MUEST_T9. 1. MUESTREO de CONGLOMERADOS SIN SUBMUESTREO.

3. COEFICIENTE de CORRELACIÓN INTRACONGLOMERADO y su INTERPRETACIÓN.

2. ESTIMADORES, VARIANZAS y sus ~~INTERPRETACIONES~~ ESTIMACIONES. EFECTO de DISEÑO.

4. UTILIZACIÓN de ESTIMADORES de RAZÓN

---

1 – MUESTREO de CONGLOMERADOS

El muestreo de conglomerados considera una población finita con un total de M unidades elementales o últimas, agrupadas en N conglomerados ~~...~~ o unidades primarias, de forma que constituyen una partición de la población. La unidad de muestreo es el conglomerado y de la población se extrae una muestra aleatoria de n conglomerados, a partir de la cual se estiman los parámetros poblacionales.

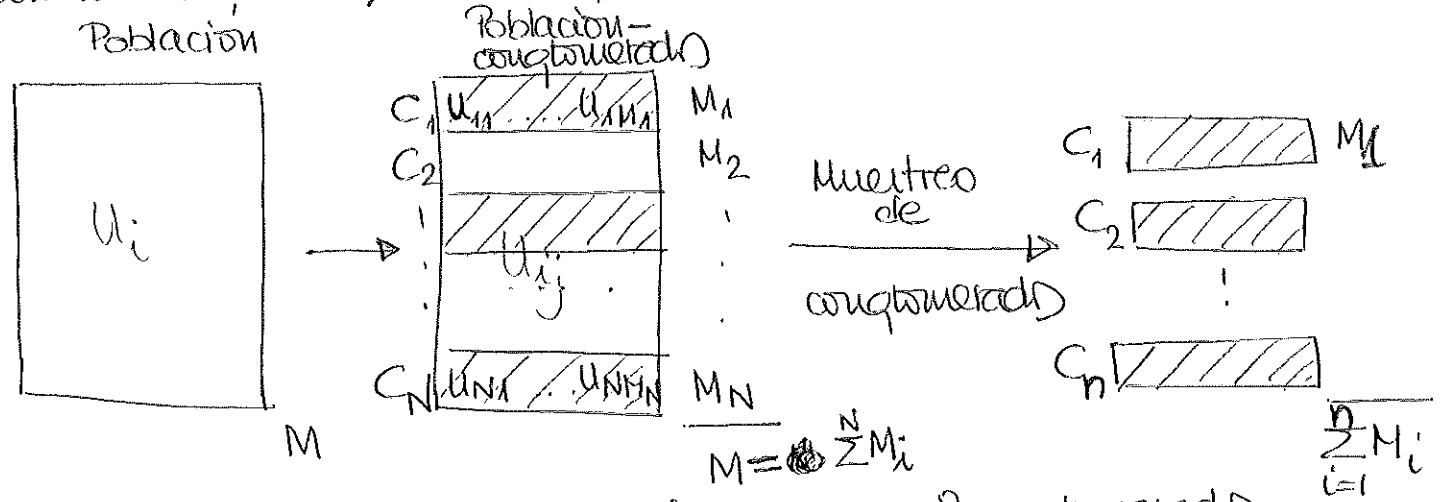
- Si de cada conglomerado muestral se observan todas las unidades elementales ⟶ Muestreo de conglomerados ~~mono~~etápico o sin submuestreo.
- Si de cada conglomerado se selecciona una muestra ⟶ muestreo polietápico.

El nº de unidades elementales de cada conglomerado se denomina tamaño del conglomerado < tamaños iguales / tamaños ≠

~~los conglomerados deberían~~

El muestreo por conglomerados debería ser de modo que los conglomerados fuesen lo más homogéneos entre sí (a la hora de constituir la muestra no se pierde información) y dentro de cada conglomerado se recoja la mayor heterogeneidad posible (similar a la de la población, población a escala).

La situación ideal es que un único conglomerado pudiese representar fielmente a la población (muestra de tamaño 1 con mínimo coste)



Unidad elemental poblacional, u_{ij} $\begin{cases} i = 1 \ldots N \rightarrow \text{nº conglomerados} \\ j = 1 \ldots M_i \rightarrow \text{tamaño muestral conglom.} \end{cases}$

Unidad elemental muestral, u_{ij} $\begin{cases} i = 1 \ldots n \rightarrow \text{nº conglom. muestr.} \\ j = 1 \ldots M_i \rightarrow \text{en muestreo monoetápico se observan todas las unid. del conglomerado} \end{cases}$

El muestreo de conglomerados sin submuestreo consiste en seleccionar aleatoriamente $\underline{\underline{n}}$ conglomerados (se pueden utilizar ≠ tipos de muestreo), y observar dentro de cada conglomerado todas sus unidades elementales, en total $\sum_{i=1}^{n} M_i$ unid. elementales.

Como cada conglomerado se observa de manera exhaustiva, sin efectuar submuestreo dentro de él, también se denomina muestreo de conglomerados monoetápico, MCM(n).

Antes de continuar, quizás sería conveniente explicar detalladamente las diferencias entre muestreo estratificado y muestreo por conglomerados.

~~Ex~~ El muestreo estratificado trata de obtener una partición de la población en estratos heterogéneos entre sí y homogéneos dentro. El muestreo por conglomerados divide a la población en conglomerados homogéneos entre sí y heterogéneos dentro.

La situación ideal del muestreo estratificado es obtener una muestra de tamaño 1 de cada estrato → L unid. elem.
La situación ideal del muestreo por conglomerados es ~~tomar~~ información de un único conglomerado ~~y obser~~ y en el caso monoetápico, observar de él todas sus unidades elem.
→ Mi unid. elementales.

Es muy frecuente que los conglomerados estén definidos como áreas geográficas (división territorial de la población), por lo que al muestreo por conglomerados también se le conoce como muestreo por áreas. Se utiliza por razones de economía en coste, en tiempo, en recursos, etc, y algunas veces porque se disminuye el sesgo al ser más fácil la supervisión.

Nótese que para efectuar m.a.s. hace falta una lista detallada con todos los ~~e~~ elementos de la población (marco), y en muestreo estratificado una lista detallada para todos y cada uno de los estratos → (mucho coste y mucho tiempo)
En la práctica, no se dispone de tales listas. Es preferible dividir la población en áreas y elaborar las listas solamente de las áreas seleccionadas.

Ventajas del muestreo por conglomerados:

+ Marco + fácil: No se necesita un marco muy específico, como en el caso de mas y de me.

+ Marco + barato: Al seleccionar previamente las áreas sobre las que elaborar el listado de unid. elem., el marco de conglomerados es más fácil de conseguir, en términos de tiempo, dinero y de efectivos.

+ Marco ya existe: Se pueden utilizar archivos ya existentes por necesidades administrativas como marco.

+ Proceso más rápido y barato: la concentración de unidades disminuye el tiempo de ~~desp~~ recogida de datos, reduce el coste por desplazamiento y permite utilizar recursos disponibles por zonas (personal y supervisores).

Inconvenientes:

- Menor precisión en las estimaciones: Aunque lo ideal es que la heterogeneidad dentro de cada conglomerado sea máxima, siempre va a existir un grado de homogeneidad inevitable dentro de los conglom.

- La eficiencia disminuye al aumentar el tamaño de los conglomerados: cuando en la práctica este tipo de muestreo es muy útil en poblaciones muy numerosas en las que se puedan construir conglomerados grandes.

2 – ESTIMADORES, VARIANZAS Y ESTIMACIONES

El muestreo por conglomerados distingue las siguientes situaciones:

- Conglomerados de igual tamaño (AQUÍ)
- Conglomerados de $\neq$ tamaño $\left\langle \begin{array}{l} \text{parecido} \\ \text{muy} \neq \end{array} \right\}$ En 4. RAZÓN.

Además, el método de muestreo elegido para seleccionar los conglomerados muestrales lleva a estimador $\neq$s, con distintas varianzas. Distinguimos:

- Muestreo sin reposición $\left\langle \begin{array}{l} \text{probab.} = (*) + \text{utilizado.} \\ \text{probab.} \neq \end{array} \right.$
- Muestreo con reposición $\left\langle \begin{array}{l} \text{probab} = \\ \text{probab} \neq \end{array} \right.$

En este epígrafe, consideramos el caso más sencillo: todos los conglomerados tienen el mismo tamaño $\overline{M} = \frac{M}{N}$ y las probab. son iguales

Sean:

N = nº conglomerados de la población

n = nº conglomerados en la muestra

$\overline{M}$ = tamaño del conglomerado $(M_1 = \ldots = M_N = \overline{M})$

$N\overline{M}$ = nº total de unid. elementales en la poblac. $(N\overline{M} = M)$

$n\overline{M}$ = nº total de unid. elementales en la muestra.

Sea $\theta = \boxed{\sum_{i=1}^{N} Y_i} = \sum_{i=1}^{N} \sum_{j=1}^{\overline{M}} Y_{ij}$, parámetro poblacional a estimar

$\downarrow$

notación anterior

N = nº unid. elementales de la poblac.

N = tamaño poblacional.

a) SIN reposición, probab IGUALES: (probab $\neq \Rightarrow \pi_i$).

El estimador lineal insesgado de θ es el estimador de Horvitz y Thompson

$$\hat{\theta}_{HT} = \sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} \frac{y_{ij}}{\pi_i} \xlongequal{\text{probab. iguales } \pi_i=\frac{n}{N}} \sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} \frac{y_{ij}}{n/N} = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} y_{ij}.$$

Es insesgado de θ:

$e_i = \begin{cases} 1 & \text{si cong. } i \in \text{muestra con probab. } \pi_i \\ 0 & \notin \quad 1-\pi_i \end{cases}$ | $e_i \to Bern(\pi_i)$

$$E[\hat{\theta}_{HT}] = E\left[\sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} \frac{y_{ij}}{\pi_i}\right] = E\left[\sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} \frac{y_{ij}}{\pi_i} e_i\right] = \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} \frac{y_{ij}}{\pi_i} E[e_i] =$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} \frac{y_{ij}}{\pi_i} \pi_i = \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} y_{ij} = \theta.$$

La expresión del estimador $\hat{\theta}_{HT}$ en los parámetros más utilizados es:

$$\theta = X = \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} X_{ij} \Rightarrow \hat{X} = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} x_{ij} = \frac{N\bar{M}}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{\bar{M}} \sum_{j=1}^{\bar{M}} x_{ij}}_{\bar{x}_i} =$$

$$= N\bar{M} \cdot \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i = N\bar{M} \cdot \bar{\bar{x}}$$

↳ el estimador del total es el nº total de unidades de la población multiplicado por el estimador de la media (estimador de expansión)

$$\theta = \bar{X} = \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} \frac{X_{ij}}{N\bar{M}} \Rightarrow \hat{\bar{X}} = \frac{\cancel{N}}{n} \cdot \frac{1}{\cancel{N}\bar{M}} \sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} x_{ij} = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{\bar{M}} \sum_{j=1}^{\bar{M}} x_{ij}}_{\bar{x}_i} = \bar{\bar{x}}$$

↳ el estimador insesgado de la media poblacional es la media muestral de las medias poblac de los conglomerados

$$\theta = P = \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} \frac{A_{ij}}{N\bar{M}} \Rightarrow \hat{P} = \frac{\cancel{N}}{n} \cdot \frac{1}{\cancel{N}\bar{M}} \sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} A_{ij} = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{\bar{M}} \sum_{j=1}^{\bar{M}} A_{ij}}_{P_i} = \frac{1}{n} \sum_{i=1}^{n} \overset{\text{pobl.}}{P_i}$$

$$\theta = A = \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} A_{ij} \Rightarrow \hat{A} = N\bar{M} \cdot \hat{P} = N\bar{M} \cdot \frac{1}{n} \sum_{i=1}^{n} P_i$$

MIRAR

VARIANZAS :

$f = \frac{n\overline{M}}{N\overline{M}} = \frac{n}{N}$ = fracción de muestreo

$\overline{M} S_x^2 = S_b^2$

Para la media :

$$V(\bar{\bar{x}}) = V\left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_i\right) \overset{SR}{=} \frac{(1-f)}{n}\cdot\frac{\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2}{N-1} = \frac{(1-f)}{n\overline{M}}\cdot\frac{\sum_{i=1}^{N}\overline{M}(\bar{X}_i-\bar{X})^2}{N-1} =$$

$$= \frac{(1-f)}{n\overline{M}}\cdot S_b^2 = \cancel{\frac{(1-f)}{n}S_x^2 \text{, donde } S_x^2 = \frac{\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2}{N-1}}$$

donde $S_b^2 = \frac{\sum^{N}\overline{M}(\bar{X}_i-\bar{X})^2}{N-1}$ = cuasivarianza poblac entre conglomerados

between

de la cual se pueden deducir las expresiones de las varianzas de los otros estimadores :

Para el total poblacional :

$$\hat{X} = N\overline{M}\,\bar{\bar{x}} \Rightarrow V(\hat{X}) = N^2\overline{M}^2 V(\bar{\bar{x}}) = N^2\overline{M}\cdot\frac{S_b^2}{n}(1-f)$$

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n}P_i \Rightarrow V(\hat{P}) = \frac{(1-f)}{n\overline{M}}\cdot S_b^2 = \frac{(1-f)}{n}\cdot\frac{\sum^{N}(P_i-P)^2}{N-1}$$

donde $S_b^2 = \frac{\overline{M}\sum(P_i-\hat{P})^2}{N-1}$

$$\hat{A} = N\overline{M}\hat{P} \Rightarrow V(\hat{A}) = N^2\overline{M}^2 V(\hat{P}) = N^2\overline{M}\cdot\frac{(1-f)}{n}S_b^2$$

ESTIMACIONES de las VARIANZAS

Como la cuasivarianza muestral entre conglomerados $\hat{S}_b^2$ es un estimador insesgado de la varianza poblacional S_b^2 :

$$E[\hat{S}_b^2] = E\left[\frac{\overline{M}}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \overline{M}\,E\left[\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}\right] = \overline{M}E[\hat{S}_x^2] =$$

$$= \overline{M}\cdot S_x^2 = S_b^2 .$$

$$\hat{V}(\bar{\bar{x}}) = \frac{(1-f)}{n\overline{M}}\cdot\hat{S}_b^2 \quad \text{con } \hat{S}_b^2 = \frac{\overline{M}}{n-1}\sum^{n}(\bar{X}_i-\bar{X})^2$$

$$\hat{V}(\hat{X}) = N^2\overline{M}\,\frac{(1-f)}{n}\hat{S}_b^2$$

$$\hat{V}(\hat{P}) = \frac{(1-f)}{n\overline{M}}\hat{S}_b^2 \quad \text{con } \hat{S}_b^2 = \frac{\overline{M}}{n-1}\sum^{n}(P_i-P)^2$$

$$\hat{V}(\hat{A}) = N^2\overline{M}\cdot\frac{(1-f)}{n}\hat{S}_b^2$$

MIRAR

DESCOMPOSICIÓN de la VARIANZA

SCT = SCD + SCE

$$\sum_{i=1}^{N}\sum_{j=1}^{\bar M}(X_{ij}-\bar X)^2=\sum_{i=1}^{N}\sum_{j=1}^{\bar M}(X_{ij}-\bar X_i+\bar X_i-\bar X)^2=\sum\sum(X_{ij}-\bar X_i)^2+\sum\sum(\bar X_i-\bar X)^2+$$

$$+\underbrace{2\sum\sum(X_{ij}-\bar X)(\bar X_i-\bar X)}_{=0}=\sum_{i=1}^{N}\sum_{j=1}^{\bar M}(X_{ij}-\bar X)^2+\sum_{i=1}^{N}\sum_{j=1}^{M}(X_i-\bar X)^2$$

$$\sum_{i=1}^{N}\sum_{j=1}^{\bar M}(X_{ij}-\bar x)^2=\sum_{i=1}^{N}\sum_{j=1}^{\bar M}(X_{ij}-\bar X_i)^2+\sum_{i=1}^{N}\sum_{j=1}^{\bar M}(\bar X_i-\bar X)^2$$

$$\downarrow\qquad\qquad\downarrow\qquad\qquad\downarrow$$

$$(N\bar M-1)S_T^2 \;=\; N(\bar M-1)S_W^2 + (N-1)S_b^2$$

$$N\bar M\sigma^2 \quad = \quad N\bar M\sigma_W^2 \qquad N\sigma_b^2$$

$$\sigma^2 \qquad\qquad \sigma_W^2 \qquad \sigma_b^2/\bar M$$

En el caso de la muestra:

$$\sum_{i=1}^{n}\sum_{j=1}^{\bar M}(X_{ij}-\bar{\bar x})^2=\sum\sum(X_{ij}-\bar X_i+\bar X_i-\bar{\bar x})^2=\sum\sum_{j=1}^{\bar M}(X_{ij}-\bar X_i)^2+\sum\sum(\bar X_i-\bar{\bar X})^2+$$

$$+2\sum\sum(X_{ij}-\bar X_i)(\bar X_i-\bar{\bar X}) \qquad\qquad \sum_{i=1}^{n}\bar M(\bar X_i-\bar X)^2$$

$$(n\bar M-1)\hat S_T^2 = \qquad n(\bar M-1)\hat S_W^2 + (n-1)\hat S_b^2$$

$\hat S_b^2$ es un estimador insesgado de S_b^2

$\hat S_W^2$ es un estimador insesgado de S_W^2

MIRAR ATRÁS

Un estimador insesgado de S^2 es:

$$S_o^2=\frac{N(\bar M-1)}{N\bar M-1}\hat S_W^2+\frac{N-1}{N\bar M-1}\hat S_b^2$$

aunque para muestras de más de 50 conglomerados se puede considerar como estimador insesgado de S^2 a:

$$\hat S_T^2=\frac{1}{n\bar M-1}\sum\sum(X_{ij}-\bar{\bar x})^2$$

b) CON reposición, probabilidades IGUALES:

El estimador lineal insesgado de Θ es el estimador de Hansen y Hurwitz: probab = , $P_i = 1/N$

$$\hat{\Theta}_{HH} = \sum_{i=1}^{n}\sum_{j=1}^{\overline{M}} \frac{y_{ij}}{nP_i} = \sum_{i=1}^{n}\sum_{j=1}^{\overline{M}} \frac{y_{ij}}{n/N} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{\overline{M}} y_{ij}.$$

cuya expresión en probabilidades iguales coincide con el estimador de Horvitz y Thompson.

Varianza:

Para la media poblacional: NO

$$V(\bar{\bar{x}}) = V\left(\frac{1}{n}\sum_{i=1}^{n}\bar{X}_i\right) = \frac{\sigma_x^2}{n} = \frac{\frac{1}{N}\sum^{N}(X_i-\bar{X})^2}{n} \to \frac{\frac{1}{N}\cdot\sum \overline{M}(X_i-\bar{X})^2}{n\overline{M}} = \frac{\sigma_b^2}{n\overline{M}}$$

donde $\sigma_b^2 = \frac{1}{N}\sum_{i=1}^{N}\overline{M}(X_i-\bar{X})^2 \equiv$ varianza entre conglomerados pobl.

(expresión similar al m.a.s. sustituyendo σ^2 por σ_b^2 y n por $n\overline{M}$, nº total de unidades elementales en la muestra).

Para el total poblacional:

$$V(\hat{X}) = V(N\overline{M}\cdot\bar{\bar{x}}) = N^2\overline{M}^2 V(\bar{\bar{x}}) = N^2\overline{M}\cdot\frac{\sigma_b^2}{n}$$

Para la proporción:

$$V(\hat{P}) = \frac{\sigma_b^2}{n\overline{M}} = \frac{\frac{\overline{M}}{N}\sum_i (P_i-P)^2}{n\overline{M}} = \sum_{i=1}^{N}\frac{(P_i-P)^2}{nN}$$

σ_p^2

Para el total de clase:

$$V(\hat{A}) = V(N\overline{M}\hat{P}) = N^2\overline{M}^2 V(\hat{P}) = N^2\overline{M}^2\frac{\sum^{N}(P_i-P)^2}{nN}$$

Estimaciones

$$\hat{V}(\bar{\bar{x}}) = \frac{\hat{S}_b^2}{n\overline{M}}$$

$$\hat{V}(\hat{X}) = N^2\overline{M}^2\cdot\hat{V}(\bar{\bar{x}})$$

$$\hat{V}(\hat{P}) = \frac{\hat{S}_b^2}{n\overline{M}}$$

$$V(\hat{A}) = N^2\overline{M}^2\hat{V}(\hat{P}).$$

3_ COEF. de CORRELACIÓN INTRACONGLOMERADOS Y SU INTERPRETACIÓN. EFECTO de DISEÑO

El coeficiente de correlación intraconglomerados es una medida de la homogeneidad dentro de los conglomerados. Interesa que tenga un valor muy pequeño, pues en muest. por conglomerados lo ideal es la heterogeneidad dentro de los conglomerados.

Se define como el coef. de correlación lineal entre todos los pares de unidades elementales pertenecientes a cada conglomerado, para todos los conglomerados.

Para cada i, δ_{jk}

$$\delta_{jz} = \frac{cov(X_{ij}, X_{iz})}{\underbrace{\sigma(X_{ij})\,\sigma(X_{iz})}_{\text{son iguales}}} = \frac{E\left[(X_{ij}-E[X_{ij}])(X_{iz}-E[X_{iz}])\right]}{\sigma^2} \xrightarrow{\text{nº pares } = N\cdot\bar{M}\cdot(\bar{M}-1)}$$

$$\longrightarrow \frac{\sum_{i=1}^{N}\sum_{j\neq z}^{\bar{M}}(X_{ij}-\bar{X})(X_{iz}-\bar{X})}{N\bar{M}(\bar{M}-1)\,\sigma^2} = \frac{\sum_{i=1}^{N}\sum_{j\neq z}^{\bar{M}}(X_{ij}-\bar{X})(X_{iz}-\bar{X})}{(\bar{M}-1)(N\bar{M}-1)S^2}$$

$$\sigma^2 = \frac{N\bar{M}-1}{N\bar{M}}S^2$$

Utilizando el coef. de correlación intraconglomerados se puede expresar la varianza de los estimadores:

$$(SR) \longrightarrow V(\bar{\bar{x}}) = (1-f)\cdot\frac{S_b^2}{n\bar{M}} = \frac{(1-f)}{n\cancel{\bar{M}}}\cdot\frac{\cancel{\bar{M}}}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2 =$$

$$= \frac{1-f}{n(N-1)}\sum_{i=1}^{N}\left[\frac{1}{\bar{M}}\sum_{j=1}^{\bar{M}}(X_{ij}-\bar{X})\right]^2 = \frac{1-f}{n(N-1)}\frac{1}{\bar{M}^2}\sum_{i=1}^{N}\left(\sum^{\bar{M}}(X_{ij}-\bar{X})\right)^2 =$$

$$S_b^2 \underset{N\to\infty}{\simeq} S^2[1+(\bar{M}-1)\delta]$$

$$= \frac{1-f}{n(N-1)\bar{M}^2}\left[\sum_i\sum_j(X_{ij}-\bar{X})^2 + \sum_i\sum_{j\neq z}(X_{ij}-\bar{X})(X_{iz}-\bar{X})\right] =$$

$$= \frac{1-f}{n(N-1)\bar{M}^2}\left[(N\bar{M}-1)S^2 + (N\bar{M}-1)(\bar{M}-1)S^2\cdot\delta\right] =$$

$$= \frac{(1-f)S^2(N\bar{M}-1)}{n(N-1)\bar{M}^2}\left[1+(\bar{M}-1)\delta\right] \xrightarrow[N\to\infty]{} \frac{(1-f)S^2}{n\bar{M}}\left[1+(\bar{M}-1)\delta\right]$$

(SR) $\longrightarrow$ $V(\bar{\bar{x}}) = \underbrace{(1-f)\cdot\frac{S^2}{n\bar{M}}}_{S^2_{\bar{y}(n\bar{M})}}\left[1+(\bar{M}-1)\delta\right]$

que se puede expresar en función de la varianza de la media muestral obtenida con m.a.s.

$$V(\bar{\bar{x}}) = V_{mas}(\bar{\bar{x}})\left[1+(\bar{M}-1)\delta\right]$$

por lo que podemos hacer una comparación de varianzas a partir de la interpretación de δ:

$\delta>0 \Rightarrow$ Var. conglomerados es mayor que Var. mas. para muestras del mismo tamaño $n\bar{M}$.

Esta diferencia será máxima para $\delta=1$, caso más desfavorable. El caso más favorable, varianza mínima será cuando $\delta=-\frac{1}{\bar{M}-1}$, en el que la var. de conglomerados será 0.

Para $\delta=0$, ambos métodos son igual de precisos.

El término $(\bar{M}-1)$ se interpreta como el aumento de varianza debido a seleccionar n conglomerados de tamaño $\bar{M}$, en vez de hacer un m.a.s. de $n\bar{M}$ unidades elementales.

$\delta<0 \Rightarrow$ Var. conglomerados es menor que la varianza obtenida con m.a.s.

En la práctica, suele ocurrir que los elementos de cada conglomerado guarden cierto parecido entre sí, con lo que la correlación es positiva y el muestreo por conglomerados es menos preciso que el mas.

$$V_{MC}(\bar{\bar{x}}) = V_{mas}(\bar{\bar{x}})\left[1+(\bar{M}-1)\delta\right] \Rightarrow \begin{cases} \delta>0 \Rightarrow \text{Congl. peor que mas} \\ \delta=0 \Rightarrow \text{Congl. igual que mas} \\ \delta<0 \Rightarrow \text{Congl. mejor que mas} \end{cases}$$

Si llamamos n_a y n_c al tamaño de una muestra, expresado en unidades elementales para obtener una precisión dada, en el caso en el que los dos tipos de muestreo tengan la misma precisión: $V_{MC}(\bar{\bar{x}}) = V_{MAS}(\bar{x})[1+(\bar{M}-1)\delta] = V_{MAS}(\bar{x})$

por lo que: $$(1-f)\frac{S^2}{n_c}[1+(\bar{M}-1)\delta] = (1-f)\frac{S^2}{n_a}$$

$$\Rightarrow n_c = n_a[1+(\bar{M}-1)\delta],$$

es decir, la cantidad $[1+(\bar{M}-1)\delta]$ es aquella por la que hay que multiplicar el tamaño muestral de mas para obtener el tamaño muestral de conglomerados, se llama <u>efecto de diseño</u>.

Podemos interpretar los valores de δ utilizando toda la información anterior:

Si $V_{MC}(\bar{\bar{x}}) \simeq V_{MAS}(\bar{x})[1+(\bar{M}-1)\delta]$, se puede escribir como

$$(1-f)\cdot\frac{S_b^2}{n\bar{M}} \simeq (1-f)\frac{S^2}{n\bar{M}}[1+(\bar{M}-1)\delta], \text{ de donde:}$$

$$S_b^2 \simeq S^2[1+(\bar{M}-1)\delta] \longrightarrow S_b^2 - S^2 \simeq (\bar{M}-1)S^2\cdot\delta$$

$$\delta \simeq \frac{S_b^2 - S^2}{(\bar{M}-1)S^2} \in \left[-\frac{1}{\bar{M}-1}, 1\right]$$

$$\delta \simeq \frac{\frac{S_b^2}{S^2} - 1}{(\bar{M}-1)}$$

- $\delta = -\frac{1}{M-1} \Rightarrow S_b^2 = 0$
  $\Rightarrow V(\bar{\bar{x}}) = 0$

  $\Rightarrow$ toda la variabilidad procede de dentro de los conglomerados, que son homogéneos entre sí
  $\Rightarrow$ 1 solo conglomerado proporciona toda la información $\to$ caso ideal.

- $-\frac{1}{M-1} < \delta < 0 \Rightarrow V_{MC}(\bar{\bar{x}}) < V_{MAS}(\bar{x})$ ó $n_c < n_a$

  ⇒ El muestreo por conglom. es + preciso que el m.a.s.

  → Caso excepcional, conglomerados heterogéneos dentro

- $\delta = 0 \Rightarrow S_b^2 = S^2$, $V_{MC}(\bar{\bar{x}}) = V_{MAS}(\bar{x})$ ó $n_c = n_a$.

  → Los dos tipos de muestreo son igual de precisos.

  → La variabilidad entre conglomerados coincide con la variabilidad entre las unidades elementales.

- $0 < \delta < 1 \Rightarrow [1+(\bar{M}-1)\delta] > 1 \Rightarrow V_{MC}(\bar{\bar{x}}) > V_{MAS}(\bar{x})$ ó $n_c > n_a$

  → Existe homogeneidad dentro de los conglomerados (caso + habitual), y $S_b^2 > S^2$.

  → Aunque el m.c. es menos preciso, se suele utilizar por motivos de coste.

- $\delta = 1 \Rightarrow S_b^2 \simeq \bar{M} S^2$, $S_W^2 = 0$ y $n_c = \bar{M} n_a$

  → Caso más desfavorable, no hay variabilidad dentro de los conglomerados ⇒ se recomienda observar 1 sola observación de cada conglomerado.

  $$V(\bar{\bar{x}}) \simeq (1-f)\cdot\frac{S^2}{n\bar{M}}[1+(\bar{M}-1)\cdot 1] = \underbrace{(1-f)\frac{S^2}{n}}_{\text{m.a.s. con n unidades elementales}}$$

  muy superior a la var$_{MAS}$ obtenida con $n\bar{M}$ unid. elementales

---

(✓) En la EPA, $[1+(\bar{M}-1)\delta] \simeq 2$, pero reduce costes

$$(CR) \rightarrow \delta = \frac{\sum_{i=1}^{N} \sum_{j \neq z}^{\bar{M}} (X_{ij}-\bar{X})(X_{iz}-\bar{X})}{N\bar{M}(\bar{M}-1)\ \sigma^2}$$

$$\text{donde} \quad \sigma^2 = \frac{1}{N\bar{M}} \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} (X_{ij} - \bar{\bar{X}})^2$$

por lo que podemos expresar la varianza del estimador,

$$V(\bar{\bar{x}}) = \frac{\frac{1}{N}\sum_{i=1}^{N}(\bar{X}_i - \bar{X})^2}{n} = \frac{1}{nN}\sum_{i=1}^{N}\left[\frac{1}{\bar{M}}\sum_{j=1}^{\bar{M}} X_{ij} - \frac{\bar{M}\cdot\bar{X}}{\bar{M}}\right]^2 =$$

$$= \frac{1}{nN\bar{M}^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{\bar{M}} X_{ij} - \sum_{j=1}^{\bar{M}} \bar{X}\right)^2 = \frac{1}{nN\bar{M}^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{\bar{M}} (X_{ij}-\bar{X})\right)^2 =$$

$$= \frac{1}{nN\bar{M}^2}\left[\sum_{i=1}^{N}\sum_{j=1}^{\bar{M}}(X_{ij}-\bar{X})^2 + \sum_{i=1}^{N}\sum_{j\neq z}^{\bar{M}}(X_{ij}-\bar{X})(X_{iz}-\bar{X})\right] =$$

$$= \frac{1}{nN\bar{M}^2}\left[N\bar{M}\sigma^2 + N\bar{M}(\bar{M}-1)\sigma^2\cdot\delta\right] =$$

$$= \frac{\cancel{N}\cancel{\bar{M}}\sigma^2}{n\cancel{N}\bar{M}^{\cancel{2}}}\left[1 + (\bar{M}-1)\delta\right] = \frac{\sigma^2}{n\bar{M}}\left[1 + (\bar{M}-1)\delta\right]$$

$$(CR) \rightarrow V(\bar{\bar{x}}) = \frac{\sigma^2}{n\bar{M}}\left[1 + (\bar{M}-1)\delta\right]$$

~~de~~ expresión de la que se deducen las mismas conclusiones que en el caso de SIN reposición:

- $\delta \in \left[-\frac{1}{\bar{M}-1}, 1\right]$
- $V_{MC}(\bar{\bar{x}}) = V_{MAS}(\bar{x})\left[1 + (\bar{M}-1)\delta\right] \Rightarrow \begin{cases} \delta < 0 \rightarrow \text{congl. Mejor mas} \\ \delta = 0 \rightarrow \text{iguales} \\ \delta > 0 \rightarrow \text{congl. PEOR mas.} \end{cases}$

y cuanto más se acerque a los límites, más se acentúe la ganancia / pérdida en precisión del MC respecto al MAS

4_ UTILIZACIÓN de ESTIMADORES de RAZÓN

Hasta ahora hemos considerado el caso más sencillo, todos los conglomerados tienen el mismo tamaño ($M_i = M$).
Pero lo habitual es que los conglomerados tengan $\neq$ tamaño

a) Si M_i son similares,

podemos considerar $\overline{M} = \sum_{i=1}^{N} \frac{M_i}{N}$ como la media de todos los tamaños de los conglomerados y utilizar las fórmulas estudiadas hasta ahora.

En el caso de muestreo sin reposición con probab. iguales

$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} \overline{X}_i = \frac{1}{n\overline{M}} \sum_{i=1}^{n} X_i \quad , \text{ estimador insesgado de } \overline{X}$$

$$V(\bar{\bar{x}}) = \frac{(1-f)}{n\overline{M}} S_b^2 = \frac{1-f}{n\overline{M}} \sum_{i=1}^{N} \frac{(X_i - \overline{X})^2}{(N-1)\overline{M}} = \frac{1-f}{n\overline{M}^2} \sum_{i=1}^{N} \frac{(X_i - \overline{X})^2}{N-1}$$

$$\hat{V}(\bar{\bar{x}}) = \frac{(1-f)}{n\overline{M}} \hat{S}_b^2 = \ldots\ldots = \frac{1-f}{n\overline{M}^2} \sum_{i=1}^{n} \frac{(X_i - \bar{\bar{X}})^2}{n-1}$$

b) Si M_i son muy distintos,

La varianza del estimador tomando la media muestral de los tamaños puede ser muy grande si los tamaños de los conglomerados difieren mucho entre sí, porque también habrá una variabilidad alta entre los totales $X_i = \sum_{j=1}^{M_i} X_{ij}$.

En este caso, la precisión puede mejorar utilizando el estimador de razón $\hat{X}_R$, que aunque es sesgado, suele ser más acurado.

$$\hat{\bar{X}}_R = \bar{\bar{x}}_R = \hat{R} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i} \quad , \text{ estimador de } R = \frac{\sum^{N} X_i}{\sum^{N} M_i} = \bar{\bar{X}}$$

Por ser un estimador de razón, su varianza aproximada (para n sufic. grande) es:

$$(SR) \longrightarrow V(\hat{\bar{X}}) = V(\hat{R}) = V(\hat{R}-R) = V\left(\frac{\sum^{n} X_i}{\sum^{n} M_i} - R\right) =$$

$$= V\left(\frac{\hat{X}}{\hat{M}} - R\right) = V\left(\frac{\hat{X} - R\hat{M}}{\hat{M}}\right) = V\left(\frac{N\hat{X} - RN\hat{M}}{M^2 \; \hat{M}}\right)$$

Tirar

C. Pérez

$$\simeq \frac{N^2}{M^2} \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \sum_{i=1}^{N} (X_i - RM_i)^2$$

$$V(\hat{R}) = \frac{N^2(1-f)}{n M^2} \cdot \frac{1}{N-1} \sum_{i=1}^{N} M_i^2 (\bar{X}_i - \bar{\bar{X}})^2$$

$$\hat{V}(\hat{R}) = \frac{N^2}{M^2} \cdot \frac{1-f}{n} \cdot \frac{\sum^{n} M_i^2 (\bar{X}_i - \bar{\bar{x}})^2}{n-1}$$

Para el total: $\hat{X} = M\hat{R} \Rightarrow V(\hat{X}) = M^2 V(\hat{R})$

Para la proporción: $V(\hat{P}) = \frac{N^2}{M^2} \cdot \frac{1-f}{n} \cdot \frac{\sum_{i=1}^{N} M_i^2 (P_i - P)^2}{N-1}$

$$\hat{V}(\hat{P}) = \frac{N^2}{M^2} \cdot \frac{1-f}{n} \cdot \frac{\sum_{i=1}^{N} M_i^2 (P_i - \bar{P})^2}{n-1}$$

Para el total de clase: $\hat{A} = M\hat{P} \Rightarrow V(\hat{A}) = M^2 V(\hat{P})$.

$$(CR) \longrightarrow V(\hat{\bar{X}}) = V(\hat{R}) = \frac{N^2}{M^2} \cdot \frac{1}{n} \cdot \frac{1}{N} \sum_{i=1}^{N} M_i^2 (\bar{X}_i - \bar{\bar{X}})^2$$

$$\hat{V}(\hat{\bar{X}}) = \frac{N^2}{M^2} \cdot \frac{1}{n} \cdot \frac{1}{n-1} \sum^{n} M_i^2 (\bar{X}_i - \bar{\bar{x}})^2$$

Para el total: $\hat{X} = M\hat{\bar{X}} \Rightarrow V(\hat{X}) = M^2 V(\hat{\bar{X}})$

Para la proporción: $V(\hat{P}) = \frac{N^2}{M^2} \cdot \frac{1}{n} \cdot \frac{\sum_{i=1}^{N} M_i^2 (P_i - P)^2}{N}$

$$\hat{V}(\hat{P}) = \frac{N^2}{M^2} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{M_i^2 (P_i - \bar{P})^2}{n-1}$$

Para el total de clase: $\hat{A} = M\hat{P} \Rightarrow V(\hat{A}) = M^2 V(\hat{P})$.

Comentar en 1

En el caso de probab. desiguales, habría que acudir a la fórmula general de los estimadores lineales insesgados de Horvitz y Thompson ($\pi_i = P$(comp. i ∈ muestra)) y Hansen y Hurwitz ($P_i = P$(comp. i ∈ muestra)).

Los métodos más interesantes eran aquellos con probab. proporcionales a los tamaños M_i, en cuyo caso

(SR) → $\hat{\bar{X}} = \frac{\hat{X}_{HT}}{M} = \frac{M\bar{\bar{x}}}{M} = \bar{\bar{x}}$

↳ la expresión del estimador coincide con la de prob=

↳ los valores de las varianzas y sus estimadores dependerán del valor de π_{ij} en cada unid. de selección.

(CR) → $\hat{\bar{X}} = \frac{\hat{X}_{HH}}{M} = \frac{M\bar{\bar{x}}}{M} = \bar{\bar{x}}$

↳ la expresión es igual

$$V(\hat{X}_{HH}) = \frac{1}{nM}\sum_{i=1}^{N} M_i(\bar{X}_i - \bar{\bar{X}})^2$$

$$\hat{V}(\hat{X}_{HH}) = \frac{1}{n(n-1)}\sum(\bar{x}_i - \bar{\bar{x}})^2$$