2 Ideas básicas sobre estimación en muestreo probabilístico

Sea una población de tamaño conocido N, formada por $U = \{1, \dots, N\}$. Una muestra es un subconjunto s de U, seleccionada mediante un diseño probabilístico. El tamaño de la muestra se denota n_S.

El estadístico debe seleccionar un diseño muestral y definir el procedimiento de selección (algoritmo muestral) y el estimador correspondiente, ya que ambos están relacionados. Los algoritmos muestrales pueden clasificarse en: (i) enumerativos, (ii) de martingalas, (iii) secuenciales, (iv) por extracción individual, (v) eliminatorios y (vi) de rechazo. En general, consisten en experimentos aleatorizados que determinan qué elementos se incluyen en la muestra.

Ejemplo. Un algoritmo secuencial para muestreo aleatorio simple sin reemplazamiento: se selecciona un primer elemento entre los N con probabilidad $1/N$. Se selecciona un segundo entre los $N-1$ restantes con probabilidad $1/(N-1)$. Se repite hasta seleccionar n elementos, cada vez entre los restantes.

Tras seleccionar la muestra, se observa el valor y_k, $\forall k \in s$, que permiten calcular estimaciones de los parámetros.

2.1 Diseño muestral

Definición. Dado un algoritmo muestral, el diseño muestral es una función $p(\cdot)$ que asigna a cada muestra $s \in \Omega$ la probabilidad $P(S = s) = p(s)$ de ser seleccionada. Esta función cumple: $p(s) \geq 0$ para todo $s \in \Omega$ y $\sum_{s\in\Omega} p(s) = 1$.

Ejemplo. En el muestreo aleatorio simple sin reemplazamiento de tamaño fijo n, todas las muestras tienen igual probabilidad: $p(s) = \binom{N}{n}^{-1}$ para todo $s \in \Omega$.

Comentario. Hay que tener en cuenta que, en este caso, todas las muestras son de tamaño n, pero no siempre es así. Además, diferentes algoritmos pueden implementar el mismo diseño.

Definición (Estrategia muestral)**.** Combinación de un diseño muestral y un estimador.

Elegir una buena estrategia es clave para obtener estimaciones fiables del parámetro poblacional de interés.

2.2 Probabilidades de inclusión

Definición (Probabilidad de inclusión)**.** Sea una población $\{u_1, \dots, u_N\}$ y un diseño muestral $p(\cdot)$. Definimos la variable indicadora de inclusión del elemento k como

$$I_k = \begin{cases} 1 & \text{si } u_k \in S \\ 0 & \text{si } u_k \notin S \end{cases}$$

Definición. La probabilidad de inclusión de primer orden del elemento u_k, π_k, es la probabilidad de que dicho elemento esté en la muestra $\pi_k = P(u_k \in S) = \sum_{s \ni u_k} p(s)$.

Definición. La probabilidad de inclusión de segundo orden de los elementos u_k y u_l, π_{kl}, es la probabilidad de que ambos estén en la muestra $\pi_{kl} = P(u_k, u_l \in S) = \sum_{s \ni u_k, u_l} p(s)$.

Ejemplo (Muestreo aleatorio simple sin reemplazamiento)**.** Tenemos que $\pi_k = \dfrac{n}{N}$, para todo k y $\pi_{kl} = \dfrac{n(n-1)}{N(N-1)}$, para $k \neq l$.

Definición (Diseño probabilístico)**.** Un diseño es probabilístico si todo elemento tiene $\pi_k > 0$. Así, todos los elementos tienen posibilidad de ser seleccionados.

Definición (Diseño medible)**.** Un diseño es medible si además de $\pi_k > 0$ también cumple que $\pi_{kl} > 0$ para todo $k \neq l$. Si $k = l$, entonces $\pi_{kk} = \pi_k$, ya que $P(I_k^2 = 1) = P(I_k = 1)$.

En diseños simples (una etapa), las probabilidades de inclusión suelen ser conocidas. En diseños complejos (varias etapas), no siempre es posible conocerlas desde el inicio.

2.3 La noción de estadístico

Para estimar parámetros poblacionales a partir de la muestra, utilizamos funciones basadas en los datos muestrales.

Definición (Estadístico)**.** Sea $S \in \Omega$ una muestra aleatoria. Un estadístico $Q = Q(S)$ es una función real de S, cuya distribución se llama distribución en el muestreo de Q. Un estadístico es una variable aleatoria que depende de la muestra extraída, pero no de parámetros desconocidos.

Se buscan estadísticos llamados estimadores que no varíen mucho entre las muestras, y centrados en el valor real. Se definen: $\mathbb{E}(Q) = \sum_{s\in\Omega} p(s) \cdot Q(s)$, $\mathrm{var}(Q) = \sum_{s\in\Omega} p(s) \cdot [Q(s) - \mathbb{E}(Q)]^2$ y $\mathrm{cov}(Q_1, Q_2) = \sum_{s\in\Omega} p(s) \cdot (Q_1(s) - \mathbb{E}(Q_1)) \cdot (Q_2(s) - \mathbb{E}(Q_2))$.

2.4 Indicadores de pertenencia a la muestra

Muchos estadísticos pueden escribirse en función de los indicadores $I_k(S)$. Por ejemplo, el total muestral de una variable y: $Q(S) = \sum_{k\in S} y_k = \sum_{k\in U} I_k(S) \cdot y_k$.

Proposición. Para todo $k, l = 1, \ldots, N$:

(a) Sabemos que $I_k \sim \mathcal{B}(\pi_k)$. $\mathbb{E}[I_k] = 1 \cdot \pi_k + 0 \cdot (1 - \pi_k) = \pi_k$.

(b) $\mathrm{var}(I_k) = \mathbb{E}[I_k^2] - (\mathbb{E}[I_k])^2 = \pi_k - \pi_k^2 = \pi_k(1 - \pi_k)$.

(c) $\mathrm{cov}(I_k, I_l) = \mathbb{E}[I_k I_l] - \mathbb{E}[I_k]\mathbb{E}[I_l] = \pi_{kl} - \pi_k \pi_l$.

Proposición. Si el tamaño muestral es fijo n:

(a) $\sum_{k\in U} \pi_k = \sum_{k\in U} \mathbb{E}[I_k] = \mathbb{E}[\sum_{k\in U} I_k] = \mathbb{E}[n] = n$.

(b) $\sum_{k\in U} \sum_{l\in U, l\neq k} \pi_{kl} = \sum_k \sum_{l\neq k} \mathbb{E}[I_k I_l] = \mathbb{E}\left[\sum_k \sum_{l\neq k} I_k I_l\right] = \mathbb{E}[n(n-1)] = n(n-1)$.

(c) $\sum_{l\in U, l\neq k} \pi_{kl} = \mathbb{E}\left[I_k \sum_{l\neq k} I_l\right] = \mathbb{E}[I_k(n - I_k)] = \pi_k(n-1)$.

Ejemplo (Muestreo aleatorio simple sin reemplazo)**.** En este caso $\pi_k = \dfrac{n}{N}$, $\pi_{kl} = \dfrac{n(n-1)}{N(N-1)}$. Entonces: $\sum_k \pi_k = N \cdot \dfrac{n}{N} = n$, $\sum_k \sum_{l\neq k} \pi_{kl} = N(N-1) \cdot \dfrac{n(n-1)}{N(N-1)} = n(n-1)$ y $\sum_{l\neq k} \pi_{kl} = (N-1) \cdot \dfrac{n(n-1)}{N(N-1)} = (n-1) \cdot \dfrac{n}{N} = (n-1)\pi_k$.

2.5 Estimadores y sus propiedades básicas

Sea $\theta = (\theta_1, \ldots, \theta_j)$ un vector de parámetros poblacionales. Un estimador produce una estimación puntual $\theta(s)$ del valor real de θ.

Ejemplo. Se desea estimar el total poblacional $Y_U = \sum_{k\in U} y_k$, un estimador en un muestreo sin reemplazo con probabilidades iguales es $\theta(S) = \frac{N}{n} \sum_{k\in S} y_k = \sum_{k\in U} I_k y_k \cdot \frac{N}{n}$, mientras que la estimación es $\theta(s) = \frac{N}{n} \sum_{k\in s} y_k$.

A partir de ahora se usa s para representar tanto la muestra aleatoria como una muestra particular, para simplificar.

Definición (Distribución del estimador)**.** Conjunto de posibles valores C que puede tomar $\theta(s)$ junto con sus probabilidades $P_C = P(\theta = c) = \sum_{s\in\Omega_C} p(s)$, donde Ω_C es el conjunto de muestras que hacen que $\theta(s) = c$.

Definición (Insesgadez)**.** Un estimador θ es insesgado si su esperanza es igual al parámetro poblacional: $\mathbb{E}[\theta] = \theta$.

Definición (Error cuadrático medio)**.** Medida de calidad general del estimador: $MSE(\theta) = \mathbb{E}[(\theta - \theta)^2] = \sum_{s\in\Omega} p(s) \cdot (\theta(s) - \theta)^2$.

Dado que $\text{var}(\theta) = \sum_{s\in\Omega} p(s) \cdot (\theta(s) - \mathbb{E}[\theta])^2$, tenemos que $MSE(\theta) = \text{var}(\theta) + B(\theta)^2$. *Ver demostración en el bloque de Inferencia.*

Ejemplo. En un muestreo con probabilidades iguales y sin reemplazo, $\theta = \frac{N}{n} \sum_{k\in s} y_k$ es insesgado para $\sum_{k\in U} y_k$.

En la práctica, se prefiere aquel estimador cuyo MSE sea menor, ya que indica mayor proximidad sistemática al valor real. Por ello, suele seleccionarse un estimador insesgado con la menor varianza posible.