Método del muestreo estratificado

Sean X tal que $\mathbb{E}[X] = \mu$ y Z una variable aleatoria discreta con valores en $\{1, \dots, m\}$ y tal que las probabilidades $p_j = \mathbb{P}(Z = j)$ son conocidas para cada $j = 1, \dots, m$.

Entonces,

$$\mu = \mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Z]\big] = \sum_{j=1}^{m} p_j \mathbb{E}[X \mid Z = j]$$

Por tanto, el estimador

$$\hat{\mu}_{\text{str},n}(X, Z) = \sum_{j=1}^{m} p_j \left(\frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \right)$$

donde $x_{1j}, \dots, x_{n_j j} \overset{\text{i. i. d.}}{\sim} X \mid Z = j$ y $n_1 + \cdots + n_m = n$, es un estimador insesgado de μ.

La varianza del estimador viene dada por

$$\operatorname{Var}\big(\hat{\mu}_{\text{str},n}(X, Z)\big) = \sum_{j=1}^{m} \frac{p_j^2 \sigma_j^2}{n_j}$$

sigmaj son las varianzas de las X/Z=j

donde $\sigma_j^2 = \operatorname{Var}(X \mid Z = j)$.

El estimador $\hat{\mu}_{\text{str},n}(X, Z)$ depende del tamaño de cada uno de los estratos. Una manera simple de elegir esos tamaños es mediante una *asignación proporcional*, $n_j = p_j n$ (si, además, $p_j = 1/m$ para todo $j = 1, \dots, m$, la asignación se dice que es *sistemática*). Entonces,

$$\begin{aligned}
\operatorname{Var}\big(\hat{\mu}_{\text{str},n}(X, Z)\big) &= \sum_{j=1}^{m} \frac{p_j^2 \sigma_j^2}{n_j} \\
&= \frac{1}{n} \sum_{j=1}^{m} p_j \sigma_j^2 \qquad \text{asignación es proporcional} \\
&= \frac{1}{n} \sum_{j=1}^{m} p_j \operatorname{Var}(X \mid Z = j) \\
&= \frac{1}{n} \mathbb{E}\big[\operatorname{Var}(X \mid Z)\big] \\
&\leq \frac{1}{n} \operatorname{Var}(X) \\
&= \operatorname{Var}(\hat{\mu}_n(X))
\end{aligned}$$

No obstante, el siguiente resultado establece cuál es la asignación óptima.

Teorema 1 *Los tamaños de muestra para cada estrato dados por*

$$n_j = \frac{n p_j \sigma_j}{\sum_{k=1}^{m} p_k \sigma_k}, \quad j = 1, \dots, m$$

proporcionan la menor varianza para $\hat{\mu}_{\text{str},n}(X, Z)$ *de entre todos los posibles* $n_1, \dots, n_m$ *tales que* $n_1 + \cdots + n_m = n$.

El mínimo valor de la varianza es $\frac{1}{n}\left(\sum_{j=1}^{m} p_j \sigma_j\right)^2$.

La aplicación práctica del resultado anterior se ve dificultada por el hecho de que los σ_j no se conocen usualmente. Se puede usar un procedimiento en dos etapas:

1. Usando una asignación proporcional, se generan valores $x_{1j}, \dots, x_{n_j j} \overset{\text{i. i. d.}}{\sim} X \mid Z = j$ que se usan para estimar σ_j^2 mediante la cuasivarianza muestral:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$$
$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(x_{ij} - \hat{\mu}_j\right)^2$$

   La cantidad total n de valores generados debería ser pequeña, pero lo suficientemente grande para obtener buenas estimaciones de σ_j. Además, debe ser $n_j \geq 2$ para que esas estimaciones se puedan calcular.

2. Usando las estimaciones $\hat{\sigma}_j^2$ obtenidas, se aplica el teorema 1 para calcular el tamaño de muestra óptimo n_j de cada estrato. Debe ser $n_j \geq 2$ para poder realizar una estimación, a partir de la cuasivarianza muestral en cada estrato, de la varianza de $\hat{\mu}_{\text{str},n}(X, Z)$.