Diseño SI

Daniel Czarnievicz

2017

Estrategia de selección

El diseño SI es un diseño de muestreo directo de elementos donde n elementos son seleccionados de una población de tamaño N sin reposición de la siguiente forma:

- En la primer extracción todos los elementos tienen una probabilidad $\dfrac{1}{N}$ de ser seleccionados.
- En la segunda extracción, los restantes $N-1$ elementos tiene una probabilidad $\dfrac{1}{N-1}$ de ser seleccionados.
- En la n-ésima extracción, los restantes $N-(n-1)$ elementos tiene una probabilidad $\dfrac{1}{N-1}$ de ser seleccionados.

Cualquier secuencia ordenada de elementos tiene una probabilidad $\dfrac{(N-n)!}{N!}$ de ser seleccionada. Una secuencia especifica, s, de elementos tiene $n!$ formas distintas de ser seleccionada. Por lo tanto, el diseño muestral es:

$$p(s) = \Pr(S=s) = \begin{cases} \dfrac{1}{\binom{N}{n}} & \text{si } s \text{ tiene } n \text{ elementos} \\ \\ 0 & \text{en otro caso} \end{cases}$$

Probabilidades de inclusión

$$\star\ \pi_k = \Pr(k \in s) = \sum_{s \ni k} p(s) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{(N-1)!}{((N-1-(n-1))!} \frac{N!}{(N-n)!n!} =$$

$$= \frac{(N-1)!}{(N-n)!(n-1)!} \frac{(N-n)!n!}{N!} = \frac{(N-1)!}{(n-1)!} \frac{n(n-1)!}{N(N-1)!} \Rightarrow \boxed{\pi_k = \frac{n}{N} \quad \forall k \in U}$$

$$\star\ \pi_{kl} = \Pr(k;l \in s) = \sum_{s \ni k;l} p(s) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{(N-2)!}{((N-2-(n-2))!} \frac{N!}{(N-n)!n!} =$$

$$= \frac{(N-2)!}{(n-2)!} \frac{n(n-1)(n-2)!}{N(N-1)(N-2)!} = \frac{(N-2)!}{N(N-1)(N-2)!} \frac{n(n-1)(n-2)!}{(n-2)!} \Rightarrow$$

$$\Rightarrow \boxed{\pi_{kl} = \frac{n(n-1)}{N(N-1)} \quad \forall k \neq l \in U}$$

$$\star\ \Delta_{kl} = \mathbf{Cov}_{SI}(I_k; I_l) = \pi_{kl} - \pi_k\, \pi_l = \frac{n(n-1)}{N(N-1)} - \frac{n}{N} \frac{n}{N} = \frac{n}{N} \left(\frac{n-1}{N-1} - \frac{n}{N} \right) =$$

$$= \frac{n}{N} \frac{n-N}{N(N-1)} = \frac{f}{N} \frac{fN-N}{N-1} = \frac{f}{N} \frac{N(f-1)}{N-1} \Rightarrow \boxed{\Delta_{kl} = -\frac{f(1-f)}{N-1} \quad \forall k \neq l \in U}$$

$$\star\ \Delta_{kk} = \mathbf{Cov}_{SI}(I_k; I_k) = \mathbf{Var}_{SI}(I_k) = \pi_{kk} - \pi_k\, \pi_k = \pi_k - \pi_k^2 =$$

$$= \pi_k(1-\pi_k) \Rightarrow \boxed{\Delta_{kk} = f(1-f) \ \ \forall k \in U}$$

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

$$\star\ \hat{t}_\pi = \sum\nolimits_s y_k^{\checkmark} = \sum\nolimits_s \frac{y_k}{\pi_k} = \frac{N}{n}\sum\nolimits_s y_k \Rightarrow \boxed{\hat{t}_\pi = N\,\bar{y}_s}$$

$$\star\ \mathbf{E}_{SI}(\hat{t}_\pi) = \mathbf{E}_{SI}\left(\sum\nolimits_s y_k^{\checkmark}\right) = \mathbf{E}_{SI}\left(\sum\nolimits_s \frac{y_k}{\pi_k}\right) = \sum\nolimits_U \mathbf{E}_{SI}(I_k)\frac{y_k}{\pi_k} = \sum\nolimits_U \pi_k\,\frac{y_k}{\pi_k} = \sum\nolimits_U y_k = t_y$$

$$\star\ \mathbf{Var}_{SI}(\hat{t}_\pi) = -\frac{1}{2}\sum\sum\nolimits_U \Delta_{kl}\left(y_k^{\checkmark} - y_l^{\checkmark}\right)^2 = -\frac{1}{2}\left(-\frac{f(1-f)}{N-1}\right)\sum\sum\nolimits_U \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 =$$

$$= \frac{f(1-f)}{2(N-1)}\,\frac{1}{\pi_k^2}\sum\sum\nolimits_U \left(y_k - \bar{y}_U + \bar{y}_U - y_l\right)^2 = \frac{1-f}{2f(N-1)}\sum\sum\nolimits_U \left[(y_k - \bar{y}_U) - (\bar{y}_U - y_l)\right]^2 =$$

$$= \frac{1-f}{2f(N-1)}\sum\sum\nolimits_U \left[(y_k - \bar{y}_U)^2 - 2(y_k - \bar{y}_U)(y_l - \bar{y}_U) + (y_k - \bar{y}_U)^2\right] =$$

$$= \frac{1-f}{2f}\left[\sum\sum\nolimits_U \frac{(y_k - \bar{y}_U)^2}{N-1} + \sum\sum\nolimits_U \frac{(y_k - \bar{y}_U)(y_l - \bar{y}_U)}{N-1} + \sum\sum\nolimits_U \frac{(y_l - \bar{y}_U)^2}{N-1}\right] =$$

$$= \frac{1-f}{2f}\left[\sum\sum\nolimits_U S_{y_U}^2 - \frac{1}{N-1}\Big[\underbrace{\sum\nolimits_U (y_k - \bar{y}_U)}_{\sum_U y_k - \sum_U \bar{y}_U}\Big]\Big[\underbrace{\sum\nolimits_U (y_l - \bar{y}_U)}_{\sum_U y_l - \sum_U \bar{y}_U}\Big] + \sum\nolimits_U S_{y_U}^2\right] =$$

$$= \frac{1-f}{2f}\left[N\,S_{y_U}^2 - \frac{1}{N-1}\Big[\underbrace{\sum\nolimits_U y_k - \sum\nolimits_U \bar{y}_U}_{N\bar{y}_U - N\bar{y}_U = 0}\Big]\Big[\underbrace{\sum\nolimits_U y_l - \sum\nolimits_U \bar{y}_U}_{N\bar{y}_U - N\bar{y}_U = 0}\Big] + N\,S_{y_U}^2\right] =$$

$$= \frac{1-f}{2f}\left(2NS_{y_U}^2\right) = \frac{N}{f}(1-f)S_{y_U}^2 \Rightarrow \boxed{\mathbf{Var}_{SI}(\hat{t}_\pi) = \frac{N^2}{n}(1-f)S_{y_U}^2}$$

$$\star\ \hat{\mathbf{Var}}_{SI}(\hat{t}_\pi) = \frac{N^2}{n}(1-f)S_{y_s}^2 \quad \text{donde } S_{y_s}^2 = \frac{1}{n-1}\sum\nolimits_s (y_k - \bar{y}_s)^2 \text{ se construye para}$$

$$\text{ser insesgado de } S_{y_U}^2 = \frac{1}{N-1}\sum\nolimits_U (y_k - \bar{y}_U)^2$$

$$\star\ \mathbf{E}_{SI}\left(\hat{V}_{SI}(\hat{t}_\pi)\right) = \mathbf{E}_{SI}\left(\frac{N^2}{n}(1-f)S_{y_s}^2\right) = \frac{N^2}{n}(1-f)\mathbf{E}_{SI}\left(S_{y_s}^2\right) = \frac{N^2}{n}(1-f)S_{y_U}^2 = \mathbf{Var}_{SI}(\hat{t}_\pi)$$

El estimador $\hat{\bar{y}}_{U_\pi}$

$$\star\ \hat{\bar{y}}_{U_\pi} = \frac{\hat{t}_\pi}{N} = \frac{N\bar{y}_s}{N} \Rightarrow \boxed{\hat{\bar{y}}_{U_\pi} = \bar{y}_s}$$

$$\star\ \mathbf{E}_{SI}(\hat{\bar{y}}_{U_\pi}) = \mathbf{E}_{SI}\left(\frac{\hat{t}_\pi}{N}\right) = \frac{1}{N}\mathbf{E}_{SI}(\hat{t}_\pi) = \frac{1}{N}\,t_y = \bar{y}_U$$

$$\star\ \mathbf{Var}_{SI}(\bar{y}_s) = \mathbf{Var}_{SI}\left(\frac{\hat{t}_\pi}{N}\right) = \frac{1}{N^2}\mathbf{Var}_{SI}(\hat{t}_\pi) = \frac{1}{N^2}\frac{N^2}{n}(1-f)S_{y_U}^2 \Rightarrow \boxed{\mathbf{Var}_{SI}(\bar{y}_s) = \frac{1}{n}(1-f)S_{y_U}^2}$$

$$\star\ \hat{\mathbf{Var}}_{SI}(\bar{y}_s) = \frac{1}{n}(1-f)S_{y_s}^2$$

$$\star\ \mathbf{E}_{SI}\left(\hat{V}_{SI}(\bar{y}_s)\right) = \mathbf{E}_{SI}\left(\frac{1}{n}(1-f)S_{y_s}^2\right) = \frac{1}{n}(1-f)\mathbf{E}_{SI}(S_{y_s}^2) = \frac{1}{n}(1-f)S_{y_s}^2 = \mathbf{Var}_{SI}(\bar{y}_s)$$

Estimación de una razón

Considérese un diseño SI con $n = f\,N$, y se desea estimar la razón $R = \dfrac{t_y}{t_z}$ mediante el estimador $\hat{R} = \dfrac{\hat{t}_{y\,\pi}}{\hat{t}_{z\,\pi}}$. Luego entonces, utilizando la linealización de Taylor:

$$\star\ \hat{R} \doteq \hat{R}_0 = R + \frac{1}{t_z}\sum\nolimits_s \frac{y_k - R\,z_k}{{}^n/_N} = R + \frac{1}{t_z}\Big(\hat{t}_{y\,\pi} - R\,\hat{t}_{z\,\pi}\Big) = R + \frac{1}{\bar{z}_U}\Big(\bar{y}_s - R\,\bar{z}_s\Big)$$

$$\star\ \mathbf{AVar}_{SI}\big(\hat{R}\big) = \frac{1}{t_z^2}\left[\frac{N^2}{n}(1-f)S^2_{(y-R\,z)_U}\right]$$

$$\text{donde } S^2_{(y-R\,z)_U} = \frac{1}{N-1}\sum\nolimits_U \big(y_k - R\,z_k\big)^2 = S^2_{y_U} + R^2\,S^2_{z_U} - 2\,R\,S_{yz_U}$$

$$\star\ \hat{\mathbf{Var}}_{SI}\big(\hat{R}\big) = \frac{1}{\hat{t}^2_{z\,\pi}}\,\frac{N^2}{n}(1-f)S^2_{(y-\hat{R}\,z)_s} = \frac{1}{\bar{z}_s^2}\,\frac{1}{n}(1-f)S^2_{(y-\hat{R}\,z)_s}$$

$$\text{donde } S^2_{(y-R\,z)_s} = \frac{1}{n-1}\sum\nolimits_s \big(y_k - \hat{R}\,z_k\big)^2 = S^2_{y_s} + \hat{R}^2\,S^2_{z_s} - 2\,\hat{R}\,S_{yz_s}$$

$$\text{y } S_{yz_s} = \frac{1}{n-1}\sum\nolimits_s \big(y_k - \bar{y}_s\big)\big(z_k - \bar{z}_s\big)$$

Las anteriores se cumplen dado que:

- $\sum\nolimits_U \big(y_k - R\,z_k\big) = t_y - R\,t_z = t_y - t_y = 0$
- $\sum\nolimits_s \big(y_k - \hat{R}\,z_k\big) = \hat{t}_{y\,\pi} - \hat{R}\,\hat{t}_{z\,\pi} = \hat{t}_{y\,\pi} - \hat{t}_{y\,\pi} = 0$

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

Supongamos que en un muestreo bajo diseño SI con $n = f\,N$, se cuenta con la variable auxiliar z. Se puede entonces utilizar el estimar $\hat{t}_{yra}$:

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

$$\star\ \mathbf{AVar}_{SI}\big(\hat{t}_{yra}\big) = t_z^2\,\mathbf{Var}_{SI}\big(\hat{R}\big) = \frac{N^2}{n}(1-f)S^2_{(y-R\,z)_U} = \frac{N^2}{n}(1-f)\Big[S^2_{y_U} + R^2\,S^2_{z_U} - 2\,R\,S_{yz_U}\Big]$$

$$\star\ \hat{\mathbf{Var}}_{SI}\big(\hat{t}_{yra}\big) = t_z^2\,\hat{\mathbf{Var}}_{SI}\big(\hat{R}\big) = \frac{N^2}{n}(1-f)S^2_{(y-\hat{R}\,z)_s} = \frac{N^2}{n}(1-f)\Big[S^2_{y_s} + \hat{R}^2\,S^2_{z_s} - 2\,\hat{R}\,S_{yz_s}\Big]$$

Comparamos las varianzas de $\hat{t}_\pi$ y $\hat{t}_{yra}$, ya que $\hat{t}_\pi$ es insesgado y $\hat{t}_{yra}$ es aproximadamente insesgado:

$$\mathbf{Var}_{SI}\big(\hat{t}_\pi\big) - \mathbf{Var}_{SI}\big(\hat{t}_{yra}\big) = \frac{N^2}{n}(1-f)S^2_{y_U} - \frac{N^2}{n}(1-f)\Big[S^2_{y_U} + R^2\,S^2_{z_U} - 2\,R\,S_{yz_U}\Big] =$$

$$= -\frac{N^2}{n}(1-f)\Big[R^2\,S^2_{z_U} - 2\,R\,S_{yz_U}\Big]$$

$$\text{Luego}\quad \mathbf{Var}_{SI}\big(\hat{t}_\pi\big) \geq \mathbf{Var}_{SI}\big(\hat{t}_{yra}\big) \Leftrightarrow R^2\,S^2_{z_U} - 2\,R\,S_{yz_U} \leq 0 \Leftrightarrow \frac{t_y}{t_z}\,S^2_{z_U} - 2\,r_{yz_U}\,S_{y_U}\,S_{y_U} \leq 0 \Leftrightarrow$$

$$\Leftrightarrow 2\,r_{yz_U} \geq \frac{t_y}{t_z}\frac{S_{z_U}}{S_{y_U}} \Leftrightarrow 2\,r_{yz_U} \geq \frac{CV_{z_U}}{CV_{y_U}} \Leftrightarrow r_{yz_U} \geq \frac{CV_{z_U}}{2\,CV_{y_U}}$$

Esto implica que si $CV_{z_U} \doteq CV_{y_U}$, $\hat{t}_{yra}$ será ventajoso $\Leftrightarrow r_{yz_U} \geq {}^1/_2 \Leftrightarrow r_{yz_U} \geq {}^1/_4 \Leftrightarrow R^2 \geq {}^1/_4$ en la regresión $y_k = \beta\,z_k + \varepsilon_k$.

Tamaño muestral

Dado que el diseño SI es de tamaño fijo n, se cumple que:

$$\star\ n_S = \sum_U I_k = \sum_U \pi_k$$

$$\star\ \mathbf{E}_{SI}\left(n_S\right) = \mathbf{E}_{SI}\left(\sum_U I_k\right) = \sum_U \mathbf{E}_{SI}(I_k) = \sum_U \pi_k = \sum_U \frac{n}{N} = \frac{n}{N}\sum_U 1 = \frac{n}{N}\, N = n$$

$$\star\ \sum_{k\neq l}\sum_U \pi_{kl} = \sum\sum_U \frac{n(n-1)}{N(N-1)} = \frac{n(n-1)}{N(N-1)}\sum\sum_U 1 = \frac{n(n-1)}{N(N-1)}\,(N-1) = n(n-1)$$

$$\star\ \mathbf{Var}_{SI}\left(n_S\right) = \sum\sum_U \pi_k - \left(\sum\sum_U \pi_k\right)^2 + \sum_{k\neq l}\sum_U \pi_{kl} = n - n^2 + \sum_{k\neq l}\sum_U \pi_{kl} =$$

$$= n(1-n) + n(n-1) = n\Big[\underbrace{(1-n)+(n-1)}_{=0}\Big] \Rightarrow \boxed{\mathbf{Var}_{SI}(n_S) = 0}$$

Dado un nivel de precisión, ε, y una confianza, $1-\alpha$, n se determina mediante:

$$\star\ \varepsilon^2 \doteq z^2_{1-\alpha/2}\mathbf{Var}_{SI}(\hat{t}_y) = z^2_{1-\alpha/2}\frac{N^2}{n}(1-f)S^2_{y_U} \Rightarrow \boxed{n = \frac{z^2_{1-\alpha/2}N^2S^2_{y_U}}{\varepsilon^2 + z^2_{1-\alpha/2}N^2S^2_{y_U}}}$$

Si en cambio se trabajase con el CV_{y_U}, entonces:

$$\star\ \varepsilon = z_{1-\alpha/2}\, CV_{y_U}\sqrt{\frac{1}{n}-\frac{1}{N}} \Rightarrow \frac{\varepsilon^2}{z^2_{1-\alpha/2}\, CV^2_{y_U}} = \frac{1}{n} - \frac{1}{N} = \frac{N-n}{n\,N} \Rightarrow$$

$$\Rightarrow \frac{z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2} = \frac{nN}{N-n} \Rightarrow (N-n)\left(\frac{z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2}\right) = N\,n \Rightarrow$$

$$\Rightarrow \frac{N\,z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2} - \frac{n\,z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2} = n\,N \Rightarrow \frac{N\,z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2} = n\,N + \frac{n\,z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2} \Rightarrow$$

$$\Rightarrow \frac{N\,z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2} = n\left[N + \frac{z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2}\right] = n\left[\frac{N\,\varepsilon^2 + z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2}\right] \Rightarrow$$

$$\Rightarrow n = \frac{N\,z^2_{1-\alpha/2}\, CV^2_{y_U}}{\varepsilon^2}\,\frac{\varepsilon^2}{N\,\varepsilon^2 + z^2_{1-\alpha/2}\, CV^2_{y_U}} \Rightarrow \boxed{n = \frac{N\,z^2_{1-\alpha/2}\, CV^2_{y_U}}{N\,\varepsilon^2 + z^2_{1-\alpha/2}\, CV^2_{y_U}}}$$