

$$\begin{aligned}
 &= a \sum_{r=1}^a \left[b_{1r}^2 \right] - \left[\sum_{r=1}^a b_{1r} \right]^2 \\
 &= \sum_{r=1}^a \left(ab_{1r}^2 - b_{1r}^2 \right) \\
 &= a \sum_{r=1}^a (b_{1r} - \bar{b}_{1r})^2 \quad \text{con } \bar{b}_{1r} = \frac{\sum_{r=1}^a b_{1r}}{a}
 \end{aligned}$$

- Varianza estimada del estimador H-T.

$$\hat{V}(t_{1r}) = \sum_j \sum_r \Delta_{1r} \frac{y_{1r} y_{1j}}{N r j} \quad \text{para } \Delta_{1r} = 0 \quad \text{si están } k \text{ y } l \text{ en distinto grupo.}$$

por lo cual no se puede estimar la varianza.

- Descomposición de la varianza:

Supongamos q' la población se divide en a grupos, de tal forma q' existen n elementos por grupo

$$\begin{array}{c|c|c|c|c}
 \begin{matrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_1 1} \end{matrix} & \xrightarrow{\text{con } \frac{n_1}{S_1^2}} & \begin{matrix} Y_{12} \\ Y_{22} \\ \vdots \\ Y_{n_2 2} \end{matrix} & \xrightarrow{\text{con } \frac{n_2}{S_2^2}} & \begin{matrix} Y_{13} \\ Y_{23} \\ \vdots \\ Y_{n_3 3} \end{matrix} & \xrightarrow{\text{con } \frac{n_3}{S_3^2}} & \cdots & \cdots & \cdots & \cdots & \begin{matrix} Y_{1k} \\ Y_{2k} \\ \vdots \\ Y_{n_k k} \end{matrix} & \xrightarrow{\text{con } \frac{n_k}{S_k^2}}
 \end{array}$$

se puede definir la SCT como:

$$SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2 = (N-1) S_p^2$$

\bar{Y}_v es la media global y vóq a tener varianza grupo

$$v = 1, 2, \dots, a$$

entonces

$$SCT = \sum_{r=1}^a \sum_{i=1}^{n_r} (Y_{ri} - \bar{Y}_v)^2 \quad \text{donde } \bar{Y}_v = \frac{\sum_{i=1}^a \sum_{r=1}^{n_r} Y_{ri}}{n}$$

por lo cual

$$\begin{aligned}
 SCT &= \sum_{r=1}^a \sum_{i=1}^{n_r} [(Y_{ri} - \bar{Y}_r) + (\bar{Y}_r - \bar{Y}_v)]^2 \\
 &= \sum_{r=1}^a \sum_{i=1}^{n_r} (Y_{ri} - \bar{Y}_r)^2 + (\bar{Y}_r - \bar{Y}_v)^2 \quad \rightarrow 2 \sum_{r=1}^a \sum_{i=1}^{n_r} (Y_{ri} - \bar{Y}_r)(\bar{Y}_r - \bar{Y}_v) = 0
 \end{aligned}$$

$$SCT = \sum_{r=1}^a \sum_{i=1}^{n_r} (Y_{ri} - \bar{Y}_r)^2 + \sum_{r=1}^a \sum_{i=1}^{n_r} (\bar{Y}_r - \bar{Y}_v)^2$$

Scanned with CamScanner

$$SCT = \sum_{r=1}^g \sum_{i=1}^{N_r} (y_{ri} - \bar{y}_r)^2 + \sum_{r=1}^g N_r (\bar{y}_r - \bar{y}_G)^2$$

$$SCT = \underbrace{\sum_{r=1}^g (N_r - 1) S_r^2}_{\text{Suma de cuadrados dentro de los grupos}} + \underbrace{\sum_{r=1}^g N_r (\bar{y}_r - \bar{y}_G)^2}_{\text{Suma de cuadrados entre los grupos}}$$

SCD SCE

Ejemplo

$$\begin{array}{lll} q_{NPO} 1: & 182 & 180 \\ & 184 & 175 \Rightarrow \bar{x}_1 = 180,25 & S_1^2 = 14,92 & N_1 = 4 \\ q_{NPO} 2: & 170 & 171 \\ & 172 & 162 \Rightarrow \bar{x}_2 = 168,8 & S_2^2 = 15,77 & N_2 = 5 \\ q_{NPO} 3: & 158 & 155 \\ & 159 & 157 \Rightarrow \bar{x}_3 = 157,33 & S_3^2 = 4,33 & N_3 = 3 \\ & & & \bar{x}_{global} = 169,75 & \end{array}$$

$$SCD = 3(14,92) + 4(15,77) + 2(4,33)$$

$$SCD = 116,22$$

$$SCE = 4(180,25 - 169,75)^2 + 5(168,8 - 169,75)^2 + 3(157,33 - 169,75)^2$$

$$SCE = 905,2817$$

$$SCT = 116,22 + 905,2817$$

$$SCT = 1024$$

• Tabla ANOVA

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	F
Entre	$\sum_{r=1}^g N_r (\bar{y}_r - \bar{y}_G)^2$	$g - 1$	$\frac{SCE}{g-1}$	$\frac{CMG}{CMD}$
Dentro	$\sum_{r=1}^g (N_r - 1) S_r^2$	$N - g$	$\frac{SCD}{N-g}$	
Total	$\sum_{i=1}^N (y_i - \bar{y})^2$	$N - 1$		

sin pérdida de generalidad

$$Var(\hat{y}_i) = N(SCE)$$

• Coeficiente de Correlación Intraclass

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

Esta medida va a ser $\rho \ggg$ cuando hay mucha homogeneidad dentro del grupo, entonces $\rho \approx 1$

si $SCD = 0 \Rightarrow \rho = 1 \Rightarrow$ (Homogeneidad al interior)

si $SCD = SCT \Rightarrow \rho = 0$

En MSIS lo ideal es q' la correlación sea negativa.

• Ueff

$$U_{eff} = \frac{V_{SU}(t_{fin})}{V_{MAS}(t_{fin})}$$

$$U_{eff} = \frac{N-1}{N-n} (1 + (n-1)\rho)$$

Dado el efecto divisor se concluye que esta estrategia es:

a) igual de eficiente q' MAS si $\rho = \frac{1}{1-N}$

b) Menos eficiente q' MAS si $\rho > \frac{1}{1-N}$

c) Más eficiente q' MAS si $\rho < \frac{1}{1-N}$

► MUESTREO Q-SISTEMATICO:

$$\# Q = \binom{q}{q}$$

$$\cdot p(s) = \frac{1}{\binom{q}{q}} \quad \forall s \in Q_r$$

$$\cdot \pi_k = \frac{q}{a}$$

$$\cdot \pi_{kl} = \begin{cases} \frac{q}{a} & \text{si } k \wedge l \in S_r \\ \frac{q}{a} \frac{q-1}{a-1} & \text{en otro caso} \end{cases}$$

• Estimadores de H-T

$$\hat{t}_{y\pi} = \frac{a}{q} \sum_s t_{sr} \quad \text{con } t_{sr} = \sum_k y_k$$

$$\cdot V(\hat{t}_{y\pi}) = \frac{a^2}{q} \left(1 - \frac{q}{a}\right) S_{tsu}^2$$

$$\cdot \hat{V}(\hat{t}_{y\pi}) = \frac{a^2}{q} \left(1 - \frac{q}{a}\right) S_{tsr}^2$$

$$\text{con } S_{tsr}^2 = \frac{1}{q-1} \sum_{k=1}^q (t_{sr} - \bar{t}_s)^2$$

$$\text{con } \bar{t}_s = \frac{1}{a} \sum_{a=1}^k t_{sr}$$

→ DISEÑO DE MUESTREO CON PROBABILIDADES PROPORCIONALES.

► MUESTREO POISSON:

Este diseño es una generalización del diseño de muestreo Bernoulli, en donde las probabilidades de inclusión están dadas apriori de manera independiente para cada individuo.

- Continuidad

$$\#Q = 2^N$$

- $p(S) = \prod_{k \in S} \pi_k \prod_{k \notin S} (1 - \pi_k) \quad \forall S \subseteq Q$

Ejemplo:

Sea $U = \{e_1, e_2, e_3, e_4, e_5\}$ con $\pi_k = 0,2 \quad 0,5 \quad 0,7 \quad 0,2 \quad 0,9$

la probabilidad de $n(S) = 0$ es

$$p(S_0) = (1-0,2)(1-0,5)(1-0,7)(1-0,2)(1-0,9) = 0,006.$$

- $\pi_L = \pi_K$

- $\pi_{KL} = \begin{cases} \pi_K & \text{para } k=L \\ \pi_K \pi_L & \text{en otro caso.} \end{cases}$

- Tamaño de la muestra:

Bajo muestreo Poisson, el tamaño de la muestra $n(S)$ es una variable aleatoria tal que:

$$E(n(S)) = \sum_j \pi_K \quad V(n(S)) = \sum_j \pi_K (1 - \pi_K)$$

- Algoritmo de Selección:

La selección de una muestra con diseño de muestreo Poisson se realiza mediante un algoritmo secuencial definido de manera similar q' el algoritmo utilizado en la selección de muestras Bernoulli.

- 1) Fijar para cada $k \in U$ el valor de π_K tal que $0 < \pi_K \leq 1$
- 2) Los elementos con $\pi_K = 1$ son forzosamente incluidos en la muestra.
- 3) Obtener ξ_K para $K \in U$ como N realizaciones independientes de una variable aleatoria con distribución uniforme en el intervalo $[0, 1]$
- 4) Si $\xi_K < \pi_K$ el individuo k síma es seleccionado.

Dado q' $\xi_k \sim \text{Unif}[0,1]$ se tiene q' $P(\xi_k < \pi_k) = \pi_k$ para $k \in U$. Por tanto la inclusión de los individuos k en la muestra y el éxito para $k \neq L$ es independiente. Sin embargo, la distribución de $\xi(s)$ no es de tipo Binomial puesto q' las variables aleatorias $\xi(s)$ no son idénticamente distribuidas.

- Estimador de H-T

$$\hat{t}_{\text{HT}} = \frac{1}{\pi_k} \sum_s \xi_k$$

- $V_p(\hat{t}_{\text{HT}}) = \sum_0 (\frac{1}{\pi_k} - 1) \xi_k^2$

- $\hat{V}_p(\hat{t}_{\text{HT}}) = \sum_0 (1 - \pi_k) \left(\frac{\xi_k}{\pi_k} \right)^2$

dado q' $\Delta_{KL} = \begin{cases} \pi_{KL} - \pi_K \pi_L & \text{para } K \neq L \\ \pi_{KK} - \pi_K^2 & \text{para } K = L \end{cases}$

- Optimalidad de la estrategia:

la estrategia de muestreo q' utilice H-T es óptima cuando las probabilidades de inclusión inducidas por el diseño de muestreo utilizadas están correlacionadas positivamente con las características de interés, en otras palabras cuando $\pi_k \propto \xi_k$.

En este caso cuando la muestra es de tamaño fijo $n(s)=n$, el estimador de H-T reproduce el parámetro de interés t_H con varianza nula si $\pi_k = \frac{n \cdot \xi_k}{n}$

Suponiendo un tamaño de muestra fijo, bajo un diseño de muestra泊松, la varianza del estimador se minimiza cuando

$$\pi_k = \frac{n \cdot \xi_k}{\sum_s \xi_k}$$

Esto implica q' la característica de interés ξ_k debe ser conocida, sin embargo, como el diseño POISSON es de muestra variable esto indica q'

$$\hat{t}_{\text{HT}} = \frac{1}{\pi_k} \sum_s \xi_k = \frac{t_H}{n \cdot \pi_k} \sum_s \xi_k = \frac{t_H}{n} \sum_s 1 = \frac{t_H}{n} n(s) = \hat{t}_{\text{HT}}$$

sto indica q' la varianza del estimador est' supeditada al tamaño de la muestra $n(s)$.

Esto nos lleva a pensar q' el estimador de H-T tendría un excelente desempeño cuando $X \perp Y$ y q' induzcan muestras de tamaño fijo.

Por otra lado si se tiene información auxiliar altamente correlacionada con la característica de interés cuando esto se desconoce, la variancia de la estimación sería mínima cuando

$$\pi_k = n \frac{x_k}{\sum_j x_j}$$

► DISEÑO PPT (Probabilidad de Selección Proporcional al Tamaño - PPS)

Este es un diseño de selección con reemplazo y su tamaño de muestra es aleatorio.

La probabilidad de selección en este diseño está dada por:

- $P_k = \frac{x_k}{tx}$ donde x_k es una variable auxiliar altamente correlacionada con 'Y' aunq' no necesariamente la explique.

Ya q' si se conociera realmente a cada y_k , el estimador de Hansen-Hurwitz stimaría al total con varianza nula $\gamma^2 = \frac{1}{M-1}$.

En este caso

- $x_k > 0 \wedge k \in U$
- x_k debe ser conocido (en todos los elementos) de la población.
- Si algún $x_k = 0$ ese dato se debe ignorar.

Entonces se define, un diseño de muestreo con Probabilidad de Selección Proporcional al tamaño (PPS) de la siguiente manera:

$$P(s) = \begin{cases} \frac{m!}{n_1(s)! n_2(s)! \dots n_N(s)!} \prod \left(\frac{1}{p_k} \right)^{n_k(s)} & \text{si } \sum_i n_k(s) = m \\ 0 & \text{en otro caso} \end{cases} \quad \text{donde } \sum_s p(s) = 1$$

$$\#Q = \binom{n+m-1}{m}$$

$$\sum_s p_k = 1 \rightarrow \sum_s p_k = \sum_s \frac{x_k}{tx} = \frac{tx}{tx} = 1$$

$$\pi_k = 1 - (1 - p_k)^m = 1 - \left(1 - \frac{x_k}{tx} \right)^m$$

$$\pi_{kk} = 1 - (1 - p_k)^m - (1 - p_k)^m + (1 - p_k - p_k) =$$

$$\cdot \hat{t}_{\text{IP}} = \frac{1}{m} \sum_{k=1}^m \frac{y_k}{p_k} = \frac{1}{m} \sum_{k=1}^m \frac{y_k}{\frac{x_k}{t_x}} = \boxed{\frac{t_x}{m} \sum_{k=1}^m \frac{y_k}{x_k} = t_{\text{IP}}}$$

$$\cdot V(t_{\text{IP}}) = \frac{1}{m} \sum_j \left(\frac{y_k}{p_k} - t_{\text{IP}} \right)^2 p_k$$

$$= \frac{1}{m} \sum_k \left(\frac{t_x y_k}{x_k} - t_{\text{IP}} \right)^2 \frac{x_k}{t_x}$$

$$\therefore \boxed{= \frac{t_x}{m} \sum_{k=1}^m \frac{y_k^2}{x_k} - t_{\text{IP}}^2}$$

$$\cdot \hat{V}(t_{\text{IP}}) = \frac{1}{m(m-1)} \sum_{k=1}^m \left(\frac{y_k}{p_k} - t_{\text{IP}} \right)^2$$

$$= \frac{1}{m(m-1)} \sum_{k=1}^m \left(\frac{t_x y_k}{x_k} - t_{\text{IP}} \right)^2$$

↳ Algoritmo de selección:

→ Método Acumulativo Total:

Hansen, Horwitz y Madow (1953) diseñaron este algoritmo para ser utilizado junto con su estimador 'puri' y es un algoritmo q' consiste en m selecciones tomadas \perp , tal que s

$$1) \text{ Sea } p_k = \frac{x_k}{t_x}$$

2) Sea T_k el Total Acumulado construido de la siguiente manera:

$$T_k = \sum_{l=1}^k x_l \quad \text{con } T_0 = 0$$

$$3) \text{ Generar } q_k \sim U(0,1)$$

4) Seleccionar el k-ésimo elemento si $\frac{T_{k-1}}{T_N} < q_k < \frac{T_k}{T_N}$

5) Repetir m veces el algoritmo desde el paso 3.

Ejemplo:

$N=10$ Selección $m=4$

i	x_k	$p_k = x_k/t_x$	T_k	T_k/T_N	T_k/T_N
1	22	22/41	0,53	0	0,53
2	2	2/41	0,04	0,53	0,58
3	3	3/41	0,07	0,58	0,65
4	1	1/41	0,02	0,65	0,69
5	0,5	0,5/41	0,01	0,68	0,69
6	2,1	2,1/41	0,05	0,69	0,74
7	0,9	0,9/41	0,02	0,74	0,76
8	1,3	1,3/41	0,03	0,76	0,8
9	5	5/41	0,2	0,8	0,92
10	3,2	3,2/41	0,08	0,92	1

entonces

$$q_1 = 0,9087 \Rightarrow \text{Selección}$$

i9

$$q_2 = 0,35 \Rightarrow \text{Selección}$$

i1

$$q_3 = 0,941 \Rightarrow \text{Selección}$$

i10

$$q_4 = 0,5147 \Rightarrow \text{Selección}$$

i1

$\hookrightarrow m=4$

$$\overline{x} = 41 = T_N$$

$$\sum p_k = 1$$

→ Métodos de Lohrri: (1951)

En algunos casos, cuando la población N es muy grande el método acumulativo total resulta inefficiente, Lohrri plantea el siguiente método de selección:

Siendo $M = \max(x_1, x_2, \dots, x_N)$, los siguientes son los pasos para seleccionar un elemento:

- 1) Seleccionar un número $\eta \sim \text{Uniforme discreto } [1, N] \Rightarrow \text{aleatorio.entre(1, N)}$
- 2) Seleccionar $\eta \sim U(0, M) \Rightarrow \text{Uniforme discreto } y \leq M = \text{máximo}$
- 3) Si $\eta \leq x_i$ entonces el i -ésimo elemento es seleccionado, si $\eta > x_i$ se repite el procedimiento para seleccionar una unidad. Si el tamaño de la muestra es M , entonces el anterior procedimiento se repite M veces.

Ejemplo:

Con nuestros anteriores datos: $N = 10 \quad M = 22$

- $(\eta \sim (1, 10)) = 7$ entonces $x_7 = X_7$
- $\eta \sim (1, 22) = 13$ $x_7 = 0,9$
 $0,9 < 13 \rightarrow \eta > x_7 \rightarrow \text{NO SE SELECCIONA}$
- $(\eta \sim (1, 10)) = 3$ entonces $x_3 = 3$
- $\eta \sim (1, 22) = 7$ $\eta = 7 \rightarrow \eta > x_3 \rightarrow \text{NO SE SELECCIONA}$
- ⋮

↳ Eficiencia de la estrategia:

En este caso PPS no se refiere a contra el MAS ya q' en MAS uso TI-estimador y en PPS uso PWR estimador, entonces todos los diseños en los q' utiliza PWR estimador los compara contra MAS

$$V_{MAS}(\hat{t}\hat{\eta}_P) - V_{PPS}(\hat{t}\hat{\eta}_P) = \frac{N^2}{m} \text{cov}(x; \frac{y_k^2}{x_k})$$

Prueba

$$V_{MAS}(\hat{t}\hat{\eta}_P) - V_{PPS}(\hat{t}\hat{\eta}_P) = \frac{1}{m} \left[N \sum_{k=1}^N y_k^2 - t \sum_{k=1}^N \frac{y_k^2}{x_k} + t^2 \right]$$

$$= \frac{1}{m} \left[\sum_{k=1}^N \frac{y_k^2}{x_k} (Nx_k - tx_k) \right]$$

$$= \frac{N}{m} \left[\sum_{k=1}^N \frac{y_k^2}{x_k} (x_k - \bar{x}) \right]$$

$$= \frac{N^2}{m} \text{cov}(x; \frac{y_k^2}{x_k}) \Rightarrow \text{esta última q' que:}$$

$$\begin{aligned} \text{cov}(x, y) &= \sum_{k=1}^N (x_k - \bar{x})(y_k - \bar{y}) \\ &= \sum_{k=1}^N (y_k - \bar{y}) y_k - \sum_{k=1}^N (x_k - \bar{x}) \bar{y} = \frac{\sum_{k=1}^N (x_k - \bar{x}) y_k - \sum_{k=1}^N (x_k - \bar{x}) \bar{y}}{N} \\ &= \sum_{k=1}^N (x_k - \bar{x}) \left[1 - \frac{1}{N} \right] = \sum_{k=1}^N (x_k - \bar{x}) y_k \left(\frac{N-1}{N} \right)^{N-1} = \sum_{k=1}^N (x_k - \bar{x}) y_k \\ &\rightarrow N \text{cov}(x, y) = (x_k - \bar{x}) y_k \end{aligned}$$

Esto quiere decir q' para q' PPS sea mejor q' MASE $\text{cov}(x; \frac{y^2}{x}) > 0$,
es más den mejor usar MASE.

Aleatoriamente notes q' si $\frac{y_k}{x_k} \approx C$ siendo C una constante
entonces

$$\text{cov}(x; \frac{y^2}{x}) = \text{cov}(x; y_k \frac{y_k}{x_k}) = \text{cov}(x; Cy_k) = \text{cov}(x, y_k)$$

Por tanto una condición necesaria para q' el diseño PPS sea más eficiente
q' el MASE es q' existe una correlación positiva entre la característica de interés
y la variable auxiliar y , q' ademas $\frac{y_k}{x_k} \approx C \neq K \in U$.

Esto se debe a q' si $\frac{y_k}{x_k} \approx C \neq K \in U$, se tiene q'

$$V(\hat{t}_{pp}) = \frac{1}{m} \sum_j \left(\frac{y_k}{x_k} - \bar{t}_y \right)^2 p_k \quad \text{con } p_k = \frac{x_k}{\bar{t}_x} \quad \text{y } y_k = C x_k$$

entonces

$$= \frac{1}{m} \sum_j \left(\frac{\bar{t}_x(x_k) - \bar{t}_y}{\bar{t}_x} \right)^2 \frac{x_k}{\bar{t}_x}$$

$$= \frac{1}{m} \sum_j \left((\bar{t}_x - \bar{t}_y) \frac{x_k}{\bar{t}_x} \right)^2 ; \text{ por otro lado en } y_k = C x_k, \text{ entonces}$$

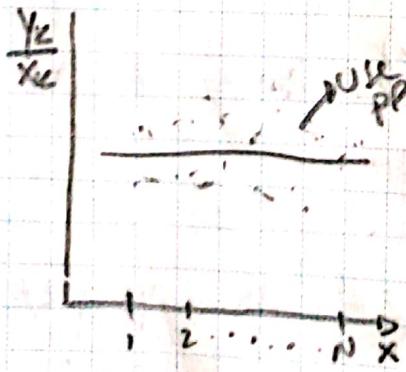
reemplazando se tiene q'

$$= \frac{1}{m} \sum_j \left(\cancel{C \bar{t}_x} - \bar{t}_y \right)^2 \frac{x_k}{\bar{t}_x}$$

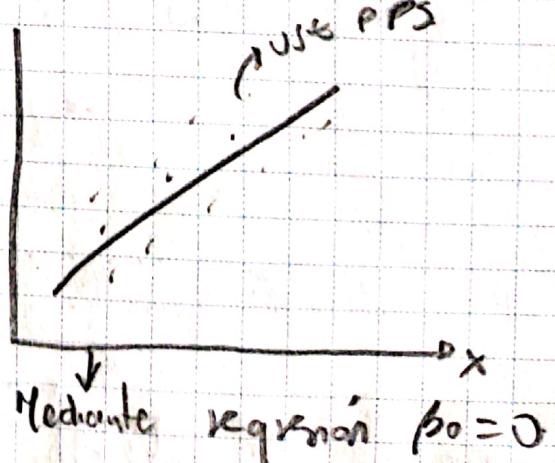
$$\bar{t}_y = \sum_k y_k = C \sum_k x_k = C \bar{t}_x$$

$$V(\hat{t}_{pp}) = 0 \text{ cuando } \frac{y_k}{x_k} \approx C \neq K \in U$$

Resumen: Si tenemos



y



► RECORDANDO

Diferencias entre muestra auxiliar:

Nombre	Con Recambio?	Tamaño de Muestra
Poisson	NO	H-T
PPT	SI	H-T
PPPT	NO	H-T

En POISSON

$$S_n = \frac{C}{X_k} \quad C = \text{constante}, \text{ entonces } V_{Poisson} = 0$$

Como hay elementos de inclusión fijos, T_{ik} puede ser ≥ 1 , en este caso T_{ik} se pone a $T_{ik}=1$

► DIFERENCIAS PPT (Probabilidad de Inclusión Proporcional al Tamaño)

T_{ik} va a ser directamente proporcional a una variable X_k , igual a una variable auxiliar.

$$T_{ik} \propto X_k \quad ; \quad X_k = \text{Variable Auxiliar}$$

- $T_{ik} = \frac{N X_k}{k} \quad ; \quad 0 < T_{ik} \leq 1 \quad y \quad T_{ik} = \text{probabilidad de inclusión fija.}$

→ Condiciones del diseño PPT:

1) El algoritmo de selección debe ser sencillo y eficiente computacionalmente

2) Todas las $T_{ik} > 0$, aunq' estos no resuelvan para estimar la variancia. Se debe comprobar q' $T_{ik} > 0$

$$\Delta_{kl} < 0 \iff \hat{V}(t_{ik}^*) > 0 \quad y \quad q'$$

$$\hat{V}(t_{ik}^*) = \sum_{l=1}^N \Delta_{kl} \frac{y_k y_l}{T_{ik} T_{il}} \quad \text{con} \quad \Delta_{kl} = T_{kl} - T_{ik} T_{il}$$

y en muestra de tamaño fijo:

$$\hat{V}_2(t_{ik}^*) = -\frac{1}{2} \left[\sum_j \sum_l \frac{\Delta_{kl}}{T_{ik}} \left(\frac{y_k}{T_{ik}} - \frac{y_l}{T_{il}} \right)^2 \right]$$

4) El algoritmo debe respetar las probabilidades de inclusión.

El MAS es un caso particular del TIPI y que se basa en la característica de información auxiliar o constante.

$$X_k = C$$

entonces

$$\Pi_k = \frac{n_k}{\sum_j n_j} \Rightarrow \Pi_k = \frac{n_k}{N} \Rightarrow \boxed{\Pi_k = \frac{n_k}{N}}$$

Note:

Puede ocurrir q' $\frac{n_k}{\sum_j n_j} > 1$ entonces hacemos $\Pi_k = 1$, esto quiere decir q' el elemento k es de inclusión fija oza.

y las demás probabilidades Π_k se calculan sobre el universo restante.

Esto s

$$\Pi_k^* = \frac{(n - n_k) X_k}{\sum_{U^*} X_k}; 0 < \Pi_k^* \leq 1; k \in U^*$$

con $n^* \Rightarrow N$ mrs de elementos de inclusión fija oza

$U^* \Rightarrow$ El universo sin incluir los elementos de inclusión fija oza

$\Pi_k^* \Rightarrow$ La probabilidad de incluir el resto de los elementos q' no tienen inclusión fija oza.

Al final del proceso se deben tener 2 grupos.

1) Un grupo de elementos de inclusión fija oza con probabilidades de inclusión $\Pi_k = 1$.

2) Un grupo de elementos tales q' $0 < \Pi_k \leq 1$ y proporcional a X_k .

Por tanto el problema se reduce a la selección de n unidades con probabilidades de inclusión tales q'

$$\sum_{k \in U} \Pi_k = n$$

Ejemplo: tengo $N=8$ con X_k respectiva y deseo $n=3$

i	X_k	Π_k	
e_1	8 000 000	$\Pi_1 = 31800000 / 10690000 = 2,24 > 1$ reformulamos	entonces $\Pi_1 = 1$
e_2	1 800 000	$\Pi_2 = (3-1)(1800000) / 2690000 = 1,3 > 1$ reformulamos	entonces $\Pi_2 = 1$
e_3	400 000	$\Pi_3 = (3-2)(400000) / 890000 = 0,45 = \Pi_3$	
e_4	300 000	$\Pi_4 = (3-2)(300000) / 890000 = 0,34 = \Pi_4$	$\sum \Pi_k = 3$
e_5	100 000	$\Pi_5 = (3-2)(100000) / 890000 = 0,11 = \Pi_5$	
e_6	50 000	$\Pi_6 = (3-2)(50000) / 890000 = 0,06 = \Pi_6$	
e_7	40 000	$\Pi_7 = (3-2)(40000) / 890000 = 0,04 = \Pi_7$	

- Estimador de $H-T$: $\hat{t}_{\pi} = \sum_k \frac{Y_k}{\pi_k}$
 - $V(\hat{t}_{\pi}) = \sum_k \sum_{k' \neq k} \Delta_{kk'} \frac{Y_k}{\pi_k} \frac{Y_{k'}}{\pi_{k'}}$
 - $\hat{V}_{\pi\pi}(\hat{t}_{\pi}) = \sum_k \sum_{k' \neq k} \frac{\Delta_{kk'}}{\pi_k \pi_{k'}} \frac{Y_k Y_{k'}}{\pi_k \pi_{k'}}$
- ↳ Tener en cuenta q' $\pi_k = C \Rightarrow Y_k \propto \pi_k \Rightarrow V(\hat{t}_{\pi}) = 0$

Los Algoritmos de Selección:

- Para $n=1$ se utiliza el método acumulativo total.
- Proceso
- 1) Define $T_0 = 0$ y $T_k = T_{k-1} + Y_k$; $k \in \cup$; (Acumulado)
 - 2) Calcular $\xi_k \sim U(0,1)$
 - 3) Si $\frac{T_{k-1}}{T_N} < \xi_k < \frac{T_k}{T_N}$ selecciona el elemento x_k donde $T_N = \sum_i Y_i$

El algoritmo de selección garantiza q' el diseño de muestreo es autentico $\pi_{PT} \approx q$:

Por definición $\pi_k = P(k \in S)$

$$\pi_k = P\left(\frac{T_{k-1}}{T_N} < \xi_k < \frac{T_k}{T_N}\right)$$

(si recordamos) $P(a < X < b) = P(X < b) - P(X < a)$

$$\pi_k = P\left(\xi_k < \frac{T_k}{T_N}\right) - P\left(\xi_k < \frac{T_{k-1}}{T_N}\right)$$

$$\pi_k = \frac{T_k - T_{k-1}}{T_N} \quad \text{pero } T_k = T_{k-1} + Y_k$$

$$\pi_k = \frac{Y_k}{T_N} \quad \text{donde } T_N = \sum_i Y_i, \text{ entonces}$$

$$\pi_k = \frac{Y_k}{T_N} = \frac{n Y_k}{T_N} \quad \text{con } n=1$$

• Para $n=2$ (Brewer)

En este escenario se debe garantizar q' las probabilidades de inclusión

$$\pi_k = \frac{2x_k}{\sum_j x_k} \quad \text{† K GU}$$

En este caso los dos elementos de la muestra son seleccionados uno x uno. Para este fin se utiliza el algoritmo de Brewer, el cual utiliza el método acumulativo total en cada una de las selecciones.

1) Selecciona el 1er elemento k con $P(k^*) = \frac{x_k}{\sum_j x_k}$ donde $c_k = \frac{x_k(T_n - x_k)}{T_n(T_n - 2x_k)}$

2) Se retira el elemento seleccionado en la extracción anterior (k^*) y el segundo elemento es seleccionado con $p_k = \frac{x_k}{T_n - x_{k^*}}$

→ Ejemplo:

$X \rightarrow 1,2, 20, 5, 1, 3, 0,5, 1,2, 2,1, 1,8$

$\Rightarrow N=8$

$$\sum_j x_k = 34,9 = T_n$$

$$p_k = x_k/T_n \rightarrow 0,03, 0,57, 0,14, 0,08, 0,01, 0,03, 0,06, 0,01$$

$$\pi_k = 0,06 \quad (1,14) \rightarrow 0,28, 0,16, 0,02, 0,06, 0,12, 0,1 \\ \rightarrow \text{Inclusión forzosa, entonces}$$

$$\pi_k = 0,06 \quad 1, 0,28, 0,16, 0,02, 0,06, 0,12, 0,1$$

→ Dado q' $n=2$ y q' hay 1 elemento de inclusión forzosa, entonces se utiliza el método acumulativo total para elegir el segundo elemento. Es decir, no aplica Brewer.

→ Ejemplo 2:

$X \rightarrow 1,2, 2,3, 2,5, 3, 0,5, 1,2, 2,1, 1,8 \Rightarrow N=8$

Este punto se omite.

$$p_k \rightarrow 0,079, 0,185, 0,1656, 0,1987, 0,083, 0,0741, 0,1391, 0,1192.$$

Este punto también se omite.

$$T_n = \sum_j x_k = 15,1 \\ n=2 \rightarrow p_k = 1.$$

$$c_k \rightarrow 0,09, 0,2401, 0,1025, 0,2642, 0,034, 0,087, 0,1658, 0,1379 \rightarrow \sum_j c_k = 1,2233$$

$$p_k^* \rightarrow 0,0713, 0,196, 0,169, 0,216, 0,028, 0,071, 0,136, 0,113 \rightarrow \sum_j p_k^* = 1$$

Con las probabilidades de selección aplicamos el método acumulativo total.

e_i	P_e^*	x_k	T_{k-1}	T_k	generamos	$\epsilon_{ik} \sim U(0,1)$
e_1	0,071	1,2	0	0,091		
e_2	0,196	2,8	0,091	0,167		$\epsilon_{ik} = 0,10$
e_3	0,169	2,5	0,167	0,436		
e_4	0,216	3	0,436	0,652		
e_5	0,018	0,5	0,652	0,680		
e_6	0,071	1,2	0,680	0,752		
e_7	0,136	2,1	0,752	0,887		
e_8	0,113	1,8	0,887	1		

aplicamos:

$$\frac{T_{k-1}}{T_N} < \epsilon_{ik} < \frac{T_k}{T_N}$$

$0,071 < 0,10 < 0,267 \rightarrow$ El elemento seleccionado es e_2

2do paso) Sacamos e_2 del sorteo y volvemos a calcular $T_{i|k}$, en este paso $P_L = \frac{x_L}{T_N - x_k}$

e_1	e_3	e_4	e_5	e_6	e_7	e_8
$x_k \rightarrow 1,2$	2,5	3	0,5	1,2	2,1	1,8
$P_L \rightarrow 0,098$	0,103	0,244	0,041	0,098	0,171	0,146

$$\sum_{i \neq k} P_L = 1$$

Con estos probabilidades de selección hacemos método acumulativo total.

e_i	x_k	T_{k-1}	T_k	generamos	$\epsilon_{ik} \sim U(0,1)$
e_1	1,2	0	0,098		
e_3	2,5	0,098	0,301		
e_4	3	0,301	0,541		

$$\epsilon_{ik} = 0,357$$

aplicamos:

$$\frac{T_{k-1}}{T_N} < \epsilon_{ik} < \frac{T_k}{T_N}$$

$0,301 < 0,357 < 0,541 \rightarrow$ El elemento seleccionado es e_4

• Para $N \geq 3$ (Método de Selección de Fronter)

$$1) \text{ Calcular } T_{ik} \left\{ \begin{array}{l} 1 \text{ si } \frac{nX_k}{n_x} \geq 1 \\ \frac{(n-n^*)X_k}{\sum_{i \neq k} X_i} \text{ en otro caso} \end{array} \right.$$

con n^* : El número de k elementos que son de inclusión fija
 $n^* = \text{Universo} - \text{inclusión fija}$.

- 2) Ordenar descendenteamente la variable T_{ik} ya calculada.
- 3) para $K=1$, el elemento $K=1$ entra en la muestra si el primer

4) Para $k \geq 2$, el elemento k -ésimo se selecciona si se satisface la siguiente desigualdad:

$$\frac{q_k}{\pi_k} \leq c_k \Rightarrow q_k \leq \frac{n - n_{k-1}}{n - \sum_{i=1}^{k-1} \pi_i} \pi_k \quad \text{donde } n_{k-1} = \text{numero de elementos seleccionados al final del paso } k-1.$$

Ejemplo

π_k	q_k	$\sum_{i=1}^k \pi_i$	c_k	Incluido? ($N=0, f_i=3$)	n_k
0,556	0,63	0,556	$\frac{\sqrt{-0}}{\sqrt{-0,556}} (0,475) = 0,5$	0	9
0,475	0,475	1,026	$\frac{\sqrt{-1}}{\sqrt{-1,026}} (0,468) = 0,47$	1	1
0,468	0,468	1,494	$\frac{\sqrt{-2}}{\sqrt{-1,494}} (0,46) = 0,53$	1	2
0,460	0,463	1,954	$\frac{\sqrt{-3}}{\sqrt{-1,954}} (0,446) = 0,44$	0	2
0,446	0,854	2,4	$\frac{\sqrt{-4}}{\sqrt{-2,4}} (0,446) = 0,51$	0	3
0,446	0,918	2,846	$\frac{\sqrt{-5}}{\sqrt{-2,846}} (0,412) = 0,57$	1	4
0,412	0,079	3,258	$\frac{\sqrt{-6}}{\sqrt{-3,258}} (0,393) = 0,45$	1	4
0,393	0,122	3,650	$\frac{\sqrt{-7}}{\sqrt{-3,650}} (0,349) = 0,26$	0	4
0,349	0,265	3,999	$\frac{\sqrt{-8}}{\sqrt{-3,999}} (0,349) = 0,31$	0	4
0,349	0,834	4,373	$\frac{\sqrt{-9}}{\sqrt{-4,373}} (0,334) = 0,51$	1	5
0,334	0,159	4,628	$\frac{\sqrt{-10}}{\sqrt{-4,628}} (0,322) = 0$	0	5
0,322	0,632	4,5			

→ Características del diseño TPT

• Si $X_k \propto Y_k$ la varianza $V(t_k Y_k) = 0$

• Es de tamaño n fijo.

• No tiene reemplazos

• $\pi_k = \begin{cases} \frac{n_k}{t_N} & \text{si } k=1, 2, \dots, K-1 \\ \frac{n_k x_k^2}{t_N} & \text{si } k=K^*, \dots, N \end{cases}$

dónde $K^* = \min\{K_0, N-n+1\}$ con K_0 equivalente al menor K para el cual se cumple que $\frac{n_k}{t_N} > 1$, $T_k = \sum_{j=1}^k x_j$ y

$$\bar{x}_k^* = \frac{T_k}{N-K^*+1}$$

por otra parte se cumple que para todos $k \neq 1 \geq 0$ $\gamma_{kk} < 0$

Como se calcula $\hat{\pi}_{k,n}$?

$$\hat{\pi}_{k,n} = \frac{1}{\pi_k} \sum_{l=1}^n \frac{\gamma_k}{\gamma_l}$$

$\hat{\gamma}(t_{k,n}) = \sum_{l=1}^n \sum_{k=1}^K \frac{\Delta_{kl}}{\pi_{kl}} \frac{\gamma_k}{\pi_k} \frac{\gamma_l}{\pi_l}$ \rightarrow Esto nos lleva a calcular las probabilidades de inclusión de segundo orden ($\hat{\gamma}(t)$) q' según el tamaño del universo puede llegar a ser inviable.

Para esto se va a aplicar el método de Reporte mínimo (escisión) donde:

- ① Permite seleccionar eficientemente los términos computacionales
- ② Permite obtener estimaciones para la varianza sin gastar muchos recursos computacionales.
(Ver Tillé 2003)

Lo que es de acuerdo de Reporte mínimo:

Si para un vector fijo de probabilidades de inclusión es posible plantear un diseño de muestras cuya Reporte contiene a lo más N muestras s, tales q' $p(s) > 0$. En tal caso, el diseño de muestras se dice de Reporte mínimo.

Pasos:

1) Ordenar el vector de probabilidades de inclusión en orden ascendente $(\pi_{(1)}, \pi_{(2)}, \dots, \pi_{(K)})'$

2) Primera iteración: $t=1$

Calcular:

$$\lambda_{(1)} = \min\{1 - \pi_{N-n}, \pi_{n+1}\}$$

Luego computar las siguientes particiones del vector de probabilidades de inclusión.

$$\pi_{k,1}^a = \begin{cases} 0 & \text{si } k \leq N-n \\ 1 & \text{si } k > N-n \end{cases}$$

$$\pi_{k,1}^b = \begin{cases} \frac{\pi_k}{1-\lambda_{(1)}} & \text{si } k \leq N-n \\ \frac{\pi_k - \lambda_{(1)}}{1-\lambda_{(1)}} & \text{si } k > N-n \end{cases}$$

3) t -ésima iteración, $t \geq 2$:

Determinar los siguientes conjuntos:
 $A(t) = \{k \mid 0 < \pi_{k,(t-1)}^b < 1\}$

$B(t) = \{k \mid \pi_{k,(t-1)}^b = 1\}$

$N^*(t) = \# A(t)$

$n^*(t) = n - \# B(t)$

Luego para elementos $k \in A(t)$ calcular

$$\lambda_t = \min\{1 - \pi_{N^*(t)-n^*(t),1}^b, \pi_{N^*(t)+1-n^*(t),1}^b\}$$

A continuación, para los elementos $k \in A(t)$ computar los siguientes componentes del vector de probabilidades de inclusiones:

$$\Pi_{k,t}^q = \begin{cases} 0 & \text{si } k \leq N_t^+ - n_t^+ \\ 1 & \text{si } k > N_t^+ - n_t^+ \end{cases}$$

$$\Pi_{k,t}^b = \begin{cases} \frac{\Pi_{k,t}^b}{1-\lambda_t} & \text{si } k \leq N_t^+ - n_t^+ \\ \frac{\Pi_{k,t}^b - \lambda_t}{1-\lambda_t} & \text{si } k > N_t^+ - n_t^+ \end{cases}$$

4) Iterar hasta obtener convergencia, es decir, hasta q' $\Pi_{k,t}^b \in \{0, 1\}$

↳ Cuándo usar o no TPT?

Para variables simétricas (en una variable y) y disponiendo de $X \approx Y$.

- Como hacer para saber si $X \approx Y$?

Se realiza un análisis de regresión y si p-value no es significativo entonces $X \approx Y$.

- En encuestas con varias variables continuas es casi imposible usar TPT/PPT/Poisson.

- En encuestas de hogares si se utiliza el tamaño (# personas) no se obtienen buenas estimaciones para variables como ingresos, gastos, ...,

- Para estimar ratios/proportiones no se usa TPT/PPT/Poisson.

- En este tipo de diseños puede ocurrir q' h' n aumento, aumento la variancia total ya q' existen ciertas configuraciones de Π_k q' no permiten disminuir la varianza \Rightarrow aumentar la muestra.

↳ Problemas del método de fuentes:

- 1) Π_k no son estrictamente proporcionales
- 2) $\sqrt{V(\hat{\Pi}_T)}$ es completamente ineficiente.

↳ Estimación de la variancia

$$\bullet \sqrt{V(\hat{\Pi}_T)} = \sum_{k \in U} \frac{b_k}{\Pi_k^2} (\hat{\gamma}_k - \bar{\gamma}_k^*)^2 \quad \text{con} \quad \bar{\gamma}_k^* = \frac{\Pi_k \sum_{l \in U} b_l \gamma_l}{\sum_l b_l}, \quad b_k = \frac{N \Pi_k (1 - \Pi_k)}{N-1}$$

$$\bullet \hat{V}(\hat{\Pi}_T) = \sum_k \frac{C_k}{\Pi_k^2} (\hat{\gamma}_k - \bar{\gamma}_k^*)^2 \quad \text{con} \quad \bar{\gamma}_k^* = \frac{\Pi_k \sum_{l \in S} C_l \gamma_l}{\sum_l C_l}; \quad C_k = (1 - \Pi_k) \frac{n}{n-1}$$

El cálculo de los errores del estimador de $H-T$ con probabilidad de inclusión desiguales, se hace difícil computacionalmente cuando $n \gg$.
Para evitar el cálculo y estimación de la variancia del estimador HT con dobles fases Deville y Tillé (2005) proponen una aproximación de la variancia y sus respectivas estimaciones para diseños de probabilidad desiguales con horizonte y diseño de probabilidades desiguales sin reemplazo que cubre casi todos los casos ... (Incluso)

↳ Fórmula mínima

$N=6$

$n=3$

$$\Pi_K = \{0.07, 0.17, 0.41, 0.61, 0.83, 0.91\}$$

Subset $D_6 \rightarrow$ (os valores módulos são os enteros de $\Pi_K(t)$)

Ponto	$P_{0,0} t=1$	$P_{0,0} t=2$	$P_{0,0} t=3$	$P_{0,0} t=4$
$\lambda - \Pi(0)$	$\lambda(1) = 0.59$	$\lambda(2) = 0.585$	$\lambda(3) = 0.471$	$\lambda = 0.238$
	Π^a	Π^b	Π^a	Π^b
0.07	0	0.171	0	0.412
0.17	0	0.415	0	1
0.41	0	1	1	1
0.61	1	0.049	0	0.118
0.83	1	0.585	1	0
0.91	1	0.780	1	0.471

$P_{0,0} t=1$

$$A(1) = \{\phi\} \quad B(1) = \{\phi\} \quad C(1) = \{e_1, e_2, e_3, e_4, e_5, e_6\}$$

$$D(1) = \{e_4, e_5, e_6\} \quad \# D(1) = n - \# B(1) = 3$$

$$\lambda(1) = \min \{1 - 0.41, 0.61\}$$

$$\lambda(1) = 0.59$$

$$\Pi^a(1) = \{0, 0, 0, 1, 1, 1\}$$

$$\Pi^b(1) = \left\{ \frac{0.07}{1-0.59}, \frac{0.17}{1-0.59}, \frac{0.41}{1-0.59}, \frac{0.61-0.59}{1-0.59}, \frac{0.83-0.59}{1-0.59} \right\}$$

$$\frac{0.91-0.59}{1-0.59} \quad \Rightarrow \quad \Pi^b(1) = \{0.171, 0.415, 1, 0.049, 0.585, 0.780\}$$

$P_{0,0} t=2$

$$A(2) = \{\phi\} \quad B(2) = \{e_3\} \quad C(2) = \{e_1, e_2, e_4, e_5, e_6\}$$

$$D(2) = \{e_8, e_9\} \quad \# D(2) = n - \# B(2) = 2$$

$$\lambda(2) = \min \{1 - 0.415, 0.780\} = 0.585$$

$$\Pi^a(2) = \{0, 0, 1, 0, 1, 1\}$$

$$\Pi^b(2) = \left\{ \frac{0.171}{1-0.585}, \frac{0.415}{1-0.585}, 1, \frac{0.049}{1-0.585}, \frac{0.585-0.585}{1-0.585}, \frac{0.780-0.585}{1-0.585} \right\}$$

$$\Pi^b(2) = \{0.412, 1, 1, 0.118, 0, 0.471\}$$

Para $t=3$

$$A(3) = \{e_5, e_6\} \quad B(3) = \{e_2, e_3\} \quad C(3) = \{e_1, e_4, e_5\}$$

$$\# D(3) = n - \# B(3) = 3 - 2 = 1$$

$$D(3) = \{e_6\}$$

$$\lambda(3) = \{1 - 0.412, 0.471\} = 0.471 = \lambda(3)$$

$$\pi^a(3) = \{0, 1, 1, 0, 0, 1\}$$

$$\pi^b(3) = \left\{ \frac{0.412}{1-0.471}, 1, 1, \frac{0.412}{1-0.471}, 0, \frac{0.471-0.412}{1-0.471} \right\}$$

$$\pi^b(3) = \{0.278, 1, 1, 0.222, 0, 0\}$$

Para $t=4$.

$$A(4) = \{e_5, e_6\} \quad B(4) = \{e_2, e_3\} \quad C(3) = \{e_1, e_4\}$$

$$\# D(4) = n - \# B(4) = 3 - 2 = 1$$

$$D(4) = \{e_1\}$$

$$\lambda(4) = \{1 - 0.222, 0.778\} = 0.778$$

$$\pi^a(4) = \{1, 1, 1, 0, 0, 0\}$$

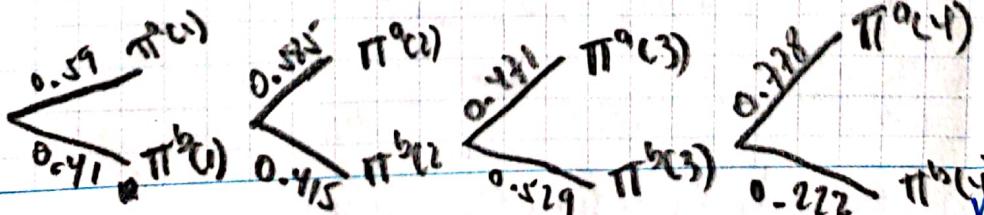
$$\pi^b(4) = \left\{ \frac{0.778 - 0.778}{1-0.778}, 1, 1, \frac{0.222}{1-0.778}, 0, 0 \right\}$$

$$\pi^b(4) = \{0, 1, 1, 1, 0, 0\}$$

Selección $\pi(t+1) = \begin{cases} \pi^a & \text{con probabilidad } \lambda(t) \\ \pi^b & \text{con probabilidad } 1 - \lambda(t) \end{cases}$

Seleción de los estados

$$\begin{cases} P(0, 0, 0, 1, 1, 1)' = 0.59 \\ P(0, 0, 1, 0, 1, 1)' = (1 - 0.59) \cdot 0.585 = 0.24 \\ P(0, 1, 1, 0, 0, 1)' = (1 - 0.59 - 0.24)(0.471) = 0.08 \\ P(1, 1, 1, 0, 0, 0)' = (1 - 0.59 - 0.24 - 0.08)(0.778) = 0.07 \\ P(0, 1, 1, 1, 0, 0)' = (1 - 0.59 - 0.24 - 0.08 - 0.07) = 0.02 \end{cases}$$



Grado de importancia
Iniciales: $\Pi(0) = \Pi$
para $t = 0, 1, 2, \dots$ hasta que la muestra es deseada, hacer:

1) Definir

$$A(t) = \{k \mid \Pi_k(t) \geq 0\}$$

$$B(t) = \{k \mid \Pi_k(t) = 1\}$$

$$C(t) = \{k \mid 0 < \Pi_k(t) < 1\}$$

2) Seleccionar un subconjunto $D(t)$ de $C(t)$ tal que
 $\# D(t) = n - \# B(t)$. $D(t)$ se selecciona con los
mayores valores de $\Pi_k(t)$ ~~los que~~ ~~que~~ menores
a \downarrow .

3) Definir

$$\Pi_k^a(t) = \begin{cases} 0 & \text{si } k \in A(t) \\ 1 & \text{si } k \in B(t) \\ \frac{\Pi_k(t)}{1 - \lambda(t)} & \text{si } k \in (C(t) \setminus D(t)) \end{cases}$$

$$\lambda(t) = \min \left\{ 1 - \max_{k \in (C(t) \setminus D(t))} \Pi_k(t); \min_{k \in D(t)} \Pi_k(t) \right\}$$

4

$$\Pi_k^b(t) = \begin{cases} 0 & \text{si } k \in A(t) \\ 1 & \text{si } k \in B(t) \\ \frac{\Pi_k(t)}{1 - \lambda(t)} & \text{si } k \in (C(t) \setminus D(t)) \\ \frac{\Pi_k(t) - \lambda(t)}{1 - \lambda(t)} & \text{si } k \in D(t) \end{cases}$$

Ejemplo

$n=3$

$$\text{Sea } X_k = \{52, 60, 75, 100, 50\}$$

$$\Pi_k = \{0.46, 0.53, 0.67, 0.9, 0.44\}$$

$$U_k = \{u_1, u_2, u_3, u_4, u_5\}$$

Paso 1: Ordena ascendente X_k : $p_{ascend}=2$

k	Π_k^a	Π_k^b	Π_k^b	Π_k^b	Π_k^b
u_5	0.44	0	0.916	0	0
u_1	0.46	0	0.991	0	0
u_2	0.53	1	0	0.298	1
u_3	0.67	1	0.298	1	0
u_4	0.90	1	0.298	1	0

Perio 1:

$$1) A(1) = \{\emptyset\} \quad C(1) = \{v_5, v_1, v_2, v_3, v_4\}$$
$$B(1) = \{\emptyset\} \quad 2) D(1) = \{v_2, v_3, v_4\}$$

$$3) \pi_{(1)}^q = \begin{cases} 0 & \text{para } 1 \\ 1 & \text{para } 2, 3, 4 \end{cases}$$

$$\lambda(1) = \min \{1 - 0.46, 0.53\}$$

$$\lambda_1 = 0.53$$

$$\pi_{(1)}^b = \left\{ \frac{0.44}{1-0.53}, \frac{0.43}{1-0.53}, \frac{0.53-0.53}{1-0.53}, \frac{0.67-0.53}{1-0.53}, \frac{0.9-0.53}{1-0.53} \right\}$$
$$= \{0.936, 0.979, 0, 0.298, 0.787\}$$

Perio 2

$$1) A(2) = \{v_2\} \quad C(2) = \{v_1, v_5, v_3, v_4\}$$
$$B(2) = \{\emptyset\} \quad 2) D(2) = \{v_5, v_1, v_4\}$$

$$3) \pi_{(2)}^q = \{v_2, v_3, v_2\} \rightarrow 0 \text{ para ellos}$$

$$\{v_5, v_1, v_4\} \rightarrow 1 \text{ para ellos}$$

$$\lambda(2) = \min \{1 - 0.67, 0.9\}$$

$$\lambda_2 = 0.33$$

$$\pi_{(2)}^b = \left\{ \frac{0.936 - 0.33}{1-0.33}, \frac{0.979 - 0.33}{1-0.33}, \frac{0.298}{1-0.33}, \frac{0.787 - 0.33}{1-0.33} \right\}$$

$$\pi_{(2)}^b = \{0.904, 0.969, 0, 0.445, 0.676\}$$

A continuación, para los elementos $k \in A(t)$ computar las siguientes particiones del vector de probabilidades de inclusión:

$$\Pi_{k(t)}^q = \begin{cases} 0 & \text{si } k \leq N_t^q - n_t^q \\ 1 & \text{si } k > N_t^q - n_t^q \end{cases}$$

$$\Pi_{k(t)}^b = \begin{cases} \frac{\Pi_{k(t)}^b}{1-\lambda_t} & \text{si } k \leq N_t^b - n_t^b \\ \frac{\Pi_{k(t)}^b - \lambda_t}{1-\lambda_t} & \text{si } k > N_t^b - n_t^b \end{cases}$$

v) Iterar hasta obtener convergencia, es decir, hasta q' $\Pi_{k(t)}^b \in \{0, 1\}$

↳ Cuándo usar o no $\Pi_{k(t)}$?

Para variables simétricas (en una variable Y) y disponiendo de $X \approx Y$.

- ¿Cómo hago para saber si $X \approx Y$?

Simplificando análisis de regresión y si p.value no es significativo entonces $X \approx Y$.

- En encuestas con varias variables continuas es casi imposible usar $\Pi_{k(t)}$ /PPT/Poisson.

- En encuestas de hogares se utiliza el tamaño (# personas), no se obtienen buenas estimaciones para variables como ingresos, gastos, ...

- Para estimar razones/proportiones no se usa $\Pi_{k(t)}$ /PPT/Poisson.

- En este tipo de diseños puede ocurrir q' fi n aumento, aumento la variancia total ya q' existen ciertas configuraciones de Π_k q' no permiten disminuir la variancia al aumentar la muestra.

↳ Problemas del método de funter:

- 1) Π_k no son estrictamente proporcionales
- 2) $V(\hat{\Pi}_k)$ es completamente ineficiente.

↳ Estimación de la variancia:

$$V(\hat{\Pi}_k) = \sum_{k \in U} \frac{b_k}{\Pi_k^2} (\bar{y}_k - \bar{y}_k^*)^2 \quad \text{con} \quad \bar{y}_k^* = \frac{\prod_k \sum_{l \in U} b_l y_{kl}}{\prod_k b_k}; \quad b_k = \frac{N \Pi_k (1 - \Pi_k)}{N-1}$$

$$\hat{V}(\hat{\Pi}_k) = \sum_k \frac{c_k}{\Pi_k^2} (\bar{y}_k - \bar{y}_k^*)^2 \quad \text{con} \quad \bar{y}_k^* = \frac{\prod_k \sum_{l \in S} c_l y_{kl}}{\prod_k c_k}; \quad c_k = (1 - \Pi_k) \frac{n}{n-1}$$

El cálculo de la varianza del estimador de HT con probabilidad de inclusión desiguales se hace difícil computacionalmente cuando $n \gg$. Para evitar el cálculo y estimación de la varianza del estimador HT con dobles fases Deville y Tillé (2005) proponen una aproximación de la varianza y su respectiva estimación para diseños de probabilidades desiguales con completa y discreta de probabilidades desiguales sin Π_k y c_k tales que $\sum_k \Pi_k = 1$ y $\sum_k c_k = 1$ (Imanol).

www.ipso.com.co

jeuplo:

$$\begin{array}{ccccc} Y & = & 108 & 125 & 154 \\ X & = & 32 & 60 & 75 \\ \bar{x} & = & 0,46 & 0,53 & 0,66 \end{array}$$

$$\begin{array}{ccccc} N & = & 204 & 104 & 154 \\ n & = & 100 & 50 & 75 \\ \bar{x} & = & 0,9 & 0,44 & 0,34 \end{array}$$

$$\begin{array}{ccccc} N & = & 52 & 3 & 37 \\ n & = & 3 & & \\ \bar{x} & = & 33,7 & & \end{array}$$

Calcular Vltqfj's

- quais os bx?

$$bx = \frac{\sqrt{(0,46)(0,54)}}{4} = 0,3105 ; \quad \frac{\sqrt{(0,53)(0,75)}}{4} = 0,31137 ; \quad \frac{\sqrt{(0,66)(0,34)}}{4} = 0,276375$$

$$\frac{\sqrt{(0,9)(0,1)}}{4} = 0,1125 ; \quad \frac{\sqrt{(0,44)(0,56)}}{4} = 0,308 \Rightarrow \sum_j bx = 1,31875$$

então us.

$$\frac{bx_1}{\pi_1} = \frac{(0,3105)(108)}{0,46} = 72,9 ; \quad \frac{(0,31137)(125)}{0,53} = 73,436 ; \quad \frac{(0,276375)(154)}{0,66} = 63,525$$

$$\frac{(0,1125)(204)}{0,9} = 25,5 ; \quad \frac{(0,308)(104)}{0,44} = 72,8 \Rightarrow \sum_j \frac{bx_i}{\pi_i} = 308,161$$

Agora

$$V^2_{\text{L}} = \frac{(0,46)(308,161)}{1,31875} = 107,49 ; \quad \frac{(0,53)(308,161)}{1,31875} = 123,85 ; \quad \frac{(0,66)(308,161)}{1,31875} = 154,231$$

$$\frac{(0,9)(308,161)}{1,31875} = 210,315 ; \quad \frac{(0,44)(308,161)}{1,31875} = 102,82$$

for low val

$$V(\hat{q}_n) = \frac{0,3105}{0,46^2} (108 - 107,49)^2 + \frac{0,31137}{0,53^2} (125 - 123,85)^2 + \frac{0,276375}{0,66^2} (154 - 154,231)^2 + \frac{0,1125}{0,9^2} (204 - 210,315)^2$$

$$+ \frac{0,308}{0,44^2} (104 - 102,82)^2$$

$$\boxed{V(\hat{q}_n) = 12,1685}$$

MUESTREO ESTRATIFICADO:

Características:

- Permite obtener estimaciones + eficientes en cuanto a $V(t_{eff})$; $V(t_{ep})$
- Tiene información auxiliar de tipo categórico.
- El dominio se establece una vez recolectados los datos, el estrato es definido de antemano.
 - En dominios hay un único diseño muestral.
 - En estratos se elabora un $p(\cdot) = \{p(S_1), p(S_2), \dots, p(S_k)\}$
- Uno de los objetivos del diseño estratificado es obtener mediciones o estimaciones precisas para cada estrato.
- El dominio tiene un tamaño de muestra denotado y el estrato tiene un tamaño de muestra conocido.

Términos
diseños muestrales.
floreo/definido de antemano
en particiones del universo

Dominio
Descomponible
1
NO
NO/son particiones de la muestra.

Estrato:
conocido.
El # de estratos
51.
51

Partición:

$$U = \bigcup_{h=1}^H U_h$$

con $h = \text{estrato}$.

$$U_i \cap U_j = \emptyset.$$

Parámetros:

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

$$t_{eff} = \sum_j t_{jk} = \sum_{h=1}^H \sum_{i=1}^{N_h} t_{ki} = \sum_{h=1}^H t_{yh}$$

$$\bar{y} = \frac{\sum_i y_i}{N} = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_h = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h \quad \rightarrow \text{Promedio ponderado de los grupos.}$$

$$S^2 = \sum_{h=1}^H \frac{N_h S_h}{N} + \sum_{h=1}^H \frac{N_h}{N} (\bar{y}_h - \bar{y})^2$$

Intervariación (dentro del estrato)

Intervariación (entre los estratos)

$$S^2 = S_{\text{dentro}}^2 + S_{\text{entre}}^2$$

$$= \frac{1}{N-1} \sum_{h=1}^H (N_h - 1) S_{yh}^2 + \frac{1}{N-1} \sum_{h=1}^H N_h (\bar{y}_h - \bar{y})^2$$

↳ Sobre la estimación:

Sea S_h la muestra aleatoria seleccionada en el estrato h con diseño de muestreo $p_{h|e}$, donde $p_{h|e}(S_h) = P(S_h = s_h)$ y sea s_h una realización de S_h , el total de muestras aleatorias dentro del estrato h es s_h

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

y en particular la muestra s seleccionada

$$s = \bigcup_{h=1}^H s_h$$

Si el tamaño de la muestra en cada estrato es n_h , entonces el tamaño de muestra seleccionada mediante un diseño de muestreo estratificado es:

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

esto quiere decir q para cada estrato $h=1, \dots, H$, existe un conjunto de posibles muestras Q_h , por lo q

$$Q^H = \bigcup_{h=1}^H Q_h$$

la cardinalidad de cada muestra Q_h depende del diseño de muestreo q se realice en cada estrato, por lo q:

$$\# Q^H = \prod_{h=1}^H \# Q_h$$

Ejemplo: Se tienen 3 estratos y en cada estrato se aplica MAS.

	Bojo	Medio	Alto
N_h	5	5	4
n_h	3	3	2
$\# Q_h$	10	10	6

$$\Rightarrow \# Q^H = 10 \times 10 \times 6 = 600$$

• Diseño Muestral:

$$p(s) = \prod_{h=1}^H p_{h|e}(s_h)$$

Problema:

Vamos a considerar:

a_1 = selección de s_1 en el estrato $h=1$

a_2 = selección de s_2 en el estrato $h=2$

⋮

a_H = selección de s_H en el estrato $h=H$

entonces

$$p(s) = p(a_1, a_2, \dots, a_n), \text{ por lo que}$$

$p(s) = p(a_1)p(a_2)\dots p(a_n)$; esto demuestra que hay independencia entre la selección de las muestras en cada estrato.

Este diseño de muestra cumple con las siguientes propiedades:

1) $p(s) > 0 \quad \forall s \in Q$.

2) $\sum_{s \in Q} p(s) = 1$

Prueba (Por inducción matemática).

- Si $h=1 \Rightarrow \sum_{s \in Q_1} p(s_1) = 1$

- Si $h=2 \Rightarrow$

$Q_1 = \{S_{11}, S_{12}, S_{13}, \dots, S_{1n}\} \rightarrow$ Cardinalidad del 1º estrato

$Q_2 = \{S_{21}, S_{22}, S_{23}, \dots, S_{2n}\} \rightarrow$ Cardinalidad del 2º estrato.

$$Q_1 \times Q_2 = \{S_{11} \cup S_{21}, S_{11} \cup S_{22}, S_{11} \cup S_{23}, \dots, S_{1n} \cup S_{21}, S_{12} \cup S_{22}, S_{12} \cup S_{23}, \dots, S_{1n} \cup S_{22}, S_{1n} \cup S_{23}, \dots, S_{1n} \cup S_{2n}\}$$

que da $p(j)$.

$$\begin{aligned} p(j) &= p(S_{11})p(S_{21}) + p(S_{11})p(S_{22}) + p(S_{11})p(S_{23}) + \dots + p(S_{11})p(S_{2n}) + p(S_{12})p(S_{21}) + p(S_{12})p(S_{22}) \\ &\quad + p(S_{12})p(S_{23}) + \dots + p(S_{12})p(S_{2n}) + \dots + p(S_{1n})p(S_{21}) + p(S_{1n})p(S_{22}) + p(S_{1n})p(S_{23}) + \dots \\ &\quad + p(S_{1n})p(S_{2n}) \end{aligned}$$

factorizando

$$\begin{aligned} p(j) &= p(S_{11})[p(S_{21}) + p(S_{22}) + \dots + p(S_{2n})] + p(S_{12})[p(S_{21}) + p(S_{22}) + \dots + p(S_{2n})] \\ &\quad + p(S_{13})[p(S_{21}) + p(S_{22}) + \dots + p(S_{2n})] + \dots + p(S_{1n})[p(S_{21}) + p(S_{22}) + \dots + p(S_{2n})] \end{aligned}$$

$$p(j) = p(S_{11}) + p(S_{12}) + p(S_{13}) + \dots + p(S_{1n})$$

$$p(j) = \frac{1}{n}$$

- Si $h=k$

Se admite para el principio de inducción que

$$\sum_{s \in Q_k} p(s) = 1$$

dónde $Q^k = \{Q_1 * Q_2 * Q_3 * \dots * Q_n\}$

$$Q_k = \{S_{11}, S_{12}, \dots, S_{1n}\}$$

$$Q_k = \{S_{21}, S_{22}, \dots, S_{2n}\}$$

$$Q_{k+1} = \{S_{11,1}, S_{11,2}, \dots, S_{1n,h}\}$$

En ese entonces se tiene que:

$$\sum_{s \in S} p(s_{k,n}) = p(S_{k,n,1}) \left[\sum_{s \in S} p(s) \right] + p(S_{k,n,2}) \left[\sum_{s \in S} p(s) \right] + \cdots + p(S_{k,n,h_n}) \left[\sum_{s \in S} p(s) \right]$$

$$\sum_{s \in S} p(s_{k,n}) = p(S_{k,n,1}) + p(S_{k,n,2}) + \cdots + p(S_{k,n,h_n})$$

$$\sum_{s \in S} p(s_{k,n}) = 1$$

→ Estimando en el muestreo estratificado:

- Si $\hat{t}_{y,n}$ estima independiente el total de la característica de interés t_y del strato n , entonces un estimador independiente para t_y está dado por:

$$\hat{t}_y = \sum_{n=1}^H \hat{t}_{y,n}$$

sto porque

$$\begin{aligned} E(\hat{t}_y) &= E\left(\sum_{n=1}^H \hat{t}_{y,n}\right) \\ &= \sum_{n=1}^H E(\hat{t}_{y,n}) \\ &= \sum_{n=1}^H t_{y,n} \\ &= H \cdot t_y \end{aligned}$$

además

$$\begin{aligned} V(\hat{t}_y) &= V\left(\sum_{n=1}^H \hat{t}_{y,n}\right) \Rightarrow V(x_1 + x_2 + \cdots + x_H) = \sum_{i=1}^H V(x_i) + \sum_{i=1}^H \sum_{j=1, j \neq i}^H \text{cov}(x_i, x_j) \\ &= \sum_{n=1}^H V(\hat{t}_{y,n}) + \sum_{n=1}^H \sum_{j=1, j \neq n}^H (\hat{t}_{y,n}, \hat{t}_{y,j}) \end{aligned}$$

Por independencia.

$$V(\hat{t}_y) = \sum_{n=1}^H V(\hat{t}_{y,n})$$

- Estimador de Horvitz-Thompson:

$$- \hat{t}_{y,\pi} = \sum_{n=1}^H \hat{t}_{y,n,\pi} ; \quad \hat{t}_{y,n,\pi} = \sum_{k \in S_n} \frac{y_k}{\pi k}$$

$$- V_{\text{est}}(\hat{t}_{y,\pi}) = \sum_{n=1}^H V_{p.}(\hat{t}_{y,n,\pi})$$

$$- \hat{V}_{\text{est}}(\hat{t}_{y,\pi}) = \sum_{n=1}^H \hat{V}_{p.}(\hat{t}_{y,n,\pi})$$

→ DISEÑO EST-MAS:

Para tomar una muestra fija en estrato extraído, denotemos como n_1, \dots, n_H , un diseño de muestras se dice EST-MAS, si la probabilidad de seleccionar una muestra de tamaño n es dada por:

$$p(\mathbf{J}) = \begin{cases} \prod_{h=1}^H \frac{1}{n_h} & \text{si } \sum_{h=1}^H n_h = n \\ 0 & \text{en otro caso.} \end{cases}$$

notese q $\sum_{\mathbf{J} \in \Omega} p(\mathbf{J}) = 1$ porque $\# \Omega = \prod_{h=1}^H \binom{n_h}{n}$

• Algoritmo de Selección:

En la selección de muestras EST-MAS, se debe:

- 1) Separar la población en H subgrupos o estados mediante la caracterización poblacional de información auxiliar.
- 2) En cada estado realizar una selección MAS, mediante los métodos Coordenado Negativo o ran domíl y Reverso.
- 3) Realizar cada una de las H selecciones de manera independiente.

• Bajo EST-MAS:

$$- T_{kL} = \frac{n_h}{N_h} \quad \text{si } k \in U_h$$

$$- T_{kL} = \begin{cases} \frac{n_h}{N_h} & \text{si } k=L, k \in U_h \\ \frac{n_h}{N_h} \frac{n_h-1}{N_h-1} & \text{si } k, L \in U_h \\ \frac{n_h}{N_h} \frac{n_i}{N_i} & \text{si } k \in U_h, L \in U_i, i \neq h \end{cases}$$

$$- \Delta_{kL} = \begin{cases} \frac{n_h}{N_h} \frac{N_h-n_h}{N_h} & \text{si } k=L, k \in U_h \\ - \frac{n_h}{N_h^2} \frac{(N_h-n_h)}{(N_h-1)} & \text{si } k, L \in U_h \\ 0 & \text{si } k \in U_h, L \in U_i, i \neq h \end{cases}$$

$$- \text{Entonces: } \hat{t}_{j_{h,n}} = \sum_{k \in S_n} \frac{\gamma_k}{T_{kL}} \Rightarrow \hat{t}_{j_{h,n}} = \frac{N_h}{n_h} \sum_{k \in S_h} \frac{\gamma_k}{n_h}$$

$$- V_p(\hat{t}_{j_{h,n}}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{j_{h,n}}^2$$

$$- \hat{V}_p(\hat{t}_{j_{h,n}}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{j_{h,n}}^2$$

$$\text{con } S_{j_{h,n}}^2 = \frac{1}{N_h-1} \sum_{k \in S_h} (\gamma_k - \bar{\gamma}_{j_{h,n}})^2; n=1, \dots$$

$$\text{con } S_{j_{h,n}}^2 = \frac{1}{N_h-1} \sum_{k \in S_h} (\gamma_k - \bar{\gamma}_{j_{h,n}})^2; h=1, \dots, H$$

• Estimación de la Media Poblacional Bajo EST-MAS:

$$- \hat{Y}_{UH} = \frac{1}{n_h} \sum_{k \in S_h} Y_k \Rightarrow \text{Media en el estrato } h$$

- También se puede escribir como:

$$\hat{Y}_{UH} = \frac{1}{N_h} \sum_{k \in S_h} \frac{Y_k}{f_k}$$

$$- V(\hat{Y}_{UH}) = \frac{1}{N_h^2} V\left(\sum_{k \in S_h} \frac{Y_k}{f_k}\right) = \frac{1}{N_h} \left(\frac{n_h^2}{N_h} \left(1 - \frac{n_h}{N_h}\right) S_{Y_{UH}}^2\right)$$

$$V(\hat{Y}_{UH}) = \frac{1}{N_h} \left(1 - \frac{n_h}{N_h}\right) S_{Y_{UH}}^2$$

$$- \hat{V}(\hat{Y}_{UH}) = \frac{1}{N_h} \left(1 - \frac{n_h}{N_h}\right) S_{Y_{SH}}^2$$

$$- \hat{Y}_{U, \pi} = \frac{1}{N} \sum_{h=1}^H n_h \hat{Y}_{UH} = \hat{Y}_{U, \pi, \text{EST-MAS}}$$

$$- V_{\text{EST-MAS}}(\hat{Y}_{U, \pi}) = \frac{V_{\text{EST-MAS}}(\hat{Y}_{U, \pi})}{N^2} = \frac{1}{N^2} \sum_{h=1}^H \frac{n_h}{N_h} \left(1 - \frac{n_h}{N_h}\right) S_{Y_{UH}}^2$$

$$- V_{\text{EST-MAS}}(\hat{Y}_{U, \pi}) = \frac{\hat{V}_{\text{EST-MAS}}(\hat{Y}_{U, \pi})}{N^2} = \frac{1}{N^2} \sum_{h=1}^H \frac{n_h}{N_h} \left(1 - \frac{n_h}{N_h}\right) \hat{S}_{Y_{UH}}^2.$$

- Intervalo de Confianza para $\hat{Y}_{U, \pi}$

$$IC(\hat{Y}_{U, \pi}) = \hat{Y}_{U, \pi} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{\text{EST-MAS}}(\hat{Y}_{U, \pi})} \Rightarrow \text{Cn: } ① n_h >>> \\ ② \text{Número de estratos} >>>$$

en caso contrario se usa

$$IC(\hat{Y}_{U, \pi}) = \hat{Y}_{U, \pi} \pm t_{n-h, 1-\alpha/2} \sqrt{\hat{V}_{\text{EST-MAS}}(\hat{Y}_{U, \pi})}$$

• Asignación del tamaño de la muestra:

① Asignación Proporcional Basada EST-MAS

$$n_h = n \frac{n_h}{N}$$

ya q' un diseño EST tiene asignación proporcional

$$\frac{n_h}{N_h} = \frac{n}{N}; h=1, 2, \dots, H$$

en este caso particular:

$$- t_{\text{MAS}}^{\hat{\eta}} = \frac{N}{n} \sum_{h=1}^H t_{ph}^{\hat{\eta}}$$

$$- V_{\text{EST-MAS}}(t_{\text{MAS}}^{\hat{\eta}}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \underbrace{\sum_{h=1}^H \frac{nh}{n} S_{ph}^2}_{\text{Intrahasta}}$$

Este expresión + d - porceto $\geq V(\text{MAS})$ ya q

$$V_{\text{MAS}}(t_{\text{MAS}}^{\hat{\eta}}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_p^2$$

$$\approx \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \underbrace{\sum_{h=1}^H \frac{nh}{n} S_{ph}^2}_{S_{\text{p.h.}}^2} + \underbrace{\sum_{h=1}^H \frac{nh}{n} (\bar{y}_h - \bar{y}_0)^2}_{S_{\text{p.h.}}^2}$$

por lo q $V(t_{\text{MAS}}^{\hat{\eta}})$ con desviación proporcional $< V(t_{\text{MAS}}^{\hat{\eta}})$.

$$- V_{\text{EST-MAS}}(t_{\text{MAS}}^{\hat{\eta}}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \sum_{h=1}^H \frac{nh}{n} S_{ph}^2.$$

② Aproximación de Neyman:

$$n_h = n \frac{N_h S_{ph}}{\sum_{h=1}^H N_h S_{ph}}$$

donde $S_{ph}^2 = \sqrt{S_{ph}^2} \rightarrow$ la desviación standar para el estrato h estimada previamente.

③ Aproximación óptima por costo:

El costo es frecuentemente considerado como:

$$C = C_0 + \sum_{h=1}^H C_h n_h$$

con (C_0 = Presupuesto de la investigación)

(C_h = Costo de una encuesta en el estrato h)

(n_h = Tamaño de la muestra en el estrato h)

Luego si se quiere distribuir la selección del elementos entre los estratos de manera q' se minimice la varianza de $t_{\text{MAS}}^{\hat{\eta}}$, se puede demostrar q':

$$n_h = \frac{C}{\sqrt{C_h}} \frac{N_h S_{ph}}{\sum_{h=1}^H N_h \sqrt{C_h} S_{ph}} + n.$$

• Estimación en Domos: para EJ-MAS

• TOTAL

En el estrato h

$$\hat{y}_{dh} = \frac{N_h}{n_h s_h} \sum y_{hdh}$$

$$V(\hat{y}_{dh}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2$$

$$\hat{v}(\hat{y}_{dh}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2$$

En la Población

$$= \sum_{h=1}^H \frac{N_h}{n_h} \sum y_{hdh}$$

$$= \sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2$$

$$= \sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2$$

• MEDIA

En el estrato h

$$\hat{y}_{dh} = \frac{1}{n_h} \left[\frac{N_h}{n_h} \sum y_{hdh} \right]$$

$$V(\hat{y}_{dh}) = \frac{1}{N_h^2 n_h} \left[\frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2 \right]$$

$$\hat{v}(\hat{y}_{dh}) = \frac{1}{N_h^2 n_h} \left[\frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2 \right]$$

En la Población

$$= \frac{1}{N_d} \left[\sum_{h=1}^H \frac{N_h}{n_h} \sum y_{hdh} \right]$$

$$= \frac{1}{N_d^2} \left[\sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2 \right]$$

$$= \frac{1}{N_d^2} \left[\sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{ydh}^2 \right]$$

• TAMAÑO

En el estrato h

$$\hat{N}_{dh} = \frac{N_h}{n_h} \sum_{sh} 2dk = \sum_{sh} \frac{2dk}{N_h}$$

$$V(\hat{N}_{dh}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dkh}^2$$

$$\hat{v}(\hat{N}_{dh}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dsh}^2$$

En la Población

$$= \sum_{h=1}^H \frac{N_h}{n_h} \sum_{sh} 2dk$$

$$= \sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dkh}^2$$

$$= \sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dsh}^2$$

$$\text{con } S_{2dkh}^2 = \frac{n_h}{n_h-1} \hat{p}(1-\hat{p})$$

$$\text{donde } \hat{p} = N_{dh}/N_h$$

$$\text{donde } \hat{p} = N_{dh}/N_h$$

• PROPORCIÓN

En el estrato h

$$\hat{p}_{dh} = \frac{1}{n_h} \hat{N}_{dh} = \frac{1}{n_h} \sum 2dk = \frac{N_h}{n_h}$$

$$V(\hat{p}_{dh}) = \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dkh}^2$$

$$\hat{v}(\hat{p}_{dh}) = \frac{1}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dsh}^2$$

En la Población

$$= \frac{\hat{N}_{dh}}{N} = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{n_h} \sum 2dk$$

$$= \frac{1}{N^2} \sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dkh}^2$$

$$= \frac{1}{N^2} \sum_{h=1}^H \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{2dsh}^2$$

• Deff:

$$D_{eff} \cong \frac{\sum_{h=1}^H w_h s_{ph}^2}{\sum_{h=1}^H w_h [s_{ph}^2 + (\bar{y}_{uh} - \bar{y}_u)^2]}$$

$$\cong \frac{\text{Varianza dentro de los estratos}}{\text{Varianza total}}$$

$$\text{con } w_h = \frac{n_h}{n} \frac{N_h}{N}$$

↳ DISEÑO EST-PPT

$$p_{kh} = \frac{x_{kh}}{t_{kh}} \quad \text{por lo q' } \sum_{kh} p_{kh} = 1 \quad \text{q' entonces } \sum_{h=1}^H \sum_{k=1}^{M_h} p_{kh} = H$$

En cada estrato U_h de tamaño N_h se selecciona una muestra s_h con reemplazo de tamaño m_h , por tanto la cardinalidad del roporte en el estrato U_h está dada por

$$\# Q_h = \binom{N_h + m_h - 1}{m_h}$$

El roporte general estratificado, se define como la unión de los roportes en cada uno de los estratos U_h .

$$Q^* = \left\{ \bigcup_{h=1}^H S_h \mid S_h \in Q_h \right\}$$

• Estimador de Hansen-Horwitz

$$\hat{t}y_{hp} = \frac{t_{kh}}{m_h} \sum_{i=1}^{m_h} \frac{y_{ki}}{x_{ki}}$$

$$V(\hat{t}y_{hp}) = \frac{1}{m_h} \sum_{h=1}^H p_h \left(\frac{y_h}{p_h} - \hat{t}y_h \right)^2$$

$$\hat{V}(\hat{t}y_{hp}) = \frac{1}{m_h(m_h-1)} \sum_{h=1}^H \sum_{k=1}^{m_h} \left(\frac{y_{ki}}{p_{ki}} - \hat{t}y_{hp} \right)^2$$

Por lo que

$$\hat{t}y_{hp} = \sum_{h=1}^H \frac{t_{kh}}{m_h} \sum_{i=1}^{m_h} \frac{y_{ki}}{p_{ki}}$$

$$V_{EST-PPT}(\hat{t}y_{hp}) = \sum_{h=1}^H \frac{1}{m_h} \sum_{h=1}^H p_h \left(\frac{y_h}{p_h} - \hat{t}y_h \right)^2$$

$$\hat{V}_{EST-PPT}(t_{ijp}) = \sum_{h=1}^H \frac{1}{M_h(M_h-1)} \sum_{\substack{i=1 \\ i \in S}}^{M_h} \sum_{j=1}^{M_h} (Y_{ki} - t_{ijp})^2$$

► MUESTREO EN VARIAS ETAPAS:

Para mantener un equilibrio entre los costos financieros y las bondades de la estrategia de diseño, se debe aprovechar la homogeneidad dentro de los conglomerados y de esta manera no realizar un censro dentro de cada conglomerado sino proceder a realizar una submuestra dentro del conglomerado seleccionado. Como el comportamiento estructural de la característica de interés al interior de los conglomerados es homogéneo, entonces una estimación del total del conglomerado tendrá una variancia pequeña.

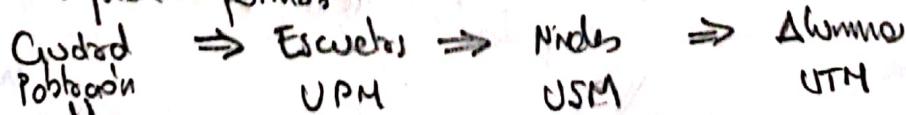
El principio básico básico del muestreo en varias etapas se puede definir como el proceso jerárquico q' realiza las siguientes pasos:

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

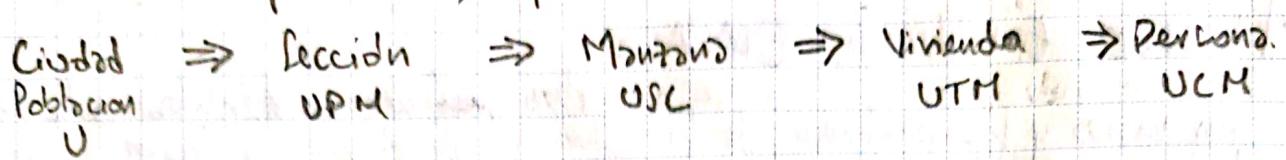
Notese q' se ha introducido el concepto de Unidad de Muestreo refiriéndose a conglomerados de elementos o elementos. Si el diseño de muestras tiene 3 etapas, por ejemplo: si se quieren obtener las estimaciones acerca del comportamiento de los alumnos en determinada ciudad y no se dispone de un marco de muestreo de los alumnos, es posible:

- En una primera etapa, levantar un marco de muestreo de todas y cada uno de los escuelas en la ciudad y realizar una selección de una muestra de escuelas mediante algún diseño de muestreo, una vez q' (a) escuelas son seleccionadas, ...
- En una segunda etapa de levanto un marco de muestreo de niveles académicos dentro de las escuelas (cursos o clases) y se procede a seleccionar una muestra de niveles, de tal forma q' q
- En la tercera etapa (y última) se levanta un marco de muestreo de elementos y se realiza la selección de la muestra de alumnos q' serán observados y medidos.

Es interesante observar como la población en el estrato de la naturaleza, se divide gracias al comportamiento jerárquico, q' en este caso particular tiene la siguiente forma:



No siempre las unidades finales de muestreo son elementos, es así como es posible planear un diseño en 2 etapas de conglomerados refinándose a q' los U's son los conglomerados, o planear un diseño en 4 etapas tal que:



El principio básico de una estrategia de muestreo en varias etapas es construir estimaciones desde abajo hasta arriba. Pero para q' los resultados de la estimación basada en el diseño de muestreo sean aplicables, se deben satisfacer los siguientes supuestos:

- **INVARIANZA:** Sugiere q' la probabilidad de selección de una muestra de unidades de muestreo (conglomerados o elementos) no depende del diseño de muestreo de la anterior etapa.
- **INDEPENDENCIA:** Interpretado como q' el nro-muestreo de cualquier unidad de muestreo se lleva a cabo de manera independiente contos otras unidades de muestreo, en la misma etapa o en etapas superiores o inferiores.

► MUESTREO EN DOS ETAPAS: (BITÁCORA)

Supongamos q' la población de elementos U se divide en N_I UPM's q' definen una partición de la población, llamadas también conglomerados y denotados como

$$U_I = \{U_1, U_2, \dots, U_{N_I}\}$$

El i -ésimo conglomerado U_i , $i=1, 2, \dots, N_I$ es de tamaño N_i .

- Una muestra S_I de unidades primarias de muestreo es seleccionada de U_I de acuerdo con un diseño de muestreo $P_I(S_I)$. Nótese q' S_I representa la muestra aleatoria de conglomerados tal q' $Pr(S_I = S_I) = P_I(S_I)$.
- Para cada conglomerado U_i , $i=1, 2, \dots, N_I$, seleccionado en la muestra S_I se selecciona una muestra de elementos, S_i , de acuerdo a un diseño de muestreo $p_i(S_i)$. Nótese q' S_i representa la muestra aleatoria de elementos tal q' $Pr(S_i = s_i) = p_i(s_i)$.

- La invarianza significa q' $Pr(S_i = s_i | S_I = s_I) = Pr(S_i = s_i)$ por lo q' $p_i(\cdot | S_I) = p_i(\cdot)$

- La independencia significa q' q:

$$Pr\left(\bigcup_{i \in S_I} S_i | S_I\right) = \prod_{i \in S_I} Pr(S_i | S_I)$$

En términos del reporte, es posible hablar también de 3 clases a saber:

- En la primera etapa existe un reporte Q_1 conteniendo todas las posibles muestras realizadas de las UPM.
- En la segunda etapa existe un reporte Q^i para cada $i \in U_1$, es decir, para cada UPM en la etapa anterior.
- En general, el reporte Q contiene todos las posibles muestras de elementos mediante un diseño bietáptico solo dado por:

$$Q = \bigcup_{r=1}^{\# Q_1} \bigcup_{i \in S_1^r} S_i^r, \text{ con } S_i^r \in Q^i$$

$$= \left\{ \bigcup_{i \in S_1^r} S_i^r, \text{ con } S_i^r \in Q^i, r = 1, \dots, \# Q_1 \right\}$$

donde S_1^r denota la r -ésima posible muestra de la primera etapa, y la cardinalidad de Q está dada por:

$$\# Q = \prod_{i \in U_1} \# Q^i$$

y la muestra de elementos (UMS) viene dada por:

$$S = \bigcup_{i \in S_1} S_i, \text{ con } S_i \in Q^i$$

contenido de la muestra aleatoria dado por

$$n(S) = \sum_{i \in S} n_i$$

→ Propiedades.

El diseño de muestras bietáptico cumple con:

$$1) p(s) \geq 0 \quad \forall s \in Q$$

$$2) \sum_{s \in Q} p(s) = 1$$

Ejemplo:

Nuestra población de ejemplos U_1 dada por

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

$$\begin{aligned} U_1 &= \{Yves, Jean\} \\ U_2 &= \{Emile, Charles\} \\ U_3 &= \{Coline\} \end{aligned}$$

Supongamos que se selecciona una muestra S_1 de UPM de $n_1=2$ mediante un MAS, tal que:

$$p_1(S_1) = \begin{cases} 0.5 & \text{si } S_1 = \{U_1, U_2\} \\ 0.4 & \text{si } S_1 = \{U_1, U_3\} \\ 0.1 & \text{si } S_1 = \{U_2, U_3\} \end{cases}$$

Ahora supongamos que dentro de cada UPM se seleccionan un solo elemento de acuerdo a los siguientes criterios de muestreo:

$$P_1(S_1|S_1) = \begin{cases} 0,5 & \text{si } S_1 = \{\text{Yes}\} \\ 0,5 & \text{si } S_1 = \{\text{Ken}\} \end{cases}$$

$$P_2(S_2|S_2) = \begin{cases} 0,9 & \text{si } S_2 = \{\text{Eric}\} \\ 0,1 & \text{si } S_2 = \{\text{Sharon}\} \end{cases}$$

$$P_3(S_3|S_2) = \begin{cases} 1,0 & \text{si } S_3 = \{\text{Leslie}\} \end{cases}$$

Es decir, el tamaño de la muestra final es $n=2$ y el reporte de la primera etapa está dado por:

$$Q_1 = \{(U_1, U_2), (U_1, U_3), (U_2, U_3)\}$$

y los reportes de la segunda etapa están dados por

$$Q^1 = \{\text{Yes}, \text{Ken}\}, \quad Q^2 = \{\text{Eric}, \text{Sharon}\}, \quad Q^3 = \{\text{Leslie}\}$$

Dado lo anterior, el reporte Q general está dado por:

$$Q = \bigcup_{i \in S_1} \bigcup_{j \in S_2} \bigcup_{k \in S_3} \{U_i, U_j, U_k\}$$

donde $\bigcup_{i \in S_1} = \{(\text{Yes}, \text{Eric}), (\text{Yes}, \text{Sharon}), (\text{Ken}, \text{Eric}), (\text{Ken}, \text{Sharon})\}$

$$\bigcup_{i \in S_2} = \{(\text{Eric}, \text{Leslie}), (\text{Sharon}, \text{Leslie})\}$$

$$\bigcup_{i \in S_3} = \{(\text{Yes}, \text{Leslie}), (\text{Ken}, \text{Leslie})\}$$

También,

	$P_1(S_1)$	$P_2(S_2)$	$P_3(S_3)$	$p(S) = P_1(S_1) \cdot P_2(S_2) \cdot P_3(S_3)$
Yes, Eric	(0,5)	(0,9)	0,5	0,225
Yes, Sharon	(0,5)	(0,1)	0,5	0,025
Ken, Eric	(0,5)	(0,9)	0,5	0,225
Ken, Sharon	(0,5)	(0,1)	0,5	0,025
Eric, Leslie	(0,9)	(1,0)	0,1	0,090
Sharon, Leslie	(0,1)	(1,0)	0,1	0,010
Yes, Leslie	(0,5)	(1,0)	0,4	0,200
Ken, Leslie	(0,5)	(1,0)	0,4	0,200

$$\sum p(S) = 1$$

→ Parámetros de interés

- Total Poblacional

$$t_y = \sum_{k \in U} y_k = \sum_{i=1}^{N_i} \sum_{k \in U_i} y_{ki} = \sum_{i=1}^{N_i} t_{yi}$$

donde $t_{yi} = \sum_{k \in U_i} y_{ki}$, es el total de la i-éxima UPM, $i = 1, 2, \dots, N_i$

- Media Poblacional

$$\bar{y}_i = \frac{\sum_{k \in U_i} y_{ki}}{N_i} = \frac{1}{N_i} \sum_{i=1}^{N_i} \sum_{k \in U_i} y_{ki} = \frac{1}{N_i} \sum_{i=1}^{N_i} N_i \bar{y}_i$$

donde $\bar{y}_i = \frac{1}{N_i} \sum_{k \in U_i} y_{ki}$ es la media de la i-éxima UPM, $i = 1, 2, \dots, N_i$.

↳ Estimaciones de los parámetros:

- Estimador de Horvitz Thompson:

- En la primera etapa:

$$\Delta_{ij} = \begin{cases} T_{hi} - T_{hi}T_{li} & \text{si } i, j \in U_i \\ T_{hi}(1 - T_{li}) & \text{si } i, j \notin U_i \end{cases}$$

- En la segunda etapa:

$$\Delta_{kl|i} = \begin{cases} T_{kli} - T_{kli}T_{li} & \text{si } k \neq l \\ T_{kli}(1 - T_{li}), & \text{si } k = l. \end{cases}$$

- En general la probabilidad de Inclusión de primer orden del k-émino elementos de U está dada por:

$$\pi_{ki} = T_{kli} T_{li}$$

- la probabilidad de inclusión de segundo orden está dada por:

$$\pi_{kl|i} = \begin{cases} T_{hi} T_{kli} & \text{si } k = l \in U_i \\ T_{hi} T_{kli} & \text{si } k \neq l \in U_i \\ T_{hi} T_{kli} T_{li} & \text{si } k \in U_i \wedge l \in U_j, (i \neq j) \end{cases}$$

- Bajo muestreo en 2 etapas el estimador de HT es igualado para el total poblacional y tiene la siguiente forma:

$$\hat{t}_{y\pi} = \sum_{i \in S} \sum_{k \in U_i} \frac{y_{ki}}{T_{hi} \pi_{ki}} = \sum_{i \in S} \frac{\hat{t}_{yi\pi}}{T_{hi}}$$

con T_{kli} = Probabilidad de inclusión del elemento k en la segunda etapa.

T_{hi} = Probabilidad de inclusión de la UPM que contiene el elemento k en la primera etapa.

www.ipos.com.co

- con variantas dentro por:

$$V_{\text{var. prop.}}(t_{ij}^{\hat{n}}) = \underbrace{\sum_{U_i} \sum_{U_j} \Delta_{ij} \frac{t_{ij} t_{ij}}{T_{ij} T_{ij}}} + \underbrace{\sum_{i \in U_1} \frac{V_{i(t_i)}}{T_{ii}}}$$

VARIANZA UPM = V_1 VARIANZA USM = V_2

dónde

$$V_{i(t_i)} = \sum_{U_i} \sum_{U_j} \Delta_{ij} \frac{y_{ij} - \bar{y}_i}{T_{ij} T_{ii}}$$

Ejemplo: MAS-MAS.

$$V_{\text{MAS-MAS}}(t_{ij}^{\hat{n}}) = \underbrace{\frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S_{y_{ij}}^2}_{\text{VARIANZA UPM} = V_1} + \underbrace{\frac{N_I}{n_I} \sum_{U_i} V_{i(t_i)}}_{\text{VARIANZA USM} = V_2}$$

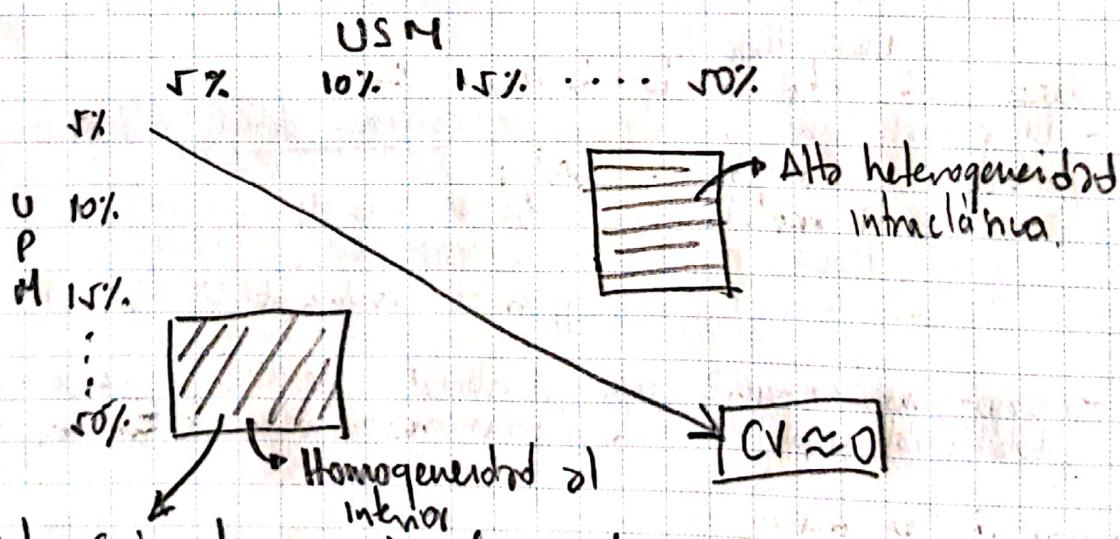
$$\text{dónde } V_{i(t_i)} = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S_{y_{ij}}^2$$

y $S_{y_{ij}}^2$ = Variancia de los totales de las UPM.

Si existe alta correlación intradáctica las variantas al interior van a ser pequeñas, cuando esto ocurre $V_1 > V_2$

la correlación intradáctica es la homogeneidad al interior de cada cosecha.

Ejemplo: Si creamos una matriz UPM + USM analizando los CV's por cada fracción de muestra, podemos analizar lo siguiente



Este sería el ejemplo de muestra cuando hay alta correlación intradáctica (alta fracción de muestra en UPM y bajo en USM), por ejemplo en un estudio de hogares en cada barrio hay homogeneidad, entonces necesito poco muestra dentro de un barrio y una muestra más grande de barrios q son los heterogéneos.

Es bueno tener en cuenta q' con el muestreo en etapa 1 se sacrifica eficiencia pero se gana en costos.

De otro lado:

$$- \hat{V}_{biológico}(\hat{t}_{ijn}) = \sum_{S_I} \sum_{I \in S_I} \Delta_{IIij} \frac{\hat{t}_{ijn} \hat{t}_{ijn}}{\Pi_{IIj} \Pi_{IIj}} + \sum_{I \in S_I} \frac{\hat{V}_1(\hat{t}_{ijn})}{\Pi_{IIj}}$$

$$\hat{V}(UPM) = \hat{V}_1 + \hat{V}(USM) = \hat{V}_2$$

donde $\hat{t}_{ijn} = \sum_{k \in S_I} \frac{y_k}{\Pi_{IIkj}}$

Es necesario notar q'

\hat{V}_1 no estimá independiente a V_1 y tampoco \hat{V}_2 estimá independiente a V_2 , pero $\hat{V}_1 + \hat{V}_2$ si estimá independiente a $V_1 + V_2$.

↳ DISEÑO MAS-MAS:

- Probabilidad de Inclusión de Primer Orden:

$$\Pi_{IIj} = \frac{N_I}{N_I}$$

$$\Pi_{Ik} = \frac{N_I}{N_I} \frac{N_i}{N_i}$$

- Probabilidad de Inclusión de Segundo Orden:

$$\Pi_{IIij} = \frac{N_I(N_I - 1)}{N_I(N_I - 1)}$$

- Estimación del total de la i -éxima UPM

$$\hat{t}_{ijn} = \frac{N_i}{n_i} \sum_{k \in S_I} y_k = N_i \bar{y}_{ui}$$

- Estimación del total de la característica de interés

$$\hat{t}_{ijn} = \frac{N_I}{n_I} \sum_{I \in S_I} \hat{t}_{ijn} = \frac{N_I}{n_I} \sum_{I \in S_I} \frac{N_i}{n_i} \sum_{k \in S_I} y_k = \sum_{k \in S_I} \frac{y_k}{\Pi_{IIkj} \Pi_{Ikj}}$$

- Varianza del Estimador

$$V_{MAS^2}(\hat{t}_{ijn}) = \frac{N_I^2}{n_I^2} \left(1 - \frac{N_I}{N_I}\right) S_{Y_{U_I}}^2 + \frac{N_I}{n_I} \sum_{I \in S_I} \frac{N_i^2}{n_i^2} \left(1 - \frac{N_i}{N_i}\right) S_{Y_{U_I}}^2$$

- Estimación de la variancia del estimador:

$$\hat{V}_{\text{Var}}(\hat{t}_{\text{UH}}) = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S_{y_{\text{UH}}}^2 + \frac{N_I}{n_I} \sum_{i \in U_I} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S_{y_{\text{UH}}|i}^2$$

donde $S_{y_{\text{UH}}|i}^2$ es la estimación de la variancia de los totales estimados \hat{t}_{UH}^i , $i \in U_I$ de todos y cada una de las UPM i .

$S_{y_{\text{UH}}|i}^2$ es la estimación de la variancia del estimador dentro de cada UPM, similamente $S_{y_{\text{UH}}}^2$ y $S_{y_{\text{UH}}}^2$.

- Tamaño de la muestra

Cada UPM contiene exactamente $N_i = M$ elementos

o USM. El sub-muestreo es tal que se selecciona una muestra de exactamente $n_i = m$ USM. Por tanto el tamaño de la población estará dado por:

$$N = N_I M \quad y \quad n = n_I m$$

respectivamente. De tal forma q' el estimador de t_Y se puede escribir como:

$$\hat{t}_Y = \frac{N_I}{n_I} \frac{M}{m} \sum_{i \in U_I} \sum_{k \in S_i} y_{ik}$$

y su varianza

$$V_{\text{Var}}(\hat{t}_Y) = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S_{y_{\text{UH}}}^2 + \frac{N_I^2 M^2}{n_I m} \left(1 - \frac{m}{M}\right) \bar{S}_{y_{\text{UH}}}^2$$

$$\text{donde } \bar{S}_{y_{\text{UH}}}^2 = \frac{1}{N_I} \sum_{i \in U_I} S_{y_{\text{UH}}|i}^2$$

para encontrar los valores óptimos de N_I y M q' serán utilizados en la primera y segunda etapa de muestreo de tal forma q' obtenga una función de costo de minimizar la varianza del estimador. Por tanto:

Al considerar la siguiente función de costo:

$$C = C_1 N_I + C_2 n_I m$$

donde C_1 es el costo del levantamiento del marco de muestras en cada UPM seleccionada en la muestra S_I y C_2 es el costo de recolectar la característica de interés para los elementos i unidades secundarias seleccionadas por el sub-muestreo.

los valores óptimos de n_1 y M que minimizan la varianza del estimador restringido al costo total de la muestra, esto dado por:

$$n_1 = \frac{C}{C_1 + C_2 M} \quad \text{y} \quad M = \frac{N_1^2}{S_{\text{ejec}}^2 - M \bar{S}_{\text{ejec}}^2} \sqrt{\frac{C_1/C_2}{S_{\text{ejec}}^2 - M \bar{S}_{\text{ejec}}^2}}$$

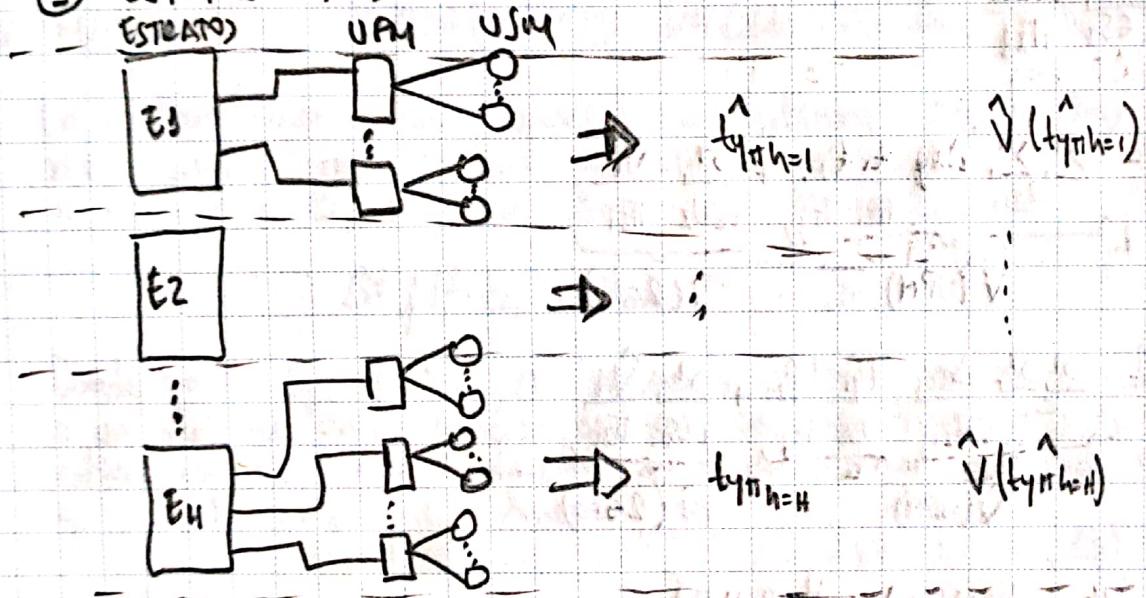
Estos resultados surgen de minimizar la varianza del estimador con restricción de costo

$$L(n_1, M, \lambda) = \frac{N_1^2}{n_1} \left(1 - \frac{n_1}{N_1} \right) S_{\text{ejec}}^2 + \frac{N_1^2 M^2}{n_1 M} \left(1 - \frac{M}{n_1} \right) \bar{S}_{\text{ejec}}^2 + \lambda (C n_1 + G_2 n_1 M - C)$$

La DISEÑO MAS-MAS ESTRATIFICADO

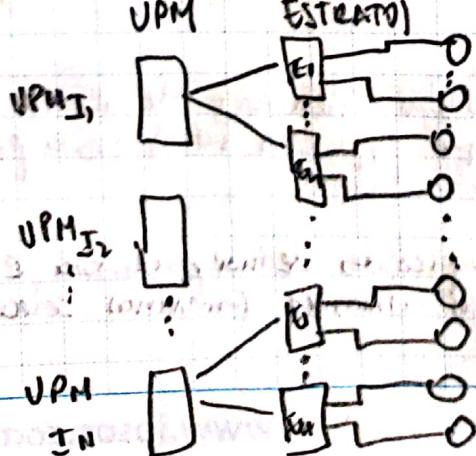
Hay 2 maneras de hacerlo.

① ESTMAS - MAS



$$\hat{t}^h_i = \sum_{h=1}^H t^h_i h_i \quad \hat{V}(t^h_i) = \sum_{h=1}^H \hat{V}(t^h_i h_i)$$

② MAS - ESTMAS



por lo que:

$$\hat{t}_{\text{yH}} = \sum_{h=1}^H \hat{t}_{yh} = \sum_{h=1}^H \left[\frac{N_h}{n_h} \sum_{i \in S_h} \frac{n_i}{N_i} \sum_{k \in U_i} \hat{\gamma}_k \right]$$

con Varianza:

$$V_{\text{HES-HM}} = \sum_{h=1}^H V(\hat{t}_{yh}) = \sum_{h=1}^H \left[\frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{yH}^2 + \frac{N_h}{n_h} \sum_{i \in S_h} \frac{n_i^2}{N_i} \left(1 - \frac{n_i}{N_i}\right) S_{yi}^2 \right]$$

y estimación de la varianza

$$\hat{V}_{\text{HES-HM}} = \sum_{h=1}^H \hat{V}(\hat{t}_{yh}) = \sum_{h=1}^H \left[\frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \hat{S}_{yH}^2 + \frac{N_h}{n_h} \sum_{i \in S_h} \frac{n_i^2}{N_i} \left(1 - \frac{n_i}{N_i}\right) \hat{S}_{yi}^2 \right]$$

→ DISEÑO MULTIESTAPLOJ:

- $\hat{t}_{yH} = \sum_{i \in S_H} \frac{t_{yi}}{T_{Bi}}$

- $V_{\text{HE}}(\hat{t}_{yH}) = \underbrace{\sum_{ij} \Delta_{ij} \frac{t_{ij} t_{ji}}{T_{Bi} T_{Bj}}}_{V(\text{UPM})} + \underbrace{\sum_{i \in U_H} \frac{V_i}{T_{Bi}}}_{V(\text{Resto})}$

- $\hat{V}_{\text{HE}}(\hat{t}_{yH}) = \underbrace{\sum_{ij} \Delta_{ij} \frac{\hat{t}_{ij} \hat{t}_{ji}}{T_{Bi} T_{Bj}}}_{\hat{V}(\text{UPM})} + \underbrace{\sum_{i \in S_H} \hat{V}_i}_{\hat{V}(\text{Resto})}$

→ ESTIMADOR DE Hansen - Horwitz

Dentro de cada UPM seleccionada el sorteo aleatorio con reemplazo se forma una sub-muestra (con o sin reemplazo). Aunque existe un perdido de eficiencia cuando el muestreo sea con reemplazo, ésto se compensa con una ganancia logística en el proceso de estimación de los varianzales requeridos para la característica de interés. El proceso general de muestreo es el siguiente:

- En la primera etapa se selecciona una muestra aleatoria de acuerdo a un diseño de muestreo con reemplazo, tal que P_i con $i \in U_1$ es la probabilidad de selección de la i-ésima UPM.
- En las siguientes etapas, se mantienen las propiedades de independencia entre sí sin importar si el diseño dentro de las unidades primarias seleccionadas sea con o sin reemplazo.

- Si una UPM es seleccionada en más de una ocasión, se debe realizar tanto los submuestreos como veces haya sido seleccionada en la primera etapa.
- Bajo un diseño de muestras multietápico, el estimador H-H para el total t_y , su varianza y su varianza estimada están dadas por:

$$\hat{t}_{yp} = \frac{1}{M_I} \sum_{v=1}^{M_I} \frac{\hat{t}_{yiv}}{p_{yiv}}$$

$$V(\hat{t}_{yp}) = \frac{1}{M_I} \sum_{i=1}^{N_I} \left(\frac{\hat{t}_{yi}}{p_{yi}} - \hat{t}_y \right)^2 + \frac{1}{M_I} \sum_{i=1}^{N_I} \frac{V_i}{p_{yi}}$$

$$\hat{V}(\hat{t}_{yp}) = \frac{1}{M_I(M_I-1)} \sum_{v=1}^{M_I} \left(\frac{\hat{t}_{yiv}}{p_{yiv}} - \hat{t}_{yp} \right)^2$$

ESTIMACION DE PARAMETROS DIFERENTES AL TOTAL:

La metodología q' se propone para estimar parámetros poblacionales es reescribirlos como función de todos los poblacionales. Así si el parámetro a estimar es B , se debe llevar a la siguiente forma:

$$B = f(t_1, t_2, \dots, t_q)$$

Donde cada t_q , $q=1, \dots, Q$ representa un total de las características de interés o un total de una función de las características de interés. El principio de estimación de este parámetro es b' en obtener estimadores (t_q) $q=1, \dots, Q$ tal q' T es' estimado por $\hat{B} = f(\hat{t}_1, \hat{t}_2, \dots, \hat{t}_q)$

Notese q' la función f puede ser lineal o no. Un resultado muy conocido de la inferencia estadística clásica nos indica q' si la función f es una función lineal entonces B , basta la forma:

$$B = a_0 + \sum_{q=1}^Q a_q t_q$$

Por tanto, un estimador inscrito de B es' dado por la siguiente expresión:

$$\hat{B} = a_0 + \sum_{q=1}^Q \hat{a}_q \hat{t}_q$$

Si en la estimación se utilizó el estimador de H-T entonces

$$\hat{\beta} = \alpha_0 + \sum_{q=1}^Q \alpha_q \sum_{k=1}^S \frac{Y_{qk}}{T_{qk}}$$

$$\hat{\beta} = \alpha_0 + \sum_{q=1}^Q \frac{1}{T_{qk}} \sum_{k=1}^S \alpha_q Y_{qk}$$

entonces se puede reescribir

$$\hat{\beta} = \alpha_0 + \sum_S \frac{E_k}{T_{qk}} \quad \text{donde } E_k = \sum_{q=1}^Q \alpha_q Y_{qk}$$

En cuanto a la varianza se tiene que:

$$V(\hat{\beta}) = V\left(\alpha_0 + \sum_S \frac{E_k}{T_{qk}}\right)$$

$$= V\left(\sum_S \frac{E_k}{T_{qk}}\right)$$

$$V(\hat{\beta}) = \sum_U \sum_{qk} \Delta_{qk} \frac{E_k T_{qk}}{T_{qk} T_{qk}}$$

por lo cual

$$\hat{V}(\hat{\beta}) = \sum_S \sum_{qk} \Delta_{qk} \frac{E_k T_{qk}}{T_{qk} T_{qk}}$$

↳ Aproximación de una función por polinomios: de Taylor:

se puede aproximar mediante un polinomio, entonces este estará definido por:

$$f(x) = f(a) + \frac{f'(a)}{1!} (x-a) + \frac{f''(a)}{2!} (x-a)^2 + \dots + \frac{f^{(n)}(a)}{n!} (x-a)^n + \dots$$

Ej: Sea $f(x) = \operatorname{sen} x$ y sea $x=0$ $\operatorname{sen}(0)=0$

$$f'(x) = \cos x$$

$$f''(x) = -\operatorname{sen} x$$

$$f'''(x) = -\cos x$$

$$\cos(0) = 1$$

$$-\operatorname{sen}(0) = 0$$

$$-\cos(0) = -1$$

Entonces

$$\operatorname{sen}(x) = 0 + x + \frac{0}{2!} x^2 - \frac{1}{3!} x^3 + \dots$$

por lo q' se puede aproximar:

$$\operatorname{sen}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \dots$$

→ Aplicación en estimadores:

Mediante esta técnica es posible aproximar la variancia de los estimadores q' no son funciones lineales de totales.

Lohr (2000) plantea los siguientes pasos para construir un estimador linealizado de la variancia de una función de totales, con $B = f(t_1, t_2, \dots, t_a)$

- Expresar el estimador del parámetro, \hat{B} , como una función de estimadores de totales insesgados, así: $\hat{B} = f(\hat{t}_1, \hat{t}_2, \dots, \hat{t}_a)$

Ejemplo: si $B = f(t_1, t_2) = \frac{t_1}{t_2}$ entonces $\hat{B} = \frac{\hat{t}_1}{\hat{t}_2}$

- Determinar todos los derivados parciales de f con respecto a cada total estimado \hat{t}_q y evaluar el resultado en las cantidades poblacionales t_q . Así:

$$\alpha_q = \left. \frac{\partial f(\hat{t}_1, \hat{t}_2, \dots, \hat{t}_a)}{\partial t_q} \right|_{\hat{t}_1=t_1, \hat{t}_2=t_2, \dots, \hat{t}_a=t_a}$$

- Aplicar el teorema de Taylor para funciones vectoriales para linearizar la estimación \hat{B} con $q = (t_1, t_2, \dots, t_a)'$. En el caso anterior se vio q' $\nabla B = (1, 1, \dots, 1)$ por lo que se tiene que:

$$\hat{B} = f(t_1, t_2, \dots, t_a) + \left. \frac{\partial f(\hat{t}_1, \dots, \hat{t}_a)}{\partial f(\hat{t}_1)} \right|_{\hat{t}_1=t_1} + \dots + \left. \frac{\partial f(\hat{t}_1, \dots, \hat{t}_a)}{\partial f(\hat{t}_a)} \right|_{\hat{t}_a=t_a} + \begin{cases} \hat{t}_1 - t_1 \\ \hat{t}_2 - t_2 \\ \vdots \\ \hat{t}_a - t_a \end{cases}$$

por lo que:

$$\hat{B} = f(t_1, t_2, \dots, t_a) + \sum_{q=1}^Q \left. \frac{\partial f(t_1, \dots, t_a)}{\partial t_q} \right|_{\hat{t}_1=t_1, \dots, \hat{t}_a=t_a} (\hat{t}_q - t_q)$$

entonces:

$$\hat{B} = f(t_1, t_2, \dots, t_a) + \sum_{q=1}^Q \alpha_q (\hat{t}_q - t_q)$$

$$= f(t_1, t_2, \dots, t_a) + \sum_{q=1}^Q \alpha_q \hat{t}_q - \underbrace{\sum_{q=1}^Q \alpha_q t_q}_{\text{constante.}}$$

$$= f(t_1, t_2, \dots, t_a) - \sum_{q=1}^Q \alpha_q t_q + \sum_{q=1}^Q \alpha_q \hat{t}_q$$

iv) Sea una nueva variable E_k con $k \in S$, el nivel de cada elemento observado en la muestra aleatoria.

$$E_k = \sum_{q=1}^Q \alpha_q Y_{qk}$$

v) Si los estimadores de \hat{Y}_q son estimadores de HT, se tiene que:

$$\text{Var}(\hat{B}) = \sum_j \sum_k \frac{\Delta E_k E_k}{T_k T_L}$$

para encontrar una estimación de la varianza de \hat{B} , no es posible utilizar directamente los valores E_k , porque estos dependen de los totales poblacionales, pues las derivadas α_q se evalúan en los totales poblacionales q son desconocidas. Por consiguiente los valores E_k se aproximan reemplazando los totales desconocidos por los estimadores de los mismos. Hasta es la aproximación de la variable linearizada dada por:

$$e_k = \sum_{q=1}^Q \hat{\alpha}_q Y_{qk} \quad \text{donde } \hat{\alpha}_q \text{ es un estimador de } \alpha_q$$

por lo q

$$\hat{\Delta}V(\hat{B}) = \sum_j \sum_k \frac{\Delta e_k e_k}{T_k T_L}$$

Ejemplo:

con $q=2$

$$R = f(t_1, t_2) = \frac{t_1}{t_2} \quad \text{entonces } \hat{R} = \frac{\hat{t}_1}{\hat{t}_2}$$

se define entonces

$$\alpha_q = \left(\frac{\partial f(\hat{t}_1, \hat{t}_2)}{\partial \hat{t}_1}, \frac{\partial f(\hat{t}_1, \hat{t}_2)}{\partial \hat{t}_2} \right) \Bigg|_{\begin{array}{l} \hat{t}_1 = t_1 \\ \hat{t}_2 = t_2 \end{array}}$$

las derivadas parciales para \hat{R} serían:

$$\left(\frac{1}{\hat{t}_2}, -\frac{t_1}{\hat{t}_2^2} \right) \Bigg|_{\begin{array}{l} \hat{t}_1 = t_1 \\ \hat{t}_2 = t_2 \end{array}} = \left(\frac{1}{t_2}, -\frac{t_1}{t_2^2} \right) = \alpha_q$$

por lo q

$$\hat{R} = \frac{t_1}{t_2} + \left(\frac{1}{t_2}, -\frac{t_1}{t_2^2} \right) \begin{pmatrix} \hat{t}_1 - t_1 \\ \hat{t}_2 - t_2 \end{pmatrix}$$

$$\hat{R} \approx \frac{t_1}{t_2} + \frac{\hat{t}_1 - t_1}{t_2} - \frac{t_1(\hat{t}_1 - t_1)}{t_2^2}$$

$$Y \text{ por su parte } \hat{A}V(\hat{\beta}) = \sum_k \sum_l \Delta_{kl} \frac{\hat{e}_k \hat{e}_l}{T_k T_l}$$

con lo que

$$\hat{e}_k = \left(\frac{1}{t_2}, -\frac{t_1}{t_2^2} \right) \begin{pmatrix} Y_k \\ Z_k \end{pmatrix} \Rightarrow \hat{e}_k = \frac{Y_k}{t_2} - \frac{t_1 Z_k}{t_2^2}$$

$$\hat{e}_k = \frac{1}{t_2} \left(Y_k - \frac{t_1}{t_2} Z_k \right) \Rightarrow \hat{e}_k = \frac{1}{t_2} (Y_k - \hat{\beta} Z_k)$$

Como \hat{e}_k es una función de $\hat{\beta}$ y $\hat{\beta}$ es definido sobre el universo de tiene q' realizar una estimación de \hat{e}_k para estimar la varianza.

$$\hat{e}_k = \frac{1}{t_2} (Y_k - \hat{\beta} Z_k)$$

por lo q'

$$\hat{A}V(\hat{\beta}) = \sum_k \sum_l \frac{\Delta_{kl}}{T_k T_l} \frac{\hat{e}_k \hat{e}_l}{T_k T_l}$$

con

$$\hat{\beta} = \frac{\hat{Y} \hat{Z}}{\hat{t}_2 n}$$

• En la práctica.

i) Calcula $\hat{e}_k = \frac{1}{t_2} (Y_k - \hat{\beta} Z_k)$

ii) EN MAS:

$$\hat{t}_2 = \frac{N}{n} \sum_k Z_k \quad \hat{Y} = \frac{N}{n} \sum_k Y_k$$

$$\hat{A}V(\hat{\beta}) = \frac{N^2}{n} \left(1 - \frac{n}{N} \right) S_{\text{resid}}^2$$

Ejemplo: $N=31 \quad n=5$

$$\hat{Y} = \frac{1}{5} (2 - (0,4822)(4)) = 0,00058$$

$$\hat{t}_2 = \frac{31}{5} (19,7) = \underline{\underline{122,14}}$$

$$1 \quad 1,2 \quad \frac{1}{122,14} (1 - (0,4822)(1,2)) = 0,0034498$$

$$\hat{Y} = \frac{31}{5} (9,5) = \sqrt{8,9}$$

$$2 \quad 5 \quad \frac{1}{122,14} (2 - (0,4822)(5)) = -0,003336$$

$$\hat{\beta} = \frac{\sqrt{8,9}}{122,14} = 0,4822$$

$$1,5 \quad 3,5 \quad \frac{1}{122,14} (1,5 - (0,4822)(3,5)) = -0,00154$$

$$3 \quad 6 \quad \frac{1}{122,14} (3 - (0,4822)(6)) = 0,0002748 \Rightarrow S_{\text{resid}}^2 = 0,00000666$$

$$4 \quad \hat{A}V(\hat{\beta}) = \frac{31^2}{5} \left(1 - \frac{5}{31} \right) (0,00000666) = 0,001098$$

↳ Razón Poblacional en MAS:

$$\bullet \hat{R} = \frac{\bar{y}_s}{\bar{z}_s}$$

$$\bullet \text{AVMAS}(\hat{R}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{EU}^2$$

$$\bullet \hat{\text{AVMAS}}(\hat{R}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_{EGS}^2$$

↳ Razón Poblacional en MAS²

$$\hat{R} = \frac{\sum_{i \in S} N_i \bar{y}_{Si}}{\sum_{i \in S} N_i \bar{z}_{Si}}$$

$$\text{AVAR}_{MAS^2}(\hat{R}) = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S_{EU_I}^2 + \frac{N_I}{n_I} \sum_{i \in W_I} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S_{EU_i}^2$$

$$\hat{\text{AVAR}}_{MAS^2}(\hat{R}) = \frac{N_I^2}{n_I} \left(1 - \frac{n_I}{N_I}\right) S_{EGS_I}^2 + \frac{N_I}{n_I} \sum_{i \in S_I} \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S_{EGS_i}^2$$

↳ Estimación de un promedio:

Un estimador del promedio poblacional \bar{y}_v , definido como una razón jerárquica:

$$\tilde{y}_s = \frac{\hat{t} \hat{y}_H}{\hat{N}_H} \quad \text{con} \quad \hat{N}_H = \sum_S \frac{1}{T_{HS}}$$

$$\text{AVAE}(\tilde{y}_s) = \frac{1}{N^2} \sum_U \sum_L \Delta_{KL} \left(\frac{y_K - \bar{y}_U}{T_{KU}} \right) \left(\frac{y_L - \bar{y}_U}{T_{LU}} \right)$$

$$\hat{\text{AVAE}}(\tilde{y}_s) = \frac{1}{\hat{N}^2} \sum_S \sum_L \frac{\Delta_{KL}}{T_{HL}} \left(\frac{y_K - \tilde{y}_S}{T_{KU}} \right) \left(\frac{y_L - \tilde{y}_S}{T_{LU}} \right)$$

↳ Estimación de un promedio en un dominio

$$\hat{q}_{sd} = \frac{\hat{t}y_{sd}}{\hat{N}_{sd}} \quad \text{con} \quad \hat{t}y_{sd} = \sum_S \frac{y_{sd}}{\pi_{kL}} \quad y \quad \hat{N}_{sd} = \sum_S \frac{2}{\pi_{kL}}$$

$$AV(\hat{q}_{sd}) = \frac{1}{N_g^2} \sum_0 \sum_L \Delta_{KL} \left(\frac{y_{dkL} - \bar{y}_{ud}}{\pi_{kL}} \right) \left(\frac{y_{dkL} - \bar{y}_{sd}}{\pi_{kL}} \right)$$

$$\hat{AV}(\hat{q}_{sd}) = \frac{1}{\hat{N}_g^2} \sum_S \sum_L \frac{\Delta_{KL}}{\pi_{kL}} \left(\frac{y_{dkL} - \hat{q}_{sd}}{\pi_{kL}} \right) \left(\frac{y_{dkL} - \hat{q}_{sd}}{\pi_{kL}} \right)$$