Ejercicios del Sarndal

Pedro Leal

Problema 2.1

En la planificación de un estudio de red de oficinas, se propuso el siguiente esquema de muestreo secuencial para seleccionar una muestra aleatoria de dos intervalos horarios no adyacentes de los ocho intervalos 9–10, 10–11, ..., 16–17 (etiquetados 1, 2, ..., 8):

Seleccionar el primer intervalo con probabilidad uniforme de los ocho intervalos.

Seleccionar, sin reemplazo, el segundo intervalo con probabilidad uniforme de los intervalos no adyacentes al seleccionado en el primer paso.

a) Determine las probabilidades de inclusión de primer orden.

b) Determine las probabilidades de inclusión de segundo orden. ¿Es el diseño inducido por el esquema de muestreo medible?

c) Determine las covarianzas de los indicadores de pertenencia a la muestra.

d) Verifique que el Resultado 2.6.2 se cumple en esta aplicación.

Solución 2.1

(a) Probabilidades de inclusión de primer orden

La probabilidad π_i de que el intervalo i esté en la muestra es:

$$\pi_i = P(\text{seleccionar } i \text{ primero}) + P(\text{seleccionar } i \text{ segundo})$$

La probabilidad de seleccionar el intervalo i-esimo para todo i es $\frac{1}{8}$ dado que distribuye como uniforme, pero para ver las probabilidades de que el i-esimo sea escogido de segundas cambia y para ello es mas claro por medio de la siguiente tabla

Cuadro 1: Probabilidad de selección en segunda etapa para cada i

i	Conjunto de posibles j dado i	$P(\text{de cualquier } j \in A_j)$
1	$A_j = \{3,4,5,6,7,8\}$	$\frac{1}{6}$
2	$A_j = \{4,5,6,7,8\}$	$\frac{1}{5}$
3	$A_j = \{1,6,7,8\}$	$\frac{1}{5}$
4	$A_j = \{1,2,6,7,8\}$	$\frac{1}{5}$
5	$A_j = \{1,2,3,7,8\}$	$\frac{1}{5}$
6	$A_j = \{1,2,3,4,8\}$	$\frac{1}{5}$
7	$A_j = \{1,2,3,4,5\}$	$\frac{1}{5}$
8	$A_j = \{1,2,3,4,5,6\}$	$\frac{1}{6}$

Para el caso del primer intervalo se tiene

$$\begin{aligned}\pi_1 &= \underbrace{\frac{1}{8}}_{P(1)} + \underbrace{0 \cdot \frac{1}{5}}_{P(1|2)} + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(1|3)} + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(1|4)} + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(1|5)} + + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(1|6)} + + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(1|7)} + \underbrace{\frac{1}{8}\cdot\frac{1}{6}}_{P(1|8)} \\ &= \frac{1}{8}\left(1 + \frac{5}{5} + \frac{1}{6}\right) = \frac{1}{8}\cdot\frac{13}{6} = \frac{13}{48}\end{aligned}$$

En el caso del segundo,

$$\begin{aligned}\pi_2 &= \underbrace{\frac{1}{8}}_{P(2)} + \underbrace{0\cdot\frac{1}{6}}_{P(2|1)} + \underbrace{0\cdot\frac{1}{5}}_{P(2|3)} + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(2|4)} + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(2|5)} + + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(2|6)} + + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(2|7)} + \underbrace{\frac{1}{8}\cdot\frac{1}{6}}_{P(2|8)} \\ &= \frac{1}{8}\left(1+\frac{4}{5}+\frac{1}{6}\right) = \frac{1}{8}\cdot\frac{59}{30} = \frac{59}{240}\end{aligned}$$

Y en el ultimo caso especifico,

$$\begin{aligned}\pi_3 &= \underbrace{\frac{1}{8}}_{P(3)} + \underbrace{\frac{1}{8}\cdot\frac{1}{6}}_{P(3|1)} + \underbrace{0\cdot\frac{1}{5}}_{P(3|2)} + \underbrace{0\cdot\frac{1}{5}}_{P(3|4)} + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(3|5)} + + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(3|6)} + + \underbrace{\frac{1}{8}\cdot\frac{1}{5}}_{P(3|7)} + \underbrace{\frac{1}{8}\cdot\frac{1}{6}}_{P(3|8)} \\ &= \frac{1}{8}\left(1+\frac{3}{5}+\frac{2}{6}\right) = \frac{1}{8}\cdot\frac{58}{30} = \frac{58}{240}\end{aligned}$$

Observando la simetría del problema se puede afirmar:

$$\begin{aligned}\pi_1 = \pi_8 &= \frac{13}{48} \\ \pi_2 = \pi_7 &= \frac{59}{240} \\ \pi_3 = \pi_4 = \pi_5 = \pi_6 &= \frac{58}{240}\end{aligned}$$

Note que

$$\sum_{i\in s}\pi_i = 2\left(\frac{13}{48}\right) + 2\left(\frac{59}{240}\right) + 4\left(\frac{58}{240}\right) = 2 = N$$

(b) Probabilidades de inclusión de segundo orden

$$\pi_{ij} = P(\text{seleccionar } i \text{ y } j) = \frac{1}{8}\left(\frac{1}{k_i}+\frac{1}{k_j}\right)$$

donde k_i = número de intervalos no adyacentes a i.

Casos:

- **Pares adyacentes:** $\pi_{ij} = 0$ (ej. π_{12}).
- **Pares no adyacentes:**
  - **Bordes entre sí (1 y 8):**
  $$\pi_{18} = \frac{1}{8}\left(\frac{1}{6}+\frac{1}{6}\right) = \frac{1}{24} \approx 0{,}0417$$
  - **Borde y central (1 y 3):**
  $$\pi_{13} = \frac{1}{8}\left(\frac{1}{6}+\frac{1}{5}\right) = \frac{11}{240} \approx 0{,}0458$$
  - **Centrales entre sí (3 y 5):**
  $$\pi_{35} = \frac{1}{8}\left(\frac{1}{5}+\frac{1}{5}\right) = \frac{1}{20} = 0{,}05$$

Matriz completa π_{ij}:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	π_1	0	$\frac{11}{240}$	$\frac{11}{240}$	$\frac{11}{240}$	$\frac{11}{240}$	$\frac{11}{240}$	$\frac{1}{24}$
2	0	π_2	0	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{11}{240}$
3	$\frac{11}{240}$	0	π_3	0	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{11}{240}$
4	$\frac{11}{240}$	$\frac{1}{20}$	0	π_4	0	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{11}{240}$
5	$\frac{11}{240}$	$\frac{1}{20}$	$\frac{1}{20}$	0	π_5	0	$\frac{1}{20}$	$\frac{11}{240}$
6	$\frac{11}{240}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	0	π_6	0	$\frac{11}{240}$
7	$\frac{11}{240}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{20}$	0	π_7	0
8	$\frac{1}{24}$	$\frac{11}{240}$	$\frac{11}{240}$	$\frac{11}{240}$	$\frac{11}{240}$	$\frac{11}{240}$	0	π_8

Medibilidad:

El diseño es **no es medible** porque las probabilidades de inclusiones de intervalos adyacentes son nulos, es decir $\pi_{i,i\pm1} = 0$

(c) Covarianzas de los indicadores de pertenencia a la muestra

La covarianza entre los indicadores I_i e I_j se calcula como:

$$\mathrm{Cov}(I_i, I_j) = \pi_{ij} - \pi_i \pi_j$$

Matriz de covarianzas completa:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0,1975	−0,0666	−0,0654	−0,0654	−0,0654	−0,0654	−0,0654	−0,0113
2	−0,0666	0,0605	0,0000	−0,0123	−0,0123	−0,0123	−0,0123	−0,0666
3	−0,0654	0,0000	0,0584	0,0000	−0,0584	−0,0584	−0,0584	−0,0654
4	−0,0654	−0,0123	0,0000	0,0584	0,0000	−0,0584	−0,0584	−0,0654
5	−0,0654	−0,0123	−0,0584	0,0000	0,0584	0,0000	−0,0584	−0,0654
6	−0,0654	−0,0123	−0,0584	−0,0584	0,0000	0,0584	0,0000	−0,0654
7	−0,0654	−0,0123	−0,0584	−0,0584	−0,0584	0,0000	0,0605	−0,0666
8	−0,0113	−0,0666	−0,0654	−0,0654	−0,0654	−0,0654	−0,0666	0,1975

Explicación de los valores clave:

- **Diagonal principal (Varianzas):**

$$\mathrm{Var}(I_i) = \pi_i(1 - \pi_i)$$

Ejemplo para $i = 1$:

$$\text{Var}(I_1) = \frac{13}{48}\left(1 - \frac{13}{48}\right) = \frac{455}{2304} \approx 0{,}197$$

- **Pares adyacentes ($|i - j| = 1$):**

$$\text{Cov}(I_i, I_j) = -\pi_i \pi_j$$

Ejemplo para $(1, 2)$:

$$\text{Cov}(I_1, I_2) = -\frac{13}{48} \times \frac{59}{240} = -\frac{767}{11520} \approx -0{,}0666$$

- **Pares no adyacentes ($|i - j| > 1$):**

$$\text{Cov}(I_i, I_j) = \pi_{ij} - \pi_i \pi_j$$

Ejemplo para $(1, 3)$:

$$\text{Cov}(I_1, I_3) = \frac{11}{240} - \left(\frac{13}{48} \times \frac{29}{120}\right) = -\frac{377}{5760} \approx -0{,}0654$$

Note que la suma total de covarianzas satisface:

$$\sum_{i=1}^{8} \sum_{j=1}^{8} \text{Cov}(I_i, I_j) = 0$$

(d) Cálculo de Covarianza y Varianza por Separado

Calculamos $\text{Cov}(I_k, I_k)$ y $\text{Var}(I_k)$ por separado para verificar que sean iguales en la diagonal.

Fórmulas

Para una variable indicadora I_k:

$$\text{Cov}(I_k, I_k) = \pi_{k,k} - \pi_k \pi_k = \pi_k - \pi_k^2 = \pi_k(1 - \pi_k) = \text{Var}(I_k)$$

Cálculos específicos

- Para $k = 1$ y $k = 8$ ($\pi_1 = \pi_8 = \frac{13}{48}$):

$$\text{Cov}(I_1, I_1) = \frac{13}{48}\left(1 - \frac{13}{48}\right) = \frac{13}{48} \cdot \frac{35}{48} = \frac{455}{2304}$$

$$\text{Var}(I_1) = \frac{13}{48}\left(1 - \frac{13}{48}\right) = \frac{455}{2304}$$

$$\text{Cov}(I_1, I_1) = \text{Var}(I_1) = \frac{455}{2304}$$

- Para $k = 2$ y $k = 7$ ($\pi_2 = \pi_7 = \frac{59}{240}$):

$$\text{Cov}(I_2, I_2) = \frac{59}{240}\left(1 - \frac{59}{240}\right) = \frac{59}{240} \cdot \frac{181}{240} = \frac{10679}{57600}$$

$$\text{Var}(I_2) = \frac{59}{240}\left(1 - \frac{59}{240}\right) = \frac{10679}{57600}$$

$$\text{Cov}(I_2, I_2) = \text{Var}(I_2) = \frac{10679}{57600}$$

- Para $k = 3, 4, 5, 6$ ($\pi_3 = \pi_4 = \pi_5 = \pi_6 = \frac{29}{120}$):

$$\text{Cov}(I_3, I_3) = \frac{29}{120}\left(1 - \frac{29}{120}\right) = \frac{29}{120} \cdot \frac{91}{120} = \frac{2639}{14400}$$

$$\text{Var}(I_3) = \frac{29}{120}\left(1 - \frac{29}{120}\right) = \frac{2639}{14400}$$

$$\text{Cov}(I_3, I_3) = \text{Var}(I_3) = \frac{2639}{14400}$$

En todos los casos, $\text{Cov}(I_k, I_k) = \text{Var}(I_k) = \pi_k(1 - \pi_k)$, como se esperaba.

Problema 2.3

Considere una población U compuesta por tres subpoblaciones disjuntas U_1, U_2, y U_3 de tamaños $N_1 = 600, N_2 = 300$, y $N_3 = 100$, respectivamente. Así, U es de tamaño $N = 1,000$. Para cada elemento $k \in U$, la inclusión o no inclusión en la muestra, s, está determinada por un experimento de Bernoulli que da al elemento k la probabilidad π_k de ser seleccionado. Los experimentos son independientes.

a) Sea $\pi_k = 0{,}1$ para todo $k \in U_1$, $\pi_k = 0{,}2$ para todo $k \in U_2$, y $\pi_k = 0{,}8$ para todo $k \in U_3$. Encuentre el valor esperado y la varianza del tamaño de la muestra, n_s, bajo este diseño.

b) Suponga que π_k es constante para todo $k \in U$. Determine esta constante de modo que el valor esperado del tamaño de la muestra coincida con el tamaño de muestra esperado obtenido en el caso (a). Obtenga la varianza del tamaño de la muestra; compare con la varianza en el caso (a).

Solución 2.3

El tamaño de la muestra n_s es la suma de variables indicadoras I_k, donde $I_k = 1$ si el elemento k es seleccionado y $I_k = 0$ en caso contrario. Así, $n_s = \sum_{k \in U} I_k$. Dado que la selección de cada elemento es un experimento de Bernoulli independiente, $P(I_k = 1) = \pi_k$. El valor esperado de I_k es $E(I_k) = \pi_k$. La varianza de I_k es $V(I_k) = \pi_k(1 - \pi_k)$.

El valor esperado del tamaño de la muestra es:

$$E(n_s) = E\left(\sum_{k \in U} I_k\right) = \sum_{k \in U} E(I_k) = \sum_{k \in U} \pi_k$$

Dado que los experimentos son independientes, la varianza del tamaño de la muestra es:

$$V(n_s) = V\left(\sum_{k \in U} I_k\right) = \sum_{k \in U} V(I_k) = \sum_{k \in U} \pi_k(1 - \pi_k)$$

(a) Probabilidades de inclusión variables por subpoblación

Tenemos:

- Para $k \in U_1$: $N_1 = 600$, $\pi_k = 0{,}1$
- Para $k \in U_2$: $N_2 = 300$, $\pi_k = 0{,}2$
- Para $k \in U_3$: $N_3 = 100$, $\pi_k = 0{,}8$

El valor esperado del tamaño de la muestra $E_a(n_s)$ es:

$$\begin{aligned} E_a(n_s) &= \sum_{k \in U_1} \pi_k + \sum_{k \in U_2} \pi_k + \sum_{k \in U_3} \pi_k \\ &= N_1 \cdot 0{,}1 + N_2 \cdot 0{,}2 + N_3 \cdot 0{,}8 \\ &= 600 \cdot 0{,}1 + 300 \cdot 0{,}2 + 100 \cdot 0{,}8 \\ &= 60 + 60 + 80 \\ &= 200 \end{aligned}$$

La varianza del tamaño de la muestra $V_a(n_s)$ es:

$$\begin{aligned} V_a(n_s) &= \sum_{k \in U_1} \pi_k(1 - \pi_k) + \sum_{k \in U_2} \pi_k(1 - \pi_k) + \sum_{k \in U_3} \pi_k(1 - \pi_k) \\ &= N_1 \cdot 0{,}1(1 - 0{,}1) + N_2 \cdot 0{,}2(1 - 0{,}2) + N_3 \cdot 0{,}8(1 - 0{,}8) \\ &= 600 \cdot 0{,}1 \cdot 0{,}9 + 300 \cdot 0{,}2 \cdot 0{,}8 + 100 \cdot 0{,}8 \cdot 0{,}2 \\ &= 600 \cdot 0{,}09 + 300 \cdot 0{,}16 + 100 \cdot 0{,}16 \\ &= 54 + 48 + 16 \\ &= 118 \end{aligned}$$

Entonces, $E_a(n_s) = 200$ y $V_a(n_s) = 118$.

(b) Probabilidad de inclusión constante

Sea $\pi_k = \pi_c$ para todo $k \in U$, donde π_c es una constante. El tamaño total de la población es $N = N_1 + N_2 + N_3 = 600 + 300 + 100 = 1000$. El valor esperado del tamaño de la muestra $E_b(n_s)$ es:

$$E_b(n_s) = \sum_{k \in U} \pi_c = N \cdot \pi_c = 1000 \cdot \pi_c$$

Se nos pide que $E_b(n_s)$ sea igual a $E_a(n_s)$:

$$1000 \cdot \pi_c = 200$$
$$\pi_c = \frac{200}{1000} = 0{,}2$$

Así, la constante de probabilidad de inclusión es $\pi_c = 0{,}2$.

Ahora, calculamos la varianza del tamaño de la muestra $V_b(n_s)$ con $\pi_c = 0{,}2$:

$$\begin{aligned} V_b(n_s) &= \sum_{k \in U} \pi_c(1 - \pi_c) \\ &= N \cdot \pi_c(1 - \pi_c) \\ &= 1000 \cdot 0{,}2(1 - 0{,}2) \\ &= 1000 \cdot 0{,}2 \cdot 0{,}8 \\ &= 1000 \cdot 0{,}16 \\ &= 160 \end{aligned}$$

Entonces, $V_b(n_s) = 160$.

Comparando las varianzas: $V_a(n_s) = 118$ $V_b(n_s) = 160$ La varianza del tamaño de la muestra en el caso (b) ($V_b(n_s) = 160$) es mayor que en el caso (a) ($V_a(n_s) = 118$). Esto indica que estratificar las probabilidades de inclusión, como se hizo en (a), puede llevar a un tamaño de muestra más predecible (menor varianza) para el mismo tamaño de muestra esperado.

Problema 2.4

Una población de 1,600 individuos está dividida en 800 conglomerados (hogares) tal que hay N_a conglomerados de tamaño a ($a = 1, 2, 3, 4$) según la siguiente tabla:

a	N_a
1	250
2	350
3	150
4	50

Una muestra de individuos se selecciona de la siguiente manera: 300 conglomerados se extraen de los 800 mediante el diseño SI (Muestreo Aleatorio Simple sin reemplazo), y todos los individuos en los conglomerados seleccionados deben ser entrevistados. Si n_s denota el número total de individuos a ser entrevistados, calcule $E(n_s)$ y $V(n_s)$.

Solución 2.4

Sea M el número total de conglomerados en la población y m el número de conglomerados seleccionados. Aquí, $M = 800$ y $m = 300$. Sea M_i el tamaño (número de individuos) del conglomerado i-ésimo en la población. La población de conglomerados tiene la siguiente distribución de tamaños:

- 250 conglomerados de tamaño 1 ($M_i = 1$)
- 350 conglomerados de tamaño 2 ($M_i = 2$)
- 150 conglomerados de tamaño 3 ($M_i = 3$)
- 50 conglomerados de tamaño 4 ($M_i = 4$)

El número total de conglomerados es $250 + 350 + 150 + 50 = 800$, lo cual coincide con M. El número total de individuos en la población es $N_{total} = \sum_{i=1}^{M} M_i$:

$$N_{total} = (250 \times 1) + (350 \times 2) + (150 \times 3) + (50 \times 4) = 250 + 700 + 450 + 200 = 1600$$

Esto coincide con la información del problema.

El tamaño promedio de conglomerado en la población es $\bar{M}_U = \frac{N_{total}}{M}$:

$$\bar{M}_U = \frac{1600}{800} = 2 \text{ individuos por conglomerado}$$

El número total de individuos a ser entrevistados, n_s, es la suma de los tamaños de los m conglomerados seleccionados: $n_s = \sum_{k \in s} M_k$, donde s es la muestra de conglomerados. Para un muestreo aleatorio simple (SI) de conglomerados, el valor esperado de n_s es (Resultado 2.8.2 de Särndal et al.):

$$E(n_s) = m\bar{M}_U$$

$$E(n_s) = 300 \times 2 = 600$$

Así, el número esperado de individuos a ser entrevistados es 600.

Para calcular $V(n_s)$, primero necesitamos la varianza poblacional de los tamaños de los conglomerados, σ_M^2.

$$\sigma_M^2 = \frac{1}{M} \sum_{i=1}^{M} (M_i - \bar{M}_U)^2$$

La suma de las desviaciones cuadradas es:

$$\begin{aligned}
\sum_{i=1}^{M} (M_i - \bar{M}_U)^2 &= N_1(1 - \bar{M}_U)^2 + N_2(2 - \bar{M}_U)^2 + N_3(3 - \bar{M}_U)^2 + N_4(4 - \bar{M}_U)^2 \\
&= 250(1-2)^2 + 350(2-2)^2 + 150(3-2)^2 + 50(4-2)^2 \\
&= 250(-1)^2 + 350(0)^2 + 150(1)^2 + 50(2)^2 \\
&= 250(1) + 350(0) + 150(1) + 50(4) \\
&= 250 + 0 + 150 + 200 \\
&= 600
\end{aligned}$$

Entonces, la varianza poblacional de los tamaños de los conglomerados es:

$$\sigma_M^2 = \frac{600}{800} = \frac{3}{4} = 0{,}75$$

La varianza del número total de individuos entrevistados n_s para un diseño SI de conglomerados es (Resultado 2.8.3 de Särndal et al., adaptado para la suma muestral en lugar del promedio muestral):

$$V(n_s) = m\sigma_M^2 \frac{M - m}{M - 1}$$

Sustituyendo los valores:

$$\begin{aligned}
V(n_s) &= 300 \times 0{,}75 \times \frac{800 - 300}{800 - 1} \\
&= 225 \times \frac{500}{799} \\
&= \frac{112500}{799} \\
&\approx 140{,}801
\end{aligned}$$

Por lo tanto, $E(n_s) = 600$ y $V(n_s) \approx 140{,}801$.

Problema 2.5

Considere una población de tamaño $N = 3, U = \{1, 2, 3\}$. Sean $s_1 = \{1, 2\}, s_2 = \{1, 3\}, s_3 = \{2, 3\}, s_4 = \{1, 2, 3\}$, con $p(s_1) = 0{,}4, p(s_2) = 0{,}3, p(s_3) = 0{,}2$, y $p(s_4) = 0{,}1$.

a) Calcule todas las π_k y todas las π_{kl}.

b) Encuentre el valor de $E(n_s)$ de dos maneras: (i) mediante un cálculo directo, usando la definición y (ii) mediante el uso de la fórmula que expresa $E(n_s)$ como una función de las π_k.

Solución 2.5

La población es $U = \{1, 2, 3\}$. Las muestras posibles y sus probabilidades son:

- $s_1 = \{1, 2\}$, $p(s_1) = 0{,}4$
- $s_2 = \{1, 3\}$, $p(s_2) = 0{,}3$
- $s_3 = \{2, 3\}$, $p(s_3) = 0{,}2$
- $s_4 = \{1, 2, 3\}$, $p(s_4) = 0{,}1$

La suma de las probabilidades es $0{,}4 + 0{,}3 + 0{,}2 + 0{,}1 = 1{,}0$.

(a) Cálculo de π_k y π_{kl}

Las probabilidades de inclusión de primer orden π_k se calculan como $\pi_k = \sum_{s:k\in s} p(s)$.

$$\begin{aligned}\pi_1 &= p(s_1) + p(s_2) + p(s_4) = 0{,}4 + 0{,}3 + 0{,}1 = 0{,}8\\ \pi_2 &= p(s_1) + p(s_3) + p(s_4) = 0{,}4 + 0{,}2 + 0{,}1 = 0{,}7\\ \pi_3 &= p(s_2) + p(s_3) + p(s_4) = 0{,}3 + 0{,}2 + 0{,}1 = 0{,}6\end{aligned}$$

Las probabilidades de inclusión de segundo orden π_{kl} se calculan como $\pi_{kl} = \sum_{s:k,l\in s} p(s)$. Para $k \neq l$:

$$\begin{aligned}\pi_{12} &= p(s_1) + p(s_4) = 0{,}4 + 0{,}1 = 0{,}5\\ \pi_{13} &= p(s_2) + p(s_4) = 0{,}3 + 0{,}1 = 0{,}4\\ \pi_{23} &= p(s_3) + p(s_4) = 0{,}2 + 0{,}1 = 0{,}3\end{aligned}$$

Para $k = l$, $\pi_{kk} = \pi_k$:

$$\begin{aligned}\pi_{11} &= \pi_1 = 0{,}8\\ \pi_{22} &= \pi_2 = 0{,}7\\ \pi_{33} &= \pi_3 = 0{,}6\end{aligned}$$

La matriz de probabilidades de inclusión de segundo orden $\Pi = (\pi_{kl})$ es:

$$\Pi = \begin{pmatrix} 0{,}8 & 0{,}5 & 0{,}4\\ 0{,}5 & 0{,}7 & 0{,}3\\ 0{,}4 & 0{,}3 & 0{,}6\end{pmatrix}$$

(b) Cálculo de $E(n_s)$

El tamaño de la muestra n_s para cada muestra posible es:

- $n_{s_1} = |\{1, 2\}| = 2$
- $n_{s_2} = |\{1, 3\}| = 2$
- $n_{s_3} = |\{2, 3\}| = 2$
- $n_{s_4} = |\{1, 2, 3\}| = 3$

(i) Cálculo directo usando la definición

El valor esperado del tamaño de la muestra $E(n_s)$ es $\sum_s n_s p(s)$.

$$\begin{aligned}E(n_s) &= n_{s_1}p(s_1) + n_{s_2}p(s_2) + n_{s_3}p(s_3) + n_{s_4}p(s_4)\\ &= (2)(0{,}4) + (2)(0{,}3) + (2)(0{,}2) + (3)(0{,}1)\\ &= 0{,}8 + 0{,}6 + 0{,}4 + 0{,}3\\ &= 2{,}1\end{aligned}$$

(ii) Usando la fórmula $E(n_s) = \sum_{k \in U} \pi_k$

$$\begin{aligned} E(n_s) &= \pi_1 + \pi_2 + \pi_3 \\ &= 0{,}8 + 0{,}7 + 0{,}6 \\ &= 2{,}1 \end{aligned}$$

Ambos métodos producen el mismo resultado, $E(n_s) = 2{,}1$.

Problema 2.6

Considere la población y el diseño del Ejercicio 2.5 anterior. Sean los valores de la variable de estudio y : $y_1 = 16, y_2 = 21$, y $y_3 = 18$. Entonces tenemos $t = 55$.

a) Calcule el valor esperado y la varianza del estimador π a partir de las definiciones de la Sección 2.7.

b) Calcule la varianza del estimador π usando el Resultado 2.8.1.

c) Calcule el coeficiente de variación del estimador π.

d) Calcule una estimación de la varianza $\hat{V}(\hat{t}_\pi)$ usando el estimador de varianza (2.8.6) para cada una de las cuatro muestras posibles. Determine el valor esperado del estimador de varianza en la situación actual usando la definición de valor esperado en la Sección 2.7.

Solución 2.6

Los valores de la variable de estudio son $y_1 = 16, y_2 = 21, y_3 = 18$. El total poblacional es $t = y_1 + y_2 + y_3 = 16 + 21 + 18 = 55$. Las probabilidades de inclusión de primer orden del Problema 2.5 son $\pi_1 = 0{,}8, \pi_2 = 0{,}7, \pi_3 = 0{,}6$. Los valores ponderados $w_k = y_k/\pi_k$ son:

$$\begin{aligned} w_1 &= y_1/\pi_1 = 16/0{,}8 = 20 \\ w_2 &= y_2/\pi_2 = 21/0{,}7 = 30 \\ w_3 &= y_3/\pi_3 = 18/0{,}6 = 30 \end{aligned}$$

El estimador π (estimador de Horvitz-Thompson) para el total t es $\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} w_k$. Los valores de $\hat{t}_{\pi,s}$ para cada muestra s son:

- $s_1 = \{1, 2\}$, $p(s_1) = 0{,}4$: $\hat{t}_{\pi,s_1} = w_1 + w_2 = 20 + 30 = 50$
- $s_2 = \{1, 3\}$, $p(s_2) = 0{,}3$: $\hat{t}_{\pi,s_2} = w_1 + w_3 = 20 + 30 = 50$
- $s_3 = \{2, 3\}$, $p(s_3) = 0{,}2$: $\hat{t}_{\pi,s_3} = w_2 + w_3 = 30 + 30 = 60$
- $s_4 = \{1, 2, 3\}$, $p(s_4) = 0{,}1$: $\hat{t}_{\pi,s_4} = w_1 + w_2 + w_3 = 20 + 30 + 30 = 80$

(a) Valor esperado y varianza de $\hat{t}_\pi$ usando definiciones

El valor esperado de $\hat{t}_\pi$ es $E(\hat{t}_\pi) = \sum_s \hat{t}_{\pi,s} p(s)$.

$$\begin{aligned} E(\hat{t}_\pi) &= (50)(0{,}4) + (50)(0{,}3) + (60)(0{,}2) + (80)(0{,}1) \\ &= 20 + 15 + 12 + 8 \\ &= 55 \end{aligned}$$

El estimador es insesgado, ya que $E(\hat{t}_\pi) = t = 55$.

La varianza de $\hat{t}_\pi$ es $V(\hat{t}_\pi) = E(\hat{t}_\pi^2) - [E(\hat{t}_\pi)]^2$. Primero, calculamos $E(\hat{t}_\pi^2) = \sum_s \hat{t}_{\pi,s}^2 p(s)$.

$$\begin{aligned} E(\hat{t}_\pi^2) &= (50^2)(0{,}4) + (50^2)(0{,}3) + (60^2)(0{,}2) + (80^2)(0{,}1) \\ &= (2500)(0{,}4) + (2500)(0{,}3) + (3600)(0{,}2) + (6400)(0{,}1) \\ &= 1000 + 750 + 720 + 640 \\ &= 3110 \end{aligned}$$

Entonces, la varianza es:

$$V(\hat{t}_\pi) = 3110 - (55)^2 = 3110 - 3025 = 85$$

(b) Varianza de $\hat{t}_\pi$ usando el Resultado 2.8.1

El Resultado 2.8.1 establece que $V(\hat{t}_\pi) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$. Sean $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$. Usando π_k y π_{kl} del Problema 2.5:

$$\begin{aligned}
\Delta_{11} &= \pi_1 - \pi_1^2 = 0{,}8 - (0{,}8)^2 = 0{,}8 - 0{,}64 = 0{,}16\\
\Delta_{22} &= \pi_2 - \pi_2^2 = 0{,}7 - (0{,}7)^2 = 0{,}7 - 0{,}49 = 0{,}21\\
\Delta_{33} &= \pi_3 - \pi_3^2 = 0{,}6 - (0{,}6)^2 = 0{,}6 - 0{,}36 = 0{,}24\\
\Delta_{12} &= \pi_{12} - \pi_1\pi_2 = 0{,}5 - (0{,}8)(0{,}7) = 0{,}5 - 0{,}56 = -0{,}06\\
\Delta_{13} &= \pi_{13} - \pi_1\pi_3 = 0{,}4 - (0{,}8)(0{,}6) = 0{,}4 - 0{,}48 = -0{,}08\\
\Delta_{23} &= \pi_{23} - \pi_2\pi_3 = 0{,}3 - (0{,}7)(0{,}6) = 0{,}3 - 0{,}42 = -0{,}12
\end{aligned}$$

La varianza es: $V(\hat{t}_\pi) = \Delta_{11}w_1^2 + \Delta_{22}w_2^2 + \Delta_{33}w_3^2 + 2\Delta_{12}w_1w_2 + 2\Delta_{13}w_1w_3 + 2\Delta_{23}w_2w_3$

$$\begin{aligned}
V(\hat{t}_\pi) &= (0{,}16)(20^2) + (0{,}21)(30^2) + (0{,}24)(30^2)\\
&\quad + 2(-0{,}06)(20)(30) + 2(-0{,}08)(20)(30) + 2(-0{,}12)(30)(30)\\
&= (0{,}16)(400) + (0{,}21)(900) + (0{,}24)(900)\\
&\quad + (-0{,}12)(600) + (-0{,}16)(600) + (-0{,}24)(900)\\
&= 64 + 189 + 216 - 72 - 96 - 216\\
&= 469 - 384\\
&= 85
\end{aligned}$$

Este resultado coincide con el obtenido en (a).

(c) Coeficiente de variación de $\hat{t}_\pi$

El coeficiente de variación $CV(\hat{t}_\pi)$ es $\frac{\sqrt{V(\hat{t}_\pi)}}{E(\hat{t}_\pi)}$.

$$CV(\hat{t}_\pi) = \frac{\sqrt{85}}{55} \approx \frac{9{,}21954}{55} \approx 0{,}1676$$

(d) Estimador de varianza (2.8.6) y su valor esperado

El estimador de varianza (2.8.6) de Särndal et al. (forma Sen-Yates-Grundy) es:

$$\hat{V}_{SYG}(\hat{t}_\pi) = \frac{1}{2}\sum_{k\in s}\sum_{l\in s, l\neq k}\frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 = \sum_{k<l\in s}\frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}(w_k - w_l)^2$$

Calculamos los coeficientes $c_{kl} = \frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}$:

$$\begin{aligned}
c_{12} &= \frac{(0{,}8)(0{,}7) - 0{,}5}{0{,}5} = \frac{0{,}56 - 0{,}5}{0{,}5} = \frac{0{,}06}{0{,}5} = 0{,}12\\
c_{13} &= \frac{(0{,}8)(0{,}6) - 0{,}4}{0{,}4} = \frac{0{,}48 - 0{,}4}{0{,}4} = \frac{0{,}08}{0{,}4} = 0{,}20\\
c_{23} &= \frac{(0{,}7)(0{,}6) - 0{,}3}{0{,}3} = \frac{0{,}42 - 0{,}3}{0{,}3} = \frac{0{,}12}{0{,}3} = 0{,}40
\end{aligned}$$

Las diferencias cuadradas $(w_k - w_l)^2$:

$$\begin{aligned}
(w_1 - w_2)^2 &= (20 - 30)^2 = (-10)^2 = 100\\
(w_1 - w_3)^2 &= (20 - 30)^2 = (-10)^2 = 100\\
(w_2 - w_3)^2 &= (30 - 30)^2 = (0)^2 = 0
\end{aligned}$$

Estimaciones de varianza $\hat{V}_s(\hat{t}_\pi)$ para cada muestra:

- $s_1 = \{1, 2\}$: $\hat{V}_{s_1} = c_{12}(w_1 - w_2)^2 = (0{,}12)(100) = 12$
- $s_2 = \{1, 3\}$: $\hat{V}_{s_2} = c_{13}(w_1 - w_3)^2 = (0{,}20)(100) = 20$

- $s_3 = \{2,3\}$: $\hat{V}_{s_3} = c_{23}(w_2 - w_3)^2 = (0{,}40)(0) = 0$
- $s_4 = \{1,2,3\}$:

$$\begin{aligned}\hat{V}_{s_4} &= c_{12}(w_1 - w_2)^2 + c_{13}(w_1 - w_3)^2 + c_{23}(w_2 - w_3)^2 \\ &= (0{,}12)(100) + (0{,}20)(100) + (0{,}40)(0) \\ &= 12 + 20 + 0 = 32\end{aligned}$$

El valor esperado del estimador de varianza es $E(\hat{V}_{SYG}(\hat{t}_\pi)) = \sum_s \hat{V}_s p(s)$.

$$\begin{aligned}E(\hat{V}_{SYG}(\hat{t}_\pi)) &= (12)(0{,}4) + (20)(0{,}3) + (0)(0{,}2) + (32)(0{,}1) \\ &= 4{,}8 + 6{,}0 + 0 + 3{,}2 \\ &= 14{,}0\end{aligned}$$

La varianza verdadera es $V(\hat{t}_\pi) = 85$. El valor esperado del estimador de varianza (2.8.6) es 14.0. Por lo tanto, este estimador de varianza es sesgado para este diseño ($14{,}0 \neq 85$). Esto es esperable, ya que el estimador Sen-Yates-Grundy (2.8.6) es insesgado para diseños de tamaño muestral fijo, y este diseño tiene tamaños de muestra variables (2 y 3).

Problema 2.12

Para estimar el ingreso promedio por hogar ($\sum_U y_k/N$) para una población de $N = 200$ hogares, se utilizó una lista de las 600 personas que pertenecen a los 200 hogares de la siguiente manera. Se extrajo una muestra SIR (Muestreo Aleatorio Simple) de tamaño $m = 10$ personas. Se identificaron los hogares de las personas seleccionadas y se recopiló información sobre el ingreso promedio del hogar, y_k/x_k, donde y_k es el ingreso total del hogar en dólares y x_k es el número de personas en el hogar. Los resultados son los siguientes:

Extracción i	Ingreso promedio del hogar (y_{k_i}/x_{k_i})
1	7,000
2	8,000
3	6,000
4	5,000
5	9,000
6	4,000
7	7,000
8	8,000
9	4,000
10	2,000

Calcule una estimación del ingreso promedio por hogar basada en el estimador pwr (probability weighted estimator), así como el cve (coeficiente de variación estimado) correspondiente.

Solución 2.12

Definición de Parámetros y Variables

- $N = 200$: Número total de hogares en la población.
- $M_0 = 600$: Número total de personas en la población.
- $m = 10$: Tamaño de la muestra de personas seleccionadas mediante SIR (Muestreo Aleatorio Simple sin reemplazo).
- y_k: Ingreso total del hogar k.
- x_k: Número de personas en el hogar k.
- El parámetro de interés es el ingreso promedio por hogar: $\bar{Y}_H = \frac{1}{N}\sum_{k=1}^{N} y_k$.
- Para la i-ésima persona seleccionada, se observa $a_i = y_{k_i}/x_{k_i}$, que es el ingreso per cápita del hogar k_i al que pertenece la persona i.

Formulación del Estimador

La muestra se toma de la población de M_0 personas. Para cada persona p en la población, asociamos el valor $a_p = y_{k(p)}/x_{k(p)}$, donde $k(p)$ es el hogar de la persona p. Consideremos la suma de estos valores a_p sobre todas las personas en la población:

$$\sum_{p=1}^{M_0} a_p = \sum_{p=1}^{M_0} \frac{y_{k(p)}}{x_{k(p)}}$$

Un hogar k tiene x_k personas. Cada una de estas x_k personas contribuye con el valor y_k/x_k a la suma anterior. Por lo tanto, la contribución del hogar k a esta suma es $x_k \cdot (y_k/x_k) = y_k$. Así, la suma de los valores a_p sobre todas las personas es igual a la suma de los ingresos totales y_k sobre todos los hogares:

$$\sum_{p=1}^{M_0} a_p = \sum_{k=1}^{N} y_k = T_y$$

El promedio poblacional de los valores a_p es $\bar{A}_P = \frac{1}{M_0}\sum_{p=1}^{M_0} a_p = \frac{T_y}{M_0}$. Se extrae una muestra aleatoria simple (SIR) de m personas. Sea a_i el valor observado para la i-ésima persona en la muestra. El promedio muestral $\bar{a}_s = \frac{1}{m}\sum_{i=1}^{m} a_i$ es un estimador insesgado de $\bar{A}_P$. Por lo tanto, un estimador insesgado para el total poblacional $T_y = M_0\bar{A}_P$ es:

$$\hat{T}_y = M_0\bar{a}_s = M_0\frac{1}{m}\sum_{i=1}^{m} a_i$$

Este estimador puede considerarse un "probability weighted estimator"(pwr) o estimador ponderado por probabilidad, ya que $\hat{T}_y = \sum_{i=1}^{m} \frac{a_i}{m/M_0}$, donde m/M_0 es la probabilidad de inclusión (aproximada o por extracción) de una persona. El estimador para el ingreso promedio por hogar $\bar{Y}_H = T_y/N$ es:

$$\hat{\bar{Y}}_H = \frac{\hat{T}_y}{N} = \frac{M_0\bar{a}_s}{N} = \frac{M_0}{N \cdot m}\sum_{i=1}^{m} a_i$$

Cálculo de la Estimación

Los valores observados de $a_i = y_{k_i}/x_{k_i}$ son: 7000, 8000, 6000, 5000, 9000, 4000, 7000, 8000, 4000, 2000. La suma de estos valores es:

$$\sum_{i=1}^{10} a_i = 7000 + 8000 + 6000 + 5000 + 9000 + 4000 + 7000 + 8000 + 4000 + 2000 = 60,000$$

El promedio muestral es:

$$\bar{a}_s = \frac{60,000}{10} = 6,000$$

La estimación del ingreso promedio por hogar es:

$$\hat{\bar{Y}}_H = \frac{M_0}{N}\bar{a}_s = \frac{600}{200} \times 6,000 = 3 \times 6,000 = 18,000$$

La estimación del ingreso promedio por hogar es de \$18,000.

Cálculo del Coeficiente de Variación Estimado (cve)

El coeficiente de variación estimado es $\widehat{CV}(\hat{\bar{Y}}_H) = \frac{\sqrt{\hat{V}(\hat{\bar{Y}}_H)}}{\hat{\bar{Y}}_H}$. Necesitamos la varianza estimada $\hat{V}(\hat{\bar{Y}}_H)$.

$$V(\hat{\bar{Y}}_H) = V\left(\frac{M_0}{N}\bar{a}_s\right) = \left(\frac{M_0}{N}\right)^2 V(\bar{a}_s)$$

Para un muestreo SIR de personas, la varianza de $\bar{a}_s$ es $V(\bar{a}_s) = \frac{S_a^2}{m}(1 - \frac{m}{M_0})$, donde $S_a^2 = \frac{1}{M_0 - 1}\sum_{p=1}^{M_0}(a_p - \bar{A}_P)^2$. Un estimador insesgado para S_a^2 es la varianza muestral $s_a^2 = \frac{1}{m-1}\sum_{i=1}^{m}(a_i - \bar{a}_s)^2$. Entonces, la varianza estimada de $\hat{\bar{Y}}_H$ es:

$$\hat{V}(\hat{\bar{Y}}_H) = \left(\frac{M_0}{N}\right)^2 \frac{s_a^2}{m}\left(1 - \frac{m}{M_0}\right)$$

Calculamos s_a^2: $\sum_{i=1}^{10}(a_i - \bar{a}_s)^2 = \sum_{i=1}^{10}(a_i - 6000)^2$

$$\begin{aligned}
&= (7000-6000)^2 + (8000-6000)^2 + (6000-6000)^2 + (5000-6000)^2 + (9000-6000)^2 \\
&\quad + (4000-6000)^2 + (7000-6000)^2 + (8000-6000)^2 + (4000-6000)^2 + (2000-6000)^2 \\
&= (1000)^2 + (2000)^2 + (0)^2 + (-1000)^2 + (3000)^2 \\
&\quad + (-2000)^2 + (1000)^2 + (2000)^2 + (-2000)^2 + (-4000)^2 \\
&= 1,000,000 + 4,000,000 + 0 + 1,000,000 + 9,000,000 \\
&\quad + 4,000,000 + 1,000,000 + 4,000,000 + 4,000,000 + 16,000,000 \\
&= 44,000,000
\end{aligned}$$

$$s_a^2 = \frac{44,000,000}{10-1} = \frac{44,000,000}{9} \approx 4,888,888{,}89$$

Ahora, calculamos $\hat{V}(\hat{\bar{Y}}_H)$:

$$\begin{aligned}
\hat{V}(\hat{\bar{Y}}_H) &= \left(\frac{600}{200}\right)^2 \frac{1}{10}\left(\frac{44,000,000}{9}\right)\left(1 - \frac{10}{600}\right) \\
&= (3)^2 \times \frac{1}{10} \times \frac{44,000,000}{9} \times \left(1 - \frac{1}{60}\right) \\
&= 9 \times \frac{1}{10} \times \frac{44,000,000}{9} \times \frac{59}{60} \\
&= \frac{1}{10} \times 44,000,000 \times \frac{59}{60} \\
&= 4,400,000 \times \frac{59}{60} \\
&= 440,000 \times \frac{59}{6} \\
&= \frac{25,960,000}{6} = \frac{12,980,000}{3} \approx 4,326,666{,}67
\end{aligned}$$

El error estándar estimado es:

$$\widehat{SE}(\hat{\bar{Y}}_H) = \sqrt{\hat{V}(\hat{\bar{Y}}_H)} = \sqrt{\frac{12,980,000}{3}} \approx \sqrt{4,326,666{,}67} \approx 2079{,}92$$

El coeficiente de variación estimado es:

$$\widehat{CV}(\hat{\bar{Y}}_H) = \frac{\widehat{SE}(\hat{\bar{Y}}_H)}{\hat{\bar{Y}}_H} = \frac{2079{,}92}{18000} \approx 0{,}11555$$

El cve es aproximadamente 0.1156 o 11.56 %.

Resumen de Resultados

- Estimación del ingreso promedio por hogar ($\hat{\bar{Y}}_H$): \$18,000.
- Coeficiente de variación estimado ($\widehat{CV}(\hat{\bar{Y}}_H)$): 0.1156 (o 11.56 %).

Problema 3.1 (Särndal)

Suponga que desea extraer una muestra Bernoulli de la población CO124 para estimar el total de la variable $IMP(= y)$ con un error estándar relativo del 10 % mediante (a) el estimador π y (b) el estimador alternativo mostrado en la ecuación (3.2.6) del libro de Särndal. Usando $\sum_U y_k = 1{,}81 \cdot 10^6$ y $\sum_U y_k^2 = 1{,}69 \cdot 10^{11}$, calcule los tamaños de muestra esperados necesarios. Comente su resultado. El error estándar relativo se define como $[V(\hat{t})]^{1/2}/t$.

Solución 3.1

Datos de la población CO124: $N = 124$ Total poblacional de y: $t_y = \sum_U y_k = 1{,}81 \times 10^6$ Suma de cuadrados de y: $\sum_U y_k^2 = 1{,}69 \times 10^{11}$ Error estándar relativo (RSE) deseado: $RSE(\hat{t}) = CV(\hat{t}) = 0{,}10$ Definición de RSE: $CV(\hat{t}) = \frac{\sqrt{V(\hat{t})}}{t_y}$

(a) Estimador π ($\hat{t}_\pi$)

Para un diseño Bernoulli, el estimador π del total es $\hat{t}_\pi = \frac{1}{\pi}\sum_{k\in s} y_k$. Su varianza poblacional es $V(\hat{t}_\pi) = \left(\frac{1}{\pi} - 1\right)\sum_U y_k^2$. Queremos $CV(\hat{t}_\pi) = 0{,}10$:

$$\frac{\sqrt{\left(\frac{1}{\pi} - 1\right)\sum_U y_k^2}}{t_y} = 0{,}10$$

$$\left(\frac{1}{\pi} - 1\right)\sum_U y_k^2 = (0{,}10 \cdot t_y)^2$$

$$\frac{1}{\pi} - 1 = \frac{(0{,}10 \cdot t_y)^2}{\sum_U y_k^2}$$

$$\frac{1}{\pi} = 1 + \frac{(0{,}10 \cdot 1{,}81 \times 10^6)^2}{1{,}69 \times 10^{11}}$$

$$\frac{1}{\pi} = 1 + \frac{(1{,}81 \times 10^5)^2}{1{,}69 \times 10^{11}} = 1 + \frac{3{,}2761 \times 10^{10}}{1{,}69 \times 10^{11}}$$

$$\frac{1}{\pi} = 1 + 0{,}19385207 \approx 1{,}193852$$

$$\pi = \frac{1}{1{,}193852} \approx 0{,}83763$$

El tamaño de muestra esperado necesario es $n_e = N\pi$:

$$n_e = 124 \times 0{,}83763 \approx 103{,}866$$

Se requeriría un tamaño de muestra esperado de aproximadamente 104.

(b) Estimador alternativo ($\hat{t}_{alt}$)

El estimador alternativo para el total en un diseño Bernoulli es $\hat{t}_{alt} = N\bar{y}_s = N\frac{\sum_{k\in s} y_k}{n_s}$. La varianza aproximada de este estimador es $V(\hat{t}_{alt}) \approx N\left(\frac{1}{\pi} - 1\right) S_y^2$, donde $S_y^2 = \frac{1}{N-1}\sum_U (y_k - \bar{y}_U)^2$. Primero, calculamos $\bar{y}_U$ y S_y^2:

$$\bar{y}_U = \frac{t_y}{N} = \frac{1{,}81 \times 10^6}{124} \approx 14596{,}774$$

$$\begin{aligned}
S_y^2 &= \frac{1}{N-1}\left(\sum_U y_k^2 - \frac{(\sum_U y_k)^2}{N}\right) \\
&= \frac{1}{123}\left(1{,}69 \times 10^{11} - \frac{(1{,}81 \times 10^6)^2}{124}\right) \\
&= \frac{1}{123}\left(1{,}69 \times 10^{11} - \frac{3{,}2761 \times 10^{12}}{124}\right) \\
&= \frac{1}{123}(1{,}69 \times 10^{11} - 2{,}64193548 \times 10^{10}) \\
&= \frac{1}{123}(16{,}9 \times 10^{10} - 2{,}64193548 \times 10^{10}) \\
&= \frac{1}{123}(14{,}25806452 \times 10^{10}) \approx 1{,}1591922 \times 10^9
\end{aligned}$$

Queremos $CV(\hat{t}_{alt}) = 0{,}10$:

$$\frac{\sqrt{N\left(\frac{1}{\pi} - 1\right) S_y^2}}{t_y} = 0{,}10$$

$$N\left(\frac{1}{\pi} - 1\right) S_y^2 = (0{,}10 \cdot t_y)^2$$

$$\frac{1}{\pi} - 1 = \frac{(0{,}10 \cdot t_y)^2}{N S_y^2}$$

$$\frac{1}{\pi} = 1 + \frac{(0{,}10 \cdot 1{,}81 \times 10^6)^2}{124 \cdot (1{,}1591922 \times 10^9)}$$

$$\frac{1}{\pi} = 1 + \frac{3{,}2761 \times 10^{10}}{1{,}4373983 \times 10^{11}} = 1 + 0{,}227913$$

$$\frac{1}{\pi} \approx 1{,}227913$$

$$\pi = \frac{1}{1{,}227913} \approx 0{,}81439$$

El tamaño de muestra esperado necesario es $n_e = N\pi$:

$$n_e = 124 \times 0{,}81439 \approx 100{,}984$$

Se requeriría un tamaño de muestra esperado de aproximadamente 101.

Comentario

El estimador alternativo $\hat{t}_{alt}$ requiere un tamaño de muestra esperado ligeramente menor (aprox. 101) en comparación con el estimador π (aprox. 104) para alcanzar el mismo error estándar relativo del 10 %. Esto sugiere que, para esta población y variable, el estimador alternativo es un poco más eficiente en términos de tamaño de muestra requerido, lo cual es consistente con la teoría que indica que $V(\hat{t}_{alt})$ puede ser menor que $V(\hat{t}_\pi)$ si la variabilidad del tamaño de muestra n_s no es demasiado grande o si no hay una fuerte correlación entre y_k y la probabilidad de ser incluido (que es constante π aquí para cada y_k, pero el efecto se ve a través de n_s).

Problema 3.2 (Särndal)

Usando un valor $\pi = 0{,}1$, se extrajo una muestra Bernoulli de la población MU284 para estimar los totales de las dos variables P85 y RMT85. Se obtuvieron los siguientes resultados:

Variable	$\sum_s y_k$	$\sum_s y_k^2$
P85	564	15,790
RMT85	4,178	878,452

Calcule estimaciones insesgadas de los dos totales, así como las correspondientes estimaciones de varianza insesgadas y los correspondientes cve's.

Solución 3.2

Población MU284, por lo tanto $N = 284$. Diseño Bernoulli con $\pi = 0{,}1$.

El estimador π insesgado del total es $\hat{t}_\pi = \frac{1}{\pi}\sum_{k\in s} y_k$. El estimador insesgado de la varianza de $\hat{t}_\pi$ es $\hat{V}(\hat{t}_\pi) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right)\sum_{k\in s} y_k^2$. El coeficiente de variación estimado (cve) es $\widehat{cve}(\hat{t}_\pi) = \frac{\sqrt{\hat{V}(\hat{t}_\pi)}}{\hat{t}_\pi}$.

Variable P85

Datos muestrales: $\sum_s y_k = 564$, $\sum_s y_k^2 = 15,790$.

- Estimación insesgada del total:

$$\hat{t}_\pi(P85) = \frac{1}{0{,}1} \times 564 = 10 \times 564 = 5,640$$

- Estimación insesgada de la varianza:

$$\begin{aligned}\hat{V}(\hat{t}_\pi(P85)) &= \frac{1}{0{,}1}\left(\frac{1}{0{,}1} - 1\right) \times 15,790 \\ &= 10 \times (10-1) \times 15,790 \\ &= 10 \times 9 \times 15,790 \\ &= 90 \times 15,790 = 1,421,100\end{aligned}$$

- Coeficiente de variación estimado (cve):

$$\widehat{SE}(\hat{t}_\pi(P85)) = \sqrt{1,421,100} \approx 1192{,}099$$

$$\widehat{cve}(\hat{t}_\pi(P85)) = \frac{1192{,}099}{5640} \approx 0{,}211365 \quad (o \quad 21{,}14\,\%)$$

Variable RMT85

Datos muestrales: $\sum_s y_k = 4,178$, $\sum_s y_k^2 = 878,452$.

- Estimación insesgada del total:

$$\hat{t}_\pi(RMT85) = \frac{1}{0{,}1} \times 4,178 = 10 \times 4,178 = 41,780$$

- Estimación insesgada de la varianza:

$$\begin{aligned}\hat{V}(\hat{t}_\pi(RMT85)) &= \frac{1}{0{,}1}\left(\frac{1}{0{,}1} - 1\right) \times 878,452 \\ &= 10 \times (10-1) \times 878,452 \\ &= 10 \times 9 \times 878,452 \\ &= 90 \times 878,452 = 79,060,680\end{aligned}$$

- Coeficiente de variación estimado (cve):

$$\widehat{SE}(\hat{t}_\pi(RMT85)) = \sqrt{79,060,680} \approx 8891{,}6073$$

$$\widehat{cve}(\hat{t}_\pi(RMT85)) = \frac{8891{,}6073}{41780} \approx 0{,}212822 \quad (o \quad 21{,}28\,\%)$$

Problema 3.3 (Särndal)

Una muestra Bernoulli de tamaño $n_s = 71$, extraída con un valor $\pi = 0{,}5$ de la población CO124, contenía 21 países de Asia (continente 1). Calcule una estimación puntual insesgada y un intervalo de confianza de aproximadamente 95 % para el porcentaje de países que pertenecen Asia.

Solución 3.3

Datos:

- Población CO124, por lo tanto $N = 124$.
- Diseño Bernoulli con probabilidad de inclusión individual $\pi = 0{,}5$.
- Tamaño de muestra obtenido $n_s = 71$.
- Número de países de Asia en la muestra $n_{s,Asia} = 21$.

Queremos estimar P_{Asia}, el porcentaje (o proporción) de países en la población CO124 que pertenecen a Asia.

Definimos una variable indicadora y_k para cada país k en la población:

$$y_k = \begin{cases} 1 & \text{si el país } k \text{ es de Asia} \\ 0 & \text{en caso contrario} \end{cases}$$

El parámetro de interés es la proporción poblacional $P_{Asia} = \frac{1}{N}\sum_{k\in U} y_k$.

Estimación puntual insesgada

Para un diseño Bernoulli, un estimador insesgado de la proporción poblacional P es la proporción muestral $\hat{P}_s = \frac{1}{n_s}\sum_{k\in s} y_k$. En este caso:

$$\hat{P}_{Asia} = \frac{n_{s,Asia}}{n_s} = \frac{21}{71}$$

Calculando el valor:

$$\hat{P}_{Asia} \approx 0{,}2957746$$

Así, la estimación puntual insesgada del porcentaje de países que pertenecen a Asia es aproximadamente 29.58 %.

Intervalo de confianza de aproximadamente 95 %

Para un diseño Bernoulli, la varianza del estimador de la proporción $\hat{P}_s$ es:

$$V(\hat{P}_s) \approx \frac{P(1-P)}{N\pi}(1-\pi)$$

(Esta es una aproximación, la varianza exacta es más compleja debido a que n_s es aleatorio. Särndal et al. (p. 70, Result 3.3.1 para la media) indica que para estimar una media (o proporción) bajo muestreo Bernoulli, $V(\hat{\bar{y}}_s) \approx \frac{S_y^2}{N\pi}(1-\pi)$. Para una variable binaria y_k, $S_y^2 \approx P(1-P)$ (específicamente $S_y^2 = \frac{N}{N-1}P(1-P)$). Usaremos la aproximación más simple $P(1-P)$ para la varianza poblacional de una variable binaria.)

Un estimador insesgado de la varianza $V(\hat{P}_s)$ bajo muestreo Bernoulli es:

$$\hat{V}(\hat{P}_s) = \frac{\hat{P}_s(1-\hat{P}_s)}{n_s - 1}\left(1 - \frac{n_s}{N}\right) \text{ si se considera } n_s \text{ como fijo y SI.}$$

Sin embargo, para diseño Bernoulli, una fórmula más apropiada para la varianza estimada del estimador de la media $\hat{\bar{y}} = \frac{1}{N\pi}\sum_s y_k$ es $\hat{V}(\hat{\bar{y}}) = \frac{1}{(N\pi)^2}\sum_s(\frac{y_k}{\pi_k} - \hat{t}_\pi/N)^2\frac{1-\pi_k}{\pi_k^2}$, que se simplifica. Alternativamente, Särndal (Resultado 3.2.2, p. 63, para el estimador alternativo de la media, que es $\bar{y}_s$):

$$\hat{V}(\bar{y}_s) \approx \frac{s_y^2}{n_s}(1 - \frac{n_s}{E(n_s)}) \text{ (aproximación)}$$

Donde $s_y^2 = \frac{1}{n_s-1}\sum_s(y_k - \bar{y}_s)^2$. Para una proporción, $s_y^2 = \frac{n_s}{n_s-1}\hat{P}_s(1-\hat{P}_s)$. Y $E(n_s) = N\pi = 124 \times 0{,}5 = 62$.

$$\hat{V}(\hat{P}_{Asia}) \approx \frac{1}{n_s - 1}\hat{P}_{Asia}(1-\hat{P}_{Asia})\left(1 - \frac{\pi N}{N}\right) \text{ (Si se usa varianza de SI con corrección para } \pi)$$

Una fórmula más directa para la varianza estimada de $\hat{P}_s = \bar{y}_s$ cuando $\pi_k = \pi$ es (Särndal et al., Ej. 3.6, p. 115, adaptado de $\hat{t}_{alt}$):

$$\hat{V}(\hat{P}_s) = \frac{1}{N^2}\hat{V}(\hat{t}_{alt}) = \frac{1}{N^2}N^2\frac{1-\pi}{n_s\pi}\hat{P}_s(1-\hat{P}_s) = \frac{1-\pi}{n_s\pi}\hat{P}_s(1-\hat{P}_s)$$

Esto parece más consistente con el estimador alternativo. Vamos a usar esta última.

$$\hat{P}_{Asia}(1-\hat{P}_{Asia}) = \frac{21}{71}\left(1 - \frac{21}{71}\right) = \frac{21}{71}\cdot\frac{50}{71} = \frac{1050}{71^2} = \frac{1050}{5041} \approx 0{,}208292$$

$$\hat{V}(\hat{P}_{Asia}) = \frac{1-0{,}5}{71 \times 0{,}5} \times 0{,}208292 = \frac{0{,}5}{35{,}5} \times 0{,}208292 \approx 0{,}0140845 \times 0{,}208292 \approx 0{,}0029336$$

El error estándar estimado es:

$$\widehat{SE}(\hat{P}_{Asia}) = \sqrt{0{,}0029336} \approx 0{,}05416$$

Un intervalo de confianza de aproximadamente 95 % para P_{Asia} es:

$$\hat{P}_{Asia} \pm z_{1-\alpha/2} \cdot \widehat{SE}(\hat{P}_{Asia})$$

Para un 95 % de confianza, $z_{1-\alpha/2} = z_{0{,}975} \approx 1{,}96$.

$$IC_{95\,\%}(P_{Asia}) = 0{,}2958 \pm 1{,}96 \times 0{,}05416$$

$$IC_{95\,\%}(P_{Asia}) = 0{,}2958 \pm 0{,}1061536$$

Límite inferior: $0{,}2958 - 0{,}1061536 = 0{,}1896464$ Límite superior: $0{,}2958 + 0{,}1061536 = 0{,}4019536$ El intervalo de confianza de aproximadamente 95 % para el porcentaje de países que pertenecen a Asia es [18.96 %, 40.20 %].

Nota sobre la varianza del estimador de proporción en Bernoulli: La varianza del estimador $\hat{P}_s = \sum_s y_k/n_s$ en un diseño Bernoulli donde $\pi_k = \pi$ para todo k es un tema que puede tener diferentes aproximaciones. Si consideramos $\hat{P}_\pi = \frac{1}{N\pi}\sum_s y_k$ como el estimador insesgado de P, su varianza es $V(\hat{P}_\pi) = \frac{1}{(N\pi)^2}V(\hat{t}_\pi) = \frac{1}{(N\pi)^2}(\frac{1}{\pi} - 1)\sum_U y_k^2$. Su estimador de varianza es $\hat{V}(\hat{P}_\pi) = \frac{1}{(N\pi)^2}\hat{V}(\hat{t}_\pi) = \frac{1}{(N\pi)^2}\frac{1}{\pi}(\frac{1}{\pi} - 1)\sum_s y_k^2$. Para y_k binaria, $\sum_s y_k^2 = \sum_s y_k = n_{s,Asia}$.

$$\hat{V}(\hat{P}_\pi) = \frac{1}{(124 \times 0{,}5)^2}\frac{1}{0{,}5}\left(\frac{1}{0{,}5} - 1\right)(21) = \frac{1}{62^2}\cdot 2\cdot(2-1)\cdot 21 = \frac{1}{3844}\cdot 2\cdot 1\cdot 21 = \frac{42}{3844} \approx 0{,}010926$$

$$\widehat{SE}(\hat{P}_\pi) \approx \sqrt{0{,}010926} \approx 0{,}1045$$

En este caso, $\hat{P}_\pi = \frac{1}{124\times 0{,}5} \times 21 = \frac{21}{62} \approx 0{,}3387$. El intervalo sería $0{,}3387 \pm 1{,}96 \times 0{,}1045 = 0{,}3387 \pm 0{,}2048 = [0{,}1339, 0{,}5435]$ o [13.39 %, 54.35 %].

El problema pide .^estimación puntual insesgada", y $\hat{P}_s = n_{s,Asia}/n_s$ es comúnmente usado y es insesgado bajo ciertas condiciones o asintóticamente. La fórmula de varianza $\frac{1-\pi}{n_s\pi}\hat{P}_s(1-\hat{P}_s)$ es una aproximación razonable para $\bar{y}_s$ bajo Bernoulli. La diferencia en los resultados del IC se debe a la elección del estimador de la proporción y su correspondiente varianza. El estimador $\hat{P}_s$ es más intuitivo como "porcentaje en la muestra". El libro de Särndal (p. 63) para el estimador alternativo de la media $\bar{y}_s$ usa la varianza $V(\bar{y}_s) \approx \frac{S_y^2}{E(n_s)}(1-\pi)$. Un estimador para esto sería $\hat{V}(\bar{y}_s) \approx \frac{s_y^2}{n_s}(1-\pi)$ (asumiendo $n_s \approx E(n_s)$) donde $s_y^2 = \frac{n_s}{n_s-1}\hat{P}_s(1-\hat{P}_s)$.

$$\hat{V}(\hat{P}_{Asia}) \approx \frac{1}{n_s-1}\hat{P}_{Asia}(1-\hat{P}_{Asia})(1-\pi) = \frac{1}{70}(0{,}208292)(0{,}5) \approx 0{,}0014878$$

$$\widehat{SE}(\hat{P}_{Asia}) \approx \sqrt{0{,}0014878} \approx 0{,}03857$$

IC: $0{,}2958 \pm 1{,}96 \times 0{,}03857 = 0{,}2958 \pm 0{,}0756 = [0{,}2202, 0{,}3714]$ o [22.02 %, 37.14 %]. Esta última parece la más consistente con las fórmulas del libro para el estimador alternativo.

Revisando el Resultado 3.2.6 de Särndal et al. (p. 63), la varianza del estimador alternativo del total $\hat{t}_{alt} = N\bar{y}_s$ es $V(\hat{t}_{alt}) \approx NS_y^2(\frac{1}{E(n_s)} - \frac{1}{N})$. No, esa es para SI. Para Bernoulli es $V(\hat{t}_{alt}) \approx NS_y^2\frac{1-\pi}{E(n_s)}$. Entonces $V(\hat{P}_s) = V(\bar{y}_s) \approx \frac{S_y^2(1-\pi)}{N\pi}$. Estimando S_y^2 por $s_y^2 = \frac{n_s}{n_s-1}\hat{P}_s(1-\hat{P}_s)$ y $N\pi$ por n_s (si π es desconocido y se estima por n_s/N, o usando $E(n_s) = N\pi$ directamente):

$$\hat{V}(\hat{P}_s) \approx \frac{n_s}{n_s-1}\hat{P}_s(1-\hat{P}_s)\frac{1-\pi}{N\pi}$$

$$\begin{aligned}\hat{V}(\hat{P}_{Asia}) &\approx \frac{71}{70}(0{,}208292)\frac{1-0{,}5}{124\times 0{,}5} \\ &= \frac{71}{70}(0{,}208292)\frac{0{,}5}{62} \\ &\approx 1{,}01428 \times 0{,}208292 \times 0{,}0080645 \\ &\approx 0{,}001703\end{aligned}$$

$$\widehat{SE}(\hat{P}_{Asia}) \approx \sqrt{0{,}001703} \approx 0{,}04127$$

IC: $0{,}2958 \pm 1{,}96 \times 0{,}04127 = 0{,}2958 \pm 0{,}08089 = [0{,}2149, 0{,}3767]$ o [21.49 %, 37.67 %]. Este último cálculo parece el más riguroso siguiendo el libro. La primera fórmula $\frac{1-\pi}{n_s\pi}\hat{P}_s(1-\hat{P}_s)$ es equivalente a $\frac{\hat{P}_s(1-\hat{P}_s)}{n_s}\frac{1-\pi}{\pi}$ que no es estándar. La fórmula correcta para la varianza estimada de una proporción muestral $\hat{p}$ bajo Bernoulli con parámetro π poblacional (no la prob de selección) y tamaño muestral n_s es $\frac{\hat{p}(1-\hat{p})}{n_s}$. Si el muestreo es Bernoulli con prob de selección π, el estimador de la proporción $\bar{y}_s$ tiene varianza $V(\bar{y}_s) \approx \frac{S_y^2(1-\pi)}{N\pi}$. Usando $s_y^2 = \frac{n_s}{n_s-1}\hat{P}(1-\hat{P})$ para S_y^2,

$$\hat{V}(\hat{P}_s) \approx \frac{n_s}{n_s-1}\hat{P}_s(1-\hat{P}_s)\frac{1-\pi}{N\pi}$$

Esta es la fórmula 0,001703 usada arriba. El IC es [21,49 %, 37,67 %].

Problema 3.5 (Särndal)

Se extrajo una muestra SI (Muestreo Aleatorio Simple sin reemplazo) de tamaño 30 para estimar el total de la variable S82 (= y) para la población MU284. Se calcularon los totales muestrales $\sum_s y_k = 1,424$ y $\sum_s y_k^2 = 70,758$. Calcule un intervalo de confianza de aproximadamente 95 % para el total de S82, el número total de escaños en los consejos municipales.

Solución 3.5

Datos:

- Población MU284, por lo tanto $N = 284$.
- Diseño SI (Muestreo Aleatorio Simple sin reemplazo).
- Tamaño de la muestra $n = 30$.

- Suma muestral de y: $\sum_{k\in s} y_k = 1,424$.
- Suma muestral de y^2: $\sum_{k\in s} y_k^2 = 70,758$.

El parámetro de interés es el total poblacional $t_y = \sum_{k\in U} y_k$.

Estimación puntual del total

Bajo muestreo aleatorio simple sin reemplazo (SI), el estimador π del total es:

$$\hat{t}_\pi = N\bar{y}_s = N\frac{1}{n}\sum_{k\in s} y_k$$

$$\hat{t}_\pi = 284 \times \frac{1}{30} \times 1,424 = 284 \times 47{,}4666... = 13470{,}5333...$$

La estimación puntual del total de escaños es aproximadamente 13,470.53.

Estimación de la varianza del estimador del total

La varianza estimada de $\hat{t}_\pi$ bajo SI es (Resultado 3.1.3 de Särndal et al.):

$$\hat{V}(\hat{t}_\pi) = N^2 \frac{S_y^2}{n}\left(1 - \frac{n}{N}\right)$$

donde S_y^2 es la varianza muestral:

$$S_y^2 = \frac{1}{n-1}\sum_{k\in s}(y_k - \bar{y}_s)^2 = \frac{1}{n-1}\left(\sum_{k\in s} y_k^2 - \frac{(\sum_{k\in s} y_k)^2}{n}\right)$$

Calculamos $\bar{y}_s$:

$$\bar{y}_s = \frac{1424}{30} \approx 47{,}46667$$

Calculamos S_y^2:

$$\begin{aligned} S_y^2 &= \frac{1}{30-1}\left(70,758 - \frac{(1,424)^2}{30}\right) \\ &= \frac{1}{29}\left(70,758 - \frac{2,027,776}{30}\right) \\ &= \frac{1}{29}(70,758 - 67,592{,}5333...) \\ &= \frac{1}{29}(3165{,}4666...) \\ &\approx 109{,}15402 \end{aligned}$$

Ahora calculamos $\hat{V}(\hat{t}_\pi)$:

$$\begin{aligned} \hat{V}(\hat{t}_\pi) &= (284)^2 \times \frac{109{,}15402}{30} \times \left(1 - \frac{30}{284}\right) \\ &= 80656 \times \frac{109{,}15402}{30} \times (1 - 0{,}1056338) \\ &= 80656 \times 3{,}6384673 \times 0{,}8943662 \\ &\approx 293490{,}74 \times 0{,}8943662 \\ &\approx 262504{,}6 \end{aligned}$$

Intervalo de confianza de aproximadamente 95 %

El error estándar estimado es:

$$\widehat{SE}(\hat{t}_\pi) = \sqrt{\hat{V}(\hat{t}_\pi)} = \sqrt{262504{,}6} \approx 512{,}352$$

Un intervalo de confianza de aproximadamente 95 % para t_y es:

$$\hat{t}_\pi \pm z_{1-\alpha/2} \cdot \widehat{SE}(\hat{t}_\pi)$$

Para un 95 % de confianza, $z_{1-\alpha/2} = z_{0,975} \approx 1{,}96$.

$$IC_{95\,\%}(t_y) = 13470{,}53 \pm 1{,}96 \times 512{,}352$$

$$IC_{95\,\%}(t_y) = 13470{,}53 \pm 1004{,}21$$

Límite inferior: $13470{,}53 - 1004{,}21 = 12466{,}32$ Límite superior: $13470{,}53 + 1004{,}21 = 14474{,}74$ El intervalo de confianza de aproximadamente 95 % para el número total de escaños en los consejos municipales es [12466.32, 14474.74]. Dado que el número de escaños debe ser un entero, podríamos redondear a [12466, 14475].

Problema 3.8 (Särndal)

Considere el Ejemplo 3.3.2. Hay 24 países en Europa. Use esta información y los datos del Ejemplo 3.3.2 para estimar la media de la variable P83 para Europa. Calcule también una estimación de la varianza así como el cve correspondiente.

Solución 3.8

Del Ejemplo 3.3.2 de Särndal et al., se nos informa que de una muestra SI de $n = 50$ países de la población CO124 (que tiene $N = 124$ países), se obtuvieron los siguientes datos para la variable P83 (población en 1983 en millones) para los países europeos en la muestra:

- Número de países europeos en la muestra: $n_d = 9$ (denotaremos el dominio Europa como d).
- Suma de P83 para los países europeos en la muestra: $\sum_{k \in s_d} y_k = 205{,}2$
- Suma de P83 al cuadrado para los países europeos en la muestra: $\sum_{k \in s_d} y_k^2 = 6,232{,}84$

Información adicional para el dominio Europa:

- Número total de países en Europa en la población CO124: $N_d = 24$.

Queremos estimar la media poblacional para el dominio Europa, $\bar{Y}_d = \frac{1}{N_d} \sum_{k \in U_d} y_k$.

Estimación de la media para el dominio Europa

Un estimador insesgado para la media del dominio $\bar{Y}_d$ cuando N_d es conocido es (Särndal et al., eq. 3.3.16):

$$\hat{\bar{Y}}_d = \frac{\hat{t}_d}{N_d} = \frac{N \bar{y}_{s_d}}{N_d}$$

donde $\bar{y}_{s_d} = \frac{1}{n_d} \sum_{k \in s_d} y_k$ es la media muestral observada para las unidades del dominio que cayeron en la muestra.

$$\bar{y}_{s_d} = \frac{205{,}2}{9} = 22{,}8$$

$$\hat{\bar{Y}}_d = \frac{124 \times 22{,}8}{24} = \frac{2827{,}2}{24} = 117{,}8$$

La estimación de la media de la variable P83 para Europa es 117.8 millones.

Estimación de la varianza de la media del dominio

La varianza estimada de $\hat{\bar{Y}}_d$ es (Särndal et al., eq. 3.3.14, dividida por N_d^2):

$$\hat{V}(\hat{\bar{Y}}_d) = \frac{1}{N_d^2} \hat{V}(\hat{t}_d) = \frac{1}{N_d^2} N^2 \left(1 - \frac{n}{N}\right) \frac{P_{s_d} S^2_{y,s_d} + Q_{s_d} \bar{y}^2_{s_d}}{n}$$

donde $P_{s_d} = n_d/n$ y $Q_{s_d} = 1 - P_{s_d}$. Y $S^2_{y,s_d} = \frac{1}{n_d-1}\sum_{k\in s_d}(y_k - \bar{y}_{s_d})^2 = \frac{1}{n_d-1}\left(\sum_{k\in s_d} y_k^2 - \frac{(\sum_{k\in s_d} y_k)^2}{n_d}\right)$.
Calculamos S^2_{y,s_d}:

$$\begin{aligned} S^2_{y,s_d} &= \frac{1}{9-1}\left(6232{,}84 - \frac{(205{,}2)^2}{9}\right) \\ &= \frac{1}{8}\left(6232{,}84 - \frac{42107{,}04}{9}\right) \\ &= \frac{1}{8}(6232{,}84 - 4678{,}56) \\ &= \frac{1}{8}(1554{,}28) = 194{,}285 \end{aligned}$$

Calculamos P_{s_d} y Q_{s_d}:

$$P_{s_d} = \frac{n_d}{n} = \frac{9}{50} = 0{,}18$$
$$Q_{s_d} = 1 - P_{s_d} = 1 - 0{,}18 = 0{,}82$$

Ahora, la varianza estimada de $\hat{t}_d$:

$$\begin{aligned} \hat{V}(\hat{t}_d) &= (124)^2\left(1 - \frac{50}{124}\right)\frac{0{,}18 \times 194{,}285 + 0{,}82 \times (22{,}8)^2}{50} \\ &= 15376\,(1 - 0{,}4032258)\,\frac{34{,}9713 + 0{,}82 \times 519{,}84}{50} \\ &= 15376 \times 0{,}5967742 \times \frac{34{,}9713 + 426{,}2688}{50} \\ &= 15376 \times 0{,}5967742 \times \frac{461{,}2401}{50} \\ &= 15376 \times 0{,}5967742 \times 9{,}224802 \\ &\approx 9176{,}129 \times 9{,}224802 \approx 84640{,}78 \end{aligned}$$

Entonces, la varianza estimada de la media del dominio:

$$\hat{V}(\hat{\bar{Y}}_d) = \frac{\hat{V}(\hat{t}_d)}{N_d^2} = \frac{84640{,}78}{24^2} = \frac{84640{,}78}{576} \approx 146{,}9458$$

Coeficiente de Variación Estimado (cve)

$$\widehat{SE}(\hat{\bar{Y}}_d) = \sqrt{\hat{V}(\hat{\bar{Y}}_d)} = \sqrt{146{,}9458} \approx 12{,}12212$$

$$\widehat{cve}(\hat{\bar{Y}}_d) = \frac{\widehat{SE}(\hat{\bar{Y}}_d)}{\hat{\bar{Y}}_d} = \frac{12{,}12212}{117{,}8} \approx 0{,}102904$$

El cve correspondiente es aproximadamente 0.1029 o 10.29 %.

Problema 3.9 (Särndal)

Bajo el diseño SI, el intervalo de confianza habitual del $100(1-\alpha)\,\%$ para la media poblacional $\bar{Y}_U$ puede escribirse como $\bar{y}_s(1 \pm A)$ con

$$A = z_{1-\alpha/2} cv_{y_s}\sqrt{\frac{1-f}{n}}$$

donde $cv_{y_s} = S_{y_s}/\bar{y}_s$ es el coeficiente de variación de y en la muestra, y $f = n/N$. Asumamos que $cv_{y_s} \approx cv_{y_U} = S_{y_U}/\bar{Y}_U$ y que f es despreciable. Sea $\alpha = 5\,\%$ ($z_{0{,}975} \approx 1{,}96$). ¿Qué tamaño de muestra n se requiere, aproximadamente, para alcanzar la precisión $A \leq 3\,\%$ si (a) $cv_{y_U} = 0{,}5$; (b) $cv_{y_U} = 1{,}0$; (c) $cv_{y_U} = 1{,}5$?

Solución 3.9

Dado:

- $A = z_{1-\alpha/2} \cdot cv_{y_s} \cdot \sqrt{\frac{1-f}{n}}$
- Asumimos $cv_{y_s} \approx cv_{y_U}$
- Asumimos $f = n/N$ es despreciable, por lo tanto $(1-f) \approx 1$.
- $\alpha = 0{,}05$, entonces $z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$.
- Precisión deseada $A \leq 0{,}03$.

Con estas asunciones, la fórmula para A se simplifica a:

$$A \approx z_{1-\alpha/2} \cdot cv_{y_U} \cdot \sqrt{\frac{1}{n}} = \frac{z_{1-\alpha/2} \cdot cv_{y_U}}{\sqrt{n}}$$

Queremos encontrar n tal que $A \leq 0{,}03$:

$$\frac{z_{1-\alpha/2} \cdot cv_{y_U}}{\sqrt{n}} \leq 0{,}03$$

Despejando n:

$$\sqrt{n} \geq \frac{z_{1-\alpha/2} \cdot cv_{y_U}}{0{,}03}$$

$$n \geq \left(\frac{z_{1-\alpha/2} \cdot cv_{y_U}}{0{,}03} \right)^2$$

Sustituyendo $z_{1-\alpha/2} = 1{,}96$:

$$n \geq \left(\frac{1{,}96 \cdot cv_{y_U}}{0{,}03} \right)^2$$

$$n \geq \left(\frac{196}{3} \cdot cv_{y_U} \right)^2$$

$$n \geq (65{,}333...)^2 \cdot (cv_{y_U})^2$$

$$n \geq 4268{,}44 \cdot (cv_{y_U})^2$$

(a) Si $cv_{y_U} = 0{,}5$

$$n \geq 4268{,}44 \cdot (0{,}5)^2 = 4268{,}44 \cdot 0{,}25 = 1067{,}11$$

Se requiere un tamaño de muestra $n \approx 1068$.

(b) Si $cv_{y_U} = 1{,}0$

$$n \geq 4268{,}44 \cdot (1{,}0)^2 = 4268{,}44 \cdot 1 = 4268{,}44$$

Se requiere un tamaño de muestra $n \approx 4269$.

(c) Si $cv_{y_U} = 1{,}5$

$$n \geq 4268{,}44 \cdot (1{,}5)^2 = 4268{,}44 \cdot 2{,}25 = 9603{,}99$$

Se requiere un tamaño de muestra $n \approx 9604$.