PROBAB. DESIGUALES

PROBAB. e_i = nº veces que aparece u_i en la muestra.

SR $\longrightarrow e_i = \{0,1\}$, $e_i \longrightarrow B(1, \pi_i)$

$$E[e_i] = \pi_i$$
$$V[e_i] = \pi_i(1-\pi_i)$$
$$cov[e_i, e_j] = \pi_{ij} - \pi_i \pi_j$$

Propiedades:

P1 – $\sum^N \pi_i = n$

P2 – $\sum_{i \neq j}^N \pi_i = n - \pi_j$

P3 – $\sum_{i \neq j} \pi_{ij} = (n-1)\pi_j$

P4 – $\sum_{i \neq j} (\pi_{ij} - \pi_i \pi_j) = -\pi_j(1-\pi_j)$

para u_j fijo

CR $\longrightarrow e_i = \{0,1 \ldots n\}$, $e_i \longrightarrow B(n, P_i)$

$$E[e_i] = nP_i$$
$$V[e_i] = nP_i(1-P_i)$$
$$E[e_i e_j] = n(n-1)P_i P_j$$
$$cov(e_i e_j) = -nP_i P_j$$

ESTIMADOR $\theta = \sum^N y_i \longrightarrow \hat{\theta} = \sum_{i=1}^n y_i w_i$

SR $\longrightarrow \hat{\theta}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$

$$\hat{X}_{HT} = \sum^n \frac{X_i}{\pi_i} \longrightarrow \hat{\bar{X}}_{HT} = \frac{1}{N}\hat{X}_{HT}$$
$$\hat{A}_{HT} = \sum^n \frac{A_i}{\pi_i} \longrightarrow \hat{P}_{HT} = \frac{1}{N}\hat{A}_{HT}$$

CR $\longrightarrow \hat{\theta}_{HH} = \sum_{i=1}^n \frac{y_i}{nP_i}$

$$\hat{X}_{HH} = \sum^n \frac{X_i}{nP_i} \longrightarrow \hat{\bar{X}}_{HH} = \frac{1}{N}\hat{X}_{HH}$$
$$\hat{A}_{HH} = \sum^n \frac{A_i}{nP_i} \longrightarrow \hat{P}_{HH} = \frac{1}{N}\hat{A}_{HH}$$

VARIANZA $\hat{\theta}$ insesgado $\Rightarrow$ Precisión $\sim$ Acuracidad

SR $\longrightarrow V(\hat{\theta}_{HT}) = \sum_{i=1}^N \frac{y_i^2}{\pi_i}(1-\pi_i) + 2\sum_{i=1}^N \sum_{j>i}^N \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)$

CR $\longrightarrow V(\hat{\theta}_{HH}) = \frac{1}{n}\left[\sum^N \frac{y_i^2}{P_i} - \theta^2\right] \sim \frac{1}{n}\sum_{i=1}^N \left(\frac{y_i}{P_i} - \theta\right)^2 P_i$.

ESTIMADOR VARIANZA insesgados

SR $\xrightarrow{\text{Divido por } \pi} \hat{V}(\hat{\theta}_{HT}) = \sum_{i=1}^n \frac{y_i^2}{\pi_i^2}(1-\pi_i) + 2\sum_{i=1}^n \sum_{j>i}^n \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}$

Yates y Grundy: YG $\longrightarrow \hat{V}(\hat{\theta}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}$

CR $\longrightarrow \hat{V}(\hat{\theta}_{HH}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^n \left(\frac{y_i}{P_i}\right)^2 - n\hat{\theta}_{HH}^2\right] \sim \underbrace{\frac{1}{n(n-1)}\sum_{i=1}^n \left(\frac{y_i}{P_i} - \hat{\theta}_{HH}\right)^2}_{\text{divido por } (n-1)P_i}$

PROBABILIDADES OPTIMAS

SR $\longrightarrow$ $\pi_i = K M_i$ $\Longrightarrow$ $\pi_i = \frac{n}{M} \cdot M_i$ (prop. a M_i)

CR $\longrightarrow$ $P_i = K M_i$ $\Longrightarrow$ $P_i = \frac{1}{M} \cdot M_i$ (prop. a M_i)

MTS. ESPECIALES SIN REPOSIC Y PPT

Una Generalidad $\rightarrow$ $u_i \sim M_i$ bolas color i. 1ª extr. sale color i $\rightarrow$ u_i muestra
$\rightarrow$ se sacan M_i bolas

$\rightarrow$ $n=1 \rightarrow \pi_i = \frac{M_i}{M}$

$n=2 \rightarrow \pi_i = \frac{M_i}{M}\left(1 + \sum_{j \neq i}^{N} \frac{M_j}{M - M_j}\right)$

Brewer (1963) $\rightarrow$ Para $n=2$;

$\rightarrow$ $n=1 \rightarrow P(u_i \in \text{muestra}) = \pi_i = \frac{K_i}{K} = \frac{P_i(1-P_i)/(1-2P_i)}{\sum^N K_i}$ $\quad P_i = \frac{M_i}{M}$ / $P_i < \frac{1}{2}$

$n=2 \rightarrow \pi_i = 2P_i$ (sin reposic.)

$\rightarrow$ $\pi_{ij} \equiv$ proporc. a P_i y a P_j

$\rightarrow$ $\hat{\theta}_{HT}$

Durbin (1967) $\rightarrow$ Para $n=2$

$\rightarrow$ $n=1$, sin repos. $P_i = \frac{M_i}{M}$

$\rightarrow$ $n=2$, π_i prop. a $K_j = P_j\left(\frac{1}{1-2P_i} + \frac{1}{1-2P_i}\right)$ / $P_j < \frac{1}{2}$.

$\rightarrow$ $\pi_i, \pi_{ij}, \hat{\theta}_{HT}$ coinciden con Brewer

Esquema mixto SCG $\rightarrow$ $u_i \sim M_i$ bolas color i.

1ª extr. sale bola color i $\Rightarrow$ sólo se saca esa bola (quedan $M_i - 1$)
$\Rightarrow$ u_i puede elegirse $\{0, 1, \ldots \min\{n, M_i\}$ veces
$\Rightarrow$ probab. gradualmente variables

Esquema mixto — SR para bolas
— CR para unidades

$e_i \rightarrow H(M, n, P_i)$

$\hat{\theta}_{SCG} = \sum \frac{Y_i}{nP_i} = \hat{\theta}_{HH}$, pero $V(\hat{\theta}_{SCG}) \neq V(\hat{\theta}_{HH})$ y

$\hat{V}(\hat{\theta}_{SCG}) = \frac{M-n}{M} \hat{V}(\hat{\theta}_{HH}) \Rightarrow$ + preciso

MUEST_T3. MUESTREO CON PROBAB. DESIGUALES.
ESTIMADORES LINEALES.
VARIANZAS de los estimadores y sus ESTIMACIONES.
~~COMPARACIÓN entre el muestreo~~
PROBABILIDADES ÓPTIMAS de SELECCIÓN.
ESTIMADORES ESPECIALES de SELECCIÓN SIN REPOSICIÓN y PROBAB. PROPORCIONALES al tamaño

---

1. MUESTREO CON PROBAB. DESIGUALES

Muestreo probabilístico en el que la probabilidad de que una unidad poblacional, u_i, $i = 1 \dots N$, forme parte de la muestra es distinta para cada unidad (en cada extracción)

$$P(u_i \in M) \neq P(u_j \in M)$$

Muestreo idóneo para unidades compuestas, en el que la importancia de cada unidad es proporcional (PPT) al nº de unid. elementales que la componen (probab. proporcionales al tamaño) ó bien con unidades elementales con distinta importancia. $\neq w_i$

Estas probabilidades de pertenecer a la muestra dependen del tipo de muestreo, con o sin reposición.

Yo creo que el nº de muestras posibles será igual que en el caso de probab. iguales, aunque la probab. de la muestra sea $\neq$ para cada muestra.

#S	NO importa orden	SÍ importa orden
SIN repos	$C_{N,n} = \binom{N}{n}$	$V_{N,n} = \binom{N}{n} n!$
CON repos	$CR_{N,n} = \binom{N+n-1}{n}$	$VR_{N,n} = N^n$

2_ MUESTREO SIN REPOSICIÓN y PROBAB. DESIGUALES

Procedimiento de selección de la muestra:

- con probabilidades desiguales,
- se seleccionan las unidades de la muestra 1 a 1 sin reposición,
- no importa el orden de aparición.

Por lo que las muestras obtenidas mediante este procedimiento se caracterizan por:

- muestras con elementos repetidos son imposibles,
- muestras no equiprobables (pq las unid. u_i tienen $\neq$ probab. de pertenecer a la muestra).

PROBABILIDADES

Dada la población $\{u_1 \ldots u_N\}$ de la que seleccionamos sin reposición una muestra de tamaño n, para cada elemento poblacional u_i definimos:

$$e_i = \begin{cases} 1 & \text{si } u_i \in \text{Muestra, con probab. } \pi_i \\ 0 & \text{si } u_i \notin \text{Muestra, con probab. } 1-\pi_i \end{cases}$$

e_i	$P(e_i)$
1	π_i
0	$1-\pi_i$

e_i es una v.a. que mide el nº de veces que cada unidad poblacional aparece en la muestra, definida a través de su función de probabilidad.

$e_i \longrightarrow B(1, \pi_i)$ por lo que:

$E[e_i] = \pi_i = 1\cdot\pi_i + 0(1-\pi_i)$

$E[e_i^2] = \pi_i = 1^2\cdot\pi_i + 0^2\cdot(1-\pi_i) = \pi_i$

$V[e_i] = pq = \pi_i(1-\pi_i) = E[e_i^2] - E[e_i]^2 = \pi_i - \pi_i^2$

Para cada $i \neq j$ consideramos la ocurrencia conjunta a través de la v.a. auxiliar

$$e_i \cdot e_j = \begin{cases} 1 & \text{si } (u_i, u_j) \in \text{Muestra con probab } \pi_{ij} \\ 0 & \text{si } (u_i, u_j) \notin \text{Muestra con probab } 1-\pi_{ij} \end{cases}$$

$E[e_i \cdot e_j] = 1\cdot\pi_{ij} + 0(1-\pi_{ij}) = \pi_{ij}$

$cov[e_i \cdot e_j] = E[e_i \cdot e_j] - E[e_i]E[e_j] = \pi_{ij} - \pi_i \pi_j$

ATRÁS

ESTIMADOR lineal insesgado

En el caso de que la característica poblacional que se quiere estudiar sobre los elementos poblacionales $u_i, i=1 \dots N$, sea el parámetro $\Theta = \sum_{i=1}^{N} Y_i$

un buen estimador lineal es

$$\hat{\Theta} = \sum_{i=1}^{n} w_i Y_i$$

Para que este estimador sea insesgado $E[\hat{\Theta}] = \Theta$, en el caso de m.a.p.d.s.r., las ponderaciones $w_i = \frac{1}{\pi_i}$ s.r.

$$E[\hat{\Theta}] = E[\sum^{n} w_i Y_i] = E[\sum^{N} w_i Y_i e_i] = \sum^{N} w_i Y_i E[e_i] \overset{\{0,1\}}{=}$$
$$= \sum^{N} w_i Y_i \pi_i$$

$$E[\hat{\Theta}] = \Theta \iff w_i \pi_i = 1 \iff w_i = \frac{1}{\pi_i}, \forall i.$$

Luego el estimador lineal insesgado para Θ en el caso de muestreo sin reposición es el estimador de Horvitz y Thompson (1952), cuya expresión es:

$$\hat{\Theta}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$$

Particularizando para los parámetros poblacionales más comunes:

Parámetro		Estimador Horvitz y Thompson
Total poblacional	$\Theta = X = \sum_{i=1}^{N} X_i$ $\quad Y_i = X_i$	$\hat{\Theta}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i}$
Media poblacional	$\Theta = \bar{X} = \sum_{i=1}^{N} \frac{X_i}{N}$ $\quad Y_i = X_i/N$	$\hat{\bar{X}}_{HT} = \frac{1}{N}\sum_{i=1}^{n} \frac{X_i}{\pi_i} = \frac{1}{N} \cdot \hat{X}_{HT}$
Total de clase	$\Theta = A = \sum_{i=1}^{N} A_i$ $\quad Y_i = A_i$	$\hat{A}_{HT} = \sum_{i=1}^{n} \frac{A_i}{\pi_i}$
Proporción de clase	$\Theta = P = \sum_{i=1}^{N} \frac{A_i}{N}$ $\quad Y_i = A_i/N$	$\hat{P}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{A_i}{\pi_i} = \frac{1}{N} \cdot \hat{A}_{HT}$

VARIANZA del estimador de Horvitz y Thompson

La precisión de un estimador se medirá a partir de tres conceptos:

- error de muestreo, $\sigma(\hat{\theta}) = \sqrt{V(\hat{\theta})}$
- acuracidad, $ECM(\hat{\theta}) = E[\hat{\theta}-\theta]^2 = V(\hat{\theta}) + b(\hat{\theta})^2$
- sesgo, $b(\hat{\theta}) = E[\hat{\theta}] - \theta$

En el caso de que el estimador sea insesgado, $E[\hat{\theta}]=\theta$ resulta que $b(\hat{\theta})=0$ y por tanto $ECM(\hat{\theta}) = V(\hat{\theta})$, por lo que bastará con medir la varianza del estimador para estudiar su precisión y su acuracidad.

En el caso de muestreo sin reposición, la varianza del estimador de Horvitz y Thompson es:

$$V(\hat{\theta}_{HT}) = V\left(\sum_{i=1}^{n} \frac{y_i}{\pi_i}\right) = V\left(\sum_{i=1}^{N} \frac{y_i}{\pi_i} e_i\right) = \sum_{i=1}^{N} V\left(\frac{y_i}{\pi_i} e_i\right) +$$

$$+ 2\sum_{i=1}^{N}\sum_{j>i} cov\left(\frac{y_i}{\pi_i} e_i, \frac{y_j}{\pi_j} e_j\right) =$$

$$= \sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2 V(e_i) + 2\sum_{i=1}^{N}\sum_{j>i} \frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j} cov(e_i, e_j) =$$

$V(e_i) = \pi_i(1-\pi_i)$
$cov(e_i, e_j) = \pi_{ij} - \pi_i\pi_j$

$$= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2}\cdot \pi_i(1-\pi_i) + 2\sum_{i=1}^{N}\sum_{j>i} \frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j}(\pi_{ij} - \pi_i\cdot\pi_j)$$

$$V(\hat{\theta}_{HT}) = \sum_{i=1}^{N} \frac{y_i^2 \cancel{\pi_i}}{\pi_i \cancel{\pi_i}}(1-\pi_i) + 2\sum_{i=1}^{N}\sum_{j>i} \frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)$$

ESTIMACIÓN de la VARIANZA del estimador de Horvitz y Thompson

La expresión de $V(\hat{\theta}_{HT})$ depende de todas las unid. poblacionales, $i=1\ldots N$, y solamente tenemos información de las unidades muestrales, $i=1\ldots n$. Por ello, se hace necesario estimar la varianza del estimador en función de la información muestral.

Un estimador insesgado de $V(\hat{\theta}_{HT})$ es:

$$\hat{V}(\hat{\theta}_{HT}) = \sum_{i=1}^{n} \frac{y_i^2}{\pi_i^2}(1-\pi_i) + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j}\cdot\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}$$

Efectivamente, es insesgado, $E[\hat{V}(\hat{\theta}_{HT})] = V(\hat{\theta}_{HT})$.

$$E[\hat{V}(\hat{\theta}_{HT})] = E\left[\sum_{i=1}^{n} \frac{y_i^2}{\pi_i^2}(1-\pi_i) + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j}\,\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right] =$$

$$= E\left[\sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2}(1-\pi_i)\,e_i\right] + 2E\left[\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j}\,\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\,e_ie_j\right]$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2}(1-\pi_i)\underbrace{E[e_i]}_{=\pi_i} + 2\sum_{i=1}^{N}\sum_{j>i} \frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j}\cdot\frac{\pi_{ij}-\pi_i\pi_j}{\cancel{\pi_{ij}}}\cdot\underbrace{E[e_ie_j]}_{\pi_{ij}}$$

$$= V(\hat{\theta}_{HT}).$$

Este estimador, aunque es el más utilizado en la práctica, es algo inestable (puede proporcionar valores negativos). Otro estimador más estable se debe a Yates y Grundy:

$$\hat{V}(\hat{\theta}_{HT}) = \sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\cdot\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}} \quad \leftarrow Y-G$$

También es insesgado

Dem: $E[\sum^{n}(\quad)] = E[\sum^{N}(\quad)\cdot e_ie_j] = \sum^{N}(\quad)E[e_ie_j] =$

$$= \sum^{N}(\quad)\cdot\pi_{ij} =$$

$$= \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\cdot\frac{\pi_i\pi_j-\pi_{ij}}{\cancel{\pi_{ij}}}\cdot\cancel{\pi_{ij}} =$$

Deshaciendo el cuadrado:

$$= \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{y_i^2}{\pi_i^2}+\frac{y_j^2}{\pi_j^2}\right)(\pi_i\pi_j-\pi_{ij}) - 2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{y_iy_j}{\pi_i\pi_j}(\pi_i\pi_j-\pi_{ij}) =$$

$$= \sum_{i=1}^{N}\frac{y_i^2}{\pi_i^2}\sum_{j\neq i}^{N}(\pi_i\pi_j-\pi_{ij}) + 2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{y_iy_j}{\pi_i\pi_j}(\pi_{ij}-\pi_i\pi_j) =$$

Como $\sum_{j\neq i}^{N}(\pi_{ij}-\pi_{ij}) = \pi_i(1-\pi_i)$

$$= \sum_{i=1}^{N}\frac{y_i^2}{\pi_i^{\cancel{2}}}\cancel{\pi_i}(1-\pi_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{y_i}{\pi_i}\cdot\frac{y_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j) = V(\hat{\theta}_{HT})$$

3_MUESTREO CON REPOSICIÓN y PROBAB. DESIGUALES

Procedimiento de selección de la muestra
- con probab. desiguales
- las muestras con elementos repetidos son posibles,
- cada elemento poblacional puede estar 0, 1, ... n veces en la muestra.

PROBABILIDADES

Para cada unidad poblacional u_i, $i = 1 \dots N$, definimos la v.a. auxiliar $e_i \equiv$ nº veces que aparece u_i en la muestra.

$e_i \longrightarrow B(n, P_i)$, donde $P_i = P(u_i \in$ muestra en cada extrac.)

$E[e_i] = nP_i$

$V[e_i] = nP_i(1-P_i)$

La probab. de una muestra cualquiera seguirá el modelo multinomial, donde cada elemento u_i va a estar en la muestra e_i veces, $\sum_{i=1}^{N} e_i = n$.

$(e_1, e_2 \dots e_N) \longrightarrow \text{Multinomial}(n, P_1 \dots P_N)$ $\sum_{i=1}^{N} P_i = 1$

$$P(e_1 \dots e_N) = \frac{N!}{e_1! \dots e_N!} P_1^{e_1} \dots P_N^{e_N}, \quad e_i = 0, 1 \dots N \quad \sum e_i = n$$

Para la ocurrencia conjunta, definimos la v.a. auxiliar $e_i \cdot e_j =$ "nº veces que aparecen (u_i, u_j) en la muestra".

~~$e_i \cdot e_j = 0, 1 \dots$~~ ?

A partir de la f. generatriz de momentos de la distr. multinomial se deducen la esperanza y la covarianza:

$E[e_i \cdot e_j] = n(n-1) P_i P_j$

$$\text{cov}[e_i \cdot e_j] = E[e_i \cdot e_j] - E[e_i]E[e_j] = n(n-1)P_iP_j - nP_i \cdot nP_j =$$
$$= \cancel{n^2 P_i P_j} - nP_iP_j - \cancel{n^2 P_i P_j} = -nP_iP_j$$

Ya tenemos definido el vector esperanza y la matriz de var-cov: $E[e_1 \dots e_N] = (nP_1 \dots nP_N)$

$$\Sigma(e_1 \dots e_N) = \begin{pmatrix} nP_1(1-P_1) & -nP_1P_2 & \dots & -nP_1P_N \\ -nP_2P_1 & nP_2(1-P_2) & & \\ \vdots & & \ddots & \\ -nP_NP_1 & & & nP_N(1-P_N) \end{pmatrix}$$

ESTIMADOR lineal insesgado

Para el parámetro poblacional $\Theta = \sum_{i=1}^{N} Y_i$, utilizamos como estimador lineal $\hat{\Theta} = \sum_{i=1}^{N} Y_i W_i$

Para que sea insesgado, $E[\hat{\Theta}] = \Theta$

$$E[\hat{\Theta}] = E[\sum^{n} Y_i W_i] = E[\sum^{N} Y_i W_i e_i] = \sum^{N} Y_i W_i E[e_i]$$

en el caso de muestreo con reemplazamiento $E[e_i] = nP_i$, por lo que

$$E[\hat{\Theta}] = \sum_{i=1}^{N} Y_i W_i n P_i = \underbrace{\sum_{i=1}^{N} Y_i}_{=\Theta} \iff W_i = \frac{1}{nP_i}, \forall i = 1 \dots N$$

Luego el estimador lineal insesgado para Θ en el caso de muestreo con reposición es el estimador de Hansen y Hurwitz (1943):

$$\boxed{\hat{\Theta}_{HH} = \sum_{i=1}^{n} \frac{Y_i}{nP_i}}$$

cuya expresión se puede particularizar para los parámetros poblacionales más utilizados:

Parámetro		Estimador Hansen y Hurwitz, $\hat{\Theta}_{HH}$	
Total poblacional	$\Theta = \sum_{i=1}^{N} X_i = X$	$Y_i = X_i \longrightarrow$	$\hat{X}_{HH} = \sum_{i=1}^{n} \frac{X_i}{nP_i}$
Media poblacional	$\Theta = \sum_{i=1}^{N} \frac{X_i}{N} = \bar{X}$	$Y_i = \frac{X_i}{N} \longrightarrow$	$\hat{\bar{X}}_{HH} = \frac{1}{N} \sum_{i=1}^{n} \frac{X_i}{nP_i} = \frac{1}{N}\hat{X}$
Total de clase	$\Theta = \sum_{i=1}^{N} A_i = A$	$Y_i = A_i \longrightarrow$	$\hat{A}_{HH} = \sum_{i=1}^{n} \frac{A_i}{nP_i}$
Proporción poblacional	$\Theta = \sum_{i=1}^{N} \frac{A_i}{N} = P$	$Y_i = \frac{A_i}{N} \longrightarrow$	$\hat{P}_{HH} = \frac{1}{N} \sum_{i=1}^{n} \frac{A_i}{nP_i} = \frac{1}{N}\hat{A}_{HH}$

VARIANZA del estimador de Hansen y Hurwitz

$$V(\hat{\theta}_{HH}) = V\left(\sum_{i=1}^{n} \frac{y_i}{nP_i}\right) = V\left(\sum_{i=1}^{N} \frac{y_i}{nP_i} e_i\right) = \sum_{i=1}^{N} \frac{y_i^2}{n^2P_i^2} \underbrace{V(e_i)}_{=nP_i(1-P_i)} +$$

$$+ \sum_{i\neq j}^{N} \frac{y_i}{nP_i} \cdot \frac{y_j}{nP_j} \underbrace{cov(e_i \cdot e_j)}_{-nP_iP_j} =$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{n^{\cancel{2}}P_i^{\cancel{2}}} \cdot \cancel{n}\cancel{P_i}(1-P_i) + \sum_{i\neq j}^{N} \frac{y_i}{\cancel{nP_i}} \cdot \frac{y_j}{\cancel{nP_j}} (-\cancel{n}\cancel{P_i}\cancel{P_j}) =$$

$$= \sum_{i=1}^{N} \frac{y_i^2(1-P_i)}{nP_i} - \sum_{i\neq j}^{N} \frac{y_iy_j}{n} = \sum_{i=1}^{N} \frac{y_i^2}{nP_i} - \frac{1}{n}\sum_{i=1}^{N} y_i^2 - \sum_{i\neq j}^{N} \frac{y_iy_j}{n} =$$

Como $\left(\sum_{i=1}^{N} y_i\right)^2 = \sum_{i=1}^{N} y_i^2 + \sum_{i\neq j}^{N} y_iy_j$

$$= \sum_{i=1}^{N} \frac{y_i^2}{nP_i} - \frac{1}{n}\left[\sum_{i=1}^{N} y_i^2 + \sum_{i\neq j}^{N} y_iy_j\right] = \sum_{i=1}^{N} \frac{y_i^2}{nP_i} - \frac{1}{n}\Big(\underbrace{\sum_{i=1}^{N} y_i}_{=\theta}\Big)^2$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{nP_i} - \frac{1}{n}\theta^2 = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{y_i^2}{P_i} - \theta^2\right) =$$

$$V(\hat{\theta}_{HH}) = \frac{1}{n}\left[\sum_{i=1}^{N}\left(\frac{y_i}{P_i}\right)^2 \cdot P_i - \theta^2\right]$$

Podemos llegar a una formulación equivalente, ya que:

$$\sum_{i=1}^{N}\left(\frac{y_i}{P_i} - \theta\right)^2 P_i = \sum_{i=1}^{N}\left(\frac{y_i}{P_i}\right)^2 P_i + \theta^2 \underbrace{\sum_{i=1}^{N} P_i}_{=1} - 2\theta \underbrace{\sum_{i=1}^{N} \frac{y_i}{\cancel{P_i}} \cdot \cancel{P_i}}_{=\theta} =$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{P_i} - \theta^2$$

Otra expresión equivalente de la varianza del estimador de Hansen y Hurwitz es:

$$V(\hat{\theta}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{P_i} - \theta\right)^2 P_i$$

Otra manera equivalente sería:

$$V(\hat{\theta}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\frac{y_i}{P_i} - \frac{y_j}{P_j}\right)^2 P_iP_j$$

ESTIMACIÓN de la Varianza del estimador de Hansen y Hurwitz

Un estimador insesgado para $V(\hat{\theta}_{HH})$ viene dado por:

$$\hat{V}(\hat{\theta}_{HH}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 - n\hat{\theta}_{HH}^2\right]$$

Es un estimador insesgado de $V(\hat{\theta}_{HH})$, porque $E[\hat{V}(\hat{\theta}_{HH})] = V(\hat{\theta}_{HH})$

$$E[\hat{V}(\hat{\theta}_{HH})] = E\left[\sum^{n}\quad\right] = E\left[\sum^{N}\quad\right] =$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N}\left(\frac{y_i^2}{p_i^2}\right)\cdot \underbrace{E[e_i]}_{\substack{\shortparallel \\ np_i}} - nE[\hat{\theta}_{HH}^2]\right] =$$

$V(\hat{\theta}) = E[\hat{\theta}^2] - E[\hat{\theta}]^2$

$$= \frac{\cancel{n}}{\cancel{n}(n-1)}\left[\sum_{i=1}^{N}\frac{y_i^2}{p_i} - [V[\hat{\theta}_{HH}] + \underbrace{E[\hat{\theta}_{HH}]}_{\theta}{}^2]\right] =$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{N}\frac{y_i^2}{p_i} - V(\hat{\theta}_{HH}) - \theta^2\right] = \frac{1}{n-1}\left[\underbrace{\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i} - \theta^2\right)}_{n\cdot V(\hat{\theta}_{HH})} - V(\hat{\theta}_{HH})\right] =$$

$$= \frac{n-1}{n-1}V(\hat{\theta}_{HH}) = V(\hat{\theta}_{HH})$$

Una expresión equivalente es:

$$\hat{V}(\hat{\theta}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{\theta}_{HH}\right)^2$$

ya que

$$\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{\theta}_{HH}\right)^2 = \sum_{i=1}^{n}\left(\frac{y_i^2}{p_i^2} + \hat{\theta}_{HH}^2 - 2\hat{\theta}_{HH}\cdot\frac{y_i}{p_i}\right) =$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 + n\hat{\theta}_{HH}^2 - 2\hat{\theta}_{HH}\underbrace{\sum_{i=1}^{n}\frac{y_i}{p_i}}_{\substack{\shortparallel \\ n\cdot\hat{\theta}_{HH}}} =$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 + n\hat{\theta}_{HH}^2 - 2n\hat{\theta}_{HH}^2 =$$

$$= \sum^{n}\left(\frac{y_i}{p_i}\right)^2 - n\hat{\theta}_{HH}^2$$

MUEST_T3

4. PROBABILIDADES ÓPTIMAS de SELECCIÓN

En muchas ocasiones, es conveniente asignar a las unidades probabilidades de pertenecer a la muestra teniendo en cuenta el tamaño de la unidad. Es el caso de unidades de muestreo compuestas de unidades elementales.

Para cada unidad poblacional u_i, $i=1\dots N$, defino $M_i \equiv$ tamaño de la unidad $\rightarrow$ nº de unidades elementales que componen u_i. $\sum_{i=1}^{N} M_i = M$

Muestreo SIN reposición:

$$\theta = \sum_{i=1}^{N} Y_i \longrightarrow \hat{\theta}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \quad \text{tq } \pi_i = P(u_i \in \text{muestra})$$

Si fuera $\pi_i = n \cdot \frac{y_i}{\theta}$, el valor del estimador sería

$$\hat{\theta}_{HT} = \sum_{i=1}^{n} \cancel{y_i} \cdot \frac{\theta}{n\cancel{y_i}} = \frac{n\theta}{n} = \theta \Rightarrow V(\hat{\theta}_{HT}) = 0$$

En este caso ideal (θ descon.) el estimador proporcionaría siempre el valor del parámetro. Este resultado sugiere utilizar probabilidades π_i proporcionales a una variable conocida, en general de tamaño M_i, que se supone correlacionada con Y_i.

$$\pi_i = K \cdot M_i \Rightarrow \overset{(P1)}{n} = \sum_{i=1}^{N} \pi_i = K \sum_{i=1}^{N} M_i = KM \longrightarrow K = \frac{n}{M}$$

$$\pi_i = \frac{nM_i}{M}$$

En la práctica, las unidades de muestreo suelen ser conglomerados de tamaño M_i, aunque a veces este modelo se utiliza con unidades simples, donde M_i son las distintas ponderaciones que dan mayor o menor importancia a las unidades poblacionales.

Muestreo CON reposición:

$$\theta = \sum_{i=1}^{N} Y_i \longrightarrow \hat{\theta}_{HH} = \sum_{i=1}^{n} \frac{y_i}{nP_i} \quad \text{tq} \quad P_i = P(u_i \in \text{muestra})$$

En el hipotético caso de que

$P_i = \frac{Y_i}{\theta}$, el valor del estimador sería

$$\hat{\theta}_{HH} = \sum_{i=1}^{n} \frac{y_i}{n \cdot \frac{y_i}{\theta}} = \frac{n\theta}{n} = \theta \Rightarrow V(\hat{\theta}_{HH}) = 0.$$

Luego en este caso también es razonable asignar probabilidades de selección P_i proporcionales al tamaño de las unidades poblacionales M_i.

$$P_i = K \cdot M_i \Rightarrow 1 = \sum_{i=1}^{N} P_i = K \sum_{i=1}^{N} M_i = KM \Rightarrow K = \frac{1}{M}$$

$$P_i = \frac{M_i}{M}$$

SIN

5. ESTIMADORES ESPECIALES DE SELECCIÓN SIN REPOSICIÓN Y PROBAB. PROPORCIONALES AL TAMAÑO (ppt)

De una población $\{u_1 \dots u_N\}$ de unidades compuestas (o no equiprobables) con tamaños (o ponderaciones) $\{M_1 \dots M_N\}$ tq $\sum_{i=1}^{N} M_i = M$.

En el caso del muestreo SIN reposición tiene sentido que las probabilidades Π_i de pertenecer a la muestra sean proporcionales a los tamaños M_i.

- Modelo de urna generalizada
- Método de Brewer
- Método de Durbin
- Esquema mixto de Sánchez-Crespo y Gabeiras

a) Esquema de urna generalizado

Cada unidad u_i aparece representada por M_i bolas del mismo color en la urna. Se realiza la 1ª extracción, sale bola de color i $\Rightarrow u_i \in$ muestra

$\Rightarrow$ se retiran todas las bolas de color i.

$\Rightarrow$ Quedan $M - M_i$ bolas

$\hookrightarrow$ Siguiente extracción.

$$n = 1 \rightarrow \pi_i = P(u_i \in M) = \frac{M_i}{M}$$

$$n=2 \rightarrow \pi_i = P(u_i \in M) = P(u_i \in 1^a ext) + P(u_i \notin 1^a ext \cap u_i \in 2^a)$$

$$= P(u_i \in 1^a extr) + \sum_{j\neq i} P(u_j \in 1^a extr) \cdot P(u_i \in 2^a ext / u_j \in 1^a)$$

$$= \frac{M_i}{M} + \sum_{j\neq i}^{N} \frac{M_j}{M} \cdot \frac{M_i}{M - M_j} = \frac{M_i}{M}\underbrace{\left(1 + \sum_{j \neq i} \frac{M_j}{M - M_j}\right)}_{K}$$

etc.

$\pi_{ij} = P(u_i \in 1^a \cap u_j \in 2^a) + P(u_j \in 1^a \cap u_i \in 2^a)$ proporcional a M_i

b) ~~Esquema~~ Método de Brewer (1963)

Propuso para $n = 2$. La primera unidad se extrae sin reposición y con probab. proporcional a $k_i = P_i \frac{(1-P_i)}{(1-2P_i)}$,

siendo $P_i = \frac{M_i}{M} < 1/2$. (para que $k_i > 0$)

La segunda extracción también es sin reposición y con probabilidad proporcional a P_i.

Sea $K = \sum^N k_i = \sum^N P_i \frac{(1-P_i)}{1-2P_i} = \frac{1}{2}\left(\sum_{i=1}^{N} \frac{P_i}{1-2P_i} + 1\right)$

$$\pi_i = P(u_i \in \text{Muestra}) = P(u_i \in 1^a) + P(u_i \notin 1^a \cap u_i \in 2^a) =$$

$$= P(u_i \in 1^a) + \sum_{j\neq i} P(u_j \in 1^a) \cdot P(u_i \in 2^a / u_j \in 1^a) = \frac{k_i}{K} + \sum_{j\neq i}^{N} \frac{k_j}{K} \cdot \frac{P_i}{1-P_j} =$$

$$= \frac{k_i}{K} + \frac{P_i}{K}\sum_{j\neq i}^{N} \frac{\frac{P_j(1-P_j)}{(1-2P_j)}}{(1-P_j)} = \frac{P_i}{K}\left[\frac{P_i(1-P_i)}{1-2P_i} + \sum_{j\neq i}^{N} \frac{P_j}{1-2P_j}\right] =$$

$$= \frac{P_i}{K}\left[1 + \frac{P_i}{1-2P_i} + \sum_{j \neq i}^{N} \frac{P_j}{1-2P_j}\right] = \frac{P_i}{K}\underbrace{\left[1 + \sum_{i=1}^{N} \frac{P_i}{1-2P_i}\right]}_{= 2K} = \frac{P_i}{K} \cdot 2K = 2P_i$$

La probabilidad π_{ij} de que las unidades u_i y u_j pertenezcan a la muestra será:

$$\pi_{ij} = P((u_i,u_j) \in \text{Muestra}) = P(u_i \in 1^a \cap u_j \in 2^a) + P(u_j \in 1^a \cap u_i \in 2^a) =$$

$$= P(u_i \in 1^a)\cdot P(u_j \in 2^a / u_i \in 1^a) + P(u_j \in 1^a)\cdot P(u_i \in 2^a / u_j \in 1^a) =$$

$$= \frac{K_i}{K}\cdot\frac{P_j}{1-P_i} + \frac{K_j}{K}\cdot\frac{P_i}{1-P_j} = \frac{1}{K}\left[\frac{P_i \cancel{(1-P_i)}}{1-2P_i}\cdot\frac{P_j}{\cancel{1-P_i}} + \frac{P_j \cancel{(1-P_j)}}{1-2P_j}\cdot\frac{P_i}{\cancel{1-P_j}}\right] =$$

$$= \frac{1}{K}\cdot P_i P_j\left[\frac{1}{1-2P_i} + \frac{1}{1-2P_j}\right] \longrightarrow \text{proporcional a } P_i.$$

$\Rightarrow$ proporcional a π_i.

Como es muestreo SIN reemplazamiento se utiliza el estimador de Horvitz y Thompson que para $n=2$ tiene como expresión:

$$\hat{\theta}_{HT} = \frac{Y_i}{\pi_i} + \frac{Y_j}{\pi_j} = \frac{Y_i}{2P_i} + \frac{Y_j}{2P_j} = \frac{1}{2}\left(\frac{Y_i}{P_i} + \frac{Y_j}{P_j}\right)$$

$V(\hat{\theta}_{HT})$ se obtiene sustituyendo π_i y π_{ij} en la fórmula general.

Método de Durbin (1967)

Para $n=2$, la primera unidad se extrae sin reposición con probabilidad $P_i = \frac{M_i}{M}$ y la segunda con probabilidad proporcional a:

$$K_j = P_j\left(\frac{1}{1-2P_i} + \frac{1}{1-2P_j}\right) \;/\; P_j < 1/2 \quad \left(P_j = \frac{M_j}{M}\right).$$

Resulta inmediato comprobar que los valores π_i y π_{ij} coinciden con los obtenidos por el mét. de Brewer, luego tb. coinciden el estimador y su varianza.

Esquema mixto de Sánchez-Crespo y Gabeiras (1987)

Es un esquema de urna con M_i bolas del color i representando a la unidad u_i.

En la 1ª extracción sale bola de color i ~~⇒ se retira solo~~ se retira solamente esa bola (quedan $M_i - 1$ bolas color i)

⇒ las unidades tienen probabilidades gradualmente variables

⇒ u_i puede ser elegida $0, 1 \ldots \min\{n, M_i\}$ veces

Es un método mixto porque < sin reposición (bola no se repone) / con reposición (u_i puede)

Consideramos la v.a. auxiliar $e_i \equiv$ nº veces que aparece u_i en la muestra,

$$e_i \longrightarrow H(M, n, \underline{P_i}) \begin{cases} E[e_i] = nP_i \\ V[e_i] = \frac{M-n}{M-1} n P_i (1-P_i) \\ cov(e_i, e_j) = -\frac{M-n}{M-1} n P_i P_j . \end{cases}$$

$$P(e_1 \ldots e_N) = \frac{\binom{M_1}{e_1} \cdots \binom{M_N}{e_N}}{\binom{M}{n}} \Big/ \sum e_i = n$$

El estimador insesgado de θ coincide con el de Hansen y Hurwitz, pero su varianza no: ($P(u_i \in$ Muestra$) =$ en cada extracc.)

⊛ $$\hat{\theta}_{SCG} = \sum \frac{y_i}{nP_i} = \hat{\theta}_{HH}$$

$$V(\hat{\theta}_{SCG}) = \frac{M-n}{M-1} V(\hat{\theta}_{HH}) \quad , \quad \hat{V}(\hat{\theta}_{SCG}) = \frac{M-n}{M} \hat{V}(\hat{\theta}_{HH})$$

⇒ $\hat{\theta}_{SCG}$ es más preciso que $\hat{\theta}_{HH}$.

Existe una versión mejorada de Gabeiras, en la que se retira de la urna b bolas en cada extracción, $b = \frac{\min\{M_i\}}{n-1}$ de modo que todas las unidades siguen estando representadas en la siguiente extracción

---

$$P(u_i \in 1^a) = \frac{M_i}{M} = P_i$$

$$P(u_i \in 2^a) = P(u_i \in 1^a \cap u_i \in 2^a) + P(u_i \notin 1^a \cap u_i \in 2^a) = \sum_{j \neq i}^{N} P(u_j \in 1^a) \cdot P(u_i \in 1^a / u_j \in 2^a) =$$

$$= \frac{M_i}{M} \cdot \frac{M_i - 1}{M-1} + \sum_{j \neq i} \frac{M_j}{M} \cdot \frac{M_i}{M-1} = \frac{M_i}{M} \left[\frac{M_i - 1}{M-1} + \sum_{j \neq i} \frac{M_j}{M-1} \right] =$$ ATRÁS →