ALGUNAS APORTACIONES PARA EL CALCULO SIMPLIFICADO DE LOS ERRORES DE MUESTREO DURANTE EL PERIODO 1967–1970

J.L. Sánchez – Crespo

INSTITUTO NACIONAL DE ESTADISTICA

INTRODUCCION

Cuando fue promulgada la Ley de Estadística de 1945 lo que hoy se conoce en el mundo entero con el nombre de "Muestreo probabilístico" puede decirse que había nacido sólo una década antes, al publicar J. Neyman su famoso trabajo sobre el muestreo estratificado.

En el período 1945–1970 la teoría del muestreo de poblaciones finitas continúa su desarrollo en múltiples direcciones. Sin pretender, ni muchos menos, ser exhaustivos en la exposición de esta evolución, nos limitaremos a mencionar algunas de sus principales líneas.

En el quinquenio 1945–1950 se desarrolla la teoría del muestreo sistemático. Se inicia con las muestras interpenetrantes lo que hoy se conoce como muestreo reiterado, y comienzan las investigaciones de métodos para el tratamiento de la no–respuesta, que darían lugar a numerosas publicaciones en años posteriores.

En la década 1950–1960 se inicia la construcción de estimadores insesgados de la razón, la selección controlada, los dominios de estudio, modelos para estudiar cambios en ocasiones sucesivas, la investigación sobre los límites de estratos, la estratificación de doble entrada, la construcción de estimadores compuestos que aprovechan la capacidad de los ordenadores, técnicas para simplificar el cálculo de los errores de muestreo en diseños complejos, modelos para el tratamiento del error total y las encuestas de evaluación.

Finalmente, en la década 1960–1970, se continúan las investigaciones iniciadas en años anteriores dedicándose especial atención a los modelos y encuestas de evaluación, a cuyo fin inicial de medir la acuracidad de censos y encuestas se añadió el de mejorar su eficiencia. Podemos, pues, señalar como una línea fundamental en la evolución de esta década la investigación de los errores ajenos al muestreo y de técnicas que permitan simplificar el cálculo de los errores de muestreo en los cada vez más complejos diseños.

Los primeros textos de muestreo fueron debidos a Yates (1949), Deming (1950), Hansen, Hurwitz y Madow (1953), y Cochran (1953), a los que han seguido otros muchos en años posteriores. Es de destacar que ya en 1950 fuese editado por el Instituto Nacional de Estadística el primer texto en lengua española debido a E. Cansado.

Como las aportaciones a la teoría del muestreo señaladas en esta introducción han sido incorporadas a los textos existentes, hemos pensado tratar en este artículo sólo algunas relativas a la simplificación en el cálculo de los errores de muestreo durante el período 1967–1970 que, por ser tan recientes, aún no figuran en los textos actuales.

TECNICAS PARA SIMPLIFICAR EL CALCULO DE LOS ERRORES DE MUESTREO.

La creciente aplicación del muestreo para la realización de encuestas a gran escala ha llevado consigo el empleo de diseños muestrales cada vez más complejos,que ha imposibilitado, en la práctica, la utilización de los métodos tradicionales para estudiar los errores de muestreo. Este hecho ha provocado un gran esfuerzo en la investigación de métodos simplificados durante los últimos veinticinco años. Las muestras interpenetrantes debidas a Mahalanobis (1946) y el método de reiteraciones de Deming (1960) son un ejemplo en este sentido, cuando se eligen las unidades primarias con reemplazamiento. Paralelamente, Yates (1949) estableció una regla práctica y aplicable cuando las unidades primarias de la muestra eran obtenidas con probabilidades iguales. Esta regla fue generalizada por Durbin (1953) para el caso de probabilidades desiguales.

La selección sin reemplazamiento y probabilidades desiguales ha recibido gran atención durante las dos últimas décadas, especialmente para el caso de 2 unidades por estrato. Las aportaciones de Narain (1951); Murthy (1957); Des Raj (1956); Hartley y Rao (1962); Rao, Hartley y Cochran (1962), y Stuart (1964) han sido recogidas en los textos de muestreo. La dificultad para el cálculo de la probabilidad conjunta de inclusión en la muestra de cada par de unidades ha hecho que estos métodos no tuviesen aceptación en la práctica, a pesar de la disponibilidad de los ordenadores.

METODOS DE DURBIN

Durbin, en 1967, ha propuesto dos métodos de selección de unidades de muestreo y estimación de varianzas para diseños polietápicos estratificados, en los que la selección de dos unidades primarias por estrato se realiza sin reemplazamiento y con probabilidades estrictamente proporcionales a sus tamaños. Presentamos a continuación estos métodos en forma sucinta:

Método I

Sea, en un determinado estrato, M_i el tamaño de la unidad primaria i–ésima y $P_i = \frac{M_i}{\Sigma M_i}$ su probabilidad de selección con $\Sigma P_i = 1$. Se elige la primera unidad, U_i, con probabilidad P_i y la segunda con probabilidad proporcional a

$$P_j \left[\frac{1}{1 - 2 P_i} + \frac{1}{1 - 2 P_j}\right] \quad \text{con } i \neq j$$

La probabilidad total de que U_i y U_j figuren en la muestra será:

$$f_{ij} = P_i \cdot P(j/i) + P_j \cdot P(i/j)$$

con $$P(j/i) = \lambda_i P_j \left[\frac{1}{1 - 2P_i} + \frac{1}{1 - 2 P_j}\right]$$

y $$\sum_{j \neq i} P(j/i) = 1$$

de donde sumando los dos miembros para $j \neq i$

$$1 = \lambda_i \sum_{j \neq i} \frac{P_j}{1-2P_i} + \lambda_i \sum_{j \neq i} \frac{P_j}{1-2P_j} = \frac{\lambda_i}{1-2P_i} \sum_{j \neq i} P_j + \lambda_i \sum_{j \neq i} \frac{P_j}{1-2P_j}$$

pero

$$P_i + \sum_{j \neq i} P_j = 1 \qquad \sum_{j \neq i} P_j = 1 - P_i$$

luego

$$1 = \lambda_i \frac{1 - P_i}{1 - 2P_i} + \lambda_i \sum_{j \neq i} \frac{P_j}{1 - 2P_i}$$

y como

$$\sum_{k=1}^{N} \frac{P_k}{1 - 2P_k} = \frac{P_i}{1 - 2P_i} + \sum_{j \neq i} \frac{P_j}{1 - 2P_j}$$

tenemos

$$1 = \lambda_i \left[\frac{1 - P_i}{1 - 2P_i} - \frac{P_i}{1 - 2P_i}\right] + \lambda_i \sum_{k}^{N} \frac{P_k}{1 - 2P_k}$$

$$1 = \lambda_i \left[1 + \sum_{k}^{N} \frac{P_k}{1 - 2P_k}\right]$$

de donde se deduce para λ_i un valor que no depende de i:

$$\lambda = \frac{1}{1 + \sum_{k}^{N} \frac{P_k}{1 - 2P_k}}$$

y por lo tanto

$$f_{ij} = P_i \lambda P_j \left[\frac{1}{1 - 2P_i} + \frac{1}{1 - 2P_j}\right] + P_j \lambda P_i \left[\frac{1}{1-2P_j} + \frac{1}{1-2P_i}\right] =$$

$$= 2 \lambda P_i P_j \left[\frac{1}{1 - 2P_i} + \frac{1}{1 - 2P_j}\right]$$

por ser iguales los dos sumandos $P_i\ P(j/i)$ y $P_j\ P(i/j)$

la probabilidad de selección de la unidad i en segundo lugar es igual a la de su selección en el primero, es decir P_i, de donde la probabilidad total de que la mencionada unidad pertenezca a la muestra será

$$f_i = 2P_i \quad \text{lo que obliga a que} \quad P_i \leqslant 1/2$$

para aplicar este método. Si el máx. de P_i es $1/2$ se incluye con certeza la unidad mayor y se selecciona la segunda entre las restantes.

La selección de las dos unidades puede hacerse utilizando el método de Lahiri (1951) con el que se opera en la forma siguiente:

Siendo M_i el tamaño de la i–ésima unidad primaria y M_o un valor tal que $M_o \geqslant$ Máx. M_i se extrae un par de número aleatorios (i, k) tales que

$$1 \leqslant i \leqslant N\ ;\ 0 \leqslant k \leqslant M_o\ .$$

Si $k \leqslant M_i$ se acepta U_i, y en otro caso se rechaza, obteniéndose otro par (i, k) repitiéndose el procedimiento.

Para la segunda unidad se sigue idéntico método con

$$M'_j = M_j \left[\frac{1}{(M - 2M_i)} + \frac{1}{(M - 2M_j)}\right] \quad j \neq i$$

donde $M = \Sigma\ M_i$ y $M'_o \geqslant$ máx. M'_j

Para un estimador lineal

$$\theta^* = \sum_{h=1}^{L} (y_{hi} + y_{hj})$$

donde y_{hi}, y_{hj} son las contribuciones de las dos unidades primarias a través de varias etapas puede demostrarse (Cochran 1963) que

$$\hat{V}_s(\theta^*) = \sum_h^L \left(\frac{f_{hi}\, f_{hj}}{f_{hij}} - 1\right)(y_{hi} - y_{hj})^2 + \sum_h^L (f_{hi}\, S^2_{hi} + f_{hj}\, S^2_{hj})$$

es un estimador insesgado de $V(\theta^*)$.

Si el muestreo hubiese sido realizado con reemplazamiento

$$\hat{V}_c(\theta^*) = \sum_{h=1}^{L} (y_{hi} - y_{hj})^2$$

sería un estimador insesgado de $V_c(\theta^*)$ (Cochran 1963).

La complicación de $\hat{V}_s(\theta^*)$ derivada de la consideración de las f_i y f_{ij} es evidente.

El agrupamiento de unidades primarias en cada estrato permite simplificaciones importantes.

Se procede del modo siguiente:

a) En cada grupo se incluye el menor número posible de unidades, de forma que el máximo de P_i sea menor o igual que la mitad del total de las P_i en el grupo.

b) Se seleccionan 2 unidades del estrato con reemplazamiento y probabilidad proporcional a P_i

c) Si las unidades pertenecen a distintos grupos se aceptan ambas.

Si pertenecen al mismo grupo, se acepta la primera y se rechaza la segunda, procediéndose a una selección más en el grupo con el Método I, sustituyendo P_i por $P_i = \dfrac{P_i}{\sum_G P_i}$ designando por G el grupo.

Durbin demuestra en su artículo los resultados que permiten establecer la siguiente regla para estimar varianzas, aplicable tanto a la primera etapa como a las etapas posteriores

1º Si las unidades de un estrato proceden de grupos distintos se aplica la fórmula del muestreo con reemplazamiento

2º Si proceden del mismo grupo se calcula la contribución a la varianza co-

mo si cada estrato estuviese formado solamente por el grupo que contiene las unidades.

Método II

Con este método aproximado se procede de la siguiente forma:

1°.– Relación de unidades según tamaño y formación de pares adyacentes. (No necesaria en la aplicación).

2°.– Selección de 2 unidades con reemplazamiento y probabilidades proporcionales al tamaño.

3°.– Si las unidades coinciden, se acepta ésta y se toma el otro miembro del par. Si no coinciden se aceptan ambas.

Método de las pseudo – reiteraciones

Una desventaja importante del muestreo reiterado reside en la posible dificultad de obtener el número suficiente de reiteraciones independientes para que la varianza estimada tenga una estabilidad adecuada en el muestreo.

El Bureau of the Census de los Estados Unidos utiliza un plan de muestreo, para la Current Population Survey, en el que sólo se obtiene una unidad primaria de cada estrato. Para estimar varianzas se forman pares de estratos, disponiéndose sólo de dos reiteraciones independientes. Para evitar esta dificultad se desarrolló por el citado Bureau un método denominado de pseudo–reiteraciones con semi muestras, publicado en 1963. El método ha sido aplicado utilizando 40 pseudo–reiteraciones y asegurando que cada unidad primaria apareciese en 20 de ellas.

El método de pseudo–reiteraciones equilibradas con semi muestras es una modificación del anteriormente citado, que permite obtener L semi-muestras equilibradas, del conjunto total de 2^L, que para el propósito de estimar varianzas contiene toda la información disponible en el conjunto total. Este método ha sido publicado por Mc Carthy en 1969.

A continuación presentamos un resumen de ambos métodos, para dos selecciones independientes por estrato.

Método de pseudo–reiteraciones con semi muestras

Sea $\bar{\bar{X}}_{st} = \sum_{h}^{L} W_h \bar{X}_h$ el estimador de la media poblacional $\bar{X}$ donde L es el número de estratos y

$$\bar{X}_h = \frac{X_{h1} + X_{h2}}{2}$$

El estimador insesgado de la varianza es

$$\hat{V}(\bar{X}_{st}) = \sum_h^L W_h^2 \frac{S_h^2}{2} = \frac{1}{4} \sum_h^L W_h^2 (X_{hi} - X_{h2})^2 = \frac{1}{4} \sum_h^L W_h^2 d_h^2$$

Una pseudo-reiteración con semi-muestra consiste en la elección en cada estrato del índice 1 ó del 2, con lo cual el número de reiteraciones sería el de variaciones con repetición de 2 elementos tomados de L en L, es decir 2^L.

Sea $\bar{Z} = \sum_h^L W_h X_{hi}$ (con i = 1 ó i = 2 para cada estrato) un estimador de $\bar{X}$ basado en semi muestras, puede demostrarse que $E[\bar{Z}]$ sobre las L^2 es igual a a $\bar{X}_{st}$ y que $E[\bar{Z} - \bar{X}_{st}]^2 = V(\bar{X}_{st})$

Consideremos ahora la desviación de un valor $\bar{Z}_1$ respecto a $\bar{X}_{st}$

$$\bar{Z}_1 - \bar{X}_{st} = \sum_h^L W_h X_{hi} - \sum_h^L W_h \cdot \frac{X_{hi} + X_{h2}}{2} =$$

$$= \frac{1}{2} \sum_{h=1}^L W_h d_h \qquad \text{y en general}$$

$$\bar{Z} - \bar{X}_{st} = \frac{1}{2} (\Sigma \pm W_h d_h)$$

Elevando el cuadrado y teniendo en cuenta que los signos de las desviaciones quedan determinados por la elección del apropiado signo en cada estrato, tenemos:

$$(\bar{Z} - \bar{X}_{st})^2 = \frac{1}{4} \Sigma W_h^2 d_h^2 + \frac{1}{2} \Sigma \pm W_h W_k d_h d_k$$

de donde para una semi muestra

$$E(\bar{Z} - \bar{X}_{st})^2 = \hat{V}(\bar{X}_{st}) \text{ por ser } E(d_h d_k) = 0$$

y para todas las posibles muestras:

$$E\ [\bar{Z} - \bar{X}_{st}]^2 = V(X_{st})$$

El método se aplica a la estimación de varianzas en la forma siguiente:

a) Se selecciona, con reemplazamiento, de las 2^L posibles semi-muestras una muestra de k semi-muestras obteniéndose las medias $\bar{Z}_1, \bar{Z}_2 \ldots \bar{Z}_k$.

b) Se calcula

$$\sum_{i}^{k} \frac{(\bar{Z}_i - \bar{X}_{st})^2}{k}$$

Gurney (1962) y Mc. Carthy (1966) han encontrado, bajo ciertas hipótesis, que un valor de K entre 20 y 40 es satisfactorio en la práctica.

Pseudo – reiteraciones equilibradas con semi-muestras (Mc. Carthy 1969)

La variabilidad entre las estimaciones de la varianza, con el método anterior, proceden de las contribuciones entre estratos debidos a los productos cruzados $d_h\ d_k$. Estos términos se compensan para el conjunto de las 2^L posibles semi-muestras.

Mc. Carthy muestra, con un ejemplo que expondremos a continuación, como es posible seleccionar un subconjunto de semi-muestras en la que se compensen los citados productos cruzados.

Ejemplo:

Reiteración	Estratos 1	2	3	$\bar{Z}_i - \bar{X}_{st}$
1	X_{11}	X_{21}	X_{31}	$\frac{1}{2}(+W_1 d_1 + W_2 d_2 + W_3 d_3)$
2	X_{11}	X_{22}	X_{32}	$\frac{1}{2}(+W_1 d_1 - W_2 d_2 - W_3 d_3)$
3	X_{12}	X_{22}	X_{31}	$\frac{1}{2}(-W_1 d_1 - W_2 d_2 + W_3 d_3)$
4	X_{12}	X_{21}	X_{32}	$\frac{1}{2}(-W_1 d_1 + W_2 d_2 - W_3 d_3)$

siendo $d_h = X_{h1} \quad X_{h2}$

Elevando al cuadrado cada desviación tenemos

$$(\bar{Z}_1 - \bar{X}_{st})^2 = \frac{1}{4}\sum_h^L W_h^2 d_h^2 + W_1 W_2 d_1 d_2 + W_1 W_3 d_1 d_3 + W_2 W_3 d_2 d_3$$

$$(\bar{Z}_2 - \bar{X}_{st})^2 = \frac{1}{4}\sum_h^L W_h^2 d_h^2 - W_1 W_2 d_1 d_2 - W_1 W_3 d_1 d_3 + W_2 W_3 d_2 d_3$$

$$(\bar{Z}_3 - \bar{X}_{st})^2 = \frac{1}{4}\sum_h^L W_h^2 d_h^2 + W_1 W_2 d_1 d_2 - W_1 W_3 d_1 d_3 - W_2 W_3 d_2 d_3$$

$$(\bar{Z}_4 - \bar{X}_{st})^2 = \frac{1}{4}\sum_h^L W_h^2 d_h^2 - W_1 W_2 d_1 d_2 + W_1 W_3 d_1 d_3 - W_2 W_3 d_2 d_3$$

luego

$$\frac{\sum_i^4 (\bar{Z}_i - \bar{X}_{st})}{4} = \frac{1}{4}\sum_h^L W_h^2 d_h^2$$

Para que esta propiedad se cumpla para cualquier L y k es suficiente que las columnas de la matriz de signos en las desviaciones sean ortogonales entre si

+	+	+
+	−	−
−	−	+
−	+	−

Este método, presentado con un caso lineal muy sencillo, tiene su principal aplicación para la estimación de varianzas en diseños y métodos de estimación complejos, para los que no existe una teoría que permita la estimación de varianzas. El autor menciona como ejemplo la estimación de la varianza para el coeficiente de regresión múltiple y algunos problemas en los dominios de estudio.

Es de esperar continúe la investigación necesaria que permita llegar a conocer mejor las características de las estimaciones de varianzas obtenidas con este método.

REFERENCIAS

Cansado, E. (1950). Conferencias sobre muestreo. I.N.E. Madrid.

Cocharan, W.G. (1953). Sampling Techniques. Wley. New York.

Cochran, W.C. (1963). Sampling Techniques. Wiley. New York. 2nd Ed.

Deming, W.E. (1950). Some Theory of Sampling. Wiley. New York.

Deming, W.E. (1960). Sample Design in Business Research. Wiley. New York.

Des Raj (1956a). Some estimators in sampling with varying probabilities without replacement. Jour. Amer. Stat. Assoc. 51, 269—284.

Des Raj (1956b). A note on the determination of optimum probabilities in sampling without replacement. Sankhya, 17, 197—200.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. Jour. Roy. Stat. Soc. Ser. B 262—269.

Durbin, J. (1967). Design of Multi—Stage Surveys for the Estimation of Sampling errors. Applied Stat., 16, 152—164.

Hansen, Hurwitz y Madow. (1953). Sample Survey Methods and Theory. Vol. 1 y 2. Wiley. New York.

Hartley y Rao (1962). Sampling with unequal probabilities and without replacement. Ann. Math. Stat. 33, 350—374.

Lahiri, D.B. (1951. A method of sample selection providing umbiased ratio estimates. Bull. Inst. Internat. Statist. 33. n^{o} 133—140.

Mahalanobis, P.C. (1946). Recent experiment in statistical sampling in the Indian Statistical Instituto. Jours. Roy. Stat. Soc. 109, 325—370.

McCarty. (1969). Pseudo—replication: Haf Samples. Bull. Inst. Internat. Statist. Vol. 37, 239—264.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. Sankhya, 18, 379—390.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. Jour. Ind. Soc. Agric. Stat., 3, 169—174.

Neyman, J. (1934). On the two different asêcts pf the reŕesentative method :

the method of stratified sampling and the method of purposive selection. Jour. Roy. Stat. Soc., 97, 558—606.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962). A simple procedure of unequal probability sampling without replacement. Jour. Roy. Stat. Soc., B, 24, (in press).

Stuart, A. (1964). Multistage sampling with preliminary stratification of first stage units. Bull. Inst. Internat. Statist. Vol. 32, n° 3.

U.S. Bureau of the Census (1963). The Current Population Survey a Report on Methodology. Tech. Paper, n° 7. Washington. D.C.

Yates, F. (1949). Sampling Methods for Censuses and Surveys Charles Griffin. London.