Diseño STSI

Daniel Czarnievicz

2017

Estrategia de selección

El mecanismo de selección consiste en implementar un diseño SI para cada estrato en la población. Esto implica que para cada estrato se obtendrá una muestra de tamaño n_{s_h} fijo.

Probabilidades de inclusión

$$\star \; \pi_k = P(k \in s) = P(k \in s_h) = \frac{n_h}{N_h} \;\; \forall k \in U; \;\; h = 1; \ldots ; H$$

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

$$\star \; \hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{\pi_h} = \sum_{h=1}^{H} \sum\nolimits_{s_h} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum\nolimits_{s_h} \frac{N_h \, y_k}{n_h} = \sum_{h=1}^{H} \sum\nolimits_{s_h} N_h \bar{y}_{s_h}$$

$$\star \; V_{STSI}(\hat{t}_\pi) = \sum_{h=1}^{H} V_{SI_h}(\hat{t}_{\pi_h}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}(1 - f_h) S^2_{y_{U_h}}$$

$$\text{donde } f_h = \frac{n_h}{N_h} \text{ y } S^2_{y_{U_h}} = \frac{1}{N_h - 1} \sum\nolimits_{U_h} \Big(y_k - \bar{y}_{U_h} \Big)^2$$

$$\star \; \hat{V}_{STSI}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}_{SI_h}(\hat{t}_{\pi_h}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}(1 - f_h) S^2_{y_{s_h}} \text{ donde } S^2_{y_{s_h}} = \frac{1}{n_h - 1} \sum\nolimits_{s_h} \Big(y_k - \bar{y}_{s_h} \Big)^2$$

Tamaño muestral

Asignación proporcional de la muestra

Dado un n, se asignan los n_h de forma proporcional al tamaño del estrato. Por lo tanto:

$$\star \; n_h = n \, \frac{N_h}{N}$$

De esta forma, la varianza del estimador π es:

$$\star \; V_{STSI,\,prop}(\hat{t}_\pi) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) S^2_{y_{U_h}} = \sum_{h=1}^{H} \frac{N_h^2}{n \, \frac{N_h}{N}} \left(1 - \frac{n \, \frac{N_h}{N}}{N_h} \right) S^2_{y_{U_h}} =$$

$$= \sum_{h=1}^{H} \frac{N \, N_h}{n} \left(1 - \frac{n}{N} \right) S^2_{y_{U_h}} = \frac{N}{n} \big(1 - f \big) \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}}$$

Nótese que con asignación proporcional, $\pi_k = P(k \in s) = \dfrac{n_h}{N_h} = \dfrac{n}{N} \;\; \forall k \in U$, por esto este tipo de diseños suele llamarse *diseños autoponderados*.

Asignación óptima de la muestra

Supongamos que el costo de relevamiento es lineal en el tamaño de la muestra por estrato, siendo c_h el costo de relevar cada observación en el estrato h, y c_0 un costo fijo:

$$C = c_0 + \sum_{h=1}^{H} c_h\, n_h \text{ con } c_h > 0\ \forall h$$

El objetivo es, dado un tamaño de muestra n, encontrar la mejor asignación por estrato (es decir, la que minimice la varianza del estimador π), sujeto a una restricción de costo. El mismo problema podría verse de forma opuesta, es decir, minimizar los costos, dado que se desea una determinada varianza. Es decir:

$$\min_{n_h} \left\{ V_{STSI}(\hat{t}_\pi) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}(1-f_h) S^2_{y_{U_h}} \right\} \quad \text{o} \quad \min_{n_h} \left\{ C = c_0 + \sum_{h=1}^{H} c_h\, n_h \right\}$$

$$\text{s.a. } C \text{ fijo} \qquad\qquad \text{s.a. } V_{STSI}(\hat{t}_\pi) \text{ fija}$$

En cualquiera de los dos casos el problema se resuelve haciendo mínimo el producto $V_{STSI}(\hat{t}_\pi) \times C$, y esto equivale a:

$$\min_{n_h} \left\{ \left(\sum_{h=1}^{H} \frac{N_h^2\, S^2_{y_{U_h}}}{n_h} \right) \left(\sum_{h=1}^{H} c_h\, n_h \right) \right\}$$

Utilizando la desigualdad de Cauchy-Schwartz:

$$\left(\sum_{h=1}^{H} \frac{N_h^2\, S^2_{y_{U_h}}}{n_h} \right) \left(\sum_{h=1}^{H} c_h\, n_h \right) \geq \left[\sum_{h=1}^{H} \left(\frac{N_h\, S_{y_{U_h}}}{\sqrt{n_h}} \right) \left(\sqrt{c_h}\, \sqrt{n_h} \right) \right]^2$$

donde la igualdad se cumple si:

$$\frac{N_h\, S_{y_{U_h}}}{\sqrt{n_h}} = \lambda \sqrt{c_h}\, \sqrt{n_h} \Rightarrow \frac{N_h\, S_{y_{U_h}}}{\sqrt{c_h}} = \lambda\, n_h$$

Luego entonces, teniendo en cuenta que $n = \sum_{h=1}^{H} n_h$, obtenemos:

$$\sum_{h=1}^{H} \frac{N_h\, S_{y_{U_h}}}{\sqrt{n_h}} = \lambda\, n \Rightarrow \lambda = \frac{1}{n} \sum_{h=1}^{H} \frac{N_h\, S_{y_{U_h}}}{\sqrt{n_h}}$$

Llegamos entonces a que:

$$\left.\begin{array}{c} \dfrac{N_h\, S_{y_{U_h}}}{\lambda\, \sqrt{c_h}} = n_h \\ \\ \text{y} \\ \\ \lambda = \dfrac{1}{n} \displaystyle\sum_{h=1}^{H} \dfrac{N_h\, S_{y_{U_h}}}{\sqrt{n_h}} \end{array}\right\} \Rightarrow \boxed{n_h = n \frac{N_h\, S_{y_{U_h}}}{\sqrt{c_h}\, \sum_{h=1}^{H} \left(\frac{N_h\, S_{y_{U_h}}}{\sqrt{c_h}} \right)}}$$

Por lo tanto, el tamaño de muestra en cada estrato será mayor cuanto:

1. mayor sea N_h

2. mayor sea $S_{y_{U_h}}$
3. menor es c_h

Ahora podemos calcular el tamaño de muestra total a tomar, partiendo de cualquiera de las especificaciones del problema:

1. $\min\limits_{n_h} \left\{ V_{STSI}(\hat{t}_\pi) = \sum\limits_{h=1}^{H} \frac{N_h^2}{n_h}(1-f_h)S^2_{y_{U_h}} \right\}$ s.a. C fijo

   Sustituyendo el n_h hallado, obtenemos que:

$$C = c_0 + \sum_{h=1}^{H} c_h \, n \, \frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h} \sum\limits_{h=1}^{H} \left(\frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h}} \right)}$$

   De donde podemos despejar n para obtener:

$$\star \;\; n = (C - c_0) \left[\frac{\sum\limits_{h=1}^{H} \left(\frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h}} \right)}{\sum\limits_{h=1}^{H} \sqrt{c_h} \, N_h \, S_{y_{U_h}}} \right]$$

2. $\min\limits_{n_h} \left\{ C = c_0 + \sum\limits_{h=1}^{H} c_h \, n_h \right\}$ s.a. $V_{STSI}(\hat{t}_\pi)$ fija

   La varianza se fija según la precisión que se desee:

$$\varepsilon^2 = z^2_{1-\alpha/2} \, V_{STSI}(\hat{t}_\pi) \Rightarrow \varepsilon^2 = z^2_{1-\alpha/2} \left[\sum_{h=1}^{H} \frac{N_h^2 \, S^2_{y_{U_h}}}{n_h} - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right]$$

   Remplazando por los n_h hallados:

$$\frac{\varepsilon^2}{z^2_{1-\alpha/2}} = \sum_{h=1}^{H} \frac{N_h^2 \, S^2_{y_{U_h}}}{n \frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h} \sum\limits_{h=1}^{H} \left(\frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h}} \right)}} - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \Rightarrow$$

$$\Rightarrow \frac{\varepsilon^2}{z^2_{1-\alpha/2}} = \frac{1}{n} \sum_{h=1}^{H} \left[N_h \, S_{y_{U_h}} \sqrt{c_h} \sum_{h=1}^{H} \frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h}} \right] - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \Rightarrow$$

$$\Rightarrow \frac{\varepsilon^2}{z^2_{1-\alpha/2}} + \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} = \frac{1}{n} \sum_{h=1}^{H} \left[N_h \, S_{y_{U_h}} \sqrt{c_h} \sum_{h=1}^{H} \frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h}} \right] \Rightarrow$$

$$\Rightarrow \frac{1}{n} = \frac{\frac{\varepsilon^2}{z^2_{1-\alpha/2}} + \sum\limits_{h=1}^{H} N_h \, S^2_{y_{U_h}}}{\sum\limits_{h=1}^{H} \left[N_h \, S_{y_{U_h}} \sqrt{c_h} \sum\limits_{h=1}^{H} \frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h}} \right]} \Rightarrow$$

$$\Rightarrow n = \frac{\sum\limits_{h=1}^{H} \left[N_h \, S_{y_{U_h}} \sqrt{c_h} \sum\limits_{h=1}^{H} \frac{N_h \, S_{y_{U_h}}}{\sqrt{c_h}} \right]}{\frac{\varepsilon^2}{z^2_{1-\alpha/2}} + \sum\limits_{h=1}^{H} N_h \, S^2_{y_{U_h}}} \Rightarrow$$

$$\Rightarrow \boxed{n = \frac{z^2_{1-\alpha/2}\left[\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}\,\sqrt{c_h}\right]\left[\sum\limits_{h=1}^{H}\frac{N_h\, S_{y_{U_h}}}{\sqrt{c_h}}\right]}{\varepsilon^2 + z^2_{1-\alpha/2}\sum\limits_{h=1}^{H} N_h\, S^2_{y_{U_h}}}}$$

Si el costo por dato relevado es fijo $c_h = c$, entonces:

$$\star\ n_h = n\,\frac{N_h\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}$$

$$n = \frac{z^2_{1-\alpha/2}\left[\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}\,\sqrt{c}\right]\left[\sum\limits_{h=1}^{H}\frac{N_h\, S_{y_{U_h}}}{\sqrt{c}}\right]}{\varepsilon^2 + z^2_{1-\alpha/2}\sum\limits_{h=1}^{H} N_h\, S^2_{y_{U_h}}} = \frac{z^2_{1-\alpha/2}\left[\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}\right]\left[\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}\right]}{\varepsilon^2 + z^2_{1-\alpha/2}\sum\limits_{h=1}^{H} N_h\, S^2_{y_{U_h}}} \Rightarrow$$

$$\Rightarrow \boxed{n = \frac{z^2_{1-\alpha/2}\left[\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}\right]^2}{\varepsilon^2 + z^2_{1-\alpha/2}\sum\limits_{h=1}^{H} N_h\, S^2_{y_{U_h}}}}$$

De esta forma, la varianza del estimador π es:

$$\star\ V_{STSI,\,opt}(\hat{t}_\pi) = \sum_{h=1}^{H}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)S^2_{y_{U_h}} = \sum_{h=1}^{H}\frac{N_h^2}{n\,\frac{N_h\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}}\left(1-\frac{n\,\frac{N_h\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}}{N_h}\right)S^2_{y_{U_h}} =$$

$$= \sum_{h=1}^{H}\frac{N_h}{\frac{n}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}}\left(1-\frac{n\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}\right)S_{y_{U_h}} = \left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)\sum_{h=1}^{H}\frac{N_h}{n}\left(1-\frac{n\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}\right)S_{y_{U_h}} =$$

$$= \frac{1}{n}\left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)\left[\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\left(1-\frac{n\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}\right)\right] =$$

$$= \frac{1}{n}\left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)\left[\sum_{h=1}^{H}\left(N_h\, S_{y_{U_h}} - \frac{N_h\, S_{y_{U_h}}\, n\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}\right)\right] =$$

$$= \frac{1}{n}\left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)\left[\sum_{h=1}^{H} N_h\, S_{y_{U_h}} - n\sum_{h=1}^{H}\frac{N_h\, S^2_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}\right] =$$

$$= \frac{1}{n}\left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)\left[\sum_{h=1}^{H} N_h\, S_{y_{U_h}} - \frac{n}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}}\sum_{h=1}^{H} N_h\, S^2_{y_{U_h}}\right] =$$

$$= \frac{1}{n}\left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)^2 - \sum_{h=1}^{H} N_h\, S^2_{y_{U_h}}$$

Si $n_h > N_h$ para algún estrato (supongamos que para el estrado H), entonces dicho estrato debe censarse (tomar $n_H = N_H$). Luego nos queda una población de tamaño $N - N_H$ en $H - 1$ estratos. Para esta nueva población se toma una muestra de tamaño $n - N_H$. El total, t_y, será:

$$t_y = \sum_{h=1}^{H-1} t_{y_h} + t_{y_H}$$

Se busca estimar. $t_y^* = \sum\limits_{h=1}^{H-1} t_{y_h}$ para una población de tamaño $N^* = N - N_H$ tomando una muestra de tamaño $n^* = n - N_H$ bajo un $STSI$. Por lo tanto,

$$\star\ \hat{t}_\pi^* = \sum_{h=1}^{H-1}\sum\nolimits_{s_h} y_k^{\checkmark} = \sum_{h=1}^{H-1} \hat{t}_{\pi_h}$$

$$\star\ V_{STSI,\,opt}(\hat{t}_\pi^*) = \frac{1}{n}\left(\sum_{h=1}^{H-1} N_h\, S_{y_{U_h}}\right)^2 - \sum_{h=1}^{H-1} N_h\, S^2_{y_{U_h}}$$

Si $n_h = 1$ entonces no puede calcularse $\hat{V}_{STSI}\hat{t}_\pi$. Si $N_h > 1$, entonces deberán tomarse un $n_h = 2$ dado que el aumento en el costo no debería ser significativo. Si esto no ocurre entonces se procede de la siguiente forma. Supongamos que el estrato h tiene igual CV que el estrato $h+1$, entonces:

$$CV_{y_{U_h}} = CV_{y_{U_{h+1}}} \Rightarrow \frac{S_{y_{U_h}}}{\bar{y}_{U_h}} = \frac{S_{y_{U_{h+1}}}}{\bar{y}_{U_{h+1}}}$$

Por lo tanto, $\hat{CV}_{y_{U_h}} = \frac{S_{y_{s_{h+1}}}}{\bar{y}_{s_{h+1}}} \Rightarrow$ aproximamos $S_{y_{U_h}}$ por $\hat{S}_{y_{U_h}} = \frac{S_{y_{s_{h+1}}}}{\bar{y}_{s_{h+1}}}\,\hat{t}_{y_h}$ donde $\hat{t}_{y_h} = \bar{y}_{s_h}$ ya que $N_h = 1$.

Si $n_h = 1$ para muchos estratos, entonces se deben colapsar estratos para poder calcular la estimación de la varianza.

Asignación x-óptima de la muestra

El problema con la asignación óptima es que $S_{y_{U_h}}$ es generalmente desconocida. Si se cuenta con una variable auxiliar x conocida para todos los elementos de la población y altamente correlacionada con la variable y, esta puede utilizarse para lograr una asignación x-óptima, de la siguiente forma:

$$\star\ n_h = n\,\frac{N_h\, S_{x_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{x_{U_h}}}$$

Asignación proporcional a t_y

En este caso se requiere que $y_k \geq 0 \; \forall k \in U$. Si esto se cumple, los tamaños muestrales se determinan mediante:

$$\star \; n_h = n \frac{\sum_{U_h} y_k}{\sum_U y_k} = n \frac{t_{y_h}}{t_y}$$

Luego si $CV_h = \dfrac{S_{y_{U_h}}}{\bar{y}_{U_h}} = CV \; \forall h = 1; \ldots ; H$ se tiene que:

$$n_h = n \frac{t_{y_h}}{t_y} = \frac{N_h \, \bar{y}_{U_h}}{\sum\limits_{h=1}^{H} N_h \, \bar{y}_{U_h}} = \frac{N_h \, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h \, S_{y_{U_h}}}$$

Asignación proporcional a t_x

De nuevo, esta última estrategia no puede aplicarse dado que se desconocen los datos sobre y_k. Por lo tanto, debe usarse una variable auxiliar x_k tal que $y_k \doteq a + b \, x_k \;\; \forall k \in U$. Si esto se cumple, entonces $CV_{y_{U_h}} \doteq CV_{x_{U_h}} \;\; \forall h = 1; \ldots ; H$, por lo que podemos utilizar:

$$\star \; n_h = n \frac{\sum_{U_h} x_k}{\sum_U x_k} = n \frac{t_{x_h}}{t_x}$$

Construcción de los estratos

Cuando solo se trabaja con una variable de estudio, y, los estratos deben construirse considerando la distribución de dicha variable, o de alguna variable auxiliar x.

Sean y_0 y y_H los valores extremos de y en la población. El problema es hallar los valores $y_1; \ldots ; y_{H-1}$ tales que $V_{STSI,\, opt}(\hat{t}_\pi)$ sea mínima. Si ignoramos el FCPF y multiplicamos por ${}^1/_N$ el problema puede verse de la siguiente forma:

$$\min_{y_1;\cdots;y_{H-1}} \left\{ V_{STSI,\, opt}(\hat{t}_\pi) \right\} = \min_{y_1;\cdots;y_{H-1}} \left\{ \frac{1}{n} \left(\sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right)^2 - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right\} \approx \min_{y_1;\cdots;y_{H-1}} \left\{ \sum_{h=1}^{H} w_h \, S_{y_{U_h}} \right\}$$

Supongamos estratos pequeños y numerosos de forma tal que $f_Y(y)$ es aproximadamente constante por estrato. Luego entonces:

$$w_h = \frac{N_h}{N} = \int\limits_{y_{h-1}}^{y_h} f_Y(t) \, \mathrm{d}t \doteq f_h (y_h - y_{h-1})$$

donde f_h es el valor aproximadamente constante dentro de cada estrato. Además,

$$S_{y_{U_h}} \doteq \frac{1}{\sqrt{12}} (y_h - y_{h-1})$$

Consideremos:

- $z_h = \displaystyle\int\limits_{y_0}^{y_h} \sqrt{f(t)} \, \mathrm{d}t$
- $z_h - z_{h-1} = \displaystyle\int\limits_{y_{h-1}}^{y_h} \sqrt{f(t)} \, \mathrm{d}t \doteq \sqrt{f_h} (y_h - y_{h-1})$

Sustituyendo obtenemos que:

$$\sum_{h=1}^{H} w_h\, S_{y_{U_h}} \doteq \sum_{h=1}^{H} \Big[f_h (y_h - y_{h-1}) \Big] \Big[\frac{1}{\sqrt{12}} (y_h - y_{h-1}) \Big] \Rightarrow$$

$$\Rightarrow \sqrt{12} \sum_{h=1}^{H} w_h\, S_{y_{U_h}} \doteq \sum_{h=1}^{H} f_h \Big(y_h - y_{h-1} \Big)^2 = \sum_{h=1}^{H} \Big[\sqrt{f_h} \Big(y_h - y_{h-1} \Big) \Big]^2 \doteq \sum_{h=1}^{H} \Big(z_h - z_{h-1} \Big)^2$$

Por lo tanto, el problema se transforma en:

$$\min_{z_h} \left\{ \sum_{h=1}^{H} \Big(z_h - z_{h-1} \Big)^2 \right\}$$

$$\text{s.a } \sum_{h=1}^{H} \Big(z_h - z_{h-1} \Big) = z_H - z_0 \text{ fijo}$$

Como $z_H - z_0$ es fijo, la solución se encuentra haciendo $z_h - z_{h-1}$ constante. Por lo tanto, dada $f(y)$, la regla es escoger los límites y_h para $h = 1; \ldots ; H - 1$, de forma que estos determinen valores de $\sqrt{f_h}(y_h - y_{h-1})$ aproximadamente constantes.

Elección del número de estratos

Supongamos que los estratos pueden ser construidos en función de la variable de estudio, y, donde $y \sim Unif(a;\, a + d)$. Antes de la estratificación tenemos que: $S^2_{y_U} = \dfrac{d^2}{12}$. Si se forman H estratos de igual tamaño, $S^2_{y_U} = \dfrac{(^d/_H)^2}{12} = \dfrac{d^2}{12H^2}$. Además, $w_h = \dfrac{N_h}{N} = \dfrac{^N/_H}{N} = \dfrac{1}{H}$.
Si consideramos una asignación óptima, entonces:

$$\star\ n_h = n \frac{N_h\, S_{y_{U_h}}}{\sum\limits_{h=1}^{H} N_h\, S_{y_{U_h}}} = n \frac{(^N/_H)(^d/_{\sqrt{12}\,H})}{H(^N/_H)(^d/_{\sqrt{12}\,H})} = \frac{n}{H}$$

Vemos entonces que la asignación óptima es igual a la proporcional. Si ignoramos el FCPF, entonces tenemos que:

$$V_{STSI}(\hat{t}_\pi) \doteq \frac{1}{n} \left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}} \right)^2 = \frac{1}{n} \left(\sum_{h=1}^{H} \frac{N}{H} \frac{d}{\sqrt{12}H} \right)^2 = \frac{1}{n} \left(H \frac{N}{H} \frac{d}{\sqrt{12}H} \right)^2 =$$

$$= \frac{1}{n} \frac{N^2}{12} \frac{d^2}{H^2} = \frac{1}{H^2} \frac{N^2}{n} \frac{d^2}{12} \doteq \frac{1}{H^2} V_{SI}(\hat{t}_\pi) = \frac{V_{SI}(\hat{t}_\pi)}{H^2}$$

A esto se concluye que, bajo estas condiciones, la varianza disminuye con el cuadrado de la cantidad de estratos. Si no se cuenta con y, se requiere una correlación de más de 0,9 entre y y la variable auxiliar x para que la varianza se reduzca con más de 6 estratos.

Efecto diseño

Para calcular el Deff conviene utilizar la descomposición de la varianza dentro y entre los estratos:

$$(N-1) S^2_{y_U} = \sum\nolimits_U \Big(y_k - \bar{y}_U \Big)^2 = \sum_{h=1}^{H} \sum\nolimits_{U_h} \Big(y_k - \bar{y}_U \Big)^2 =$$

$$= \sum_{h=1}^{H} \sum\nolimits_{U_h} \left(y_k - \bar{y}_{U_h} \right)^2 + \sum_{h=1}^{H} N_h \left(\bar{y}_{U_h} - \bar{y}_U \right)^2 =$$

$$= \sum_{h=1}^{H} (N_h - 1) S^2_{y_{U_h}} + \sum_{h=1}^{H} N_h \left(\bar{y}_{U_h} - \bar{y}_U \right)^2 \Rightarrow$$

$$\Rightarrow N S^2_{y_U} \doteq \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} + \sum_{h=1}^{H} N_h \left(\bar{y}_{U_h} - \bar{y}_U \right)^2$$

Luego entonces:

$$V_{SI}(\hat{t}_\pi) - V_{STSI,\,prop}(\hat{t}_\pi) = \frac{N^2}{n}(1-f) S^2_{y_U} - \frac{N}{n}(1-f) \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} =$$

$$= \frac{N}{n}(1-f) \left[N \, S^2_{y_U} - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right] \doteq$$

$$\doteq \frac{N}{n}(1-f) \left[\sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} + \sum_{h=1}^{H} N_h \left(\bar{y}_{U_h} - \bar{y}_U \right)^2 - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right] =$$

$$= \frac{N}{n}(1-f) \sum_{h=1}^{H} N_h \left(\bar{y}_{U_h} - \bar{y}_U \right)^2$$

Por lo tanto:

$$\boxed{V_{STSI,\,prop}(\hat{t}_\pi) \doteq V_{SI}(\hat{t}_\pi) - \frac{N}{n}(1-f) \sum_{h=1}^{H} N_h \left(\bar{y}_{U_h} - \bar{y}_U \right)^2}$$

Esto implica que la varianza del diseño estratificado con asignación proporcional es aproximadamente igual a la varianza del SI, menos un término positivo, el cual aumenta conforme más heterogéneos en media sean los estratos. Por lo tanto, la estratificación reduce la varianza del estimador π. Si $\bar{y}_{U_h} = \bar{y}_U \;\; \forall h \Rightarrow V_{SI}(\hat{t}_\pi) = V_{STSI,\,prop}(\hat{t}_\pi)$.

Por su parte:

$$V_{STSI,\,prop}(\hat{t}_\pi) - V_{STSI,\,opt}(\hat{t}_\pi) = \left[\frac{N}{n}(1-f) \sum_{h=1}^{H} N_h \, S^2_{y_U} \right] - \left[\frac{1}{n} \left(\sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right)^2 - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right] =$$

$$= \left[\frac{N}{n} \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} - \sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right] - \left[\frac{1}{n} \left(\sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right)^2 - \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right] =$$

$$= \left[\frac{N}{n} \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right] - \left[\frac{1}{n} \left(\sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right)^2 \right] =$$

$$= \left[\frac{N}{n} \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right] - 2 \left[\frac{1}{n} \left(\sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right)^2 \right] + \left[\frac{1}{n} \left(\sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right)^2 \right] =$$

$$= \left[\frac{N}{n} \sum_{h=1}^{H} N_h \, S^2_{y_{U_h}} \right] - 2 \left[\frac{1}{n} \left(\sum_{h=1}^{H} N_h \, S_{y_{U_h}} \right)^2 \right] + \left[\frac{N^2}{n} \left(\sum_{h=1}^{H} \frac{N_h \, S_{y_{U_h}}}{N} \right)^2 \right] =$$

$$= \left[\frac{N}{n}\sum_{h=1}^{H} N_h\, S^2_{y_{U_h}}\right] - 2\left[\frac{N}{n}\left(\sum_{h=1}^{H}\frac{N_h\, S_{y_{U_h}}}{N}\right)\left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)\right] + \left[\frac{N^2}{n}\left(\sum_{h=1}^{H}\frac{N_h\, S_{y_{U_h}}}{N}\right)^2\right] =$$

$$= \left[\frac{N}{n}\sum_{h=1}^{H} N_h\, S^2_{y_{U_h}}\right] - 2\left[\frac{N}{n}\,\bar{S}_U\left(\sum_{h=1}^{H} N_h\, S_{y_{U_h}}\right)\right] + \left[\frac{N^2}{n}\,\bar{S}^2_U\right] =$$

$$= \frac{N}{n}\left[\sum_{h=1}^{H} N_h\, S^2_{y_{U_h}} - 2\,\bar{S}_U\sum_{h=1}^{H} N_h\, S_{y_{U_h}} + N\,\bar{S}^2_U\right] =$$

$$= \frac{N}{n}\left[\sum_{h=1}^{H} N_h\, S^2_{y_{U_h}} - 2\,\bar{S}_U\sum_{h=1}^{H} N_h\, S_{y_{U_h}} + \sum_{h=1}^{H} N_h\,\bar{S}^2_U\right] =$$

$$= \frac{N}{n}\left[\sum_{h=1}^{H}\left(N_h\, S^2_{y_{U_h}} - 2\,\bar{S}_U\, N_h\, S_{y_{U_h}} + N_h\,\bar{S}^2_U\right)\right] =$$

$$= \frac{N}{n}\left[\sum_{h=1}^{H} N_h\left(S^2_{y_{U_h}} - 2\,\bar{S}_U\, S_{y_{U_h}} + \bar{S}^2_U\right)\right] = \frac{N}{n}\left[\sum_{h=1}^{H} N_h\left(S_{y_{U_h}} - \bar{S}_U\right)^2\right]$$

Por lo tanto:

$$\boxed{V_{STSI,\,prop}(\hat{t}_\pi) = V_{STSI,\,opt}(\hat{t}_\pi) + \frac{N}{n}\left[\sum_{h=1}^{H} N_h\left(S_{y_{U_h}} - \bar{S}_U\right)^2\right]}$$

Con lo anterior podemos comparar $V_{STSI,\,opt}(\hat{t}_\pi)$ con $V_{SI}(\hat{t}_\pi)$

$$\left.\begin{aligned} V_{STSI,\,prop}(\hat{t}_\pi) &\doteq V_{SI}(\hat{t}_\pi) - \frac{N}{n}(1-f)\sum_{h=1}^{H} N_h\left(\bar{y}_{U_h} - \bar{y}_U\right)^2 \\ V_{STSI,\,prop}(\hat{t}_\pi) &= V_{STSI,\,opt}(\hat{t}_\pi) + \frac{N}{n}\left[\sum_{h=1}^{H} N_h\left(S_{y_{U_h}} - \bar{S}_U\right)^2\right] \end{aligned}\right\} \Rightarrow$$

$$\Rightarrow V_{SI}(\hat{t}_\pi) - \frac{N}{n}(1-f)\sum_{h=1}^{H} N_h\left(\bar{y}_{U_h} - \bar{y}_U\right)^2 \doteq V_{STSI,\,opt}(\hat{t}_\pi) + \frac{N}{n}\left[\sum_{h=1}^{H} N_h\left(S_{y_{U_h}} - \bar{S}_U\right)^2\right] \Rightarrow$$

$$\Rightarrow V_{SI}(\hat{t}_\pi) \doteq V_{STSI,\,opt}(\hat{t}_\pi) + \frac{N}{n}\left[\sum_{h=1}^{H} N_h\left(S_{y_{U_h}} - \bar{S}_U\right)^2\right] + \frac{N}{n}(1-f)\sum_{h=1}^{H} N_h\left(\bar{y}_{U_h} - \bar{y}_U\right)^2 \Rightarrow$$

$$\Rightarrow \boxed{V_{SI}(\hat{t}_\pi) \doteq V_{STSI,\,opt}(\hat{t}_\pi) + \frac{N}{n}\left[\sum_{h=1}^{H} N_h\left(S_{y_{U_h}} - \bar{S}_U\right)^2 + (1-f)\sum_{h=1}^{H} N_h\left(\bar{y}_{U_h} - \bar{y}_U\right)^2\right]}$$

De esto se desprende que el diseño $STSI$ con asignación óptima reduce la varianza respecto del diseño SI, tanto más cuanto más heterogéneos en media y/o en desvío sean los estratos.

Si en lugar de la aproximación de la descomposición de la varianza se utiliza el resultado exacto, entonces:

$$V_{SI}(\hat{t}_\pi) = V_{STSI,\,prop}(\hat{t}_\pi) + \frac{N^2(1-f)}{n(N-1)}\left[\sum_{h=1}^{H} N_h\left(\bar{y}_{U_h} - \bar{y}_U\right)^2 - \frac{1}{N}\sum_{h=1}^{H}\left(N - N_h\right)S^2_{y_{U_h}}\right]$$

Luego entonces,

$$V_{SI}(\hat{t}_\pi) < V_{STSI}(\hat{t}_\pi) \Leftrightarrow \sum_{h=1}^{H} N_h\left(\bar{y}_{U_h} - \bar{y}_U\right)^2 < \frac{1}{N}\sum_{h=1}^{H}\left(N - N_h\right)S^2_{y_{U_h}}$$

Esto puede ocurrir en la medida en que las medias por estrato sean aproximadamente iguales, lo cual implica que $\sum_{h=1}^{H} N_h \left(\bar{y}_{U_h} - \bar{y}_U\right)^2 \doteq 0$. Si la asignación óptima coincide con la proporcional (las varianzas por grupos son todas iguales), el diseño SI es más eficiente que el $STSI$, incluso con asignació óptima.

Estratificación a posteriori

Supongamos que en la población $U = \{1; \ldots; k; \ldots; N\}$ se consideran H estratos de tamaños conocidos N_h, pero, a priori, no se dispone de información que permita clasificar los elementos del marco en los estratos. Se toma una muestra, s, de tamaño fijo, n, bajo un diseño SI, y se clasifica a posteriori.

Se utiliza el siguiente estimador:

$$\star\ \hat{t}_{post} = \sum_{h=1}^{H} N_h \, \bar{y}_{s_h} = \sum_{h=1}^{H} N_h \sum\nolimits_{s_h} \frac{y_k}{n_h}$$

Si $n_h = 0$ para algún estrato, entonces $\hat{t}_{post}$ no se puede calcular. Bajo el supuesto de que el tamaño de muestra es lo suficientemente grande como para asegurar 20 o más observaciones en cada post-estrato[1] la estratificación a posteriori es casi tan eficiente como un diseño $STSI$ con asignación proporcional.

Nótese que $n_S \sim MHG$, y dado que $n_S = \sum_{h=1}^{H} n_h$, sus márgenes no son independientes. Consideremos las variables:

$$z_{hk} = \begin{cases} 1 & \text{si} \quad k \in U_h \\ 0 & \text{si} \quad k \notin U_h \end{cases}$$

Luego entonces, $N_h = \sum\nolimits_U z_{hk}$ y $n_h = \sum\nolimits_s z_{hk}$. Por lo tanto:

$$\star\ E_{SI}(n_s) = E_{SI}\left(\sum\nolimits_s z_{hk}\right) = \sum\nolimits_U E_{SI}(I_k)\, z_{hk} = \frac{n}{N} \sum\nolimits_U z_{hk} = \frac{N}{n} N_h = n\, w_h$$

$$\star\ V_{SI}(n_s) = V_{SI}\left(\sum\nolimits_s z_{hk}\right) = \sum\nolimits_U V_{SI}(I_k)\, z_{hk} + \sum\sum\nolimits_U COV_{SI}(I_k; I_l)\, z_{hk}\, z_{hl} =$$

$$= \sum\nolimits_U f(1-f) z_{hl}^2 + \sum_{k \neq l}\sum\nolimits_U -\frac{f(1-f)}{N-1}\, z_{hk}\, z_{hl} =$$

$$= f(1-f)N_h + \sum_{k \neq l}\sum\nolimits_U -\frac{f(1-f)}{N-1}\, z_{hk}\, z_{hl} =$$

$$= f(1-f)N_h - \frac{f(1-f)}{N-1} 2\binom{N_h}{2} = f(1-f)N_h - \frac{f(1-f)}{N-1} N_h(N_h-1) =$$

$$= f(1-f)N_h\left(1 - \frac{N_h - 1}{N-1}\right) = \frac{n}{N}\left(\frac{N-n}{N}\right) N_h \left(\frac{N-N_h}{N-1}\right) =$$

$$= \frac{N-n}{N-1} n\, w_h(1-w_h) \doteq (1-f) n\, w_h(1-w_h)$$

[1] Esto se sugiere para lograr estimaciones estables de $\bar{y}_{U_h}$.

Ahora podemos calcular la esperanza del estimador $\hat{t}_{post}$, y su varianza, haciendo uso de la Ley de Esperanzas Iteradas[2], y de la Ley de Varianzas Iteradas[3]:

$$\star\ E_{SI}\left(\hat{t}_{post}\right) = E_{n_S}\left[E_{SI}\left(\hat{t}_{post}|n_S\right)\right] \doteq E_{n_S}\left[E_{SI}\left(\sum_{h=1}^{H} N_h \sum\nolimits_{s_h} \frac{y_k}{n_h}\middle| n_h > 0\right)\right] =$$

$$= E_{n_S}\left[\sum_{h=1}^{H} N_h \sum\nolimits_{U_h} E_{SI}\left(I_k \frac{y_k}{n_h}\middle| n_h > 0\right)\right] = E_{n_S}\left[\sum_{h=1}^{H} N_h \sum\nolimits_{U_h} E_{SI}(I_k|n_h)\frac{y_k}{n_h}\right] =$$

$$= E_{n_S}\left[\sum_{h=1}^{H} N_h \sum\nolimits_{U_h} \pi_k \frac{y_k}{n_h}\right] = E_{n_S}\left[\sum_{h=1}^{H} N_h \sum\nolimits_{U_h} \frac{n_h}{N_h}\frac{y_k}{n_h}\right] = E_{n_S}\left[\sum_{h=1}^{H} \sum\nolimits_{U_h} y_k\right] =$$

$$= E_{n_S}\left[\sum_{h=1}^{H} t_{y_h}\right] = E_{n_S}\left(t_y\right) = t_y$$

$$\star\ V_{SI}\left(\hat{t}_{post}\right) = V_{n_S}\left[\underbrace{E_{SI}\left(\hat{t}_{post}|n_s\right)}_{t_y}\right] + E_{n_s}\left[V_{SI}\left(\hat{t}_{post}|n_s\right)\right] =$$

$$= \underbrace{V_{n_S}\left(t_y\right)}_{=0} + E_{n_s}\left[V_{SI}\left(\hat{t}_{post}|n_s\right)\right] = E_{n_s}\left[V_{SI}\left(\hat{t}_{post}|n_s\right)\right] =$$

$$= E_{n_s}\left[V_{SI}\left(\sum_{h=1}^{H} N_h \sum\nolimits_{s_h} \frac{y_k}{n_h}\middle| n_s\right)\right] = E_{n_s}\left[\sum_{h=1}^{H} \frac{N_h^2}{n_h}(1-f_h)S_{y_{u_h}}^2\right] =$$

$$E_{n_s}\left[\sum_{h=1}^{H} \frac{N_h^2}{n_h} S_{y_{u_h}}^2 - \sum_{h=1}^{H} N_h\, S_{y_{u_h}}^2\right] = \sum_{h=1}^{H} E_{n_s}\left(\frac{1}{n_h}\right) N_h^2\, S_{y_{u_h}}^2 - \sum_{h=1}^{H} N_h\, S_{y_{u_h}}^2 =$$

$$= \sum_{h=1}^{H}\left[\frac{1}{E(n_h)}\left(1 + \frac{V(n_h)}{E^2(n_h)}\right)\right] N_h^2\, S_{y_{u_h}}^2 - \sum_{h=1}^{H} N_h\, S_{y_{u_h}}^2 =$$

$$= \sum_{h=1}^{H}\left[\frac{1}{n\, w_h}\left(1 + \frac{\left(\frac{N-n}{N-1}\right) n\, w_h(1-w_h)}{n^2\, w_h^2}\right)\right] N_h^2\, S_{y_{u_h}}^2 - \sum_{h=1}^{H} N_h\, S_{y_{u_h}}^2 \doteq$$

$$\doteq \sum_{h=1}^{H}\left[\frac{1}{n\, w_h}\left(1 + \frac{(1-f)(1-w_h)}{n\, w_h}\right)\right] N_h^2\, S_{y_{u_h}}^2 - \sum_{h=1}^{H} N_h\, S_{y_{u_h}}^2 =$$

$$= \left(\frac{N^2}{N^2}\right)\left(\frac{1-f}{1-f}\right)\left(\frac{n}{n}\right)\left[\sum_{h=1}^{H}\left(\frac{1}{n\, w_h} + \frac{(1-f)(1-w_h)}{n^2\, w_h^2}\right) N_h^2\, S_{y_{u_h}}^2 - \sum_{h=1}^{H} N_h\, S_{y_{u_h}}^2\right] =$$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}\left(\frac{n}{(1-f)n\, w_h} + \frac{(1-f)n(1-w_h)}{(1-f)n^2\, w_h^2}\right)\underbrace{\frac{N_h^2}{N^2}}_{w_h^2} S_{y_{u_h}}^2 - \sum_{h=1}^{H}\left(\frac{n}{N^2(1-f)}\right) N_h\, S_{y_{u_h}}^2\right] =$$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}\left(\frac{1}{(1-f)w_h} + \frac{1-w_h}{n\, w_h^2}\right) w_h^2\, S_{y_{u_h}}^2 - \sum_{h=1}^{H} \frac{n}{N}\left(\frac{1}{1-f}\right)\frac{N_h}{N}\, S_{y_{u_h}}^2\right] =$$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}\left(\frac{w_h}{1-f} + \frac{1-w_h}{n}\right) S_{y_{u_h}}^2 - \sum_{h=1}^{H} f\left(\frac{1}{1-f}\right) w_h\, S_{y_{u_h}}^2\right] =$$

[2]**Ley de Esperanzas Iteradas**: $E(X) = E_A\left[E(X|A)\right]$

[3]**Ley de Varianzas Iteradas**: $V(X) = V_A\left[E(X|A)\right] + E_A\left[V(X|A)\right]$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}\left(\frac{w_h}{1-f}+\frac{1-w_h}{n}\right)S^2_{y_{u_h}} - \sum_{h=1}^{H}\left(\frac{f}{1-f}\right)w_h\,S^2_{y_{u_h}}\right] =$$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}\left(\frac{1}{1-f}\right)w_h\,S^2_{y_{u_h}} + \sum_{h=1}^{H}\left(\frac{1-w_h}{n}\right)S^2_{y_{u_h}} - \sum_{h=1}^{H}\left(\frac{f}{1-f}\right)w_h\,S^2_{y_{u_h}}\right] =$$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}\left(\frac{1}{1-f}\right)w_h\,S^2_{y_{u_h}} - \sum_{h=1}^{H}\left(\frac{f}{1-f}\right)w_h\,S^2_{y_{u_h}} + \sum_{h=1}^{H}\left(\frac{1-w_h}{n}\right)S^2_{y_{u_h}}\right] =$$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}\left[\left(\frac{1}{1-f}-\frac{f}{1-f}\right)w_h\right]S^2_{y_{u_h}} + \sum_{h=1}^{H}\frac{1-w_h}{n}\,S^2_{y_{u_h}}\right] =$$

$$= \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}w_h\,S^2_{y_{U_h}} + \frac{1}{n}\sum_{h=1}^{H}(1-w_h)\,S^2_{y_{U_h}}\right]$$

$$\star\ \hat{V}_{SI}\left(\hat{t}_{post}\right) = \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}w_h\,S^2_{y_{s_h}} + \frac{1}{n}\sum_{h=1}^{H}(1-w_h)\,S^2_{y_{s_h}}\right]$$

Podemos ver la varianza de la sigiente forma:

$$V_{SI}\left(\hat{t}_{post}\right) = \frac{N^2}{n}(1-f)\left[\sum_{h=1}^{H}w_h\,S^2_{y_{U_h}} + \frac{1}{n}\sum_{h=1}^{H}(1-w_h)\,S^2_{y_{U_h}}\right] =$$

$$= \frac{N^2}{n}(1-f)\sum_{h=1}^{H}\frac{N_h}{N}\,S^2_{y_{U_h}} + \frac{N^2}{n^2}(1-f)\sum_{h=1}^{H}(1-w_h)\,S^2_{y_{U_h}} =$$

$$= \frac{N}{n}(1-f)\sum_{h=1}^{H}N_h\,S^2_{y_{U_h}} + \frac{1-f}{f^2}\sum_{h=1}^{H}(1-w_h)\,S^2_{y_{U_h}} =$$

$$= V_{STSI,\,prop}(\hat{t}_\pi) + \frac{1-f}{f^2}\sum_{h=1}^{H}(1-w_h)\,S^2_{y_{U_h}}$$

Por lo tanto, la varianza de $\hat{t}_{post}$ es igual a la varianza del $\hat{t}_\pi$ en un diseño $STSI$ con asignación proporcional, más un factor que representa el aumento de varianza debido a la aleatoriedad de los n_h.