Revista Integración
Escuela de Matemáticas
Universidad Industrial de Santander
Vol. 28, No. 1, 2010, pág. 51–65

Relación entre algunas clases de estimadores lineales

Germán Moreno[a,*], Artur José Lemonte[b]

[a] Universidad Industrial de Santander, Escuela de Matemáticas, A.A. 678, Bucaramanga, Santander, Colombia.

[b] Universidade de São Paulo, Departamento de Estatística, São Paulo, Brasil.

Resumen. Se propone una forma diferente de representar la información de una realización muestral. En este nuevo enfoque, una realización muestral está representada por un objeto matemático común que permite identificar la forma en que la información muestral está siendo usada en la construcción de los estimadores lineales. Se presenta la relación existente entre las principales clases de estimadores lineales propuestos en la literatura: Neyman, Horvitz y Thompson, Godambe y Stanek, Singer y Lencina.
Palabras claves: Muestreo, estimador lineal, estimador de Neyman, Horvitz y Thompson, Godambe y Stanek, Singer y Lencina.
MSC2000: 62D05, 62F10, 62J05.

Relationship Between some Classes of Linear Estimators

Abstract. We propose a different way of representing the information of a sample realization. In this new approach, a sample realization is represented by a common mathematic object that allows to identify the manner in which the all sample information is being used in the construction of linear estimators. We show the relationship between the major classes of linear estimators proposed in the literature: Neyman, Horvitz and Thompson, Godambe, and Stanek, Singer and Lencina.
Keywords: Sampling, linear estimator, estimator of Neyman, Horvitz and Thompson, Godambe and Stanek, Singer and Lencina.

1. Introducción

Inicialmente, consideramos una población (universo o población objetivo) finita o infinita, como el conjunto de unidades sobre el cual deseamos obtener información. Admitimos que la población finita es un conjunto de N elementos distintos identificables por etiquetas $j = 1, \ldots, N$. Sea θ una variable aleatoria definida sobre la población finita de N

* Autor para correspondencia: *E-mail*: gmorenoa@uis.edu.co.
Recibido: 8 de enero de 2010, Aceptado: 13 de marzo de 2010.

unidades y θ_j el valor de θ asociado con el elemento j de la población. Formalmente, una población finita puede ser representada por el vector $(1, \ldots, N)^\top$, y el vector de respuesta poblacional como $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^\top$.

Generalmente tenemos interés en ciertas cantidades de la población (parámetros); por ejemplo, la media, $\mu = N^{-1} \sum_{j=1}^{N} \theta_j$, el total, $T = \sum_{j=1}^{N} \theta_j$, el valor de un elemento de la población, θ_j, la varianza, $\sigma_\theta^2 = N^{-1} \sum_{j=1}^{N} (\theta_j - \mu)^2$, etc. Uno de los objetivos del análisis estadístico consiste en obtener "buenas" aproximaciones a estas cantidades poblacionales observando únicamente algunos de los elementos de la población y usando algún *estadístico* (estimador), esto es, una función de la muestra. Para cada parámetro pueden existir diferentes estimadores, pero usualmente se escoge el estimador que posea mejores propiedades que los restantes, como ausencia de sesgo, eficiencia, convergencia, robustez y suficiencia, entre otras. Además existen diversas formas para las funciones de los estimadores. Aquí nos concentramos en estimadores lineales de las observaciones muestrales.

Una muestra de tamaño n ($n \leq N$) es un subconjunto de elementos retirados de la población finita cuyas características serán medidas y utilizadas para hacer inferencias. Las n posiciones en la muestra son identificadas por $i = 1, \ldots, n$. Además, suponemos que al seleccionar la j-ésima unidad en la muestra observamos también su valor θ_j, posiblemente con error de medida.

En la mayoría de los casos existen dos tipos de información disponibles en una muestra obtenida de una población finita:

(i) los valores de las unidades muestrales, y

(ii) las descripciones de la realización muestral.

Una realización muestral habitualmente tiene en cada posición de la muestra dos elementos: la etiqueta de la unidad que ocupa esa posición y el valor observado para esa unidad. Examinando qué etiquetas corresponden a qué posiciones, podemos reconstruir que unidades fueron seleccionadas en la muestra y en qué orden fueron seleccionadas, independientemente de los valores observados en cada unidad muestral.

Un problema inherente exclusivamente al muestreo en población finita es: ¿qué hacer con la información sobre la realización muestral? En este sentido, la respuesta cubre el abanico completo de posibilidades, que va desde no usar esa información [Neyman (1934)] hasta usar toda la información contenida en la realización muestral [Godambe (1955)]. Aquí describiremos esa gama completa de alternativas propuestas en la literatura.

En la sección 2 definiremos los elementos matemáticos usados para representar una realización muestral, que facilita la identificación de la etiqueta y el orden de selección en la

muestra. En la sección 3 expondremos las cuatro clases de estimadores lineales propuestos en la literatura: Neyman (1934), Horvitz y Thompson (1952), Godambe (1955) y Stanek, Singer y Lencina (2004). Finalizamos el artículo mostrando las relaciones existentes entre estas cuatro clases de estimadores lineales.

2. Notación y terminología

Una realización muestral de tamaño n puede ser representada por el vector $(U_1, \ldots, U_n)^\top$, donde U_i, $i = 1, \ldots, n$, representa la unidad que ocupa la i-ésima posición en la muestra con $U_i \in \{1, \ldots, N\}$. Para muestreo con substitución el número de posibles realizaciones muestrales es N^n, y sin substitución es $K = \frac{N!}{(N-n)!}$.

Por ejemplo, para $N = 15$, $n = 2$, una realización muestral $(U_1, U_2)^\top = (5, 14)^\top$ significa que la unidad con etiqueta 5 en la población fue seleccionada en la primera posición de la muestra y la unidad con etiqueta 14 en la población fue seleccionada en la segunda posición de la muestra.

2.1. Otras formas de representación de una realización muestral

Una representación alternativa de una realización muestral está dada por una matriz de selección $\boldsymbol{S}$, de tamaño nN, que contiene ceros y unos. Cada elemento de la matriz de selección $\boldsymbol{S} = (s_{ij})$ está asociado con una unidad en la población y una posición en la muestra, de tal forma que $s_{ij} = 1$ si la unidad j de la población es seleccionada en la posición i en la muestra, y $s_{ij} = 0$ en caso contrario. Cada línea de $\boldsymbol{S}$ representa una posición en la muestra y contiene un 1 y $(n-1)$ ceros, con 1 indicando cuál es la unidad de la población que ocupa aquella posición en la muestra. Cada columna en la matriz de selección representa una unidad en la población, y los unos en la columna j indican qué posición o posiciones está ocupando la unidad j en la muestra.

Cuando el muestreo es sin substitución, cada unidad puede aparecer máximo una vez en la muestra, luego para $j = 1, \ldots, N$,

$$\sum_{i=1}^{n} s_{ij} \leq 1.$$

Por otro lado, cuando el muestreo es con substitución, cada unidad puede aparecer hasta n veces en la muestra, para $j = 1, \ldots, N$,

$$\sum_{i=1}^{n} s_{ij} \leq n.$$

Una realización del vector aleatorio $(U_1, \ldots, U_n)^\top$ puede ser obtenida como

$$U_i = \sum_{j=1}^{N} j S_{ij}, \quad i = 1, \ldots, n.$$

Por ejemplo, cuando $N = 3$, $n = 2$ y el muestreo es sin substitución, existen seis posibles realizaciones muestrales:

- $\boldsymbol{s}^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, correspondiendo a la realización muestral $(U_1, U_2) = (1, 2)$, es decir, la unidad 1 fue seleccionada en la posición 1 y la unidad 2 en la posición 2; de forma similar, para
- $\boldsymbol{s}^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, correspondiendo a la realización muestral $(U_1, U_2) = (1, 3)$;
- $\boldsymbol{s}^{(3)} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, correspondiendo a la realización muestral $(U_1, U_2) = (2, 1)$;
- $\boldsymbol{s}^{(4)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, correspondiendo a la realización muestral $(U_1, U_2) = (2, 3)$;
- $\boldsymbol{s}^{(5)} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$, correspondiendo a la realización muestral $(U_1, U_2) = (3, 1)$ y
- $\boldsymbol{s}^{(6)} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, correspondiendo a la realización muestral $(U_1, U_2) = (3, 2)$.

En muestreo aleatorio simple, todas las 6 realizaciones muestrales tienen la misma probabilidad de acontecer, $p = 1/6$.

Ahora bien, si el muestreo es con substitución, existen 9 posibles realizaciones muestrales con

$$(U_1, U_2) \in \{(1,1),\ (1,2),\ (1,3),\ (2,1),\ (2,2),\ (2,3),\ (3,1),\ (3,2),\ (3,3)\}.$$

Para la realización muestral $(U_1, U_2) = (2, 2)$ la matriz de selección es

$$\boldsymbol{s}^{(4)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Una forma clásica de introducir el muestreo en población finita es a través del modelo de permutación aleatoria. Una matriz de permutación aleatoria $\boldsymbol{R}$ es una matriz de tamaño $N \times N$ que permuta todos los elementos de la población; $r_{ij} = 1$ indica que la unidad con etiqueta j en la población está ubicada en la i-ésima posición de las unidades permutadas, y $r_{ij} = 0$ en caso contrario. Sin pérdida de generalidad se toman los elementos en las primeras n posiciones de $\boldsymbol{R}$ como la muestra, y aquellas en las restantes $(N-n)$ posiciones

como la parte no muestreada o unidades remanentes. En consecuencia, podemos escribir la matriz de selección aleatoria $\boldsymbol{S}$ de las unidades muestrales como

$$\boldsymbol{S} = (\boldsymbol{I}_n \ \vdots \ \boldsymbol{0}_{n\times(N-n)})\boldsymbol{R}.$$

Las realizaciones muestrales también pueden ser descritas por el vector $(P_1, \ldots, P_N)^\top$, donde P_j, $j = 1, \ldots, N$, indica qué posición en la muestra está siendo ocupada por la unidad j, y $P_j = 0$ significa que la unidad no fue seleccionada en la muestra,

$$P_j = \sum_{i=1}^{n} iS_{ij}.$$

Algunas veces será útil tener un modo simple de indicar qué unidades fueron incluidas en una muestra, independientemente del orden de selección. Tal información es representada por el conjunto muestral $V = \{V_1, \ldots, V_n\}$; el conjunto contiene los rótulos de todas las unidades incluidas en la muestra. Una forma alternativa de representar la misma información es dada por las variables aleatorias

$$Q_j = \sum_{i=1}^{n} S_{ij},$$

para $j \in \{1, \ldots, N\}$, donde para muestreo aleatorio simple sin substitución $Q_j = 1$ si la unidad j está incluida en la muestra, y 0 en caso contrario.

Ejemplo 2.1. Para ilustrar estas definiciones, sean $N = 5$, $n = 3$, y con muestreo aleatorio sin substitución supóngase que la unidad 2 fue seleccionada en la posición 1, la unidad 5 en la posición 2 y la unidad 4 en la posición 3. Entonces, siguiendo las definiciones tenemos:

(a) $\boldsymbol{S} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$, que corresponde con las primeras n filas de la matriz R_1 ó R_2, siendo $\boldsymbol{R}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$ y $\boldsymbol{R}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$;

(b) $(U_1\ U_2\ U_3) = (2\ 5\ 4)$;

(c) $(P_1\ P_2\ P_3\ P_4\ P_5) = (0\ 1\ 0\ 3\ 2)$;

(d) $(Q_1\ Q_2\ Q_3\ Q_4\ Q_5) = (0\ 1\ 0\ 1\ 1)$;

(e) $V = \{2, 4, 5\}$ es el conjunto de etiquetas de las unidades incluidas en la muestra independientes del orden de muestreo; y

(f) $n = 3$.

Nótese que (a), (b) y (c) contienen toda la información sobre la realización muestral: las unidades incluidas en la muestra y el orden de su selección; (d) y (e) solo tienen la información de las unidades incluidas en la muestra, y (f) no tiene la información sobre la realización muestral.

2.2. Valores de las unidades muestrales

Sea Y_i la variable aleatoria que representa el valor de la respuesta de la unidad que ocupa la i-ésima posición en la muestra. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ es el vector de respuestas muestrales; si no existe error de medida, entonces $\boldsymbol{Y}$ puede ser observado después de realizar la muestra, y puede ser escrito como

$$\boldsymbol{Y} = \boldsymbol{S\theta} = (\boldsymbol{I}_n \ \vdots \ \boldsymbol{0}_{n\times(N-n)})\boldsymbol{R\theta}.$$

Si existen errores de medida exógenos y endógenos, la respuesta en la i-ésima posición de la muestra es

$$Y_i = \sum_{j=1}^{N} S_{ij}(\theta_j + E_j) + E_i^*, \quad i = 1, \ldots, n, \tag{1}$$

siendo E_j el error de medida endógeno asociado a la unidad con etiqueta j en la población, y E_i^* el error de medida exógeno asociado a la i-ésima condición de medida. En forma matricial tenemos

$$\boldsymbol{Y} = \begin{pmatrix} \sum_{j=1}^{N} S_{1j}\theta_j + \sum_{j=1}^{N} S_{1j}E_j + E_1^* \\ \sum_{j=1}^{N} S_{2j}\theta_j + \sum_{j=1}^{N} S_{2j}E_j + E_2^* \\ \vdots \\ \sum_{j=1}^{N} S_{nj}\theta_j + \sum_{j=1}^{N} S_{nj}E_j + E_n^* \end{pmatrix} = \boldsymbol{S\theta} + \boldsymbol{SE} + \boldsymbol{E}^*, \tag{2}$$

siendo $\boldsymbol{E} = (E_1, \ldots, E_N)^\top$ el vector de errores endógenos y $\boldsymbol{E}^* = (E_1^*, \ldots, E_n^*)^\top$ el vector de errores exógenos.

3. Estimación lineal

Un parámetro de la población es cualquier función $\beta(\boldsymbol{\theta})$. Nos concentramos en parámetros lineales que tienen la forma

$$\beta(\boldsymbol{\theta}) = \sum_{j=1}^{N} f_j \theta_j.$$

Cuando $\beta(\boldsymbol{\theta})$ es la media poblacional μ, $f_j = \dfrac{1}{N}$, para $j = 1, \ldots, N$; cuando es el total poblacional T, entonces $f_j = 1$, para $j = 1, \ldots, N$. Si $f_j = 1$ y $f_k = 0, \quad k \neq j$, entonces $\beta(\boldsymbol{\theta}) = \theta_j$ es el parámetro individual de la unidad j en la población.

Un estimador de $\beta(\boldsymbol{\theta})$ es llamado lineal si es una función lineal de los valores observados en la muestra, es decir, si puede ser escrito como

$$\widehat{\beta} = \boldsymbol{\alpha}^{\top} \boldsymbol{Y}. \tag{3}$$

Una pregunta importante para construir el estimador (3) es: ¿cuál es la información de la realización muestral que debe ser considerada para construir el vector de coeficientes $\boldsymbol{\alpha}$?

Existen en la literatura por lo menos tres aproximaciones diferentes para responder esta pregunta, a saber: (1) no permitir que $\boldsymbol{\alpha}$ dependa de la realización muestral; (2) permitir que $\boldsymbol{\alpha}$ dependa apenas de una parte de la realización muestral, esto es, solo las etiquetas, o en otras palabras, que dependa solo de si las unidades están incluidas o no en la muestra; y (3) permitir que $\boldsymbol{\alpha}$ dependa de toda la realización muestral, las etiquetas y el orden de selección. Nótese que (1) es un caso particular de (2), y este a su vez es un caso particular de (3).

3.3. Estimadores lineales que ignoran la realización muestral

Neyman (1934) definió la clase de estimadores lineales que tienen la forma

$$\widehat{\beta} = \sum_{i=1}^{n} \alpha_i Y_i,$$

donde $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\top}$ es un vector fijo. Así, la Clase de Estimadores Lineales de Neyman (CELN), incluye todos los estimadores lineales cuyo vector de coeficientes no cambia independientemente de la realización muestral. Un estimador lineal de Neyman está definido mediante n coeficientes. Formalmente:

$$CELN = \left\{ \widehat{\beta} \mid \widehat{\beta} = \boldsymbol{\alpha}^{\top} \boldsymbol{Y}, \quad \boldsymbol{\alpha} \in \mathbb{R}^n \right\}.$$

Ejemplo 3.1. Para $N = 10$, $n = 3$, con muestreo aleatorio sin substitución, sean $s^{(1)}, \ldots, s^{(720)}$ las 720 posibles realizaciones muestrales. Un ejemplo de un estimador en la CELN sería la media muestral, dada por

$$\overline{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3) = \frac{1}{3}Y_1 + \frac{1}{3}Y_2 + \frac{1}{3}Y_3, \quad \forall \boldsymbol{S} = \boldsymbol{s} \in \left\{ s^{(1)}, \ldots, s^{(720)} \right\};$$

así $\boldsymbol{\alpha} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^\top$ no cambia, independientemente de la realización muestral.

3.4. Estimadores lineales que incorporan un poco de información sobre la realización muestral

En Stanek, Singer y Lencina (2004) se define un modelo donde se puede realizar la estimación de parámetros y la predicción de variables aleatorias de una forma unificada. El modelo consiste en un conjunto expandido de variables aleatorias siguiendo la distribución de probabilidad del modelo de permutación aleatoria que mantiene el registro de la muestra realizada: las etiquetas de las unidades incluidas en la muestra y las posiciones en la permutación realizada.

Clase de estimadores lineales Stanek-Singer-Lencina

Inicialmente, se define un vector aleatorio expandido para la muestra, $\boldsymbol{Y}_{SSL}$, de tamaño $nN \times 1$, como

$$\boldsymbol{Y}_{SSL} = \begin{pmatrix} S_{11}\theta_1 \\ \vdots \\ S_{n1}\theta_1 \\ \vdots \\ S_{1N}\theta_N \\ \vdots \\ S_{nN}\theta_N \end{pmatrix},$$

y la clase de estimadores son funciones lineales del vector aleatorio expandido, de la forma

$$\widehat{\beta}_{SSL} = \boldsymbol{L}_{SSL}^\top \boldsymbol{Y}_{SSL}.$$

Stanek, Singer y Lencina demuestran que su clase de estimadores es más general que la CELN, dado que está basada en Nn variables aleatorias mientras que la CELN está basada en solo n variables aleatorias, siendo cada una el resultado de sumar N de las variables aleatorias en $\boldsymbol{Y}_{SSL}$. Esta clase de estimadores también puede ser definida como

$$\widehat{\beta}_{SSL} = \sum_{i=1}^{n} \alpha_i(U_i) Y_i.$$

El coeficiente que multiplica el valor que ocupa la i-ésima posición depende de una porción de la información sobre la realización muestral: la unidad que ocupa la posición i en la muestra; es decir, los coeficientes de SSL están definidos como

$$\alpha_i(U_i) = \sum_{j=1}^{N} \alpha_{ij} S_{ij} = \begin{cases} \alpha_{i1}, & \text{si la unidad 1 ocupa la posición } i, \\ \vdots & \vdots \\ \alpha_{iN}, & \text{si la unidad } N \text{ ocupa la posición } i. \end{cases}$$

Entonces, la clase de estimadores lineales de Stanek-Singer-Lencina (CELSSL) está definida como

$$CELSSL = \left\{ \widehat{\beta} \mid \widehat{\beta} = \sum_{i=1}^{n} \sum_{j=1}^{N} \alpha_{ij} S_{ij} Y_i, \ \alpha_{ij} \in \mathbb{R} \right\}.$$

Dado que $Y_i = \sum_{t=1}^{N} S_{it}\theta_t$, $i = 1, \ldots, n$, el estimador también puede ser escrito como

$$\widehat{\beta}_{SSL} = \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{t=1}^{N} \alpha_{ij} S_{ij} S_{it} \theta_t,$$

pero $S_{ij}S_{it} = 1$ cuando $j = t$ y $S_{ij}S_{it} = 0$ en caso contrario, entonces puede ser simplificado como

$$\widehat{\beta}_{SSL} = \sum_{i=1}^{n} \sum_{j=1}^{N} \alpha_{ij} S_{ij} \theta_j,$$

que es la forma usada en Stanek, Singer y Lencina (2004). Además, nótese que

$$\widehat{\beta}_{SSL} = \sum_{i=1}^{n} \sum_{j=1}^{N} \alpha_{ij} S_{ij} \theta_j = \sum_{j=1}^{N} \theta_j \left(\sum_{i=1}^{n} \alpha_{ij} S_{ij} \right) = \sum_{j=1}^{N} \theta_j \alpha_j \left(P_j \right),$$

con

$$\alpha_j \left(P_j \right) = \sum_{i=1}^{n} \alpha_{ij} S_{ij} = \begin{cases} \alpha_{1j}, & \text{si la unidad } j \text{ ocupa la posición 1}, \\ \vdots & \vdots \\ \alpha_{nj}, & \text{si la unidad } j \text{ ocupa la posición } n, \\ 0, & \text{si la unidad } j \text{ no está en la muestra.} \end{cases}.$$

En consecuencia, este estimador también puede ser escrito como

$$\widehat{\beta}_{SSL} = \sum_{j \in V} \sum_{i=1}^{n} \alpha_{ij} S_{ij} \theta_j,$$

donde V es el conjunto de unidades incluidas en la muestra.

La clase de estimadores de SSL se localiza entre los estimadores lineales que no usan información sobre la realización muestral (Sección 3.3) y aquellos que usan toda la información disponible en la realización muestral (Sección 3.5).

Clase de estimadores lineales de Horvitz-Thompson

Una subclase especial de la CELSSL es la clase de estimadores lineales definidos por Horvitz y Thompson (1952), donde cada estimador es tal que el coeficiente $\alpha_j(P_j)$ no cambia de posición para $j = 1, \ldots, N$. El estimador de Horvitz-Thompson utiliza para su cálculo los valores observados en la muestra, y la etiqueta de los individuos seleccionados en la muestra representada por los pesos de muestreo de cada unidad. Los estimadores lineales de Horvitz-Thompson tienen la forma

$$\widehat{\beta}_{HT} = \sum_{j \in V} \alpha_j(P_j)\theta_j, \quad \alpha_j(P_j) = \lambda_j, \ j = 1, \ldots, N,$$

que también pueden ser escritos como

$$\widehat{\beta}_{HT} = \sum_{j \in V} \lambda_j \theta_j = \sum_{j=1}^{N} \lambda_j Q_j \theta_j;$$

para muestreo sin substitución Q_j toma el valor 1 si la unidad j está incluida en la muestra, y 0 caso contrario.

Formalmente, la clase de estimadores lineales de Horvitz-Thompson (CELHT) está definida como

$$CELHT = \left\{ \widehat{\beta} \mid \widehat{\beta} = \sum_{j \in V} \lambda_j \theta_j, \ \lambda_j \in \mathbb{R}, j = 1, \ldots, N \right\},$$

luego un estimador lineal de Horvitz-Thompson está definido a través de N parámetros.

Ejemplo 3.2. Sea $N = 3$, $n = 2$. Un ejemplo de estimador lineal de HT puede ser,

$$\widehat{\beta} = \begin{cases} N\theta_1, & \text{si la unidad 1 está en la muestra,} \\ 0, & \text{caso contrario,} \end{cases}$$

que puede ser escrito como

$$\widehat{\beta} = NQ_1\theta_1 + 0Q_2\theta_2 + 0Q_3\theta_3.$$

Ejemplo 3.3. El estimador Horvitz-Thompson del total poblacional

$$\theta_h = \sum_{j=1}^{N_h} \theta_{hj}$$

de la variable θ en el estrato h viene dado por

$$\widehat{\beta}_h = \sum_{j=1}^{n_h} \omega_{hj} Y_{hj},$$

donde $\omega_{hj} = 1/\pi_{hj}$ son los pesos muestrales de la j-ésima unidad en el estrato h y π_{hj} es su probabilidad de inclusión. En un muestreo aleatorio simple con n_h unidades seleccionadas del total N_h, tenemos $\omega_{hj} = N_h/n_h, j = 1, \ldots, n_h$ y

$$\widehat{\beta}_h = \sum_{j=1}^{n_h} N_h Y_{hj}/n_h = N_h \overline{Y}_h.$$

Los estimadores lineales de HT no pertenecen a la CELN, dado que no pueden escribirse como una función lineal de $\boldsymbol{Y}$ con coeficientes fijos para toda realización muestral.

3.5. Estimadores lineales que incorporan toda la información sobre la realización muestral: la clase de estimadores de Godambe

Godambe (1955) definió "el tipo más general de estimador lineal" que incorpora toda la información suministrada por la realización muestral: las unidades que entraron en la muestra y el orden en que fueron seleccionadas. Los estimadores lineales de Godambe tienen la forma

$$\widehat{\beta}_G = \sum_{i=1}^{n} \alpha_i(\boldsymbol{S}) Y_i = \boldsymbol{\alpha}(\boldsymbol{S})^\top \boldsymbol{Y},$$

indicando que puede existir un vector diferente de coeficientes para cada posible realización muestral. para muestreo sin substitución, esto significa que existen potencialmente $k = \frac{n!}{(n-n)!}$ vectores diferentes de coeficientes.

Si usamos muestreo sin substitución, y $\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(K)}$ son todas las posibles realizaciones muestrales, entonces los coeficientes de Godambe pueden ser escritos como

$$\alpha_i(\boldsymbol{S}) = \sum_{k=1}^{K} \alpha_i^{(k)} \mathbf{1}_{\left\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\right\}},$$

donde $\mathbf{1}_{\left\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\right\}}$ es una función indicadora con valor 1 si $\boldsymbol{S} = \boldsymbol{s}^{(k)}$, y 0 en caso contrario.

Los estimadores lineales de Godambe también pueden ser escritos como funciones de los valores individuales de la población de la siguiente forma:

$$\widehat{\beta}_G = \sum_{i=1}^{n} \alpha_i(\boldsymbol{S}) Y_i = \sum_{k=1}^{K} \sum_{i=1}^{n} \alpha_i^{(k)} \mathbf{1}_{\left\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\right\}} Y_i = \sum_{k=1}^{K} \sum_{i=1}^{n} \alpha_i^{(k)} \mathbf{1}_{\left\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\right\}} \left(\sum_{j=1}^{N} \theta_j s_{ij}^{(k)} \right).$$

Nótese que

$$\begin{aligned} \widehat{\beta}_G &= \sum_{k=1}^{K} \mathbf{1}_{\left\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\right\}} \sum_{j=1}^{N} \theta_j \sum_{i=1}^{n} \alpha_i^{(k)} s_{ij}^{(k)} \\ &= \sum_{k=1}^{K} \mathbf{1}_{\left\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\right\}} \sum_{i=1}^{n} \sum_{j=1}^{N} \alpha_j^{(k)} Q_j \theta_j = \sum_{j \in V^{(k)}} \delta_j^{(k)} \theta_j, \end{aligned}$$

siendo $V^{(k)}$ el conjunto de unidades que son incluidas en la muestra cuando la realización muestral es $s^{(k)}$.

Formalmente, la clase de estimadores lineales de Godambe (CELG) está definida como

$$CELG = \left\{ \widehat{\beta} \mid \widehat{\beta} = \alpha(\boldsymbol{S})^{\top} \boldsymbol{Y},\ \alpha(\boldsymbol{S}) \in \mathbb{R}^n,\ \ \forall \boldsymbol{S} \right\}.$$

Un estimador lineal de Godambe está definido a través de nK parámetros.

Ejemplo 3.4. Sean $N = 3$, $n = 2$, y $s^{(1)}, \ldots, s^{(6)}$ las seis posibles realizaciones muestrales ya descritas. Un ejemplo de un estimador en la CELG puede ser

$$\widehat{\beta} = \begin{cases} \theta_1, & \text{si la unidad 1 está en la muestra,} \\ \bar{Y} = \frac{1}{2}(Y_1 + Y_2), & \text{en caso contrario.} \end{cases}$$

Los coeficientes de este estimador tienen la forma

$$\alpha(\boldsymbol{S})^{\top} = \begin{cases} \left(\frac{1}{2} \ \ \frac{1}{2}\right), & \text{si } \boldsymbol{S} = \boldsymbol{s}^{(4)}, \boldsymbol{s}^{(6)}; \text{es decir,} \\ & \text{la unidad 1 no está en la muestra,} \\ (1\ 0), & \text{si } \boldsymbol{S} = \boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}; \\ & \text{la unidad 1 está en la posición 1 de la muestra,} \\ (0\ 1), & \text{si } \boldsymbol{S} = \boldsymbol{s}^{(3)}, \boldsymbol{s}^{(5)}; \\ & \text{la unidad 1 está en la posición 2 de la muestra.} \end{cases}$$

Este estimador no pertenece a la CELSSL porque los coeficientes que corresponden a Y_2 y a Y_3 son $1/2$ ó 0, pero no dependen de la posición que ellos ocupen en la muestra y sí de que la unidad 1 esté o no incluida en la muestra.

4. Relación entre las clases de estimadores

Una forma de mostrar la relación entre estas cuatro clases de estimadores lineales es comparando el conjunto de variables aleatorias involucradas en los estimadores lineales, como veremos en el siguiente teorema.

Teorema 4.1. *Sean*

(a) $\boldsymbol{Y}_N = \boldsymbol{S}\boldsymbol{\theta}$, *con* $\boldsymbol{S} = \begin{pmatrix} S_{11} & \cdots & S_{1N} \\ \vdots & \ddots & \vdots \\ S_{n1} & \cdots & S_{nN} \end{pmatrix}$;

(b) $\boldsymbol{Y}_{HT} = \boldsymbol{D}_Q\boldsymbol{\theta}$, *con* $\boldsymbol{D}_Q = \begin{pmatrix} Q_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Q_N \end{pmatrix}$;

(*c*) $\boldsymbol{Y}_{SSL} = \boldsymbol{D}_S\boldsymbol{\theta}$, *con* $\boldsymbol{D}_S = \begin{pmatrix} S_{11} & \cdots & 0 \\ \vdots & & \vdots \\ S_{n1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & S_{1N} \\ \vdots & & \vdots \\ 0 & \cdots & S_{nN} \end{pmatrix}$;

(*d*) $\boldsymbol{Y}_G = \boldsymbol{D}_G\boldsymbol{\theta}$, *con* $\boldsymbol{D}_G = (\boldsymbol{I}_n \otimes \boldsymbol{T})\boldsymbol{S}$ *y* $\boldsymbol{T} = \left(\mathbf{1}_{\{\boldsymbol{S}=\boldsymbol{s}^{(1)}\}} \cdots \mathbf{1}_{\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\}}\right)^\top$, *con* $\otimes$ *producto de Kronecker entre matrices.*

Entonces los estimadores lineales en cada clase se pueden escribir como

(*a*) $\widehat{\beta}_N = \boldsymbol{L}_N^\top \boldsymbol{Y}_N$,

(*b*) $\widehat{\beta}_{HT} = \boldsymbol{L}_{HT}^\top \boldsymbol{Y}_{HT}$,

(*c*) $\widehat{\beta}_{SSL} = \boldsymbol{L}_{SSL}^\top \boldsymbol{Y}_{SSL}$,

(*d*) $\widehat{\beta}_G = \boldsymbol{L}_G^\top \boldsymbol{Y}_G$.

Demostración.

(a) $\boldsymbol{L}_N^\top \boldsymbol{Y}_N = \boldsymbol{L}_N^\top \boldsymbol{S}\boldsymbol{\theta} = \sum_{i=1}^{n} L_{Ni}\left(\sum_{j=1}^{N} S_{ij}\theta_i\right) = \sum_{i=1}^{n} L_{Ni}Y_i = \sum_{i=1}^{n} \alpha_i Y_i = \widehat{\beta}_N$, siendo el vector $\boldsymbol{L}_N^\top = (\alpha_1, \ldots, \alpha_n)$.

(b) $\boldsymbol{L}_{HT}^\top \boldsymbol{Y}_{HT} = \boldsymbol{L}_{HT}^\top \boldsymbol{D}_Q\boldsymbol{\theta} = \boldsymbol{L}_{HT}^\top \begin{pmatrix} Q_1\theta_1 \\ \vdots \\ Q_N\theta_N \end{pmatrix} = \sum_{j=1}^{N} L_{HTj}Q_j\theta_j = \sum_{j=1}^{N} \lambda_j\theta_j$, con $\lambda_j = L_{HTj}Q_j = L_{HTj}\left(\sum_{i=1}^{n} S_{ij}\right)$; recuérdese que $\sum_{i=1}^{n} S_{ij} = 1$ cuando la unidad con etiqueta j en la población está incluida en alguna posición de la muestra, y 0 en caso contrario; así, $\sum_{j=1}^{N} \lambda_j\theta_j = \widehat{\beta}_{HT}$.

(c) $\boldsymbol{L}_{SSL}^\top \boldsymbol{Y}_{SSL} = \boldsymbol{L}_{SSL}^\top \boldsymbol{D}_S\boldsymbol{\theta} = \boldsymbol{L}_{SSL}^\top \begin{pmatrix} S_{11}\theta_1 \\ \vdots \\ S_{n1}\theta_1 \\ \vdots \\ S_{1N}\theta_N \\ \vdots \\ S_{nN}\theta_N \end{pmatrix} = \sum_{i=1}^{n}\sum_{j=1}^{N} L_{SSLij}S_{ij}\theta_j$; nótese que

podemos definir el vector $\boldsymbol{L}_{SSL}^{\top} = (\alpha_{11}, \ldots, \alpha_{n1}, \ldots, \alpha_{1N}, \ldots, \alpha_{nN})$, así obtenemos que

$$\boldsymbol{L}_{SSL}^{\top}\boldsymbol{Y}_{SSL} = \sum_{i=1}^{n}\sum_{j=1}^{N}\alpha_{ij}S_{ij}\theta_j = \widehat{\beta}_{SSL}.$$

(d) $\boldsymbol{L}_G^{\top}\boldsymbol{Y}_G = \boldsymbol{L}_G^{\top}\left(\boldsymbol{I}_n \otimes \begin{pmatrix} \mathbf{1}_{\{\boldsymbol{S}=\boldsymbol{s}^{(1)}\}} \\ \vdots \\ \mathbf{1}_{\{\boldsymbol{S}=\boldsymbol{s}^{(K)}\}} \end{pmatrix}\right)\boldsymbol{S\theta} = \sum_{i=1}^{n}\sum_{k=1}^{K} L_{Gik}\mathbf{1}_{\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\}}Y_i = \widehat{\beta}_G$; recuérdese que $\mathbf{1}_{\{\boldsymbol{S}=\boldsymbol{s}^{(k)}\}} = 1$ cuando la realización muestral es $S = s^{(k)}$, y 0 en caso contrario; dado que existen nK valores en el vector $\boldsymbol{L}_G$, entonces $(nK - n)$ serán eliminados en la suma después de obtener la muestra, y solo n serán diferentes de cero. Luego los estimadores de Godambe usan toda la información muestral: las unidades incluidas en la muestra y la posición en que la unidad quedó en ella. ☑

En otras palabras, este teorema indica que los estimadores en la CELN son combinación lineal de n variables aleatorias en $\boldsymbol{Y}_N$; en la CELHT son combinación lineal de N variables aleatorias en $\boldsymbol{Y}_{HT}$; en la CELSSL son combinación lineal de nN variables aleatorias en $\boldsymbol{Y}_{SSL}$, y en la CELG son combinación lineal de nK variables aleatorias en $\boldsymbol{Y}_G$.

5. Conclusiones

Mostramos que los estimadores en la CELN son un caso particular de la CELHT, los estimadores en la CELHT son un caso particular de la CELSSL y estos a su vez son un caso particular de la CELG. En consecuencia, de las definiciones dadas anteriormente y de los ejemplos presentados, es claro que las clases de estimadores lineales tienen la relación

$$N \subset HT \subset SSL \subset G,$$

relación que también puede ser observada por el número de parámetros que usa el respectivo estimador lineal,

$$n < N < nN < nK,$$

donde $K = \frac{N!}{(N-n)!}$. Por ejemplo, si $N = 10$ y $n = 2$, entonces el número de parámetros en cada clase es $2 < 10 < 20 < 180$, respectivamente.

Por otro lado, Godambe (1955) mostró que para poblaciones finitas no existe un estimador lineal no viciado de varianza uniformemente mínima para el total poblacional en la clase general de estimadores lineales, aquí denotada como CELG. Moreno (2009) mostró que en la CELG no existe un estimador lineal no viciado de varianza uniformemente mínima para el valor latente de la unidad seleccionada en la i-ésima posición de la

muestra en presencia simultánea de errores de medida endógenos y exógenos. Entonces está abierta la posibilidad de definir otra clase de estimadores lineales, $CELX$, que use r parámetros y tal que $nN < r < nK$ y que cumpla la relación

$$SSL \subset X \subset G.$$

Hasta donde tenemos conocimiento no se ha definido la CELX con las condiciones definidas anteriormente y tampoco sabemos si existen situaciones prácticas donde efectivamente se necesite ese tipo de estimadores lineales.

Referencias

[1] Bolfarine H., and Zacks S., "Prediction theory for finite populations", Springer Verlag, New York, 1992.

[2] Cochran W. G., "Sampling Techniques", 3rd Edition, Wiley, New York, 1977.

[3] Godambe V.P., A unified theory of sampling from finite populations, *Journal of The Royal Statistics Society*, Series B, 17 (1955), 269-278.

[4] Horvitz D.G., and Thompson D.J., A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, 47 (1952), 663-685.

[5] Moreno G., "Modelos mistos para populações finitas com erros de medida endógenos e exógenos", Tesis de doctorado, Instituto de Matemática y Estatística, Universidade de São Paulo, IME-USP, 2009.

[6] Neyman J., On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of The Royal Statistical Society*, Series A, 109 (1934), 558-606.

[7] Stanek III E.J., Singer J.M., y Lencina V.B., A unified approach to estimation and prediction under simple random sampling, *Journal of Statistical Planning and Inference*, 121 (2004), 325-338.