approx M^2 \frac{\sigma'^2_y}{n\,\overline{m}}$$

- Muestreo en conglomerados:

$$V_2(\hat{\tau}) = N^2\left(1-\frac{n}{N}\right)\frac{\sigma'^2_{\tau_y}}{n} \approx N^2 \frac{\sigma'^2_{\tau_y}}{n}$$

y se puede demostrar equivalente, usando ρ, el coeficiente de correlación interconglomerado a:

$$V_2(\hat{\tau}) = M^2 \frac{\sigma'^2_y}{n\,\overline{m}}(1+\rho(\overline{m}-1))$$

Haciendo en cociente:

$$\frac{V_2(\hat{\tau})}{V_1(\hat{\tau})} = 1+\rho(\overline{m}-1)$$

Si $\rho >> 0$ entonces el muestreo en conglomerados es peor que el ASSR .

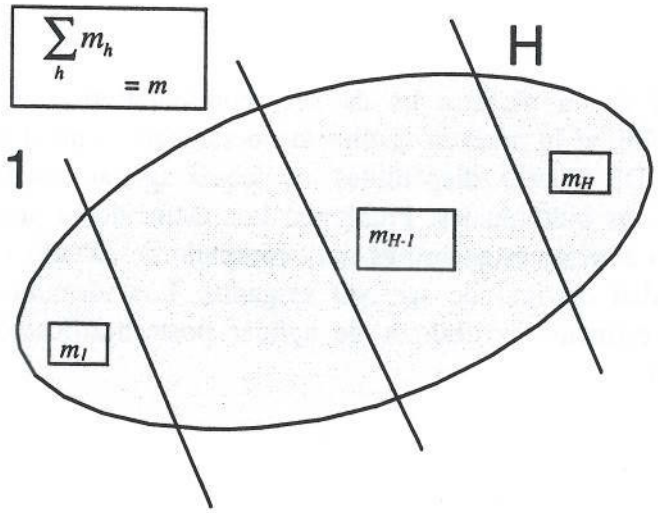
Los dos tipos de muestreo combinados tienen muestran distintas propiedades respecto al ASSR en los estimadores que facilitan:

- Estratificado: reducción error estandar respecto ASSR, por tanto incremento de precisión.
- Conglomerado: incremento error estandar respecto ASSR, por tanto decremento de la precisión.

Las propiedades contrapuestas de los dos esquemas de muestreo combinados sobre el error de los estimadores finales tiene por efecto un cierto control de la pérdida de precisión de los estimadores debido al efecto de los conglomerados.

Por otro lado, el muestreo ASSR suele dar menos precisión (más error estandard) en los estimadores que un muestreo estratificado proporcional (tasa muestreo f constante por estrato, como en el presente caso). La justificación se presenta brevemente a continuación:

Sea Y población con σ^2, m el tamaño total de una muestra tomada sobra una población de tamaño M.



La varianza total puede descomponerse como una varianza intraestrato más una varianza interestrato.

$$\sigma^2 = \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^{H} \frac{N_h}{N} (\mu_h - \mu)^2 = \sigma^2_{\text{int}\, ra} + \sigma^2_{\text{int}\, er}$$

Para $\bar{y}$, el estimador del valor medio de la variable de interes se puede determinar:

- ASSR,

$$V_1(\bar{y}) = (1-f)\frac{\sigma'^2}{m}$$

- Estratificado proporcional ,

$$V_2(\bar{y}) = (1-f)\frac{\sigma'^2_{\text{int}\, ra}}{m}$$

Haciendo cociente de los errores en ambos tipos de muestreo:

$$\frac{V_2(\bar{y})}{V_1(\bar{y})} = \frac{\sigma'^2_{\text{int}\, ra}}{\sigma'^2} = \frac{\sigma'^2 - \sigma'^2_{\text{int}\, er}}{\sigma'^2} = 1 - \frac{\sigma'^2_{\text{int}\, er}}{\sigma'^2} \leq 1$$

donde queda de manifiesto que si existen diferencias en los valores medios en los distintos estratos, entonces el muestreo estratificado proporcional siempre es mejor que el muestreo ASSR.

5. DETALLES TÉCNICOS (OPCIONAL)

El resultado final de la muestra ha de ser representativo a nivel de individuo y MUY PROBABLEMENTE, si la muestra resultante no satisface a nivel de individuo las cuotas por SEXO y GRUPO DE EDAD (disponibles) se deberá aplicar una POST-ESTRATIFICACIÓN para conseguir cuotas individuales. Problema: Los estimadores puntuales son muy fácilmente recalculables; pero el error estandard es muy complejo de cálculo y a menudo los estimadores pierden la propiedad original de ser NO sesgados: Las fórmulas para el cálculo del error estandard de los estimadores después de aplicar postestratificación no se incluyen en este documento técnico.