TEMA 3. El estimador de Horvitz–Thompson

1 Introducción

Los resultados obtenidos en el capítulo anterior justifican que en la práctica y en los diseños muestrales más usuales, sea el estimador de Horvitz–Thompson el más usado. En este capítulo estudiaremos con más detalle este estimador, lo aplicamos a los diseños muestrales usuales: muestreo aleatorio simple, muestreo sistemático, muestreo estratificado y muestreo por conglomerados, mostrando además cómo se aplica en la estimación de parámetros en subpoblaciones.

Supongamos que el parámetro tiene la forma:

$$\theta(\mathbf{y}) = \sum_{i=1}^{N} a_i y_i \text{ con } a_i \in R \tag{1}$$

llamamos estimador de Horvitz y Thompson para θ al estimador dado por la expresión:

$$\widehat{\theta}_{HT}(s, \mathbf{y}) = \sum_{i \in s} \frac{a_i y_i}{\pi_i}, \tag{2}$$

que se puede escribir de forma alternativa:

$$\widehat{\theta}_{HT}(s, \mathbf{y}) = \sum_{i=1}^{N} \frac{a_i y_i}{\pi_i} I_i(s)$$

en función de las variables indicadoras de cada unidad.

En muestreo en poblaciones finitas los parámetros de interés más comunes, el total $Y = \sum_{i=1}^{N} y_i$, la media $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, y una proporción, $P = \frac{1}{N}\sum_{i=1}^{N} A_i$, pueden estimarse con el estimador de Horvitz-Thompson: para el total, $\widehat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}$, para la media, $\widehat{\overline{Y}} = \frac{1}{N}\sum_{i \in s} \frac{y_i}{\pi_i}$, y para una proporción, $\widehat{P} = \frac{1}{N}\sum_{i \in s} \frac{A_i}{\pi_i}$. Para poder definir estos estimadores es evidente que $\pi_i > 0, i = 1, \ldots, N$, es decir, que el muestreo sea probabilístico.

2 Propiedades

2.0.1 Insesgadez

El primer resultado que presentamos es la insesgadez del estimador estudiado. Si bien la "insesgadez" no garantiza que el estimador sea razonable, en muestreo de poblaciones finitas es una propiedad deseable en los estimadores.

Proposición 2.1 *El estimador de Horvitz-Thompson de* $\theta(\mathbf{y}) = \sum_{i=1}^{N} a_i y_i$ *es insesgado de dicho parámetro.*

Demostración.

$$E\left(\widehat{\theta}_{HT}\right) = E\left(\sum_{i=1}^{N} \frac{a_i}{\pi_i} I_i(s) y_i\right) = \sum_{i=1}^{N} \frac{a_i}{\pi_i} y_i E\left(I_i(s)\right) = \sum_{i=1}^{N} \frac{a_i}{\pi_i} y_i \pi_i = \theta$$

2.1 Precisión

Puesto que el estimador de Horvitz-Thompson es insesgado, calcularemos su varianza, previamente al error de muestreo. En adelante denotaremos por $y_i^* = \frac{y_i}{\pi_i}$ y lo llamaremos valor expandido de y_i.

Proposición 2.2 *La varianza del estimador de Horvitz-Thompson de* θ *viene dada por la siguiente expresión:*

$$V\left(\widehat{\theta}_{HT}\right) = \sum_{i=1}^{N} a_i^2 \left(\frac{y_i}{\pi_i}\right)^2 \pi_i(1-\pi_i) + \sum_{i \neq j=1}^{N} a_i a_j \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j) =$$

$$= \sum_{i,j=1}^{N} a_i a_j \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\left(\pi_{ij} - \pi_i \pi_j\right). \tag{3}$$

Demostración.

$$V\left(\widehat{\theta}_{HT}\right) = V\left(\sum_{i=1}^{N} a_i y_i^* I_i(s)\right) = \sum_{i=1}^{N} V\left(a_i y_i^* I_i(s)\right) + \sum_{i \neq j=1}^{N} cov\left(a_i y_i^* I_i(s), a_j y_j^* I_j(s)\right) =$$

$$= \sum_{i=1}^{N} a_i^2 y_i^{*\,2} V\left(I_i(s)\right) + \sum_{i \neq j=1}^{N} a_i a_j y_i^* y_j^* cov\left(I_i(s), I_j(s)\right)$$

Sustituyendo $V(I_i(s))$ y $cov(I_i(s), I_j(s))$ por sus valores se obtiene la expresión (3). Esta expresión para $V(\widehat{\theta}_{HT})$ fue introducida por *Horvitz* y *Thompson* (1952), y la denotaremos $V_{HT}(\widehat{\theta}_{HT})$. Para diseños de tamaño fijo se puede deducir una expresión alternativa para $V(\widehat{\theta}_{HT})$.

Proposición 2.3 *Si el diseño es de tamaño fijo, el estimador de Horvitz–Thompson de (1) tiene por varianza:*

$$V\left(\widehat{\theta}_{HT}\right) = -\frac{1}{2} \sum_{i,j=1}^{N} \left(\pi_{ij} - \pi_i \pi_j\right) \left(a_i \frac{y_i}{\pi_i} - a_j \frac{y_j}{\pi_j}\right)^2 . \qquad (4)$$

Demostración.-

$$-\frac{1}{2} \sum_{i,j=1}^{N} \Delta_{ij} \left(a_i y_i^* - a_j y_j^*\right)^2 = -\frac{1}{2} \sum_{i,j=1}^{N} \Delta_{ij} \left((a_i y_i^*)^2 + (a_j y_j^*)^2 - 2 a_i a_j y_i^* y_j^*\right) =$$

$$= -\frac{1}{2} \left(\sum_{i=1}^{N} (a_i y_i^*)^2 \sum_{j=1}^{N} \Delta_{ij} + \sum_{j=1}^{N} (a_j y_j^*)^2 \sum_{i=1}^{N} \Delta_{ij} - 2 \sum_{i,j=1}^{N} a_i a_j \Delta_{ij} y_i^* y_j^* \right) .$$

Con las propiedades de la matriz de diseño

$$-\frac{1}{2} \left(\sum_{i,j=1}^{N} \Delta_{ij} \left(a_i y_i^* - a_j y_j^*\right) \right)^2 = \sum_{i,j=1}^{N} \Delta_{ij} a_i a_j y_i^* y_j^* = V\left(\widehat{\theta}_{HT}\right) .$$

Esta expresión se debe a *Yates* y *Grundy*, y se denotará por $V_{YG}\left(\widehat{\theta}_{HT}\right)$:

$$V_{YG}\left(\widehat{\theta}_{HT}\right) = -\frac{1}{2} \sum_{i,j=1}^{N} \left(\pi_{ij} - \pi_i \pi_j\right) \left(a_i \frac{y_i}{\pi_i} - a_j \frac{y_j}{\pi_j}\right)^2 .$$

En esta última expresión observamos que si $\pi_i \propto a_i y_i$, la varianza del estimador $\widehat{\theta}_{HT}$ es cero. En el caso de la media, el total o la

proporción, los a_i son iguales entre sí, por tanto eligiendo un diseño muestral con $\pi_i \propto y_i$ tendríamos un estimador con varianza cero. Este resultado ideal es inaccesible, puesto que y_i es desconocido, pero nos da idea de cómo deben elegirse los π_i, y por tanto el diseño muestral: si se tiene alguna variable auxiliar x conocida $\forall i \in U$, y que esté altamente relacionada con la variable de estudio y, elegiremos un diseño muestral tal que sus probabilidades de primer orden sean proporcionales a la variable auxiliar, $\pi_i \propto x_i$ (en la práctica se suele tomar $\pi_i = nx_i / \sum_{j=1}^{N} x_j$).

Una vez deducida la fórmula de la varianza del estimador $\widehat{\theta}_{HT}$, observamos que depende de todos los valores poblacionales y por tanto es desconocida. El siguiente paso es construir un estimador de esta varianza que permita dar un valor a partir de una muestra concreta $s \in S_d$.

Proposición 2.4 *Si $\pi_{ij} > 0 \ \forall (i,j)$, un estimador insesgado de la varianza viene dado por :*

$$\widehat{V}\left(\widehat{\theta}_{HT}\right) = \sum_{i,j \in s} \frac{\Delta_{ij}}{\pi_{ij}} a_i a_j \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{5}$$

Demostración.-

$$E\left(\widehat{V}\left(\widehat{\theta}_{HT}\right)\right) = E\left(\sum_{i,j=1}^{N} I_i I_j \frac{\Delta_{ij}}{\pi_{ij}} a_i a_j \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}\right) =$$

$$= \sum_{i,j=1}^{N} E\left(I_i I_j\right) \cdot \frac{\Delta_{ij}}{\pi_{ij}} a_i a_j \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} = \sum_{i,j=1}^{N} \Delta_{ij} a_i a_j \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} = V_{HT}\left(\widehat{\theta}_{HT}\right).$$

Proposición 2.5 *Si el diseño es de tamaño fijo y $\pi_{ij} > 0 \ \forall (i,j)$, un estimador insesgado de (4) viene dado por:*

$$\widehat{V}\left(\widehat{\theta}_{HT}\right) = -\frac{1}{2} \sum_{i,j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \left(a_i \frac{y_i}{\pi_i} - a_j \frac{y_j}{\pi_j}\right)^2. \tag{6}$$

Demostración.-

$$E\left(\widehat{V}\left(\widehat{\theta}_{HT}\right)\right) = E\left(-\frac{1}{2} \sum_{i,j=1}^{N} \Delta_{ij}^{*} (a_i y_i^{*} - a_j y_j^{*})^2 I_i I_j\right) =$$

$$= -\frac{1}{2}\sum_{i,j=1}^{N}\Delta_{ij}^{*}(a_i y_i^{*} - a_j y_j^{*})^2 E\left(I_i I_j\right) = -\frac{1}{2}\sum_{i,j=1}^{N}\Delta_{ij}(a_i y_i^{*} - a_j y_j^{*})^2 = V_{YG}\left(\widehat{\theta}_{HT}\right).$$

El estimador dado por (5) se conoce como estimador de la varianza de *Horvitz* y *Thompson* y lo denotamos por $\widehat{V}_{HT}\left(\widehat{\theta}_{HT}\right)$, mientras que el segundo dado en (6) se debe a *Yates* y *Grundy* y se denota por $\widehat{V}_{YG}\left(\widehat{\theta}_{HT}\right)$. Aunque en un diseño de tamaño fijo, ambas varianzas son iguales, no tienen porqué serlo sus respectivos estimadores, aunque los dos sean insesgados.

Resumiendo, para el parámetro total, Y, tenemos los siguientes resultados tenemos los siguientes resultados: $\widehat{Y}_{HT} = \sum_{i\in s}\frac{y_i}{\pi_i}$,

$$V_{HT}(\widehat{Y}_{HT}) = \sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2\pi_i(1-\pi_i) + \sum_{i\neq j=1}^{N}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j),$$

$$V_{YG}(\widehat{Y}_{HT}) = -\frac{1}{2}\sum_{i,j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2,$$

$$\widehat{V}_{HT}(\widehat{Y}_{HT}) = \sum_{i,j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}, \text{ y}$$

$$\widehat{V}_{YG}(\widehat{Y}_{HT}) = -\frac{1}{2}\sum_{i,j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.$$

3 Estimación por intervalos.

En el tema dedicado a estimación puntual y por intervalo se introdujo una lección en la que se definía el concepto de intervalo de confianza, como un intervalo aleatorio que contiene al valor desconocido del parámetro con una probabilidad fija. La misma idea se puede trasladar a poblaciones finitas.

En poblaciones finitas el método pivotal permite construir un intervalo de confianza para una parámetro $\theta(\mathbf{y})$ a partir de un estimador suyo, $\widehat{\theta}(s,\mathbf{y})$, cuando la distribución en el muestreo de $\widehat{\theta}$ sea aproximadamente normal de media θ, y exista un estimador "consistente", $\widehat{V}(\widehat{\theta})$ de su varianza, $V(\widehat{\theta})$. Esta situación está garantizada con el siguiente teorema, debido a Lindeberg–Hájeck.

Teorema 3.1 *El estimador de Horvitz–Thompson es asintóticamente normal.*

Así, si el parámetro θ es lineal, podemos definir su estimador Horvitz–Thompson, $\widehat{\theta}_{HT}$, que es insesgado y para tamaño de muestra grande es aproximadamente normal, el intervalo

$$\left(\widehat{\theta}_{HT} - z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\widehat{\theta}_{HT})}, \widehat{\theta}_{HT} + z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\widehat{\theta}_{HT})}\right),$$

es un intervalo de confianza para θ, a un nivel de confianza de $1-\alpha$.

4 El efecto del diseño

El diseño muestral que se considera tiene un papel fundamental en la estimación, ya que determina las probabilidades de inclusión, y en función de ellas se expresan los estimadores y sus varianzas.

Si nos planteáramos el problema de cuándo es mejor un diseño muestral frente a otro, tendríamos que comparar cuál de ellos proporciona mejores estimaciones, o lo que es lo mismo, cuál de las varianzas de estas estimaciones es menor. Pues bien, la com0paración de dos estrategias (d_1, e_1) y (d_2, e_2) se hace a partir de la razón de sus varianzas (si estamos con estimadores insesgados), conocida como EFECTO DEL DISEÑO. Así se define:

$$e.d.(d_1, d_2) = \frac{V_{d1}(e_1}{V_{d2}(e_2}$$

Si $e.d. < 1$ la primera estrategia será mejor que la segunda.

En general se toma una estrategia como base para la comparación, y esta estrategia es el $(MAS, \widehat{\theta}_{HT})$. Así por ejemplo si estamos considerando $\theta = \overline{Y}$, para un diseño d y un estimador e, el efecto del diseño será:

$$e.d. = \frac{V_d(e)}{(1-f)\frac{S^2}{n}}$$

Si este valor excede de la unidad, perderíamos precisión al considerar (d, e) frente al método de muestreo más simple que es el MAS.

5 Estimación de varias variables de estudio

Es frecuente obtener las muestras sobre variables de estudio que expliquen diferentes aspectos de la población, por lo que el muestreo sobre la población U se realiza al recoger la información de la variable $\mathcal{Y}$, que tiene dimensión q, $(y^1, \ldots, y^q)$, esto es, tantas componentes como aspectos a estudiar en la población.

En este caso, se desea hacer una estimación sobre cada variable de estudio para un parámetro del tipo,

$$\theta_i(\mathbf{y}) = \sum_{j=1}^{N} a_j^i y_j^i, \quad i = 1, \ldots, q,$$

es decir, se desea estimar el vector $(\theta_1, \ldots, \theta_q)$ a partir de la muestra obtenida.

Teorema 5.1 *El vector $(\theta_1, \ldots, \theta_q)$ tiene por estimador insesgado a $(\widehat{\theta}_1, \ldots, \widehat{\theta}_q)$, donde*

$$\widehat{\theta}_i = \sum_{j \in s} a_j^i \frac{y_j^i}{\pi_j}.$$

La matriz de covarianzas del estimador viene dada por los elementos

$$cov\left(\widehat{\theta}_r, \widehat{\theta}_s\right) = \sum_{k,l \in U} \Delta_{kl} a_k^r \frac{y_k^r}{\pi_k} a_l^s \frac{y_l^s}{\pi_l}, \quad r, s = 1, \ldots, q,$$

siendo un estimador insesgado de esta matriz, la dada por los elementos

$$\widehat{cov}\left(\widehat{\theta}_r, \widehat{\theta}_s\right) = \sum_{k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} a_k^r \frac{y_k^r}{\pi_k} a_l^s \frac{y_l^s}{\pi_l}, \quad r, s = 1, \ldots, q,$$

Estos resultados se obtienen en forma similar a los obtenidos para una variable. Los elementos diagonales en una y otra matriz son precisamente las varianzas de los estimadores $V(\widehat{\theta}_i)$ y sus correspondientes estimadores $\widehat{V}(\widehat{\theta}_i)$, ya vistos anteriormente.

Además, para diseños muestrales de tamaño fijo, las covarianzas y sus estimaciones tienen también una representación especial

$$\operatorname{cov}\left(\widehat{\theta}_r, \widehat{\theta}_s\right) = -\frac{1}{2}\sum_{k,l\in U}\Delta_{kl}\left(a_k^r\frac{y_k^r}{\pi_k} - a_l^r\frac{y_l^r}{\pi_l}\right)\left(a_k^s\frac{y_k^s}{\pi_k} - a_l^s\frac{y_l^s}{\pi_l}\right).$$

6 Estimación en subpoblaciones

Cuando sólo los individuos de una subpoblación o dominio $U_d \subset U$ son objeto de interés, la variable de estudio puede ser truncada de la forma

$$y_i^d = \begin{cases} y_i & i \in U_d \\ 0 & i \notin U_d \end{cases}$$

Esta nueva variable permite estimar un parámetro $\theta^d(\mathbf{y}) = \sum_{i\in U_d} a_i y_i$ al reescribirlo como parámetro de la población de la forma

$$\theta^d(\mathbf{y}) = \sum_{i\in U} a_i y_i^d,$$

en cuyo caso la estimación de Horvitz–Thompson proporciona el estimador

$$\widehat{\theta}_{HT}^d(\mathbf{y}) = \sum_{i\in s} a_i\frac{y_i^d}{\pi_i} = \sum_{i\in s_d} a_i\frac{y_i}{\pi_i}$$

con la varianza y la estimación de ésta correspondientes y donde s_d denota los elementos de la muestra s que están en la subpoblación U_d.

El problema de la estimación del tamaño de la subpoblación U_d, N_d, cuando éste sea desconocido puede resolverse con el estimador de Horvitz–Thompson del total de la variable x^d dada por

$$x_i^d = \begin{cases} 1 & i \in U_d \\ 0 & i \notin U_d \end{cases}$$

con el que se obtiene

$$\widehat{N_d} = \sum_{i\in s}\frac{x_i}{\pi_i} = \sum_{i\in s_d}\frac{1}{\pi_i}.$$

7 Aplicación a algunos diseños muestrales

En este capítulo vamos a estudiar los principales diseños muestrales. Estudiaremos sus características más importantes, las razones de su uso, el proceso de selección de la muestra y vamos a ver la forma que adopta el estimador de Horvitz–Thompson para estos diseños usados en la práctica.

7.1 Muestreo aleatorio simple

El método de muestreo con probabilidades iguales es el método base de los muestreos probabilísticos. Aunque en la práctica se utilizan planes de muestreo más elaborados, los conceptos fundamentales y las fórmulas importantes derivan de este método elemental. Por otra parte sirve como referencia cuando se intenta evaluar la calidad de un método cualquiera.

7.1.1 Muestreo aleatorio simple: planteamiento del método

El muestreo aleatorio simple es el diseño muestral en el cual n unidades distintas son seleccionadas de entre las N unidades poblacionales, de forma que cada posible combinación de n unidades tiene la misma probabilidad de ser elegida. Así si llamamos U a la población, este diseño muestral viene definido por el par $d = (S_d, p_d)$, donde el soporte del diseño, S_d está constituido por todas las muestras de tamaño n que se pueden obtener de U, y la distribución de probabilidad asociada es la uniforme.

Así pues el muestreo aleatorio simple que suele denotarse por MAS(N,n), se puede definir como:

$$p : S_d \to [0, 1]$$

tal que

$$p(s) = \frac{1}{\binom{N}{n}}$$

Por otra parte la probabilidad de que una unidad cualquiera u_i pertenezca a la muestra es:

$$\pi_i = \sum_{j=1}^{n} p(u_i \textit{ sea elegida en la extracción j-ésima}) =$$

$$= \frac{1}{N} + \frac{N-1}{N}\frac{1}{N-1} + ... + \frac{N-1}{N}\frac{N-2}{N-1}...\frac{N-n+1}{N-n+2}\frac{1}{N-n+1} =$$

$$\frac{1}{N} + ... + \frac{1}{N} = \frac{n}{N}.$$

La probabilidad que tienen dos unidades de pertenecer simultáneamente a la muestra viene dada por:

$$\pi_{ij} = \frac{\textit{casos favorables}}{\textit{casos posibles}} =$$

$$\frac{\textit{número de muestras que contienen simultáneamente al par (i,j)}}{\textit{número total de muestras}} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

7.1.2 Estimadores y sus varianzas

$$\widehat{Y}_{HT} = \sum_{i\in s}\frac{y_i}{n}N = N\overline{y}, \quad \widehat{\overline{Y}}_{HT} = \sum_{i\in s}\frac{y_i}{n} = \overline{y}, \quad \widehat{P}_{HT} = \sum_{i\in s}\frac{A_i}{n} = p.$$

Sustituyendo en la expresión de la varianza del estimador Horvitz–Thompson para un parámetro lineal $\theta = \sum_{i=1}^{N} a_i y_i$, Δ_{ij}, π_{ij} y π_i, $\forall i, j$ por los valores anteriores, se tiene, en el caso del total, usando $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 = \frac{1}{2N(N-1)}\sum_{i,j=1}^{N}(y_i - y_j)^2$,

$$V\left(\widehat{Y}_{HT}\right) = V\left(N\overline{y}\right) = \frac{1}{2}\frac{f(1-f)}{N-1}\left(\frac{N}{n}\right)^2\sum_{i,j=1}^{N}(y_i - y_j)^2 = N^2\frac{1-f}{n}S^2,$$

En el caso de la media se obtiene de forma análoga $V(\overline{y}) = \frac{1-f}{n}S^2$ y puesto que en una población dicotómica $S^2 = \frac{PQ}{N-1}N$, en el caso de una proporción se tiene (siendo $P = \frac{1}{N}\sum_{i=1}^{N}A_i$, y $Q = 1 - P$) $V(p) = \frac{N-n}{N-1}\frac{PQ}{n}$.

En cuanto a los estimadores de estas varianzas, notando $s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \overline{y})^2$, se obtiene:

$$\widehat{V}\left(N\overline{y}\right) = N^2 \frac{1-f}{n} s^2, \quad \widehat{V}\left(\overline{y}\right) = \frac{1-f}{n} s^2, \quad \widehat{V}\left(p\right) = \frac{N-n}{N-1} \frac{pq}{n-1}.$$

7.1.3 Intervalos de confianza

Para el total:

$$\left(N\overline{x} - k\frac{Ns}{\sqrt{n}}\sqrt{1-f}\ ,\ N\overline{x} + k\frac{Ns}{\sqrt{n}}\sqrt{1-f} \right),$$

Para la media:

$$\left(\overline{x} - k\frac{s}{\sqrt{n}}\sqrt{1-f}\ ,\ \overline{x} + k\frac{s}{\sqrt{n}}\sqrt{1-f} \right),$$

Para una proporción:

$$\left(\widehat{P} - k\sqrt{\frac{\widehat{P}\widehat{Q}}{n-1}(1-f)}\ ,\ \widehat{P} + k\sqrt{\frac{\widehat{P}\widehat{Q}}{n-1}(1-f)} \right).$$

donde k se obtiene a partir de la desigualdad de Tchebichev o bien a partir de los valores críticos de una Normal según el método utilizado para la construcción del intervalo.

7.2 Muestreo estratificado

7.2.1 Introducción.

En el muestreo estratificado la población de N unidades es primero dividida en subpoblaciones de $N_1,\ N_2,\ \ldots,\ N_L$ unidades, respectivamente. Estas subpoblaciones no se superponen y juntas forman la totalidad de la población, es decir:

$$\sum_{h=1}^{L} N_h = N$$

Estas subpoblaciones reciben el nombre de estratos. Una vez que han sido determinados los estratos se extrae una muestra de cada uno. La muestra final está compuesta por el conjunto de estas submuestras.

El proceso de muestreo se realiza de modo independiente en cada estrato, lo que permite la aplicación simultánea de métodos de muestreo diferentes de acuerdo con la información de que se disponga, el costo y las razones que motivaron la estratificación.

Si se toma una muestra aleatoria simple de cada estrato, el procedimiento se conoce como *muestreo estratificado aleatorio.*

7.2.2 Diseño muestral. Estimadores y sus varianzas

En la población U estratificada en L estratos U_1, ..., U_L, con el diseño muestral $d_h = (S_h, p_h)$ definido en cada estrato, independientemente de los demás estratos, se obtiene una muestra $s = (s_1, \ldots, s_L)$ con probabilidad $p(s) = p_1(s_1)p_2(s_2)\cdots p_L(s_L)$.

Las probabilidades de inclusión de primer orden son:

$\pi_i = \pi_i^{d_h}$, probabilidad de inclusión en la muestra s_h si el elemento u_i está en el estrato U_h.

Análogamente, las probabilidades de inclusión de segundo orden son:

$$\pi_{ij} = \begin{cases} \pi_{ij}^{d_h} \text{ si ambos están en } U_h \\ \\ \pi_i^{d_h}\pi_j^{d_k} \text{ si } u_i \text{ está en } U_h \text{ y } u_j \text{ en } U_k,\ k \neq h \end{cases}$$

Escribiendo el parámetro poblacional lineal como suma de los parámetros lineales sobre cada estrato

$$\theta(\mathbf{y}) = \sum_{i=1}^{N} a_i y_i = \sum_{h=1}^{L} \sum_{i \in U_h} a_i y_i = \sum_{h=1}^{L} \theta_h(y_{h1}, \ldots, y_{hN_h}),$$

si tenemos un estimador insesgado, según el diseño, en cada estrato, podemos obtener un estimador insesgado del parámetro poblacional sumando dichos estimadores. Así, si en cada estrato el diseño correspondiente verifica $\pi_i^h > 0$ y $\pi_{ij}^h > 0\ \forall i, j$, tenemos:

$$\widehat{\theta}_{HT} = \sum_{h=1}^{L} \sum_{i \in s_h} a_i \frac{y_i}{\pi_i},$$

$$V\left(\widehat{\theta}_{HT}\right) = \sum_{h=1}^{L} \sum_{i \in U_h} \Delta_{ij} a_i a_j \frac{y_i}{\pi_i}\frac{y_j}{\pi_j},\ V_{YG}\left(\widehat{\theta}_{HT}\right) = -\frac{1}{2}\sum_{h=1}^{L} \sum_{i \in U_h} \Delta_{ij}\left(a_i\frac{y_i}{\pi_i} - a_j\frac{y_j}{\pi_j}\right)^2,$$

$$\widehat{V}\left(\widehat{\theta}_{HT}\right)=\sum_{h=1}^{L}\sum_{i\in s_h}\frac{\Delta_{ij}}{\pi_{ij}}a_ia_j\frac{y_i}{\pi_i}\frac{y_j}{\pi_j},\ \widehat{V}_{YG}\left(\widehat{\theta}_{HT}\right)=-\frac{1}{2}\sum_{h=1}^{L}\sum_{i\in s_h}\frac{\Delta_{ij}}{\pi_{ij}}\left(a_i\frac{y_i}{\pi_i}-a_j\frac{y_j}{\pi_j}\right)^2.$$

Para el caso particular en que cada estrato se realice un muestreo $MAS(N_h,n_h)$ se dice que el diseño muestral es un muestreo estratificado aleatorio y las probabilidades de inclusión son: $\pi_i^h=\frac{n_h}{N_h}$, $\pi_{ij}^h=\frac{n_h}{N_h}\frac{(n_h-1)}{(N_h-1)}$, y los estimadores adoptan la forma:

$$\widehat{Y}_{st}=\sum_{h=1}^{L}N_h\overline{y}_h,\quad \widehat{\overline{Y}}_{st}=\sum_{h=1}^{L}\frac{N_h}{N}\overline{y}_h,\quad \widehat{P}_{st}=\sum_{h=1}^{L}\frac{N_h}{N}p_h,$$

con varianzas dadas por:

$$V(\widehat{Y}_{st})=\sum_{h=1}^{L}N_h^2\frac{1-f_h}{n_h}S_h^2,\ V(\widehat{\overline{Y}}_{st})=\frac{1}{N^2}V(\widehat{Y}_{st}),\ V(\widehat{P}_{st})=\sum_{h=1}^{L}(\frac{N_h}{N})^2\frac{N_h-n_h}{N_h-1}\frac{P_hQ_h}{n_h},$$

y con las estimaciones de éstas siguientes:

$$\widehat{V}(\widehat{Y}_{st})=\sum_{h=1}^{L}N_h^2\frac{1-f_h}{n_h}s_h^2,\ \widehat{V}(\widehat{\overline{Y}}_{st})=\frac{1}{N^2}\widehat{V}(\widehat{Y}_{st}),\ \widehat{V}(\widehat{P}_{st})=\sum_{h=1}^{L}(\frac{N_h}{N})^2\frac{N_h-n_h}{N_h-1}\frac{p_hq_h}{n_h-1},$$

donde $\overline{y}_h$, S_h^2, s_h^2, P_h y p_h denotan los valores correspondientes en el estrato U_h.

Sin más que sustituir las probabilidades de inclusión de primer y segundo orden en la fórmula de Yates-Grundy se obtienen los estimadores de las varianzas:

$$\widehat{V}\left(\widehat{Y}_{st}\right)=\sum_{h=1}^{L}N_h^2\frac{1-f_h}{n_h}s_h^2,$$

$$\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)=\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\frac{1-f_h}{n_h}s_h^2,$$

$$\widehat{V}\left(\widehat{P}_{st}\right)=\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\frac{N_h-n_h}{N_h-1}\frac{p_hq_h}{n_h-1},$$

donde $\overline{y}_h$, s_h^2 y p_h denotan los estimadores de la media, cuasivarianza y proporciones en el estrato h.

7.2.3 Intervalos de confianza

Si suponemos que el tamaño de la muestra es suficientemente grande como para que $\overline{x}_{st}$ se distribuya normalmente, obtenemos las siguientes fórmulas para los intervalos de confianza: para el total:

$$\left(N\overline{x}_{st} - z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\widehat{X}_{st})},\ N\overline{x}_{st} + z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\widehat{X}_{st})}\right),$$

para la media:

$$\left(\overline{x}_{st} - z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\overline{x}_{st})},\ \overline{x}_{st} + z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\overline{x}_{st})}\right),$$

y para una proporción:

$$\left(\widehat{P}_{st} - z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\widehat{P}_{st})},\ \widehat{P}_{st} + z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\widehat{P}_{st})}\right).$$

7.2.4 Elección del tamaño muestral en los estratos: afijación.

Se da el nombre de *afijación* a la asignación del tamaño muestral n entre los distintos estratos. Esto es, a la determinación de los valores n_h de forma que verifiquen que su suma sea n.

Pueden establecerse diversas formas de repartir la muestra, entre las que destacan cuatro especialmente importantes:

1. La *afijación uniforme*, en la que se toman todos los n_h iguales, es decir, se asigna el mismo número de unidades a cada estrato.

$$n_h = \frac{n}{L} = k, \quad \forall h,$$

2. La *afijación proporcional*, en la que n_h es proporcional al tamaño del estrato, expresado en número de unidades:

$$n_h = N_h \cdot cte \quad \forall h,$$

con

$$k = f = f_h \ \ W_h = \frac{n_h}{n} = w_h \ \ \forall h$$

3. La *afijación de mínima varianza o de Neyman*, en la que se eligen los n_h de forma que minimicen la varianza para un n fijo,

El problema consiste así en minimizar $V(\overline{x}_{st})$ sujeto a la restricción

$$\sum_{h=1}^{L} n_h = n.$$

$$n_h = \frac{N_h S_h}{\sum_{j=1}^{L} N_j S_j} n,$$

4. La *afijación óptima*, en la que se eligen los n_h de forma que minimicen la varianza para un coste fijo, C.

   Si este coste viene dado por una función del tipo $C = C_0 + \sum_{h=1}^{L} n_h C_h$, donde C_h es el coste de elegir una unidad en el estrato h y C_0 es el coste inicial, (dentro de cualquier estrato el coste es proporcional al tamaño de la muestra, pero el coste por unidad puede variar de estrato a estrato),

$$n_h = \frac{N_h S_h \frac{1}{\sqrt{C_h}}}{\sum_{j=1}^{L} N_j S_j \frac{1}{\sqrt{C_j}}} n,$$

7.3 Muestreo por conglomerados

7.3.1 Introducción

Vamos a ocuparnos ahora del caso en que las unidades de muestreo comprenden dos o más unidades últimas. En este caso se dice que cada unidad de muestreo constituye un conglomerado de unidades últimas, y que el muestreo es por conglomerados.

Existen diversas razones para el empleo de conglomerados. La razón fundamental de su uso es que para efectuar un muestreo aleatorio simple o un muestreo estratificado, hace falta disponer de una lista de todos los elementos de la población, y en la

práctica no suele disponerse de tales listas, salvo en casos particulares. Es preferible la división previa de la población en conglomerados de los cuales se selecciona cierto número, para lo cual sólo se necesita disponer de la lista de los conglomerados.

Las secciones censales y las manzanas de una ciudad se usan frecuentemente como conglomerados de hogares en los estudios de mercado y en las propios estudios realizados por las oficinas de estadística de los distintos países. Los mismos hogares son usados como conglomerados de personas, los hospitales son conglomerados de pacientes,...

Si las unidades de un conglomerado seleccionado proporcionan resultados similares, parece antieconómico medirlos todos. Una práctica común es seleccionar y medir una muestra en cada conglomerado seleccionado. Esta técnica se conoce con el nombre de muestreo bietápico o submuestreo.

Esta situación puede extenderse a más etapas dando lugar al muestreo polietápico.

7.3.2 Matriz de probabilidades. Estimadores y varianzas

Ahora, la población $U = \{1, 2, \ldots, N\}$ está dividida en conglomerados C_1, C_2, ..., C_M, cada uno de ellos constituido por unidades finales de U, $C_i = \{i_1, \ldots, i_{N_i}\}$.

Se extrae una muestra de conglomerados $s_c = \{C_{j_1}, \ldots, C_{j_g}\}$ de esta población de conglomerados $U_c = \{C_1, \ldots, C_M\}$ y se observan todas las unidades que lo componen, con lo que la muestra final está constituida por todas las unidades finales que pertenecen a los conglomerados de la muestra s_c.

Así, si llamamos π_i^c a la probabilidad de que el conglomerado C_i esté en la muestra s_c y π_{ij}^c a la probabilidad de que los conglomerados C_i y C_j pertenezcan a la muestra s_c, observamos que las probabilidades de inclusión de las unidades últimas vienen dadas por

$$\pi_k = \pi_i^c \text{ si } k \in C_i \ \forall k \in U; \quad \pi_{kl} = \begin{cases} \pi_{ij}^c & \text{si } k \in C_i, l \in C_j \\ \pi_i^c & \text{si } k, l \in C_i \end{cases}$$

Sea $\theta(\mathbf{y}) = \sum_{k\in U} a_k y_k$ un parámetro lineal, podemos construir

$$\widehat{\theta}_{HT} = \sum_{k\in s} a_k \frac{y_k}{\pi_k} = \sum_{i\in s_c} \sum_{k\in C_i} a_k \frac{y_k}{\pi_k} = \sum_{i\in s_c} \frac{1}{\pi_i^c} \sum_{k\in C_i} a_k y_k,$$

$$V\left(\widehat{\theta}_{HT}\right) = \sum_{i,j\in U_c} \Delta_{ij}^c \frac{\sum_{k\in C_i} a_k y_k}{\pi_i^c} \frac{\sum_{l\in C_j} a_l y_l}{\pi_j^c},$$

$$\widehat{V}\left(\widehat{\theta}_{HT}(s_c)\right) = \sum_{i,j\in s_c} \frac{\Delta_{ij}^c}{\pi_{ij}^c} \frac{\sum_{k\in C_i} a_k y_k}{\pi_i^c} \frac{\sum_{l\in C_j} a_l y_l}{\pi_j^c},$$

y en el caso de ser el diseño de tamaño fijo:

$$\widehat{V}\left(\widehat{\theta}_{HT}(s_c)\right) = -\frac{1}{2} \sum_{i,j\in s_c} \frac{\Delta_{ij}^c}{\pi_{ij}^c} \left(\frac{\sum_{k\in C_i} a_k y_k}{\pi_i^c} - \frac{\sum_{l\in C_j} a_l y_l}{\pi_j^c} \right)^2.$$

En el caso particular de que los conglomerados se elijan con $MAS(M, g)$ se tiene $\pi_i^c = \frac{g}{M}$ y $\pi_{ij}^c = \frac{g(g-1)}{M(M-1)}$, $i \neq j$,
lo que permite realizar las siguientes estimaciones:

$$\widehat{Y}_c = \sum_{i\in s_c} \frac{M}{g} \sum_{k\in C_i} y_k = M\overline{T}_c,$$

siendo $\overline{T}_c$ la media, realizada sobre la muestra de conglomerados, de los totales calculados para ellos:

$$\overline{T}_c = \frac{1}{g} \sum_{i\in s_c} y^{C_i} \text{ con } \quad y^{C_i} = \sum_{k\in C_i} y_k.$$

Además

$$\widehat{\overline{Y}}_c = \frac{M}{N}\overline{T}_c; \quad \widehat{P}_c = \frac{M}{Ng} \sum_{i\in s_c} \sum_{k\in C_i} A_i.$$

Si observamos que $\overline{T}_c$ se comporta como la media muestral de la variable y^{C_i} en un $MAS(M, g)$ se obtienen rápidamente las fórmulas para la varianza y su estimación:

$$V\left(\widehat{Y}_c\right) = M^2\left(\frac{1}{g} - \frac{1}{M}\right)\frac{1}{M-1}\sum_{i\in U_c}\left(y^{C_i} - \frac{Y}{M}\right)^2,$$

$$\widehat{V}\left(\widehat{Y}_c\right) = M^2\left(\frac{1}{g} - \frac{1}{M}\right)\frac{1}{g-1}\sum_{i\in s_c}\left(y^{C_i} - \overline{T}_c\right)^2.$$