DISEÑOS MUESTRALES POR CONGLOMERADOS

DISEÑOS EN UNA Y DOS ETAPA

Muestreo por Conglomerados

⊗ Los diseños muestrales presentados hasta el momento asumen:

a) La existencia de una lista o marco muestral de las unidades

b) La seleción de unidades elementales u_k en forma directa

⊗ Existen otros diseños que son útiles cuando:

.... no existe marco muestral que identifique a c/u de las unidades de la población, o su construcción resulta imposible o es muy costoso

..... las unidades u_k se encuentran distribuidas en un área muy grande lo que llevaría a una muestra muy dispersa geográficamente y resulta muy costoso alcanzarlas

El Conglomerado (Unidades Primarias)

- ⊗ Conglomerado o Cluster: Agrupamiento natural o inducido de las unidades u_k de la Población U

Población	Unidad	Conglomerado
Habitantes	Persona	Hogar o Vivienda
Viviendas	Vivienda	Manzana, Segmento Censal
Alumnos	Alumno	Colegio, Nivel o Grado
Pacientes Hospitales	Paciente	Hospitales
Pasajeros Aeropuerto	Pasajero	Avión y/o Tramo Horario
Clientes Bancarios	Clientes	Entidad Bancaria
Arboles en bosque	Arbol	Parcelas o áreas
Plantación	Plantas	Surcos

- ⊗ Por lo general son entidades espaciales o geográficas con identidad física propia y definida sin ambigüedad
- ⊗ En MPF al conglomerado se los denomina también UPM (Unidad Primaria de Muestreo)

Muestreo por Conglomerados en 1 etapa

$$U_{UP} = \{C_1, C_2, ..., C_M\} \quad \# C_i = N_i \quad i = 1, ..., M \quad U = \bigcup_{i \in U} C_i$$

⊗ Los C_i son una partición disjunta de U, $C_i \bigcap C_j = \varnothing \quad \forall i, j \quad i \neq j$

⊗ Cada u_k, de ahora "unidades elementales" o "secundarias" pertenece solo a un C_i o UPM

$$d^{UP} = \left(\Omega_{UP}, P_{d^{UP}}\right) \qquad \Omega_{UP} \text{ soporte de Conglomerados}$$

⊗ Las probabilidades

$$\pi_i^{UP} = \sum_{s_{UP/i \in s_{UP}}} p_{d^{UP}}(s_{UP}) \quad \text{para } C_i$$

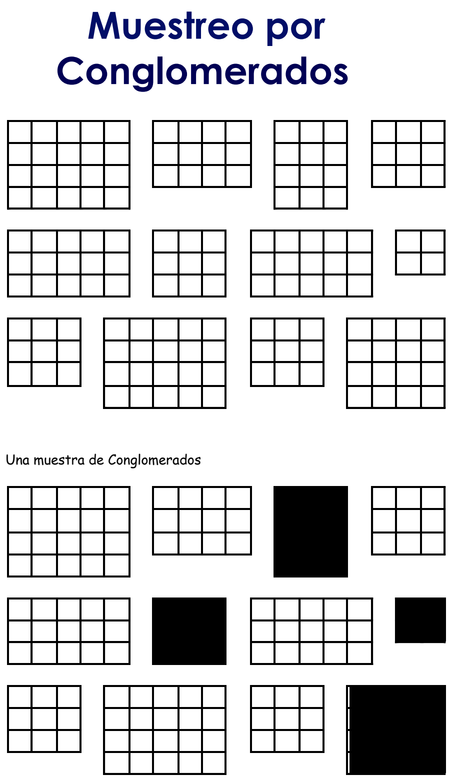
$$\pi_{ij}^{UP} = \sum_{s_{UP/i,j \in s_{UP}}} p_{d^{UP}}(s_{UP}) \quad \text{para } C_i \text{ y } C_j$$

Muestreo por Conglomerados en 1 etapa

- ⊗ La muestra de UP: $s_{UP} \quad \# s_{UP} = n(s_{UP})$
- ⊗ Censo en cada $i \in s_{UP}$
- ⊗ La muestra de unidades elementales

$$s = \bigcup_{i \in s_{UP}} C_i \qquad \# C_i = N_i$$

$$\# s = n_s = \sum_{i \in s_{UP}} N_i$$



- ⊗ $\forall k \in U$ existe un y solo un $C_i \,/\, u_k \in C_i$
  pero no necesariamente debe existir una lista de las u_k antes de la selección
- ⊗ Generalmente se la confecciona o actualiza una vez que se tiene s_{UP}

Probabilidades a nivel de elementos

Como se censan los C_i de s_{UP}
$$\begin{cases} \pi_k = \Pr(k \in s) = \Pr(i \in s_{UP}) = \pi_i^{UP} & si\ k \in C_i \\ \pi_{kl} = \Pr(k,l \in s) = \Pr(i \in s_{UP}) = \pi_i^{UP} & si\ k,l \in C_i \\ \pi_{kl} = \Pr(k,l \in s) = \Pr(i,j \in s_{UP}) = \pi_{ij}^{UP} & si\ k \in C_i\ y\ l \in C_j \end{cases}$$

$$U_{UP} = \{\{u_1,u_2\},\{u_3,u_4\},\{u_5,u_6,u_7\},\{u_8\}\} = \{C_1,C_2,C_3,C_4\}$$

$$\Omega_{UP} = \Omega_2 = \{\{C_1,C_2\},\{C_1,C_3\},\{C_1,C_4\},\{C_2,C_3\},\{C_2,C_4\},\{C_3,C_4\}\}$$

$$P_d: \quad 3/24\ ,\quad 3/24\ ,\quad 5/24\ ,\quad 4/24\ ,\quad 7/24\ ,\quad 2/24$$

$$d^{UP} \text{ fijo } \#s^{UP} = 2, \text{no uniforme y } \#s = \begin{cases} 3 & \text{si } s = C_1 \cup C_4 \text{ o } s = C_2 \cup C_4 \\ 4 & \text{si } s = C_1 \cup C_2 \text{ o } s = C_3 \cup C_4 \\ 5 & \text{si } s = C_1 \cup C_3 \text{ o } s = C_2 \cup C_3 \end{cases}$$

$$\pi^{UP} = (11/24, 14/24, 9/24, 14/24)'$$

$$\pi_2^{UP} = \frac{14}{24} \qquad \pi^{UP}{}_{24} = \frac{7}{24} \qquad \pi_2 = \frac{11}{24} \qquad \pi_{25} = \frac{3}{24} \text{ pero } \pi_{21} = \frac{11}{24}$$

Estimador y Varianza del HT

Dado el total, $T_y = \sum_{k\in U} y_k = \sum_{i\in U_{UP}} T_{yi}$ con $T_{yi} = \sum_{k\in C_i} y_k$

El estimador para el total T_y es,

$$\hat{T}_{\pi y} = \sum_{i\in s_{UP}} \frac{T_{yi}}{\pi_i^{UP}} = \sum_{i\in s_{UP}} \sum_{\substack{k\in C_i \\ C_i\in s_{UP}}} \frac{y_k}{\pi_i^{UP}} = \sum_{k\in s} \frac{y_k}{\pi_k}$$

⊗ Recordar que seleccionado un C_i por el diseño d^{UP}, se lo censa

con varianza,

$$\mathrm{V}_{\mathrm{HT}}(\hat{T}_{\pi y}) = \sum_{i\in U_{UP}} \sum_{j\in U_{UP}} (\pi_{ij}^{UP} - \pi_i^{UP}\pi_j^{UP}) \frac{T_{iy}}{\pi_i^{UP}} \frac{T_{jy}}{\pi_j^{UP}}$$

y si el diseño es de tamaño fijo

$$V_{SGY}(\hat{T}_{\pi y}) = -\frac{1}{2} \sum_{i\in U_{UP}} \sum_{j\in U_{UP}} \left(\pi_{ij}^{UP} - \pi_i^{UP}\pi_j^{UP}\right) \left(\frac{T_{iy}}{\pi_i^{UP}} - \frac{T_{jy}}{\pi_j^{UP}}\right)^2$$

Estimadores de la varianza

Los estimadores de varianza son:

$$\hat{\mathrm{V}}_{\mathrm{HT}}(\hat{T}_{\pi y}) = \sum_{i \in s_{UP}} \sum_{j \in s_{UP}} \left(\frac{\pi_{ij}^{UP} - \pi_i^{UP} \pi_j^{UP}}{\pi_{ij}^{UP}} \right) \frac{T_{yi}}{\pi_i^{UP}} \frac{T_{yj}}{\pi_j^{UP}}$$

y si el diseño d^{UP}, es fijo

$$\hat{V}_{SGY}(\hat{T}_{\pi y}) = -\frac{1}{2} \sum_{i \in s_{UP}} \sum_{j \in s_{UP}} \left(\frac{\pi_{ij}^{UP} - \pi_i^{UP} \pi_j^{UP}}{\pi_{ij}^{UP}} \right) \left(\frac{T_{yi}}{\pi_i^{UP}} - \frac{T_{yj}}{\pi_j^{UP}} \right)^2$$

- ⊗ La elección de $\pi_i^{UP} \propto X_i$ es una alternativa tentadora si $Corr(T_{iy}, X_i)$ es alta o moderadamente alta

- ⊗ Una posibilidad es $X_i = N_i$ y $N_i \neq N_0 \ \forall i, i = 1, \ldots, M$

- ⊗ O sea los diseños π_{PPT} en general se emplean cuando se involucran Conglomerados

MSA de Conglomerados (MCSA)

$$U_{UP} = \{C_1, C_2, \ldots, C_M\} \qquad \# C_i = N_i$$

$$d^{UP} = (\Omega_m, P_{MSA}) \quad \#\Omega_m = m \quad P_{MSA} = \binom{M}{m}^{-1}$$

$$\pi_i^{UP} = m/M \qquad \pi^{UP}{}_{ij} = m(m-1)/M(M-1) \text{ para } i \neq j$$

$$\hat{T}_{\pi y} = M\hat{\bar{T}}_y \qquad \hat{\bar{T}}_y = \frac{\sum_{i \in s_{UP}} T_{iy}}{m} = \frac{\sum_{i=1}^{m} T_{iy}}{m}$$

$$V_{HT} = V_{SGY} = M^2\left(\frac{1}{m} - \frac{1}{M}\right)S_T^2 \qquad S_T^2 = \frac{\sum_{i=1}^{M}\left(T_{iy} - \bar{T}_y\right)^2}{M-1} \quad \bar{T}_y = \frac{\sum_{i=1}^{M} T_{iy}}{M}$$

$$\hat{V}_{HT} = \hat{V}_{SGY} = M^2\left(\frac{1}{m} - \frac{1}{M}\right)s_T^2 \qquad s_T^2 = \frac{\sum_{i=1}^{m}\left(T_{iy} - \hat{\bar{T}}_{\pi y}\right)^2}{m-1} \quad \hat{\bar{T}}_{\pi y} = \frac{\sum_{i=1}^{m} T_{iy}}{m}$$

Conglomerados y Estratos (diferencias)

Los dos diseños dividen a la población U en grupos
pero la composición interna de los grupos juega un papel importante

Muestreo por Conglomerados en una etapa:

⊗ Divide la población en M conglomerados de tamaños N_i

⊗ Se selecciona una muestra de m conglomerados y se los censa

Muestreo Estratificado:

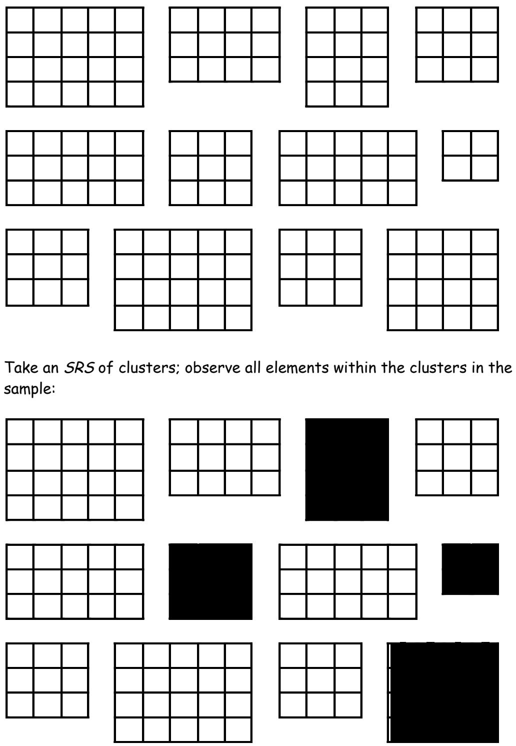
⊗ Divide a la población en H estratos de tamaños $N_h, h = 1,..,H$

⊗ Se selecciona una muestra de n elementos con n_h en cada estrato

El muestreo estratificado aumenta la precisión de las estimaciones mientras que el muestreo por conglomerados la disminuye .....

Conglomerados vs Estratos

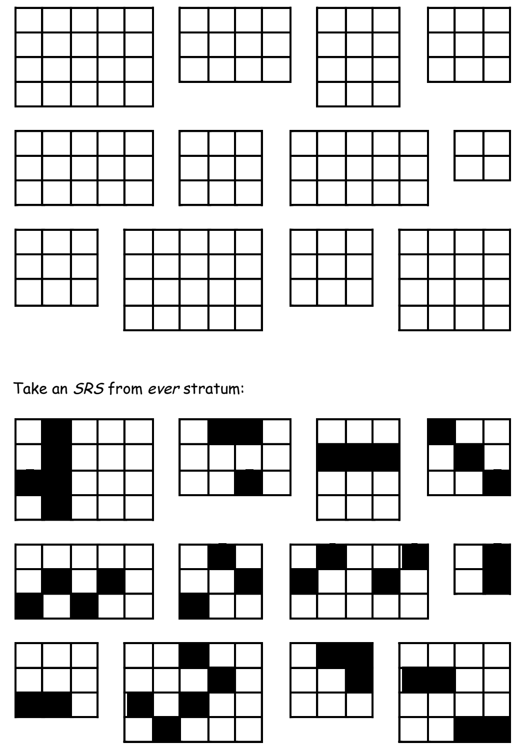
Muestreo por Conglomerados



Variabilidad de Y, por dejar

C_i sin muestrear, por $N_i \neq N_0$

Muestreo Estratificado



Variabilidad de Y

Muestra en todos los Estratos

La homogeneidad interna de los grupos tiene impacto en la V_{HT} pero de manera diferente ....

Correlación Muestral del Diseño

Dado un $d=\left(\Omega_n, P_d\right)$ fijo, por la equivalencia de Knottnerus (2003):

$$\boxed{V_{HT}\left(\hat{T}_{\pi y}\right)=\left(1+\left(n-1\right)\rho\right)\tilde{\sigma}_y^2}$$

con:

$$\rho=\frac{\sum_{k\in U}\sum_{l\in U}\pi_{kl}\left(\frac{y_k}{\pi_k}-\frac{T_y}{n}\right)\left(\frac{y_l}{\pi_l}-\frac{T_y}{n}\right)}{(n-1)\tilde{\sigma}^2} \quad \text{y} \quad \tilde{\sigma}_y^2=\sum_{k\in U}\pi_k\left(\frac{y_k}{\pi_k}-\frac{T_y}{n}\right)^2$$

$$\boxed{-\frac{1}{n-1}\leq\rho\leq 1} \qquad \text{que varía según } d, n \text{ e } Y$$

⊗ Si $d=MSA \Rightarrow \rho=0$

⊗ Si $d=MSys \Rightarrow \rho=\dfrac{\sum_r\left[\sum_{k\in s_r}\sum_{l\in s_r}\left(y_k-\bar{Y}_U\right)\left(y_l-\bar{Y}_U\right)\right]}{\left(n-1\right)\left(N-1\right)S_{yU}^2}$ donde existen r arranques posibles

⊗ Si $d^{UP}=MSA$ de conglomerados en 1 etapa: $\rho=\dfrac{\sum_{i\in U_{UP}}\left[\sum_{k\in C_i}\sum_{l\in C_i}\left(y_k-\bar{Y}_U\right)\left(y_l-\bar{Y}_U\right)\right]}{\left(n-1\right)\left(N-1\right)S_{yU}^2}$

El Grado de Homogeneidad Interna

Se define como "grado de homogeneidad" del diseño d (fijo) a la cantidad:

$$\boxed{\delta = \left(1 + \left(N-1\right)\rho\right) / N}$$

⊗ A δ se lo puede identificar como el R^2_{adj} de la regresión de Y con M variables dummy´s, $i = 1,\ldots,M$ o indicatrices

$$X_{ij} = \begin{cases} 1 \ \text{si} \ j \in C_i \\ 0 \ \text{si} \ j \notin C_i \end{cases} \text{ sobre los } N \text{ puntos de la población, } \#U = N$$

⊗ Si el diseño es un d^{UP} (fijo) sobre $U_{UP} = \left\{C_1,\ldots,C_M\right\}$ con $\#C_i = N_i$

$$\boxed{-\frac{M-1}{N-M} \leq \delta \leq 1}$$

$$\boxed{\begin{array}{l} \delta \sim 1 \Rightarrow d \text{ posee muestras internamente muy homogéneas} \\ \delta \leq 0 \Rightarrow d \text{ posee muestras internamente heterogéneas} \end{array}}$$

Bajo MSA de Conglomerados (MCSA)

Sea $U_{UP} = \left\{ C_1, \ldots., C_M \right\}$

$\#U_{UP} = M \quad \#C_i = N_i \quad N = \sum_{i \in U_{UP}} N_i$

Se supone MSA(M, m) sobre U_{UP}

Cada vez que dividimos la Población se tiene

$$(N-1) S_{yU}{}^2 = (N-M) S_{yD}^2 + S_{yE}^2$$

$$S_{yD}^2 = \frac{1}{N-M} \sum_{i \in U_{UP}} \sum_{j \in C_i} \left(Y_{ij} - \bar{Y}_i \right)^2 = \frac{\sum_{i \in U_{UP}} \left(N_i - 1 \right) S_i^2}{\sum_{i \in U_{UP}} \left(N_i - 1 \right)} \qquad S_i^2 = \frac{1}{N_i - 1} \sum_{j \in C_i} \left(Y_{ij} - \bar{Y}_i \right)^2$$

$$S_{yE}^2 = \sum_{i \in U_{UP}} N_i \left(\bar{Y}_i - \bar{Y}_U \right)^2$$

el grado de homogeneidad es quivalente a $\delta = 1 - \dfrac{S_{yD}^2}{S_{yU}^2}$

Descomposición de la Varianza

$$V_{HT}(\hat{T}_{\pi y}) = M^2\left(\frac{1}{m}-\frac{1}{M}\right)\sum_{i\in U_{UP}}\frac{\left(T_{iy}-\bar{T}_y\right)^2}{M-1} = M^2\left(\frac{1}{m}-\frac{1}{M}\right)S_{Ty}^2 \qquad \bar{T}_y = \sum_{i=1}^{M} T_{iy}\Big/M$$

definiendo $\quad Cov\left(N_i, N_i\bar{Y}_i^2\right) = \dfrac{\sum_{i\in U_{UP}}\left(N_i-\bar{N}\right)N_i\bar{Y}_i^2}{M-1}$

$\bar{N} = N/M$ promedio de unidades por Conglomerado

$$S_{Ty}^2 = \bar{N}S_y^2\left(1+\frac{N-M}{M-1}\delta\right)+Cov\left(N_i, N_i\bar{Y}_i^2\right)$$

$$V_{MCSA}(\hat{T}_{\pi y}) = KS_{Ty}^2 = K\bar{N}S_y^2\left(1+\frac{N-M}{M-1}\delta\right)+Cov$$

con $K = M^2\left(\dfrac{1}{m}-\dfrac{1}{M}\right)$

Comparación entre MCSA vs MSA

Bajo un MSA de Conglomerados el tamaño esperado de la muestra de unidades elementales es

$$E_d(n_s) = m\bar{N} = n$$

para comparar $V_{MCSA}(\hat{T}_{\pi y})$ con $V_{MSA}(\hat{T}_{\pi y}) = \bar{N}KS_y^2$

se emplea el Efecto de Diseño, $deff$:

$$deff(MCSA, MAS) = \frac{V_{MCSA}}{V_{MSA}}$$

$$deff(MCSA, MSA) = 1 + \frac{N-M}{M-1}\delta + \frac{Cov}{\bar{N}S_{yU}^2}$$

se puede ver que $deff$ depende de:

1) δ el grado de homogeneidad
2) Cov la covarianza entre N_i y $N_i\bar{y}_i^2$
3) S_{yU}^2

La Ineficiencia del MCSA sobre MSA

$\boxed{\text{CASO 1}}$: $N_i = N_0 \quad \Rightarrow \quad Cov = 0$

$$deff(MCSA, MSA) = 1 + \left(\frac{N-M}{M-1}\right)\delta$$

$V_{MCSA} \le V_{MSA} \quad \Leftrightarrow \quad \delta < 0 \quad \Rightarrow \quad deff \le 1$

$\delta < 0$ difícil en la prática

aún con $\delta = 0.08 \quad N_0 = 300 \quad deff \cong 25$

$\boxed{\text{CASO 2}}$: $Var(N_i) \ne 0 \ \Rightarrow \ Cov > 0$

usualmente es $deff \gg 1$

aún para $\delta_{\min} = -\dfrac{(M-1)}{(N-M)}$ o sea $\bar{Y}_{C_i} = \bar{Y} \quad i = 1, \ldots, M$

$$deff(MCSA, MSA) \approx \bar{N}\left(\frac{CV_N}{CV_y}\right)^2$$

$\bar{N} = N/M$ promedio de unidades por Conglomerado

$\bar{N} = 300,\ CV_y = 2,\ CV_N = 0.2 \Rightarrow deff \cong 3$

se aconseja

$\delta \ll 1$ y $N_i = N_0$ pequeño

o con pequeña variación

¿Cómo Atender esta Ineficiencia?

- $\otimes$ Sin ser la mejor solución, aumentar la muestra de UP ($m \nearrow$)
- $\otimes$ Seguramente provoca un aumento en los costos ($\$ \nearrow$)
- $\otimes$ Redefinir a los conglomerados, disminuyendo sus tamaño N_i y aumentando M $(N_i \swarrow \Rightarrow \delta \swarrow)$. No siempre es posible.
- $\otimes$ Aceptar aumentar m pero submuestrear dentro de ellos o sea no Censar el conglomerado seleccionado ($m \nearrow$ y $s_i \subset N_i$)
- $\otimes$ Se crea otra fuente de incertidumbre o variación por imponer nuevos diseños sobre los conglomerados seleccionados en la primer etapa de muestreo.

Muestreo por Conglomerados en 2 etapa

$\boxed{\text{Primera Etapa}}$ $d^{UP} = \left(\Omega_{UP}, P_{d^{UP}}\right)$ sobre $U_{UP} = \left\{C_1, \ldots, C_M\right\}$

$$s_{UP} \in \Omega_{UP} \quad , s_{UP} = \bigcup_{i \in s_{UP}} C_i \qquad \# s_{UP} = n(s_{UP})$$

$\boxed{\text{Segunda Etapa}}$ Sean $d_i = \left(\Omega_{C_i}, P_{d_i}\right)$ $\forall i \ i = 1, .., M$ independientes en cada

C_i seleccionada en la 1er etapa

Si $s_i \in \Omega_{C_i}$ son las muestras de 2da Etapa, $i = 1, .., n(s_{UP})$ con $\# s_i = n(s_i)$

$\boxed{\text{La Muestra Final}}$ está compuesta por $s = \bigcup_{i \in s_{UP}} s_i$

con diseño $P_d(s) = P_{d^{UP}}(s_{UP}) \prod_{j=1}^{n(s_{UP})} P_{d_i}(s_i / s_{UP})$

y $n(s) = \sum_{i \in s_{UP}} n(s_i)$

Condiciones

Invariante: Dado un conglomerado $C_i \in s^{UP}$, la muestra s_i es seleccionada tal que

$$P_{d_i}\left(./\, s^{UP}\right) = P_{d_i}\left(.\right)$$

"Cada vez que C_i es incluida en una muestra de 1er etapa el mismo diseño d_i deberá ser empleado irrespectivamente de los otros conglomerados elegidos"

Independencia: El submuestreos d_i se ejecutan independientemente en cada C_i

$$P\left(\bigcup_{i \in s^{UP}} s_i \,/\, s^{UP}\right) = \prod_{i \in s^{UP}} P_{d_i}\left(s_i \,/\, s^{UP}\right)$$

⊗ El no cumplimento de algunas de las condiciones lleva a los diseños muestrales denominados en "2 fases"

Las Probabilidades π_k y π_{kl}

Para $k \in U \quad \pi_k = \pi_i^{UP}\pi_{k|i} \quad \text{si} \quad k \in C_i$

$\pi_{k/i}$ probabilidad de 1er orden de la unidad k en el C_i según el diseño d_i definido en él

$\pi_{kl/i}$ probabilidad de inclusión de 2do orden para k y l según el diseño d_i aplicado en él

$$\pi_{kl} = \begin{cases} \pi_i^{UP}\pi_{k|i} & \text{si} \quad k = l \in C_i \\ \pi_i^{UP}\pi_{kl|i} & \text{si } k \text{ y } l \in C_{\mathrm{i}}, k \neq l \\ \pi_{ij}^{UP}\pi_{k|i}\pi_{l|j} & \text{si } k \in C_i \text{ y } l \in C_j,\ i \neq j \end{cases}$$

Cálculo de las π_k y π_{kl}

$$U=\{1,2,......,15\}=\{\{1,2,3\},\{4,5,6,7\},\{8,9,10\},\{11,12,13,14,15\}\}$$

$$U_{UP}=\{C_1,C_2,C_3,C_4\}\ \Omega_{UP}=\{\{C_1,C_2\},\{C_1,C_3\},\{C_2,C_3\},\{C_2,C_4\}\}$$

$$P_{UP}:\quad 3/8\qquad 2/8\qquad 2/8\qquad 1/8$$

$$\pi^{UP}=(5/8,6/8,4/8,1/8)\qquad \pi_{ij}^{UP}=\begin{pmatrix} 5/8 & 3/8 & 2/8 & 0\\ & 6/8 & 2/8 & 1/8\\ & & 4/8 & 0\\ & & & 1/8 \end{pmatrix}$$

$$d_i:MSA(N_i,2)\quad k,l\in C_i\quad \pi_{k/i}=\frac{2}{N_i}\quad \pi_{kl/i}=\frac{2*1}{N_i(N_i-1)}$$

$$\text{Ej:}\quad k=5\quad \{5\}\subset C_2\quad \pi_5=\pi_2^{UP}\pi_{5/2}=\frac{6}{8}\frac{2}{4}=\frac{3}{8}$$

$$k=7\quad \{7\}\subset C_2\quad \pi_7=\pi_2^{UP}\pi_{7/2}=\frac{1}{8}\frac{2}{4}=\frac{3}{8}$$

$$k=12\quad \{12\}\subset C_4\quad \pi_{12}=\pi_4^{UP}\pi_{12/4}=\frac{6}{8}\frac{2}{5}=\frac{3}{10}$$

$$\pi_{57}=?$$

$$\pi_{512}=?$$

Estimación para el Total Poblacional

- La estimación de T_Y a nivel población involucra estimaciones de totales en los conglomerados seleccionados
- O sea: $\hat{T}_{\pi y} = \sum_{i \in s_{UP}} \frac{\hat{T}_{i\pi y}}{\pi_i^{UP}}$
- A nivel de unidades de segunda etapa USM

$$\hat{T}_{\pi y} = \sum_{i \in s_{UP}} \frac{\hat{T}_{i\pi y}}{\pi_i^{UP}} = \sum_{i \in s_{UP}} \sum_{k \in s_i} \frac{y_k}{\pi_i^{UP} \pi_{k|i}} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} w_k y_k$$

con la ponderación o ponderador, $w_k = \frac{1}{\pi_i^{UP} \pi_{k|i}} = \frac{1}{\pi_k}$ si $k \in C_i$

con varianza según la teoría, $V_{HT}\left(\hat{T}_{\pi y}\right) = \sum_{k \in U} \sum_{l \in U} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{Y_k}{\pi_k} \frac{Y_l}{\pi_l}$

- Existe la alternativa de descomponer $V_{HT}\left(\hat{T}_{\pi y}\right)$ en componentes de 1er y 2da etapa $V_{HT}(\hat{T}_{\pi}) = V_{UP} + V_{US}$

Descomposición de V_{HT}

En todo diseño muestral en 2 Etapas existen 2 fuentes de variación por muestra dada por cada etapa de selección

⊗ Recordar que si X es una v.a y A es un evento:

$$E(X) = E_A\left(\left[E(X/A)\right]\right)$$

$$V(X) = V_A\left[E(X/A)\right] + E_A\left[V(X/A)\right]$$

si se toma $A = s^{UP}$ se tiene:

$$E_d\left(\hat{T}_{\pi y}\right) = E_{d^{UP}}\left[E\left(\hat{T}_{\pi y} / s^{UP}\right)\right]$$

$$V_I E_{II}\left(\hat{T}_{\pi y}\right) = V_{d^{UP}}\left[E\left(\hat{T}_{\pi y} / s^{UP}\right)\right] \quad E_I V_{II}\left(\hat{T}_{\pi y}\right) = E_{d^{UP}}\left[V\left(\hat{T}_{\pi y} / s^{UP}\right)\right]$$

$$V_d\left(\hat{T}_{\pi y}\right) = V_{d^{UP}}\left[E\left(\hat{T}_{\pi y} / s^{UP}\right)\right] + E_{d^{UP}}\left[V\left(\hat{T}_{\pi y} / s^{UP}\right)\right]$$

(A) + (B)

Componente (A) de V_{HT}

$$(A)\quad E\left(\hat{T}_{\pi y} / s^{UP}\right) = E\left(\sum_{i \in s_{UP}} \frac{\hat{T}_{i\pi y}}{\pi_i^{UP}} / s_{UP}\right) =$$

$$= \sum_{i \in s_{UP}} E_{d_i}\left(\frac{\hat{T}_{i\pi y}}{\pi_i^{UP}} / s_{UP}\right) = (por\ \text{invarianza})$$

$$= \sum_{i \in s_{UP}} E_{d_i}\left(\frac{\hat{T}_{i\pi y}}{\pi_i^{UP}}\right) = \sum_{i \in s_{UP}} \frac{T_{iy}}{\pi_i^{UP}}$$

$$\therefore\ V_{d^{UP}}\left[E\left(\hat{T}_{\pi y} / s^{UP}\right)\right] = V_{d^{UP}}\left(\sum_{i \in s_{UP}} \frac{T_{iy}}{\pi_i^{UP}}\right) = \boxed{\sum_{i \in U_{UP}} \sum_{j \in U_{UP}} \Delta_{ij}^{UP} \frac{T_{iy}}{\pi_i^{UP}} \frac{T_{jy}}{\pi_j^{UP}}}$$

Componente (B) de V_{HT}

$$(B) \qquad V\left(\hat{T}_{\pi y} / s_{UP}\right) = V\left(\sum_{i \in s_{UP}} \frac{\hat{T}_{i\pi y}}{\pi_i^{UP}} / s_{UP}\right) = (por\ independencia)$$

$$= \sum_{i \in s_{UP}} V_{d_i}\left(\frac{\hat{T}_{i\pi y}}{\pi_i^{UP}} / s_{UP}\right) = (por\ \text{invarianza})$$

$$= \sum_{i \in s_{UP}} V_{d_i}\left(\frac{\hat{T}_{i\pi y}}{\pi_i^{UP}}\right) = \sum_{i \in s_{UP}} \frac{1}{\left[\pi_i^{UP}\right]^2} V_i\left(\hat{T}_{i\pi y}\right)$$

$$con \quad V_i\left(\hat{T}_{i\pi y}\right) = \sum_{k \in C_i} \sum_{l \in C_i} \Delta_{kl/i} \frac{y_k}{\pi_{k/i}} \frac{y_l}{\pi_{l/i}}$$

$$\therefore \quad E_{d^{UP}}\left[V\left(\hat{T}_{\pi y} / s^{UP}\right)\right] = E_{d^{UP}}\left(\sum_{i \in s_{UP}} \frac{1}{\left[\pi_i^{UP}\right]^2} V_i\left(\hat{T}_{i\pi y}\right)\right) =$$

$$= E_{d^{UP}}\left(\sum_{i \in s_{UP}} \frac{\left[V_i\left(\hat{T}_{i\pi y}\right) / \pi_i^{UP}\right]}{\pi_i^{UP}}\right) = \boxed{\sum_{i \in U_{UP}} \frac{V_i\left(\hat{T}_{i\pi}\right)}{\pi_i^{UP}}}$$

Resumiendo

⊗ En todo diseño muestral en 2 Etapas vale

$$V_{HT}(\hat{T}_{\pi y}) = V_{UP} + V_{US}$$

⊗ V_{UP} es la componente de variabilidad atribuida a la 1er Etapa

$$V_{UP} = \sum_{i \in U_{UP}} \sum_{j \in U_{UP}} \left(\pi_{ij}^{UP} - \pi_i^{UP}\pi_j^{UP}\right) \frac{T_{iy}\ T_{jy}}{\pi_i^{UP}\pi_j^{UP}}$$

⊗ V_{US} es la componente de variabilidad asociado a la 2da Etapa

$$V_{US} = \sum_{i \in U_{UP}} \frac{V_i\left(\hat{T}_{iy}\right)}{\pi_i^{UP}}$$

⊗ $\frac{V_{UP}}{V_{HT}}$ y $\frac{V_{US}}{V_{HT}}$ miden la contribución relativa de la 1er y 2da etapas sobre la varianza del estimador

Estimación de las Componentes V_{UP} y V_{US}

⊗ Una estimación insesgada para V_{UP} es:

$$\hat{V}_{UP} = \sum_{i \in s_{UP}} \sum_{j \in s_{UP}} \left(\frac{\pi_{ij}^{UP} - \pi_i^{UP} \pi_j^{UP}}{\pi_{ij}^{UP}} \right) \frac{\hat{T}_{i\pi}}{\pi_i^{UP}} \frac{\hat{T}_{j\pi}}{\pi_j^{UP}} - \sum_{i \in s_{UP}} \frac{1}{\pi_i^{UP}} \left(\frac{1}{\pi_i^{UP}} - 1 \right) \hat{V}_i \left(\hat{T}_{i\pi} \right)$$

⊗ Una estimación insesgada para V_{US} es:

$$\hat{V}_{US} = \sum_{i \in s_{UP}} \frac{\hat{V}_i \left(\hat{T}_{i\pi} \right)}{\left(\pi_i^{UP} \right)^2}$$

⊗ Por lo tanto una estimación insesgada para V_{HT} es

$$\hat{V}(\hat{T}_\pi) = \hat{V}_{UP} + \hat{V}_{US} = \sum_{i \in s_{UP}} \sum_{j \in s_{UP}} \left(\frac{\pi_{ij}^{UP} - \pi_i^{UP} \pi_j^{UP}}{\pi_{ij}^{UP}} \right) \frac{\hat{T}_{i\pi}}{\pi_i^{UP}} \frac{\hat{T}_{j\pi}}{\pi_j^{UP}} + \sum_{i \in s_{UP}} \frac{\hat{V}_i \left(\hat{T}_{i\pi} \right)}{\pi_i^{UP}}$$

Diseños MSA en Ambas Etapas

Primera Etapa: Muestra de m conglomerados sobre M por MSA

$$d_{UP} = \left(\Omega_{UP}, P_{d_{UP}}\right) \quad \text{con} \quad P_{d_{UP}} = 1 \Big/ \binom{M}{m}$$

Segunda Etapa: Muestras de n_i unidades de N_i con diseños MSA(N_i, n_i)

$$d_{UP_i} = \left(\Omega_{UP_i}, P_{d_{UP_i}}\right) \quad \text{con} \quad P_{d_{UP_i}} = 1 \Big/ \binom{N_i}{n_i}$$

$$\pi_k = \pi_i^{UP} \pi_{k|i} = \frac{m}{M} \frac{n_i}{N_i} \quad \text{si } k \in C_{\mathrm{i}}$$

$$\pi_{kl} = \begin{cases} \pi_i^{UP} \pi_{kl|i} = \dfrac{m}{M} \dfrac{n_i\left(n_i - 1\right)}{N_i\left(N_i - 1\right)} & \text{si } k \text{ y } l \in C_{\mathrm{i}}, k \neq l \\[2ex] \pi_{ij}^{UP} \pi_{k|i} \pi_{l|j} = \dfrac{m}{M} \dfrac{n_i}{N_i} \dfrac{n_j}{N_j} & \text{si } k \in C_{\mathrm{i}} \text{ y } l \in C_{\mathrm{j}}, i \neq j \end{cases}$$

Estimadores y Varianza

$$\hat{T}_{\pi y} = \sum_{k\in s} \frac{Y_k}{\pi_k} = \sum_{\substack{i\in s_{UP} \\ k\in s_i}} \frac{Y_k}{\pi_i^{UP} \pi_{k|i}} = \frac{M}{m} \sum_{\substack{i\in s_{UP} \\ k\in s_i}} \frac{Y_k}{\pi_{k|i}} = \frac{M}{m} \sum_{i\in s_{UP}} \frac{N_i}{n_i} \sum_{k\in s_i} y_k = \frac{M}{m} \sum_{i\in s_{UP}} \hat{T}_{iy}$$

$$V_{UP} = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) S_{UP}^2 \qquad S_{UP}^2 = \frac{1}{M-1} \sum_{i\in U_{UP}} \left(T_{iy} - \bar{T}_{UP} \right)^2 \qquad \bar{T}_{UP} = \sum_{i\in U_{UP}} T_{iy} / M$$

$$V_i = N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{yC_i}^2 \qquad S_{yC_i}^2 = \frac{1}{N_i - 1} \sum_{k\in C_i} \left(y_k - \bar{y}_{C_i} \right)^2 \qquad \bar{y}_{C_i} = \sum_{k\in C_i} y_k / N_i$$

$$V_{HT}(\hat{T}_{\pi y}) = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) S_{UP}^2 + \frac{M}{m} \sum_{i\in U_{UP}} N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{yC_i}^2$$

$$\hat{V}_{HT}(\hat{T}_{\pi y}) = M^2 \left(\frac{1}{m} - \frac{1}{M} \right) s_{UP}^2 + \frac{M}{m} \sum_{i\in s_{UP}} N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) s_{yC_i}^2$$

Diseños Autoponderados

Si en la 1 Etapa se aplica un diseño d^{UP} proporcional a tamaño N_i para seleccionar una muestra s_{UP} de tamaño m de Conglomerados, se tiene,

$$\pi_i^{UP} = \frac{N_i}{N} m$$

y si en la 2 Etapa se aplican $d_i = MSA\left(N_i, n_0\right)$ n_0 constante para todo $C_i \in s_{UP}$

$$\pi_{k/i} = \frac{n_0}{N_i} \qquad \text{como consecuencia,}$$

$$\pi_k = \pi_i^{UP} \pi_{k/i} = \frac{N_i}{N} \frac{mn_0}{N_i} = \frac{mn_0}{N} = \frac{n}{N} \quad \text{para } k \in U$$

El diseño d final es de tamaño fijo $n = mn_0$ y autoponderado o sea π_k constantes

$$y \qquad \hat{T}_{HTy} = \frac{N}{n} \sum_{k \in s} y_k$$

EL Problema del tamaño de la Población

- ⊗ En un diseño por conglomerados en etapas, en la práctica $N = \# U = ?$
- ⊗ Justamente es un motivo por los que se recurre a ellos.
- ⊗ Por ejemplo si contamos con una lista de Conglomerados, $U_{UP} = \{C_1, ..., C_M\}$ con medidas de tamaño X_i para C_i, $\# C_i = N_i$, pero no se tiene una lista de unidades elementales u_k en el C_i en cuyo caso $N = \sum_{i \in U_{UP}} N_i = ?$
- ⊗ A veces es posible que para cada $C_i \in s_{UP}$ contar con N_i, pero sólo para s_{UP}
- ⊗ En un diseño en 1 Etapa en la práctica $\# s = n(s) = \sum_{i \in s_{UP}} N_i$
- ⊗ También en un diseño en 2 Etapas $s = \bigcup_{i \in s_{UP}} s_i \quad n(s) = \sum_{i \in s_{UP}} n(s_i)$
- ⊗ En ambas situaciones los tamaños de las muestras finales a nivel de unidades elementales son aleatorios.
- ⊗ Pero el diseño final será fijo o no dependiendo que los diseños en ambas etapas sean fijos o no.

Conglomerados como Unidades Geográficas

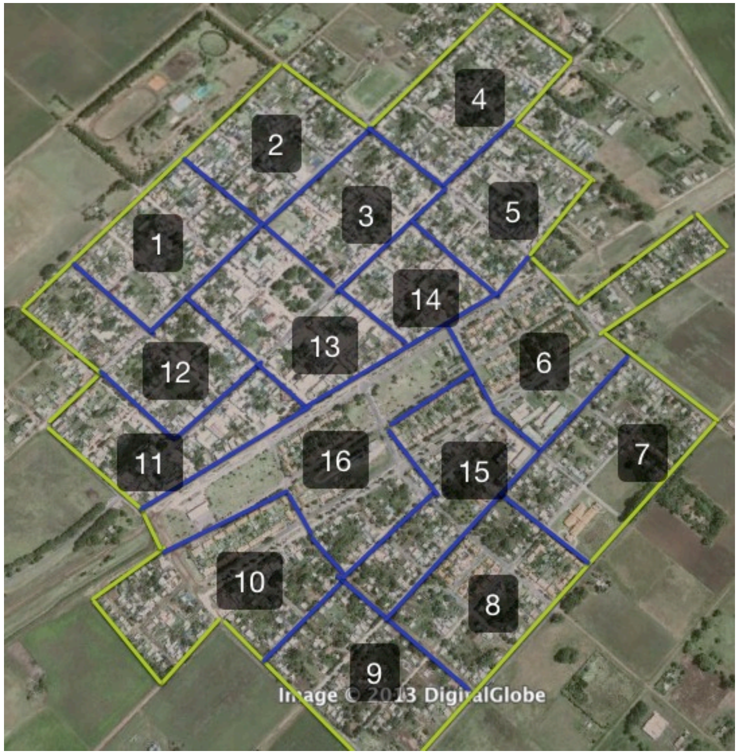
⊗ Aún cuando el diseño d^{UP} sea de tamaño fijo la muestra final de unidades elementales es variable a menos que las $N_i = N_0$ y sean conocidas

⊗ Por ejemplo, si U_{UP} es un listado de conglomerados (Radios Censales) de viviendas,

$C_i = \left\{ Radio_j, j = 1,..,N_i \right\}$,

u otro unidad cartográfica,

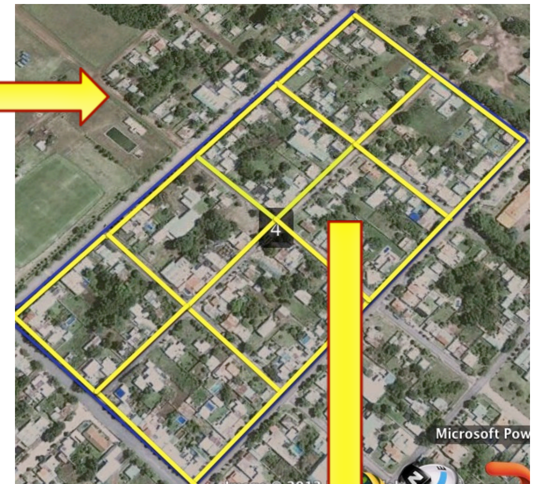
en cualquier caso los tamaños de las muestras de personas (unidades elementales) son variables



⊗ Pero que la muestra de unidades elementales sea variable, no implica que d^{UP} deje de ser fijo si lo era.

Ejemplo Típico en Encuestas Sociodemográficas

- ⊗ Listado de Radios (UP's)
- ⊗ En la práctica Estratificados
- ⊗ Primera etapa de selección
- ⊗ Diseños Π_{PPT} $\#N_i$ = total viviendas



- ⊗ Listado Manzanas o de segmentos de Viviendas

- ⊗ Selección de Viviendas
- ⊗ Eventualmente selección de Personas

¿Qué pasa con los Promedios?

Como $T_y = \sum_{k \in U} Y_k = \sum_{i \in U_{UP}} T_{iy}$ con $T_{iy} = \sum_{j \in C_i} Y_j \Rightarrow \quad \bar{Y} = \dfrac{T_y}{N}$

una estimación de $\bar{Y}$, $\quad \bar{Y}_{HT} = \dfrac{\hat{T}_{HTy}}{N}$ si N es conocido, pero en muchos d, $N = ?$

Dado un diseño d y definiendo $Z_k = 1 \ \forall k, k = 1, \ldots, N \quad \Rightarrow \ N = \sum_{k \in U} Z_k$

y por la teoría de HT $\quad \hat{N} = \sum_{k \in s} \dfrac{Z_k}{\pi_k}$ con varianza $V_{HT}(\hat{N}) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \dfrac{Z_k}{\pi_k} \dfrac{Z_l}{\pi_l}$

y una estimación, $\qquad \hat{V}_{HT}(\hat{N}) = \sum_{k \in s} \sum_{l \in s} \dfrac{\Delta_{kl}}{\pi_{kl}} \dfrac{Z_k}{\pi_k} \dfrac{Z_l}{\pi_l}$

con lo cual una alternativa para el promedio es:

$$\hat{\bar{Y}} = \frac{\hat{T}_y}{\hat{N}} \quad \text{cociente de 2 variables aleatorias}$$