Nociones básicas sobre muestreo de poblaciones finitas

Daniel Czarnievicz

2017

Muestreo sin reposición

Sea la poblacón $U = \{1; \dots ; k; \dots ; N\}$ donde N es conocido, pero los valores de y_k son desconocidos. Se busca estimar el total de la variable y, $t_y = \sum_U y_k$, o de la media poblacional $\bar{y}_U = \dfrac{t_y}{N} = \dfrac{1}{N} \sum_U y_k$. Se selecciona una *muestra* de la población, la cual se utiliza para estimar el total y la media.

Llamamos *diseño muestral* a la función $p(.)$ tal que:

$$p(s) = \Pr(S = s) = \Pr(\text{“seleccionar la muestra } s\text{”} | \text{“estrategia de selección”})$$

$p(s)$ es entonces la función de distribución de probabilidad de una variable aleatoria S con recorrido $\mathscr{S} = \{s_1; s_2; \dots\}$. $\mathscr{S}$ tiene 2^N elementos, contando $\emptyset$ y U. Dado que es una función de probabilidad, $p(s) \geq 0 \, \forall s \in \mathscr{S}$, y $\displaystyle\sum_{s \in \mathscr{S}} p(s) = 1$.

La inclusión de un elemento k en la muestra puede indicarse mediante la indicadora:

$$I_k = \begin{cases} 1 & \text{si} \quad k \in s \\ 0 & \text{si} \quad k \notin s \end{cases}$$

Definimos la *probabilidad de inclusión de primer orden* como:

$$\star \ \pi_k = \Pr(k \in s) = \Pr(I_k = 1) = \sum_{s \ni k} p(s)$$

Existen N cantidades $\pi_1; \dots ; \pi_k; \dots ; \pi_N$ asociadas a un diseño $p(.)$. Si $\pi_k \geq 0 \, \forall k \in U$ decimos que el muestreo es un *muestreo probabilístico*.

Definimos la *probabilidad de inclusión de segundo orden* como:

$$\star \ \pi_{kl} = \Pr(k; l \in s) = \Pr(I_k I_l = 1) = \sum_{\substack{s \ni k \\ s \ni l}} p(s)$$

Existen $\dfrac{N(N-1)}{2}$ cantidades $\pi_{12}; \pi_{13}; \dots ; \pi_{kl}; \dots ; \pi_{N-1,N}$ asociadas a un diseño $p(.)$, donde $\pi_{kl} = \pi_{lk}$ y $\pi_{kk} = \pi_k$. Si $\pi_{kl} \geq 0 \, \forall k, l \in U$ decimos que el diseño es *medible*. Solo en el caso de diseños medibles es posible obtener estimadores insesgados de la varianza.

Para todo diseño se cumple que:

- $\mathbf{E}_{p(s)}(I_k) = \pi_k \quad \forall k \in U$
- $\mathbf{Var}_{p(s)}(I_k) = \pi_k(1 - \pi_k) \quad \forall k \in U$
- $\Delta_{kl} = \mathbf{Cov}_{p(s)}(I_k; I_l) = \pi_{kl} - \pi_k \, \pi_l \quad \forall k \neq l \in U$

Llamamos n_S al tamaño muestral (es decir, al cardinal del conjunto s). Para todo diseño se cumple que:

- $n_S = \sum_U I_k$
- $\mathbf{E}_{p(s)}(n_S) = \sum_U \pi_k$
- $\mathbf{Var}_{p(s)}(n_S) = \sum_U \pi_k(1-\pi_k) + \sum\sum_{\substack{U \\ k\neq l}} (\pi_{kl} - \pi_k\,\pi_l) = \sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum\sum_{\substack{U \\ k\neq l}} \pi_{kl}$
- Si $p(s)$ es de tamaño fijo n, entonces:

$$\star\ E_{p(s)}(n_S) = \sum_U \pi_k = n \qquad \star\ \sum\sum_{\substack{U \\ k\neq l}} \pi_{kl} = n(n-1)$$

$$\star\ \sum_{\substack{l\in U \\ l\neq k}} \pi_{kl} = \sum_{\substack{l\in U \\ l\neq k}} \mathbf{E}_{p(s)}(I_k I_l) = \mathbf{E}_{p(s)}\left[I_k\left(\sum_U I_l - I_k\right)\right] = \mathbf{E}_{p(s)}\left(I_k \underbrace{\sum_U I_l}_{=n}\right) - \mathbf{E}_{p(s)}(I_l^2) =$$

$$= \mathbf{E}_{p(s)}(I_k n) - \mathbf{E}_{p(s)}(I_k) = n\,\mathbf{E}_{p(s)}(I_k) - \mathbf{E}_{p(s)}(I_k) = n\,\pi_k - \pi_k = (n-1)\pi_k$$

El estimador $\hat{t}_\pi$

El principio de π-expansión implica que el elemento muestral k representa $1/\pi_k$ elementos en la población. Sea el siguiente estimador de t_y:

$$\star\ \hat{t}_\pi = \sum_s y_k^{\checkmark} = \sum_s \frac{y_k}{\pi_k}$$

$$\star\ \mathbf{E}_{p(s)}(\hat{t}_\pi) = \mathbf{E}_{p(s)}\left(\sum_s \frac{y_k}{\pi_k}\right) = \sum_U \mathbf{E}_{p(s)}(I_k)\,\frac{y_k}{\pi_k} = \sum_U \pi_k\,\frac{y_k}{\pi_k} = \sum_U y_k = t_y$$

$$\star\ \mathbf{Var}_{p(s)}(\hat{t}_\pi) = \mathbf{Var}_{p(s)}\left(\sum_s y_k^{\checkmark}\right) = \sum_U \mathbf{Var}_{p(s)}(I_k)\,{y_k^{\checkmark}}^2 + \sum\sum_{\substack{U \\ k\neq l}} \mathbf{Cov}_{p(s)}(I_k; I_l)\,y_k^{\checkmark}\,y_l^{\checkmark} =$$

$$= \sum_U \Delta_{kk}\,y_k^{\checkmark}\,y_k^{\checkmark} + \sum\sum_{\substack{U \\ k\neq l}} \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} = \sum\sum_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark}$$

$$\star\ \hat{\mathbf{Var}}_{p(s)}(\hat{t}_\pi) = \sum\sum_s \Delta_{kl}^{\checkmark}\,y_k^{\checkmark}\,y_l^{\checkmark}$$

$$\star\ \mathbf{E}_{p(s)}\left(\hat{\mathbf{Var}}_{p(s)}(\hat{t}_\pi)\right) = \mathbf{E}_{p(s)}\left(\sum\sum_s \Delta_{kl}^{\checkmark}\,y_k^{\checkmark}\,y_l^{\checkmark}\right) = \sum\sum_U \mathbf{E}_{p(s)}(I_k; I_l)\Delta_{kl}^{\checkmark}\,y_k^{\checkmark}\,y_l^{\checkmark} =$$

$$= \sum\sum_U \pi_{kl}\,\frac{\Delta_{kl}}{\pi_{kl}}\,y_k^{\checkmark}\,y_l^{\checkmark} = \sum\sum_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} = \mathbf{Var}_{p(s)}(\hat{t}_\pi)$$

Si el diseño es de tamaño fijo, son válidas las siguientes expresiones:

$$\star\ \mathbf{Var}_{p(s)}(\hat{t}_\pi) = -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(y_k^{\checkmark} - y_l^{\checkmark}\right)^2$$

$$\star\ \hat{\mathbf{Var}}_{p(s)}(\hat{t}_\pi) = -\frac{1}{2}\sum\sum_s \Delta_{kl}^{\checkmark}\left(y_k^{\checkmark} - y_l^{\checkmark}\right)^2$$

$$\star\ \mathbf{E}_{p(s)}\left(\hat{\mathbf{Var}}_{p(s)}(\hat{t}_\pi)\right) = \mathbf{E}_{p(s)}\left(-\frac{1}{2}\sum\sum_s \Delta_{kl}^{\checkmark}\left(y_k^{\checkmark} - y_l^{\checkmark}\right)^2\right) =$$

$$= -\frac{1}{2}\sum\sum_U \mathbf{E}_{p(s)}\,(I_k; I_l)\,\Delta_{kl}^{\checkmark}\left(y_k^{\checkmark} - y_l^{\checkmark}\right)^2 = -\frac{1}{2}\sum\sum_U \mathbf{E}_{p(s)}(I_k; I_l)\Delta_{kl}^{\checkmark}\left(y_k^{\checkmark} - y_l^{\checkmark}\right)^2 =$$

$$= -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(y_k^{\checkmark} - y_l^{\checkmark}\right)^2 = \mathbf{Var}_{p(s)}(\hat{t}_\pi)$$

Demostración:

$$
\begin{aligned}
\mathbf{Var}_{p(s)}(\hat{t}_\pi) &= -\frac{1}{2}\sum\sum\nolimits_U \Delta_{kl}\big(y_k^{\checkmark} - y_l^{\checkmark}\big)^2 = \\
&= -\frac{1}{2}\sum\sum\nolimits_U \Delta_{kl}\big(y_k^{\checkmark 2} - 2\,y_k^{\checkmark}\,y_l^{\checkmark} + y_l^{\checkmark 2}\big) = \\
&= -\frac{1}{2}\sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark 2} + \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} - \frac{1}{2}\sum\sum\nolimits_U \Delta_{kl}\,y_l^{\checkmark 2} = \\
&= \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} - \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark 2} = \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} - \sum_{k\in U} y_k^{\checkmark 2}\sum_{l\in U}\Delta_{kl} = \\
&= \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} - \sum_{k\in U} y_k^{\checkmark 2}\sum_{l\in U}\big(\pi_{kl} - \pi_k\,\pi_l\big) = \\
&= \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} - \sum_{k\in U} y_k^{\checkmark 2}\left[\sum_{l\in U}\pi_{kl} - \sum_{l\in U}\pi_k\,\pi_l\right] = \\
&= \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} - \sum_{k\in U} y_k^{\checkmark 2}\left[n\,\pi_k - \pi_k\sum_{l\in U}\pi_l\right] = \\
&= \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark} - \sum_{k\in U} y_k^{\checkmark 2}\big(n\,\pi_k - \pi_k\,n\big) = \sum\sum\nolimits_U \Delta_{kl}\,y_k^{\checkmark}\,y_l^{\checkmark}
\end{aligned}
$$

Muestreo con reposición

Llamamos muestreo (no diseño) con reposición a los esquemas en los que los elementos son repuestos en la población luego de ser seleccionados. Por tanto, dos (o más) extracciones podrían producir el mismo elemento. Llamamos k_i al elemento seleccionado en la i-ésima extracción con $i = 1; \dots ; m$. Al vector que contiene todos los elementos seleccionados lo llamamos *muestra ordenada*: $os = \{k_1; \dots ; k_m\}$. Llamamos *multiplicidad* a la cantidad de veces que un mismo elemento fue seleccionado. Toda muestra ordenada, os, induce una muestra, s, la cual contiene a los elementos sorteados una única vez (se pierde el orden).

$$s = \{k : k = k_i \text{ para alguna extracción } i = 1; \dots ; m\}$$

Sean $p_1; \dots ; p_k; \dots ; p_N$ números positivos tales que $\sum_U p_k = 1$. Dado que los elementos se reponen en la población una vez seleccionado:

$$p_k = \Pr(\text{“seleccionar el elemento } k \text{ en la } i\text{-ésima extracción”}) \;\; \forall k \in U$$

De esta forma, la probabilidad de obtener una determinada muestra ordenada será:

$$p(os) = \Pr\big(os = \{k_1; \dots ; k_m\}\big) = p_{k_1} \times \dots \times p_{k_m} = \prod_{i=1}^{m} p_{k_i}$$

Sea la variable aleatoria r_k, la cual mide la cantidad de veces que el elemento k es extraído en las m extracciones. Por lo tanto, $r_k \sim Bin(m; p_k)$. Si N es lo suficientemente grande, $r_k \overset{a}{\sim} \text{Poisson}(m\,p_k)$.

$$\Pr(\text{“extraer } r_0 \text{ veces el elemento } k\text{”}) = \Pr(r_k = r_0) = \binom{m}{r}(p_k)^r(1-p_k)^{m-r}$$

$$\star\ \mathbf{E}(r_k) = m\,p_k \qquad \star\ \mathbf{Var}(r_k) = m\,p_k(1-p_k) \doteq m\,p_k$$

La probabilidad de que el elemento k nunca sea seleccionado en las m extracciones, y la probabilidad de que el elemento k sea seleccionado al menos una vez en las m extracciones son:[1]

$$\Pr(\text{“no selecionar } k \text{ en ninguna de las } m \text{ extraciciones”}) = \Pr(r_k = 0) = (1 - p_k)^m$$

$$\Pr(\text{“el elemento } k \text{ sea extraido”}) = \Pr(r_k \geq 1) = 1 - \Pr(r_k < 1) = 1 - \Pr(r_k = 0) = 1 - (1 - p_k)^m$$

Por lo tanto, las probabilidades de inclusión de primer y segundo orden serán:

$$\star\ \pi_k = \Pr(k \in S) = 1 - (1 - p_k)^m$$

$$\star\ \pi_{kl} = \Pr(k; l \in S) = \Pr(k \in S)\ \Pr(l \in S) = \big[1 - (1 - p_k)^m\big]\big[1 - (1 - p_l)^m\big] =$$

$$\begin{aligned} &= 1 - (1 - p_l)^m - (1 - p_k)^m + (1 - p_k)^m (1 - p_l)^m = \\ &= 1 - (1 - p_k)^m - (1 - p_l)^m + \big[(1 - p_k)(1 - p_l)\big]^m = \\ &= 1 - \big[(1 - p_k)^m + (1 - p_l)^m - (1 - p_k - p_l + p_k\, p_l)^m\big]\ \ \forall k \neq l \in U \end{aligned}$$

Si $p_k = p_l\ \ \forall k; l \in U$

$$\star\ \pi_{kl} = 1 - \big[2(1 - p_k)^m - (1 - 2p_k + p_k^2)^m\big] =$$

$$= 1 - \big[2(1 - p_k)^m - (1 - p_k)^{2m}\big]\ \ \forall k \neq l \in U$$

El estimador $\hat{t}_{pwr}$

En muestreos con reposición t_y se estima utilizando un estimador p-expandido (en lugar del estimador π-expandido):

$$\star\ \hat{t}_{pwr} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{p_k}$$

Para hallar las propiedades estadísticas del estimador $\hat{t}_{pwr}$ se definen las variables $Z_i = \dfrac{y_{k_i}}{p_{k_i}}$. De esta forma, $\hat{t}_{pwr} = \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} Z_i = \bar{Z}$. Dado que las Z_i son iid:

$$\star\ \Pr\left(Z_k = \frac{y_k}{p_k}\right) = p_k\ \ \forall i = 1; \ldots; m$$

$$\star\ \mathbf{E}(Z_i) = \sum\nolimits_U Z_i\ \Pr\left(Z_k = \frac{y_k}{p_k}\right) = \sum\nolimits_U \frac{y_k}{p_k}\ \Pr\left(Z_k = \frac{y_k}{p_k}\right) = \sum\nolimits_U \frac{y_k}{p_k}\, p_k = \sum\nolimits_U y_k = t_y$$

$$\star\ \mathbf{Var}(Z_i) = \mathbf{E}\left[(Z_i - t_y)^2\right] = \mathbf{E}(Z_i^2) - t_y^2 = \sum\nolimits_U \frac{y_k^2}{p_k} - t_y^2 = \sum\nolimits_U \frac{y_k^2}{p_k} - 2\, t_y^2 + t_y^2 =$$

$$= \sum\nolimits_U \frac{y_k^2}{p_k} - 2\, t_y \sum\nolimits_U y_k + t_y^2 \sum\nolimits_U p_k = \sum\nolimits_U \left[\frac{y_k^2}{p_k} - 2\, t_y\, y_k + t_y^2\, p_k\right] =$$

$$= \sum\nolimits_U \left[\left(\frac{y_k}{p_k}\right)^2 - 2\, t_y \left(\frac{y_k}{p_k}\right) + t_y^2\right] p_k = \sum\nolimits_U \left[\left(\frac{y_k}{p_k} - t_y\right)^2 p_k\right] = V_i$$

$$\star\ \mathbf{E}_{p(os)}\big(\hat{t}_{pwr}\big) = \mathbf{E}\left(\bar{Z}\right) = \mathbf{E}\left(\frac{1}{m} \sum_{i=1}^{m} z_i\right) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{E}(Z_i) = \frac{1}{m}(m,\, t_y) = t_y$$

[1] Si $m = 1$, entonces $\pi_k = p_k$.

$$\star\ \mathbf{Var}_{p(os)}\left(\hat{t}_{pwr}\right) = \mathbf{Var}\left(\bar{Z}\right) = \frac{1}{m}\mathbf{Var}(Z_i) = \frac{1}{m}\sum\nolimits_U \left[\left(\frac{y_k}{p_k} - t_y\right)^2 p_k\right] = \frac{V_i}{m}$$

$$\star\ \hat{V}_i = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_k}{p_k} - \hat{t}_{pwr}\right)^2$$

$$\star\ \hat{\mathbf{Var}}_{p(os)}\left(\hat{t}_{pwr}\right) = \frac{\hat{V}_i}{m} = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{y_k}{p_k} - \hat{t}_{pwr}\right)^2$$

$$\star\ \mathbf{E}_{p(os)}\left(\hat{\mathbf{Var}}_{p(os)}\left(\hat{t}_{pwr}\right)\right) = \mathbf{E}_{p(os)}\left(\hat{\mathbf{Var}}\left(\bar{Z}\right)\right) = \frac{1}{m}\,\mathbf{E}_{p(os)}\left(\hat{\mathbf{Var}}(Z_i)\right) = \frac{1}{m}\,\mathbf{Var}(Z_i) = \frac{V_i}{m} \Leftrightarrow$$

$$\Leftrightarrow \hat{V}_i = \frac{1}{m-1}\sum_{i=1}^{m}\left(Z_i - \bar{Z}\right)^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_k}{p_k} - \hat{t}_{pwr}\right)^2$$

$$\therefore \hat{\mathbf{Var}}_{p(os)}\left(\hat{t}_{pwr}\right) \text{ es insesgada para } \mathbf{Var}_{p(os)}\left(\hat{t}_{pwr}\right)$$

$$\text{Si } y_k = c\,p_k\ \ \forall k \in U \Rightarrow t_y = \sum\nolimits_U y_k = \sum\nolimits_U c\,p_k = c\underbrace{\sum\nolimits_U p_k}_{=1} = c \Rightarrow$$

$$\Rightarrow \mathbf{Var}_{p(os)}\left(\hat{t}_{pwr}\right) = \frac{1}{m}\sum\nolimits_U \left(\frac{y_k}{p_k} - t_y\right)^2 p_k = \frac{1}{m}\sum\nolimits_U \left(\frac{cp_k}{p_k} - c\right)^2 p_k = \frac{1}{m}\sum\nolimits_U (c-c)^2\,p_k = 0$$

En la práctica no es posible establecer $y_k = c\,p_k$. Pero sí es posible establecer $p_k = \dfrac{x_k}{\sum_U x_k}$ $\forall k \in U$, donde x_k es una variable auxiliar (es decir, x_k y y_k están altamente correlacionadas).

El estimador $\hat{t}_\pi$ en el diseño ordenado

$$\star\ \hat{t}_\pi = \sum\nolimits_s \check{y}_k = \sum\nolimits_s \frac{y_k}{\pi_k} = \sum\nolimits_s \frac{y_k}{1-(1-p_k)^m}$$

El cual es insesgado para t_y, y la expresión para su varianza y el estimador de su varianza son:

$$\star\ \mathbf{Var}_{p(s)}\left(\hat{t}_\pi\right) = \sum\sum\nolimits_U \Delta_{kl}\,\check{y}_k\ \check{y}_l$$

$$\star\ \hat{\mathbf{Var}}_{p(s)}\left(\hat{t}_\pi\right) = \sum\sum\nolimits_s \check{\Delta}_{kl}\,\check{y}_k\ \check{y}_l$$

No se pueden comparar los estimadores $\hat{t}_{pwr}$ y $\hat{t}_\pi$. Ambos son insesgados, pero no se pude concluir sobre sus varianzas. La varianza del $\hat{t}_{pwr}$ depende de los valores de y_k, lo cuales son desconocidos.

El estimador $\hat{t}_{alt}$

$$\star\ \hat{t}_{alt} = N\,\bar{y}_S = \frac{N}{n_S}\sum\nolimits_s y_k$$

Este estimador tiene como inconveniente que el tamaño muestral n_S es aleatorio. En general, $\hat{t}_{alt}$ tiene menor varianza que $\hat{t}_{pwr}$ o $\hat{t}_\pi$

Descomposición de la varianza

Supongamos que la población $ U = { 1; . . . ; k; . . . ; N }$ se encuentra particionada en a grupos de tamaño n, $\{S_1;\ldots;S_r,\ldots;S_a\}$. Luego entonces:

$$\sum\nolimits_U \left(y_k-\bar{y}_U\right)^2=\sum_{r=1}^{a}\sum\nolimits_{S_r}\left(y_k-\bar{y}_U\right)^2=\sum_{r=1}^{a}\left[\sum\nolimits_{S_r}\Big(\left(y_k-\bar{y}_{S_r}\right)+\left(\bar{y}_{S_r}-\bar{y}_U\right)\Big)^2\right]=$$

$$=\sum_{r=1}^{a}\left[\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)^2+2\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)\left(\bar{y}_{S_r}-\bar{y}_U\right)+\sum\nolimits_{S_r}\left(\bar{y}_{S_r}-\bar{y}_U\right)^2\right]=$$

$$=\sum_{r=1}^{a}\left[\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)^2+2\left(\bar{y}_{S_r}-\bar{y}_U\right)\underbrace{\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)}_{=0}+\sum\nolimits_{S_r}\left(\bar{y}_{S_r}-\bar{y}_U\right)^2\right]=$$

$$=\sum_{r=1}^{a}\left[\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)^2+\sum\nolimits_{S_r}\left(\bar{y}_{S_r}-\bar{y}_U\right)^2\right]=$$

$$=\sum_{r=1}^{a}\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)^2+\sum_{r=1}^{a}\sum\nolimits_{S_r}\left(\bar{y}_{S_r}-\bar{y}_U\right)^2=$$

$$=\sum_{r=1}^{a}\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)^2+\sum_{r=1}^{a}n\left(\bar{y}_{S_r}-\bar{y}_U\right)^2$$

Por lo tanto, tenemos que:

$$\underbrace{\sum\nolimits_U \left(y_k-\bar{y}_U\right)^2}_{SST}=\underbrace{\sum_{r=1}^{a}\sum\nolimits_{S_r}\left(y_k-\bar{y}_{S_r}\right)^2}_{SSW}+\underbrace{\sum_{r=1}^{a}n\left(\bar{y}_{S_r}-\bar{y}_U\right)^2}_{SSB}\Rightarrow \boxed{SST=SSW+SSB}$$

Tamaño muestral

Supongamos que se quiere estimar t_y usando el estimador $\hat{t}_y$ y el estimador de su varianza $\hat{V}_{p(s)}(\hat{t}_y)$. Supongamos que ambos estimadores son (aprox) insesgados y que es razonable suponer que:

$$\frac{\hat{t}_y-t_y}{\sqrt{V_{p(s)}(\hat{t}_y)}}\overset{a}{\sim}\mathrm{N}(0;1)$$

Buscamos un $\mathbf{E}(n_S)$ tal que para una precisión dada, $\varepsilon>0$, y un nivel de confianza dado, $0<\alpha<1$, nos permita plantear:

$$\Pr\left(|\hat{t}_y-t_y|<\varepsilon\right)\doteq 1-\alpha\Rightarrow \Pr\left(\hat{t}_y-z_{1-\alpha/2}\sqrt{\hat{V}_{p(s)}(\hat{t}_y)}<t_y<\hat{t}_y+z_{1-\alpha/2}\sqrt{\hat{V}_{p(s)}(\hat{t}_y)}\right)\doteq 1-\alpha$$

$$\Pr\left(|\hat{t}_y-t_y|<\varepsilon\right)\doteq 1-\alpha\Rightarrow \Pr\left(\frac{|\hat{t}_y-t_y|}{\sqrt{V_{p(s)}(\hat{t}_y)}}<\frac{\varepsilon}{\sqrt{V_{p(s)}(\hat{t}_y)}}\right)\doteq 1-\alpha$$

Si $\dfrac{\varepsilon}{\sqrt{\mathbf{Var}_{p(s)}(\hat{t}_y)}}\doteq z_{1-\alpha/2}\Rightarrow \varepsilon^2\doteq z^2_{1-\alpha/2}\mathbf{Var}_{p(s)}(\hat{t}_y)$, donde ε y α están fijos y, en general, $\mathbf{Var}_{p(s)}(\hat{t}_y)$ depende de n y de $S^2_{y_U}$. Luego si se cuenta con una buena estimación de $S^2_{y_U}$, se puede despejar n.

Para un tamaño de muestra fijo, si disminuimos ε, reducimos la amplitud del intervalo, con lo que reducimos la confianza, o sea, aumentamos α.

En la práctica $S^2_{y_U}$ es desconocida, pero se pueden ensayar alguna de las siguientes estrategias para obtener una aproximación:

1. Se presume algún tipo de distribución para los valores de y en la población. Luego se busca una cota para $S^2_{y_U}$.
   - Si $y \sim \text{Ber} \Rightarrow 0 \leq S^2_{y_U} \leq {}^1/_4$
   - Si $y \overset{a}{\sim} \text{N} \Rightarrow S^2_{y_U} \doteq \dfrac{y_{k_{(n)}} - y_{k_{(1)}}}{6}$ para un $\alpha \doteq 0{,}01$
2. Utilizar datos de algún relevamiento reciente o de alguna variable auxiliar para la que se pueda asumir una variabilidad similar. En estos casos se suele utilizar el CV_{y_u} dado que es más estable que la varianza. Luego se fija una precisión relativa y se determina el valor de n

$$\Pr\left(|\hat{t}_y - t_y| < t_y\,\varepsilon\right) \doteq 1 - \alpha \Rightarrow \frac{t_y\,\varepsilon}{\sqrt{\mathbf{Var}_{p(s)}(\hat{t}_y)}} = z_{1-\,^{\alpha}/_2} \Rightarrow \varepsilon = z_{1-\,^{\alpha}/_2}\left[\frac{\mathbf{Var}_{p(s)}(\hat{t}_y)}{t_y}\right]$$

3. Tomar una pequeña "muestra de iluminación" y calcular $S^2_{y_s}$ en dicha muestra. Luego se utilizar esta estimación como estimador de $S^2_{y_U}$ para calcular n.

Desarrollos de Taylor para aproximaciones en muestreo de poblaciones finitas

Supongamos que queremos estimar: $\theta = f\big(t_1;\ldots;t_q\big) = f\big(\mathbf{t}\big)$ donde $t_j = \sum_U y_{jk}\ j = 1;\ldots;q$, son los totales de las q variables poblacionales relevadas. Un estimador podría ser $\hat{\theta} = f\big(\hat{t}_{1\,\pi};\ldots;\hat{t}_{q\,\pi}\big) = f\big(\hat{\mathbf{t}}_\pi\big)$ donde $\hat{t}_j = \sum_s y^{\checkmark}_{jk}\ j = 1;\ldots;q$, son los totales de las q variables poblacionales relevadas, estimados en la muestra s.

Si $f\big(\hat{\mathbf{t}}_{j\,\pi}\big)$ es lineal tenemos que $\theta = a_0 + \sum_{j=1}^{q} a_j\,t_j = a_0 + \mathbf{a}'\mathbf{t}$. Luego entonces $\hat{\theta} = a_0 + \sum_{j=1}^{q} a_j\,\hat{t}_{\pi\,j} = a_0 + \mathbf{a}'\hat{\mathbf{t}}_\pi$ estima θ de forma tal que:

- $\mathbf{E}\big(\hat{\theta}\big) = \mathbf{E}\big(a_0 + \mathbf{a}'\hat{\mathbf{t}}_\pi\big) = \mathbf{E}(a_0) + \mathbf{E}(\mathbf{a}'\hat{\mathbf{t}}_\pi) = a_0 + \mathbf{a}'\mathbf{t} \Rightarrow \hat{\theta}$ es insesgado para θ.
- $\mathbf{Var}(\hat{\theta}) = \mathbf{Var}\big(a_0 + \mathbf{a}'\hat{\mathbf{t}}_\pi\big) = \sum_{j=1}^{q}\sum_{j=1}^{q} a_{jj}\,\mathbf{Cov}(\hat{t}_{j\,\pi};\hat{t}_{j'\,\pi}) = \mathbf{a}'\,V(\hat{\mathbf{t}}_\pi)\,\mathbf{a}$, donde

$$\mathbf{Cov}(\hat{t}_{j\,\pi};\hat{t}_{j'\,\pi}) = \sum\sum\nolimits_U \Delta_{kl}\,y^{\checkmark}_{jk}\,y^{\checkmark}_{j'k}$$

Podemos reescribir $\hat{\theta}$ de la siguiente forma:

$$\star\ \hat{\theta} = a_0 + \sum_{j=1}^{q} a_j\,\hat{t}_{j\,\pi} = a_0 + \mathbf{a}'\hat{\mathbf{t}}_\pi = a_0 + \sum_{j=1}^{q} a_j \sum\nolimits_s y^{\checkmark}_{jk} = a_0 + \sum_{j=1}^{q}\sum\nolimits_s a_j\,y^{\checkmark}_{jk} = a_0 + \sum\nolimits_s u^{\checkmark}_k$$

$$\text{con } u_k = \sum_{j=1}^{q} a_j\,y_{jk} \text{ y } u^{\checkmark}_k = \frac{u_k}{\pi_k}$$

$$\star\ \mathbf{Var}(\hat{\theta}) = \sum\sum\nolimits_U \Delta_{kl}\,u^{\checkmark}_k\,u^{\checkmark}_l$$

$$\star\ \hat{\mathbf{Var}}(\hat{\theta}) = \sum\sum_s \check{\Delta}_{kl}\, \check{u}_k\, \check{u}_l$$

Si $f(\hat{\mathbf{t}}_\pi)$ no es lineal, $\hat{\theta}$ debe aproximarse linealmente, y luego se podrán calcular $\mathbf{Var}(\hat{\theta})$ y $\hat{\mathbf{Var}}(\hat{\theta})$. La técnica aproxima $\hat{\theta}$ por un pseudo-estimador, $\hat{\theta}_0$, que es lineal en $\hat{\mathbf{t}}_\pi$. En general $\hat{\theta}_0$ dependerá de cantidades desconocidas (de ahí que se le llama pseudo-estimador). La técnica para hallar $\hat{\theta}_0$ consiste en la aproximación de Taylor de primer orden de la función f, en el entorno de un punto $\mathbf{t}$, y despreciar el término de error.

$$\hat{\theta} \doteq \hat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j\left(\hat{t}_{j\,\pi} - t_j\right) \quad \text{donde} \quad a_j = \left.\frac{\partial f}{\partial \hat{t}_{j\,\pi}}\right|_{\hat{\mathbf{t}}_\pi = \mathbf{t}}$$

En muestras grandes $\hat{\mathbf{t}}_\pi \approx \mathbf{t} \Rightarrow \hat{\theta}_0 = \hat{\theta}$ y $\mathbf{AVar}(\hat{\theta}) = \mathbf{Var}(\hat{\theta}_0)$.

$$\star\ \mathbf{AVar}(\hat{\theta}) \doteq V(\hat{\theta}_0) = V\left(\sum_{j=1}^{q} a_j\, \hat{t}_{j\,\pi}\right) = V\left(\sum_{j=1}^{q} a_j \sum_s \frac{y_{jk}}{\pi_k}\right) =$$

$$= \mathbf{Var}\left(\sum_s \check{u}_k\right) = \sum\sum_U \Delta_{kl}\, \check{u}_k\, \check{u}_l$$

Como $\mathbf{E}(\hat{\theta}_0) = \theta \Rightarrow MSE(\hat{\theta}) \doteq MSE(\hat{\theta}_0) = \mathbf{Var}(\hat{\theta}_0) = \mathbf{AVar}(\hat{\theta})$

Las cantidades u_k dependen de $a_j = \left.\frac{\partial f}{\partial \hat{t}_{j\,\pi}}\right|_{\hat{\mathbf{t}}_\pi = \mathbf{t}}$ que es desconocida ya que t_j es desconocido $\forall j$. De todas formas, la estimación puntual será $\hat{\theta}_0 = f(\hat{\mathbf{t}}_\pi)$. Para estimar la varianza se reemplaza a_j por $\hat{a}_j = \left.\frac{\partial f}{\partial \hat{t}_{j\,\pi}}\right|_{\hat{\mathbf{t}}_\pi = \hat{\mathbf{t}}_0}$, siendo $\hat{\mathbf{t}}_0$ el total observado en la muestra. Luego entonces $\hat{u}_k = \sum_{j=1}^{q} \hat{a}_j\, y_{jk}$, con lo que puede calcularse $\hat{\mathbf{Var}}(\hat{\theta}) = \sum\sum_s \check{\Delta}_{kl}\, \check{\hat{u}}_k\, \check{\hat{u}}_l$. Esto es válido ya que $\hat{u}_k$ es consistente para estimar u_k.

Estimador de una razón

El problema consiste en estimar un cociente entre totales poblacionales:

$$\star \; R = \frac{t_y}{t_z} = \frac{\bar{y}_U}{\bar{z}_U}$$

Sea el estimador:

$$\star \; \hat{R} = f\big(\hat{t}_{y\,\pi}; \hat{t}_{z\,\pi}\big) = \frac{\hat{t}_{y\,\pi}}{\hat{t}_{z\,\pi}} = \frac{\bar{y}_s}{\bar{z}_s}$$

Utilizando la linealización de Taylor:

$$\hat{R} = \hat{R}_0 = R + a_1\big(\hat{t}_{y\,\pi} - t_y\big) + a_2\big(\hat{t}_{z\,\pi} - t_z\big)$$

$$\text{donde} \quad a_1 = \left.\frac{\partial f}{\partial \hat{t}_{y\,\pi}}\right|_{\substack{\hat{t}_{y\,\pi}=t_y \\ \hat{t}_{z\,\pi}=t_z}} = \frac{1}{t_z} \quad \text{y} \quad a_2 = \left.\frac{\partial f}{\partial \hat{t}_{z\,\pi}}\right|_{\substack{\hat{t}_{y\,\pi}=t_y \\ \hat{t}_{z\,\pi}=t_z}} = -\frac{t_y}{t_z^2} = -\frac{R}{t_z}$$

Luego entonces:

$$\hat{R} \doteq \hat{R}_0 = R + \frac{1}{t_z}\big(\hat{t}_{y\,\pi} - t_y\big) - \frac{R}{t_z}\big(\hat{t}_{z\,\pi} - t_z\big) = R + \frac{\hat{t}_{y\,\pi}}{t_z} - \frac{t_y}{t_z} - \frac{R\,\hat{t}_{z\,\pi}}{t_z} + \frac{R\,t_z}{t_z} =$$

$$= R + \frac{\hat{t}_{y\,\pi}}{t_z} - R - \frac{R\,\hat{t}_{z\,\pi}}{t_z} + R = R + \frac{\hat{t}_{y\,\pi}}{t_z} - \frac{R\,\hat{t}_{z\,\pi}}{t_z} = R + \frac{1}{t_z}\big(\hat{t}_{y\,\pi} - R\,\hat{t}_{z\,\pi}\big) =$$

$$= R + \frac{1}{t_z}\sum\nolimits_s \frac{y_k - R\,z_k}{\pi_k} = R + \sum\nolimits_s \frac{u_k}{\pi_k} \quad \text{donde} \quad u_k = \frac{1}{t_z}\big(y_k - R\,z_k\big)$$

En conclusión:

$$\boxed{\hat{R} \doteq \hat{R}_0 = R + \sum\nolimits_s \frac{u_k}{\pi_k}}$$

$\hat{R} \doteq \hat{R}_0$ es aproximadamente insesgado para R.

$$\star \; \mathbf{E}(\hat{R}) \doteq \mathbf{E}(\hat{R}_0) = E\left(R + \sum\nolimits_s \frac{u_k}{\pi_k}\right) = R + \sum\nolimits_U u_k = R + \sum\nolimits_U \frac{y_k}{t_z} - R\sum\nolimits_U \frac{z_k}{t_z} =$$

$$= R + \frac{1}{t_z}\sum\nolimits_U y_k - \frac{R}{t_z}\sum\nolimits_U z_k = R + \frac{t_y}{t_z} - \frac{R}{t_z}\,t_z = \frac{t_y}{t_z} = R$$

$$\star \; \mathbf{AVar}(\hat{R}) = \mathbf{Var}(\hat{R}_0) = \sum\sum\nolimits_U \Delta_{kl}\, u_k^{\checkmark}\, u_l^{\checkmark} = \frac{1}{t_z^2}\sum\sum\nolimits_U \Delta_{kl}\, E_k^{\checkmark}\, E_l^{\checkmark}$$

$$\text{donde} \quad E_k = y_k - R\,z_k$$

$$\star \; \hat{\mathbf{Var}}(\hat{R}_0) = \sum\sum\nolimits_s \Delta_{kl}^{\checkmark}\, \hat{u}_k^{\checkmark}\, \hat{u}_l^{\checkmark} = \frac{1}{\hat{t}_z^2}\sum\sum\nolimits_s \Delta_{kl}^{\checkmark}\, e_k^{\checkmark}\, e_l^{\checkmark}$$

$$\text{donde} \quad e_k = y_k - \hat{R}\,z_k$$

$$\star \; \mathbf{E}\left(\hat{\mathbf{Var}}(\hat{R}_0)\right) = \mathbf{E}\left(\sum\sum\nolimits_s \Delta_{kl}^{\checkmark}\, \hat{u}_k^{\checkmark}\, \hat{u}_l^{\checkmark}\right) = \sum\sum\nolimits_U \Delta_{kl}\, \hat{u}_k^{\checkmark}\, \hat{u}_l^{\checkmark}$$

Por lo tanto, $\hat{\mathbf{Var}}(\hat{R}_0)$ es aproximadamente insesgado para estimar $\mathbf{Var}(\hat{R})$.

Las expresiones anteriores para $\mathbf{AVar}(\hat{R})$ y $\hat{\mathbf{Var}}(\hat{R})$ son equivalentes a:

$$\star \; \mathbf{AVar}(\hat{R}) = \frac{1}{t_z^2}\Big[\mathbf{Var}(\hat{t}_{y\,\pi}) + R^2\, V(\hat{t}_{z\,\pi}) - 2\,R\,\mathbf{Cov}\left(\hat{t}_{y\,\pi}; \hat{t}_{z\,\pi}\right)\Big]$$

$$\star \; \hat{\mathbf{Var}}(\hat{R}) = \frac{1}{\hat{t}_z^2}\Big[\hat{\mathbf{Var}}(\hat{t}_{y\,\pi}) + \hat{R}^2\, \hat{\mathbf{Var}}(\hat{t}_{z\,\pi}) - 2\,\hat{R}\,\hat{\mathbf{Cov}}\big(\hat{t}_{y\,\pi}; \hat{t}_{z\,\pi}\big)\Big]$$

El estimador $\hat{t}_{y_{ra}}$

El objetivo es estimar t_y, y se cuenta con una variable auxiliar z conocida $\forall k \in U$. Sea el "estimador de razón":

$$\star \;\; \hat{t}_{yra} = \frac{\hat{t}_{y\,\pi}}{\hat{t}_{z\,\pi}}\, t_z = \hat{R}\, t_z = \frac{t_z}{\hat{t}_{z\,\pi}}\, \hat{t}_{y\,\pi}$$

$$\star \;\; \mathbf{AVar}\left(\hat{t}_{yra}\right) = \mathbf{AVar}\left(\hat{R}\, t_z\right) = t_z^2\; \mathbf{AVar}\left(\hat{R}\right) = \sum\sum\nolimits_U \Delta_{kl} \left(\frac{y_k - R\, z_k}{\pi_k}\right)\left(\frac{y_l - R\, z_l}{\pi_l}\right)$$

$$\star \;\; \hat{\mathbf{Var}}\left(\hat{t}_{yra}\right) = \sum\sum\nolimits_s \Delta_{kl}^{\checkmark} \left(\frac{y_k - \hat{R}\, z_k}{\pi_k}\right)\left(\frac{y_l - \hat{R}\, z_l}{\pi_l}\right)$$

La lógica detrás de este estimador es la mismas que en el $\hat{t}_{alt} = \dfrac{N}{\hat{N}}\, \hat{t}_{y\,\pi}$ donde $z_k = 1 \;\; \forall k \in U$.