4.2 Teoría del muestreo aleatorio estratificado.

Sea Ω una población finita de N elementos y $\{\Omega_1, \Omega_2, \ldots, \Omega_m\}$ una partición de Ω. A cada $\Omega_i = \left\{y_{i1}, y_{i2}, \ldots, y_{iN_i}\right\}$ de tamaño N_i, con $i = 1, \ldots, m$, se le llama **estrato de la población**.

Luego, para cada $i = 1, \ldots, m$, sea $\mathcal{S}_i$ una muestra aleatoria obtenida del estrato Ω_i de tamaño n_i. Los estadísticos de la población son los siguientes:

$$t_i = \sum_{j=1}^{N_i} y_{ij} \quad \text{es el total de la población en el estrato } \Omega_i \text{ para cada } i = 1, \ldots, m$$

$$t = \sum_{i=1}^{m} t_i \quad \text{es el total de la población total}$$

$$\overline{y_{Ui}} = \frac{t_i}{N_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \quad \text{es la media de la población en el estrato } \Omega_i \text{ para cada } i = 1, \ldots, m$$

$$\overline{y_U} = \frac{t}{N} = \frac{1}{N} \sum_{i=1}^{m} t_i = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{N_i} y_{ij} \quad \text{es la media de la población total}$$

$$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(y_{ij} - \overline{y_{Ui}}\right)^2 \quad \text{es la varianza de la población en el estrato } \Omega_i \text{ para cada } i = 1, \ldots, m$$

Como en cada estrato se hace un muestreo aleatorio simple, se tiene que

1. $\overline{y_i} = \dfrac{1}{n_i} \displaystyle\sum_{j \in \mathcal{S}_i} y_{ij}$ es estimador insesgado de $\overline{y_{Ui}}$ para cada $i = 1, \ldots, m$.
2. $\widehat{t_i} = N_i \overline{y_i} = \dfrac{N_i}{n_i} \displaystyle\sum_{j \in \mathcal{S}_i} y_{ij}$ es estimador insesgado de t_i para cada $i = 1, \ldots, m$.
3. $\widehat{S_i^2} = \dfrac{1}{n_i - 1} \displaystyle\sum_{j \in \mathcal{S}_i} \left(y_{ij} - \overline{y_i}\right)^2$ es estimador insesgado de S_i^2 para cada $i = 1, \ldots, m$.

Proposición.

1. $\widehat{t} = \displaystyle\sum_{i=1}^{m} \widehat{t_i} = \sum_{i=1}^{m} N_i \overline{y_i}$ es estimador insesgado de t.
2. $\overline{y} = \dfrac{\widehat{t}}{N} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{m} t_i$ es estimador insesgado de $\overline{y_U}$.
3. $\mathrm{V}\left(\overline{y}\right) = \displaystyle\sum_{i=1}^{m} \left(1 - \frac{n_i}{N_i}\right) \left(\frac{N_i}{N}\right)^2 \frac{\widehat{S_i^2}}{n_i}$.

Demostración.

1. Se tiene que

$$\mathrm{E}\left[\,\widehat{t}\,\right] = \mathrm{E}\left[\sum_{i=1}^{m} \widehat{t_i}\right] = \sum_{i=1}^{m} \mathrm{E}\left[\,\widehat{t_i}\,\right] = \sum_{i=1}^{m} t_i = t$$

Por lo tanto, $\widehat{t}$ es un estimador insesgado de t.

2. Por el inciso anterior, se obtiene que

$$\mathrm{E}\left[\overline{y}\right] = \mathrm{E}\left[\frac{\widehat{t}}{N}\right] = \frac{1}{N}\mathrm{E}\left[\,\widehat{t}\,\right] = \frac{t}{N} = \overline{y_U}$$

Por lo tanto, $\overline{y}$ es un estimador insesgado de $\overline{y_U}$.

3. Se da que

$$\begin{aligned}
\mathrm{V}\left(\overline{y}\right) &= \mathrm{V}\left(\frac{1}{N}\sum_{i=1}^{m} t_i\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{m} \mathrm{V}\left(t_i\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{m} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{\widehat{S_i^2}}{n_i} \\
&= \sum_{i=1}^{m}\left(1-\frac{n_i}{N_i}\right)\left(\frac{N_i}{N}\right)^2\frac{\widehat{S_i^2}}{n_i}
\end{aligned}$$

Si los tamaños de las muestras dentro de los estratos son grandes o la cantidad de estratos es grande, entonces por el teorema del límite central, se tiene que

$$\frac{\overline{y}-\overline{y_U}}{\sqrt{\mathrm{V}\left(\overline{y}\right)}} \sim \mathrm{N}(0,1)$$

De esta manera, un intervalo de confianza de $(1-\alpha)100\,\%$ para $\overline{y_U}$ se puede calcular como:

$$\begin{aligned}
-z_{\frac{\alpha}{2}} \leq \frac{\overline{y}-\overline{y_U}}{\sqrt{\mathrm{V}\left(\overline{y}\right)}} &\leq z_{\frac{\alpha}{2}} \\
-z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}\left(\overline{y}\right)} \leq \overline{y}-\overline{y_U} &\leq z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}\left(\overline{y}\right)} \\
\overline{y}-z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}\left(\overline{y}\right)} \leq \overline{y_U} &\leq \overline{y}+z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}\left(\overline{y}\right)}
\end{aligned}$$

Así, $\left(\overline{y}-z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}\left(\overline{y}\right)}, \overline{y}+z_{\frac{\alpha}{2}}\sqrt{\mathrm{V}\left(\overline{y}\right)}\right)$ es el intervalo de confianza de $(1-\alpha)100\,\%$ para $\overline{y_U}$.