QÜESTIIÓ, Vol 12, nº 3, pp. 315-321, 1988

ESTIMACIÓN INSESGADA GENERALIZADA EN POBLACIONES FINITAS

M. RUIZ ESPEJO

Universidad Complutense de Madrid

Proponemos un tipo de estimador insesgado para funciones paramétricas en el contexto de poblaciones finitas. Algunos casos particulares de estos estimadores son los clásicos de Hansen-Hurwitz (1943), Horvitz-Thompson (1952), Sáchez Crespo (1977) y Murthy (1963). Además damos estrategias insesgadas uniformemente mejores que las de Sánchez-Crespo (1977) en determinadas condiciones.

Clasificación AMS: 62 D 05

Generalised unbiased estimation in finite populations.

Keywords: Generalisation, parametric functions, unbiased estimation, uniformly better precision.

1. INTRODUCCIÓN

En esta nota vamos a proponer, en poblaciones finitas, un estimador insesgado de cualquier función paramétrica descomponible en sumas, en que cada monomio contenga la variable de interés en un número de unidades inferior al tamaño muestral n. Casos particulares de este estimador propuesto son los estimadores Hansen-Hurwitz (1943), Horvitz-Thompson (1952) y Sánchez-Crespo (1977) para estimar la medida poblacional

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad ,$$

-M. Ruiz Espejo - Dept. d'Organització d'Empreses. Facultat de Ciències Econòmiques i Empresarials - Universitat Complutense - 28023 Madrid.
-Article rebut el desembre de 1988.

así como el estimador de Murthy (1963) para estimar la varianza poblacional

$$\sigma_y^2 = \frac{1}{N^2} \sum_{i \neq j=1}^{N} \sum (y_i - y_j)^2 \quad .$$

Por analogía podemos proponer por ejemplo un estimador insesgado de σ_y^2 bajo diseño de probabilidades proporcionales al tamaño con reemplazamiento (*pptr*) que es también de sencilla ejecución en la práctica. También podemos generalizar para cualquier diseño con o sin reemplazamiento al estimador Horvitz-Thompson generalizado propuesto en Ruiz (1986) para justificar la estimabilidad de funciones paramétricas polinomiales.

2. ESTIMACIÓN INSESGADA DE LA MEDIA POBLACIONAL

Ya sea el diseño ordenado o no ordenado, un estimador insesgado de la media poblacional $\bar{y}$ es

$$\hat{\bar{y}} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{E(e_i)} \cdot e_i$$

donde e_i es una variable aleatoria auxiliar que toma valores 0,1,2... hasta n como máximo según el número de veces que aparece la unidad i en la muestra $(1 \leq i \leq N)$ y siendo N el tamaño de la población finita. Para un diseño ordenado con reemplazamiento, e_i puede tomar valores superiores a 1 y por tanto su valor esperado $E(e_i)$ podría ser superior a 1. Un ejemplo de este caso es el estimador Hansen-Hurwitz (1943) para un diseño de probabilidades proporcionales al tamaño con reemplazamiento; si el tamaño de la unidad i es $p_i > 0$, con

$$\sum_{i=1}^{N} p_i = 1 \quad ,$$

entonces $e_i \tilde{=} B(n, p_i)$ (binomial) y por tanto $E(e_i) = np_i$ con lo que queda

$$\hat{\bar{y}} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \, e_i}{n.p_i}$$

Otro ejemplo del mismo caso ordenado es el estimador de Sánchez-Crespo (1977) donde e_i se destribuye hipergeométricamente, con $E(e_i) = np_i$ también. Con este diseño, el estimador coincide con el ya conocido de Hansen-Hurwitz.

En el caso de muestreo no ordenado sin reemplazamiento, entonces la variable aleatoria auxiliar e_i puede tomar los valores 0 ó 1 según no pertenezca o sí a la

muestra la unidad i. Así, e_i nunca será superior a 1. En este caso, llamando s a la muestra no ordenada a seleccionar

$$E(e_i) = 0.p\,(i \notin s) + 1.p\,(i \in s) = p\,(i \in s) = \Pi_i$$

donde Π_i es la probabilidad de inclusión de la unidad i en la muestra. Entonces, bajo este diseño, tendremos el estimador Horvitz-Thompson (1952) como caso particular de nuestro estimador general de la media poblacional que es siempre insesgado, y es

$$\hat{\bar{y}} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i e_i}{\Pi_i} \quad .$$

3. VARIANZA DE LOS ESTIMADORES

Como es bien sabido (Ruiz, 1987), no existe estimador insesgado de mínima varianza para la estimación de la media poblacional o de la varianza poblacional. Además (como se vió en Ruiz, 1988) en el caso de estimación de la media poblacional $\bar{y}$, tanto el estimador Hansen-Hurwitz (1943) como el de Sánchez-Crespo (1977) (que mejora uniformemente al estimador Hansen-Hurwitz, t_{HH}) admiten estimadores insesgados uniformemente de menor varianza. En efecto, si el tamaño muestral es $n \geq 2$, $V(t_{SC}) < V(t_{HH})$.

$$V(t_{SC}) = V\left(\sum_{i=1}^{N}\frac{y_i e_i}{n.p_i}\right) = \frac{1}{n^2}\left[\sum_{i=1}^{N}\frac{y_i^2}{p_i^2}V(e_i) + \sum_{i\neq j=1}^{N}\sum\frac{y_i y_j}{p_i p_j}\,\mathrm{Cov}\,(e_i, e_j)\right]$$

al seguir e_i una distribución hipergeométrica generalizada,

$$V(e_i) = \frac{M-n}{M-1}np_i(1-p_i)$$

y si $i \neq j$,

$$\mathrm{Cov}\,(e_i, e_j) = -\frac{M-n}{M-1}np_i\,p_j \quad ,$$

luego

$$V(t_{SC}) = \frac{M-n}{M-1} \cdot \frac{1}{n} \left[\sum_{i=1}^{N} \frac{y_i^2}{p_i}(1-p_i) - \sum_{i \neq j=1}^{N} \sum y_i y_j \right] =$$

$$= \frac{M-n}{M-1} \cdot \frac{1}{n} \left[\sum_{i=1}^{N} \frac{y_i^2}{p_i} - \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \right] =$$

$$= \frac{M-n}{M-1} \cdot \frac{1}{n} \left[\sum_{i=1}^{N} \frac{y_i^2}{p_i} - N^2 \bar{y}^2 \right] = \frac{M-n}{M-1} \cdot \frac{1}{n} \left[\sum_{i=1}^{N} \frac{y_i^2}{p_i^2} p_i - \left(\sum_{i=1}^{N} \frac{y_i}{p_i} p_i \right)^2 \right] =$$

$$= \frac{M-n}{m_1} \cdot \frac{1}{n} \sum_{i=1}^{N} \left(\frac{y_i}{p_i} - N\bar{y} \right)^2 \cdot p_i = \frac{M-n}{M-1} V(t_{HH}) < V(t_{HH}) \quad ,$$

siempre que $n \geq 2$.

Un estimador insesgado de la media poblacional $\bar{y}$ que mejora a su vez uniformemente al estimador de Sánchez-Crespo (1977), se obtiene con el mismo estimador $\hat{\bar{y}} = t_{HH} = t_{SC}$, donde e_i sigue siendo el número de veces que aparece la unidad i en la muestra, pero el diseño se modifica de modo que si en la urna existe M_i bolas con el indicador de la unidad i, siendo

$$\sum_{i=1}^{N} M_i = M \quad ,$$

podemos extraer una bola y queda seleccionada la unidad numerada en la bola extraída. Seguidamente antes de la segunda selección, se retiran de la urna un número de bolas (con el mismo indicador de la ya extraída en primer lugar) igual al máximo común divisor $m = m.c.d.\{M_i : i = 1, 2, ..., N\}$ que podemos suponer mayor que 1. De la urna con la composición de bolas resultante se obtiene la segunda extracción que indicará la segunda unidad seleccionada en la muestra. Posteriormente, y antes de la tercera extraccicón (y sucesivas) se retiran de la urna un número m de bolas con el mismo indicador que la segunda bola (y siguientes) extraída. Así actuaríamos sucesivamente hasta completar el tamaño muestral n.

De este modo quedaría un diseño que mejora uniformemente la precisión de la estrategia de Sánchez-Crespo (1977). Las comprobaciones son inmediatas. Si m es el máximo común divisor de los tamaños $M_i (i = 1, 2, ..., N)$, entonces llamamos $M_i/m = K_i$ y $M/m = K$. El último diseño propuesto es equivalente a la selección sin reemplazamiento con probabilidades iguales de K bolas de

una urna, donde K_i llevan el indicador de la unidad $i\,(i = 1, 2, ..., N)$. El estimador habitual para diseño ordenado verifica que $p_i = K_i/K = M_i/M$ y e_i sigue una distribución hipergeométrica generalizada con parámetros K, n y K_i/K. Entonces, ahora

$$E(e_i) = n\frac{K_i}{K} = np_i \quad ,$$

y por tanto nuestro estimador análogo t_{RE} es también insesgado como t_{HH} y t_{SC}, $E(t_{RE}) = \bar{y}$, y además

$$(1) \qquad V(t_{RE}) = \frac{K-n}{K-1}V(t_{HH}) < \frac{M-n}{M-1}V(t_{HH}) = V(t_{SC})$$

donde como sabemos t_{HH} y t_{SC} se obtienen de una urna con M_i bolas identificadoras de la unidad i $(= 1, 2, ..., N)$. El resultado (1) es cierto si y sólo si

$$(K-n)(M-1) < (M-n)(K-1)$$

que equivale a que $K < M$, que es cierto pues hemos supesto que $m > 1$ y $mK = M$.

4. ESTIMACIÓN INSESGADA DE LA VARIANZA POBLACIONAL

En el caso general (muestras ordenadas o no ordenadas), el estimador propuesto para la varianza poblacional σ_y^2 es

$$\hat{\sigma}_y^2 = \frac{1}{N^2}\sum_{i\neq j=1}^{N}\sum \frac{(y_i - y_j)^2}{E(e_i e_j)} \cdot e_i e_j$$

donde e_i y e_j son las mismas variables aleatorias auxiliares de las secciones anteriores. En el caso de diseño no ordenado o sin reemplazamiento,

$$E(e_i e_j) - 1.1.p\,(i, j \in s) = \Pi_{ij}$$

con lo que obtenemos el estimador insesgado de la varianza poblacional debido a Murthy (1963)

$$\hat{\sigma}_y^2 = \frac{1}{N^2}\sum_{i\neq j=1}^{N}\sum \frac{(y_i - y_j)^2}{\Pi_{ij}} e_i e_j \quad .$$

Con diseño *pptr*, idéntico al requerido para el estimador Hansen-Hurwitz para la estimación de la media poblacional, podemos proponer un nuevo estimador de la varianza poblacional. Ahora,

$$E(e_i e_j) = E(e_i)E(e_j) + \text{ Cov } (e_i, e_j) = n^2 p_i p_j - n p_i p_j = n(n-1)p_i p_j$$

y quedaría el estimador insesgado así,

$$\hat{\sigma}_y^2 = \frac{1}{N^2} \sum_{i \neq j=1}^{N} \sum \frac{(y_i - y_j)^2}{n(n-1)p_i p_j} \cdot e_i e_j \quad .$$

Del mismo modo, utilizando el diseño propuesto por Sánchez-Crespo (1977) de tipo ordenado, el estimador se obtendrá calculando

$$E(e_i e_j) = E(e_i)E(e_j) + \text{ Cov } (e_i, e_j) = n^2 p_i p_j - \tfrac{M-n}{M-1} n p_i p_j =$$
$$= n p_i p_j \left[n - \tfrac{M-n}{M-1} \right]$$

y así el estimador insesgado queda

$$\hat{\sigma}_y^2 = \frac{1}{N^2} \sum_{i \neq j=1}^{N} \sum \frac{(y_i - y_j)^2}{n p_i p_j \left[n - \frac{M-n}{M-1} \right]} e_i e_j \qquad (2)$$

que es otro nuevo estimador insesgado, caso particualr del estimador modelo propuesto. Con el diseño introducido en la sección 3, el estimador insesgado de la varianza poblacional se obtendría como en (2) sustituyendo M por $K = M/m$.

5. DISCUSIÓN DE LOS RESULTADOS

El planteamiento propuesto permite considerar con un mismo tratamiento a estimadores clásicos así como otros estimadores insesgados que surgen de modo natural por este procedimiento para otros diseños. Del mismo modo se pueden proponer estimadores insesgados de funciones paramétricas polinomiales para diseños ordenados (ya no sólo ordenados como presentó Ruiz (1986), generalizando el estimador Horvitz-Thompson (1952)).

En general, para cualquier otra función paramétrica basta descomponerla en monomios, cada uno de los cuales contiene en su forma explícita k valores $y_i (i = 1, 2, ..., k$ por simplicidad) de la variable de interés. Si este monomio no se simplifica con otro en el que intervengan los mismos valores de interés y_i $(i = 1, 2, ..., k)$, entonces en el estimador aparecerá la misma expresión monomial multiplicada por el producto $e_1 e_2 ... e_k$ y dividida por $E(e_1 e_2 ... e_k)$. Con

ello, si la esperanza última es positiva (no nula) se puede proponer un estimador insesgado de este monomio de modo análogo a como hemos expuesto en las anteriores secciones. Realizando lo mismo con los demás monomios no simplificables de la función paramétrica, podemos llegar a proponer el estimador insesgado de dicha función paramétrica ya sea para diseños ordenados o no.

6. BIBLIOGRAFÍA

[1] **Hansen, M.H. y Hurwitz, W.N.** (1943). "On the theory of sampling from finite populations". Ann. Math. Statist. 14, 333-362.

[2] **Horvitz, D.G. y Thompson, D.J.** (1952). "A generalisation of sampling without replacement from a finite universe". J. Amer. Statist. Assoc. 47, 663-685.

[3] **Murthy, M.N.** (1963). "Generalised unbiased estimation in sampling from finite populations". Sankhyā Ser. B 25, 245-262.

[4] **Ruiz, M.** (1986). "Funciones paramétricas estimables en teoría de muestras". Estadíst. Española 112-113, 69-73.

[5] **Ruiz, M.** (1987). "Sobre estimadores UMV y UMECM en poblaciones finitas". Estadíst. Española 115, 105-111.

[6] **Ruiz, M.** (1988). "El teorema de Rao-Blackwell en poblaciones finitas". Actas XVII Reunión Nacional de Estadística, I.O. e Informática, Benidorm.

[7] **Sánchez-Crespo, J.L.** (1977). "A new sampling scheme: selection with graduate variable probabilities without replacement". Bull. Internat. Statist. Inst. 47, 458-461.