**4.2 Teoría del muestreo aleatorio estratificado.**

Sea $\Omega$ una población finita de $N$ elementos y $\{\Omega_1, \Omega_2, \ldots, \Omega_m\}$ una partición de $\Omega$. A cada $\Omega_i = \{y_{i1}, y_{i2}, \ldots, y_{iN_i}\}$ 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 $\Omega_i$ de tamaño $n_i$. Los estadísticos de la población son los siguientes:

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

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

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

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

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

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

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

**Proposición.**

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

**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 un intervalo de confianza de $(1-\alpha)100\,\%$ para $\overline{y_U}$.

**Ejemplo.**

En enero y a principios de febrero de 1962, se realizaron pruebas de campo para estimar el tamaño de la manada Nelchina del caribú de Alaska. Dichas pruebas indicaron a los investigadores que algunas unidades de muestreo propuestas, como "igual tiempo de vuelo", eran difíciles de establecer en la práctica y que una unidad de muestreo conocida como "igual área" de 4 millas cuadradas serviría para este estudio. Los biólogos emplearon las estimaciones preliminares de las densidades del

caribú para dividir el área de interés en 6 estratos; cada estrato se dividía, entonces, en una retícula de unidades de muestreo de 4 millas cuadradas. Se obtuvieron los siguientes datos y estimaciones:

| Estrato | $N_i$ | $n_i$ | $\overline{y_i}$ | $\widehat{s_i^2}$ | $\widehat{t_i} = N_i\overline{y_i}$ | $\mathrm{V}(t_i) = N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{\widehat{S_i^2}}{n_i}$ |
|---|---|---|---|---|---|---|
| A | 400 | 98 | 24.1 | 5,575 | 9,640 | 6,782,040.82 |
| B | 30 | 10 | 25.6 | 4,064 | 768 | 243,840 |
| C | 61 | 37 | 267.6 | 347,556 | 16,324 | 13,751,945.51 |
| D | 18 | 6 | 179 | 22,798 | 3,222 | 820,728 |
| E | 70 | 39 | 293.7 | 123,578 | 20,559 | 6,876,006.67 |
| F | 120 | 21 | 33.2 | 9,795 | 3,984 | 5,541,171.43 |
| Total | 699 | 211 | | | 54,497 | 34,105,732.43 |

Por lo tanto, la población estimada de caribús es de 54,497 con una desviación estándar de $\sqrt{34,105,732.43} = 5840.01$. Además, un intervalo de confianza del 95 % para el número total de caribús es

$$54,497 \pm 1{,}96(5840.01) = 54,497 \pm 11,446.42$$

es decir, $(43050{,}58, 65943{,}42)$.

**Muestreo aleatorio estratificado para proporciones.**

Si se trabaja con una proporción, entonces $\widehat{p_i} = \overline{y_i}$ es un estimador insesgado para $p_i = \overline{y_{Ui}}$, para cada $i = 1, \ldots, m$. Además,

$$\widehat{S_i^2} = \frac{n_i}{n_i - 1}\widehat{p_i}\left(1-\widehat{p_i}\right) \quad \text{para cada } i = 1, \ldots, m,$$

$$\widehat{p} = \overline{y} = \frac{\widehat{t}}{N} = \frac{1}{N}\sum_{i=1}^{m}\widehat{t_i} = \frac{1}{N}\sum_{i=1}^{m}N_i\widehat{p_i} \quad \text{es estimador insesgado de } p = \overline{y_U},$$

$$\mathrm{V}\left(\widehat{p}\right) = \sum_{i=1}^{m}\left(1-\frac{n_i}{N_i}\right)\left(\frac{N_i}{N}\right)^2\frac{\frac{n_i}{n_i-1}\widehat{p_i}\left(1-\widehat{p_i}\right)}{n_i}$$

$$= \sum_{i=1}^{m}\left(1-\frac{n_i}{N_i}\right)\left(\frac{N_i}{N}\right)^2\frac{\widehat{p_i}\left(1-\widehat{p_i}\right)}{n_i-1}$$

$$\widehat{t} = \sum_{i=1}^{m}N_i\widehat{p_i} = N\widehat{p} \quad \text{y}$$

$$\mathrm{V}\left(\widehat{t}\right) = N^2\mathrm{V}\left(\widehat{p}\right) = N^2\sum_{i=1}^{m}\left(1-\frac{n_i}{N_i}\right)\left(\frac{N_i}{N}\right)^2\frac{\widehat{p_i}\left(1-\widehat{p_i}\right)}{n_i-1} = \sum_{i=1}^{m}N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{\widehat{p_i}\left(1-\widehat{p_i}\right)}{n_i-1}$$

**Ejemplo.**

El American Council of Learned Societies (ACLS) usó una muestra aleatoria estratificada de algunas sociedades ACLS en siete disciplinas, para estudiar los patrones de publicación y el uso de computadoras y bibliotecas entre los estudiosos. Los datos son los siguientes:

| Disciplina | Membresía ($N_i$) | Respuestas válidas ($n_i$) | Mujeres ( %) |
|---|---|---|---|
| Literatura | 9100 | 636 | 38 |
| Clásicos | 1950 | 451 | 27 |
| Filosofía | 5500 | 481 | 18 |
| Historia | 10850 | 611 | 19 |
| Linguística | 2100 | 493 | 36 |
| Ciencias políticas | 5500 | 575 | 13 |
| Sociología | 9000 | 588 | 26 |
| Totales | 44000 | 5835 | |

De esta manera,

$$\widehat{p} = \frac{1}{N}\sum_{i=1}^{7} N_i\,\widehat{p_i} = 0{,}2465 \quad \text{y}$$

$$\mathrm{V}\left(\widehat{t}\right) = \sum_{i=1}^{m}\left(1-\frac{n_i}{N_i}\right)\left(\frac{N_i}{44000}\right)^2 \frac{\widehat{p_i}\,\left(1-\widehat{p_i}\right)}{n_i-1} = 0{,}00005041$$

Por lo tanto, el número total estimado de mujeres que pertenecen a las sociedades es $44000 \cdot 0.2465 = 10847$, con una desviación estándar de $44000 \cdot \sqrt{0{,}00005041} = 312$.