QÜESTIIÓ, vol. 25, 1, p. 3-17, 2001

CONTRIBUCIONES AL MUESTREO SUCESIVO: ESTIMADOR PRODUCTO MULTIVARIANTE

EVA M. ARTÉS RODRÍGUEZ
Universidad de Almería*

Se considera el problema de estimar la media de una población finita para la ocasión actual, basándonos en las muestras seleccionadas en dos ocasiones. Se construye un estimador producto multivariante de doble muestreo para la parte solapada de la muestra, para el caso en que dos variables auxiliares se encuentran correlacionadas de forma negativa con la variable objeto de estudio. Se obtienen las expresiones para el estimador óptimo y su varianza. Se calcula la curva que nos proporciona la ganancia en eficiencia del estimador combinado sobre el estimador directo que no utiliza la información obtenida en la primera ocasión. Se obtienen las condiciones bajo las cuales nuestro estimador mejora en precisión al estimador combinado de producto univariante. Finalmente, se incluye un estudio empírico para analizar el buen funcionamiento del método propuesto.

Successive sampling using a multivariate product estimate

Palabras clave: Muestreo en ocasiones sucesivas, estimador producto bivariante, ganancia en eficiencia, fracción de solapamiento

Clasificación AMS (MSC 2000): 62D05

* Departamento de Estadística y Matemática Aplicada. Edificio Científico–Técnico III. 04120 Universidad de Almería. E–mail: eartes@ual.es. Telf.: 950 015172. Fax: 950 015167.

– Recibido en julio de 1999.
– Aceptado en febrero de 2001.

1. INTRODUCCIÓN

Un aspecto a destacar en el análisis de una muestra es el instante o período de tiempo al que hacen referencia los resultados muestrales. Existen dos razones fundamentales por las que ha de considerarse el factor tiempo: las características de los elementos de la población pueden modificarse a lo largo del tiempo, o bien la composición de la población puede verse modificada, debido a que nuevos individuos pueden entrar a formar parte de la misma (*nacimientos*) o dejar de hacerlo (*muertes*).

Si la composición y las características de los elementos permaneciesen inalterables, la realización de un muestreo en un instante dado sería suficiente, ya que la validez de los resultados se mantendría. En la práctica, los cambios anteriormente señalados impiden esta simplificación, y a su vez dan lugar a una serie de objetivos que pueden ser analizados mediante encuestas continuas, como son: la estimación transversal de parámetros poblacionales y de los cambios netos, estimaciones de los valores promedios de los parámetros a lo largo del tiempo, etc.

Las circunstancias de la encuesta y las características que se quieran estimar, son determinantes para elegir el tipo de diseño muestral más adecuado. Existen varias posibilidades:

1. Extraer una nueva muestra en cada ocasión (muestreo *repetido*)
2. Utilizar la misma muestra en todas las ocasiones (muestreo *panel*)
3. Realizar un reemplazamiento parcial de unidades de una ocasión a otra (muestreo en *ocasiones sucesivas*, o también llamado muestreo *rotativo* cuando los elementos tienen restringido el número de etapas en las que van a formar parte de la muestra, como es el caso de la EPA, de periodicidad trimestral, y de la mayoría de las encuestas familiares elaboradas por el INE).

Si existe una relación entre el valor de un elemento de la población en un período de tiempo, y el valor del mismo elemento en el período siguiente, entonces es posible emplear la información contenida en la muestra del período precedente, para mejorar la estimación actual del parámetro poblacional. En este sentido, para que sea posible utilizar la información muestral precedente, se debe obtener la muestra de manera que los elementos muestrales en los dos períodos sucesivos tengan algunos elementos comunes.

Algunos motivos por los que conviene utilizar el reemplazamiento parcial de unidades de la muestra son:

1. Reduce los costes, ya que utilizar una muestra completamente nueva en cada ocasión puede resultar excesivamente costoso.

2. Aumenta la precisión de los estimadores.

3. La permanencia indefinida de las mismas unidades en la muestra puede crear problemas y reducir la eficiencia de los estimadores. Por ejemplo, en las encuestas familiares de tipo *panel* se incrementan los sesgos en las estimaciones debido a la falta de colaboración de algunas familias que pertenecen al panel de hogares.

Así, el INE utiliza principalmente encuestas de muestreo *rotativo* debido a que presenta ventajas de las dos encuestas anteriores (*repetidas* y tipo *panel*).

La teoría sobre muestreo *sucesivo* desarrollada hasta el momento va dirigida a obtener el estimador óptimo combinando dos estimadores de las medias: un estimador indirecto de doble muestreo de la parte apareada de la muestra, y un estimador simple de la media de la parte no apareada.

En este contexto se ha demostrado que el estimador combinado que utiliza un estimador de razón para la parte apareada de la muestra es más preciso que el estimador usual $\bar{y}$ cuando la variable auxiliar se encuentra positivamente correlacionada con la variable objeto de estudio y, y se verifica $\rho > \frac{C_x}{2C_y}$ (*Sen, Sellers y Smith*, 1975). Si la correlación entre la variable auxiliar x y la variable de interés y es negativa, también se ha demostrado que el estimador óptimo que combina un estimador producto de doble muestreo para la parte apareada de la muestra y una media muestral simple de la parte no apareada, tiene menor varianza que el estimador usual $\bar{y}$ siempre que $\rho < -\frac{C_x}{2C_y}$ (*Artés, Rueda y Arcos*, 1998).

Con frecuencia se dispone de información, proporcionada por la encuesta en la primera ocasión, sobre varias variables auxiliares que pueden ser utilizadas para mejorar la precisión de los estimadores. En este sentido, se ha comprobado que el estimador combinado de razón multivariante mejora en precisión al estimador simple si la correlación entre las variables auxiliares y la objeto de estudio, y, es positiva y grande. Sin embargo, cuando la información complementaria se encuentra negativamente correlacionada con la variable objeto de estudio, el método de razón no resulta ser tan eficiente.

Para cubrir un amplio rango de situaciones prácticas, en este artículo se va a desarrollar la teoría en muestreo sucesivo para construir el estimador óptimo de la media en la segunda ocasión combinando un estimador producto multivariante de doble muestreo para la parte apareada de la muestra, con una media simple basada en la parte no apareada de la muestra en la segunda ocasión. Se han empleado dos variables auxiliares x_1 y x_2, por ser el caso de más frecuente aplicación.

La teoría ha sido aplicada para proporcionar estimaciones más precisas de las variables analizadas en un estudio sobre hábitos de salud y nivel de condición física en escolares llevado a cabo en los colegios de Almería capital.

2. TEORÍA

Supongamos que las muestras son de tamaño n en ambas ocasiones, que se utiliza muestreo aleatorio simple y que el tamaño de la población, N, es suficientemente grande como para poder ignorar el factor de corrección por finitud.

Sea una muestra de tamaño n seleccionada en la primera ocasión de una población de tamaño N. Al seleccionar la segunda muestra, suponemos que $n - u = m$ de las unidades de la muestra seleccionada en la primera ocasión se retienen para la segunda ocasión (muestra apareada), y las restantes u unidades son reemplazadas por una nueva selección del universo que resulta después de omitir las m unidades. Se dispone de información acerca de dos variables auxiliares x_1 y x_2 en la primera ocasión, cuyas medias denotamos por $\bar{x}_1$ y $\bar{x}_2$. Sea y la variable objeto de estudio en la segunda ocasión, que suponemos está correlacionada negativamente con x_1 y x_2.

2.1. Notación

m = tamaño muestral de aquellas unidades cuestionadas en ambas ocasiones (muestra apareada)
$u = n - m$, tamaño muestral de aquellas unidades cuestionadas sólo en la segunda ocasión (muestra no apareada)
$\bar{x}_1^m$,$\bar{x}_2^m(\bar{y}_m)$ = media muestral apareada en la primera (segunda) ocasión estimando $\bar{X}_1$,$\bar{X}_2(\bar{Y})$
$\bar{y}_u$ = media muestral no apareada en la 2^a ocasión estimando $\bar{Y}$
$C_0 = \frac{S_y}{\bar{Y}}$
$C_i = \frac{S_{x_i}}{\bar{X}_i} \qquad i = 1, 2$
$\Delta_1 = \frac{C_1}{C_0}$
$\Delta_2 = \frac{C_2}{C_0}$
ρ_{01} = correlación lineal de *Pearson* entre x_1 y y
ρ_{02} = correlación lineal de *Pearson* entre x_2 y y
ρ_{12} = correlación lineal de *Pearson* entre x_1 y x_2
$p = \frac{m}{n}$, fracción del apareamiento

2.2. El Estimador producto multivariante

Las partes apareada (m unidades) y no apareada (u unidades) de la muestra en la segunda ocasión proporcionan estimadores independientes ($\bar{y}_m$ y $\bar{y}_u$) de la media poblacional en la segunda ocasión ($\bar{Y}$).

Para la parte apareada consideramos un estimador mejor, $\bar{y}'_m$, para la media poblacional, $\bar{Y}$, utilizando un estimador producto multivariante de doble muestreo dado por

$$\bar{y}'_m = \omega_1 \frac{\bar{x}_1^m}{\bar{x}_1} \bar{y}_m + \omega_2 \frac{\bar{x}_2^m}{\bar{x}_2} \bar{y}_m$$

La extensión al caso de disponer de k ($k \geq 2$) variables auxiliares negativamente correlacionadas con y es inmediata.

Si se define $W = (\omega_1, \omega_2)$, se obtiene que

$$V\left(\bar{y}'_m\right) = \bar{Y}^2 W D W' \tag{1}$$

donde la matriz $D = (d_{ij})$, siendo

$$d_{ij} = \frac{1}{m} C_0^2 + \left(\frac{1}{m} - \frac{1}{n}\right)\left(C_i C_j \rho_{ij} + C_0 C_i \rho_{0i} + C_0 C_j \rho_{0j}\right) \quad i, j = 1, 2$$

y el valor de W se va a determinar en el sentido que maximice la precisión del estimador $\bar{y}'_m$.

En este sentido, según *Singh*(1967) y siguiendo el procedimiento utilizado por *Olkin*(1958) se obtiene el vector de pesos óptimos mediante la siguiente expresión

$$\hat{W} = \frac{eD^{-1}}{eD^{-1}e'}$$

donde $e = (1, 1)$ y D^{-1} es la matriz inversa de D. Sustituyendo en (1) obtenemos la varianza mínima para el estimador

$$V\left(\bar{y}'_m\right) = \bar{Y}^2 \hat{W} D \hat{W}'$$

Suponiendo que los pesos son uniformes para las variables auxiliares x_1 y x_2 (*Singh*, 1967) el vector de pesos óptimos vendrá dado por

$$\hat{W} = \left(\frac{1}{2}, \frac{1}{2}\right)$$

Como un ejemplo de pesos uniformes se supone

$$C_i = C \quad , \quad \rho_{0i} = \rho_0$$

y

$$\rho_{ij} = \rho \quad (i \neq j) \quad \text{para} \quad i = 1, 2 \tag{2}$$

que implica $\Delta_i = \Delta$, y proporciona una varianza para $\bar{y}'_m$ dada por

$$V_{min}\left(\bar{y}'_m\right) = \frac{S_y^2}{m}\left(1 + \frac{u}{n}\left(\frac{C^2}{C_0^2}\frac{1+\rho}{2} + 2\rho_0 \frac{C}{C_0}\right)\right) =$$

$$= \frac{S_y^2}{m}\left(1+\frac{u}{n}\Delta\left(\frac{1+\rho}{2}\Delta+2\rho_0\right)\right)$$

Un estimador de la varianza puede ser obtenido reemplazando los parámetros poblacionales en la expresión anterior por sus correspondientes estimadores muestrales.

Puesto que el estimador directo $\bar{y}_m$ basado en las m unidades tiene varianza

$$V(\bar{y}_m) = \frac{S_y^2}{m}$$

se obtiene que $\bar{y}'_m$ es más preciso que $\bar{y}_m$ si

$$\frac{2\rho_0}{1+\rho} < -\frac{1}{2}\left(\frac{C}{C_0}\right)$$

(condición análoga a la obtenida por *Singh*(1967) en muestreo simple)

Así, se puede construir un estimador de la media de la población en la segunda ocasión, $\bar{Y}$, combinando los dos estimadores independientes $\bar{y}'_m$ y $\bar{y}_u$ con pesos ω y $(1-\omega)$ respectivamente, dado por

$$\bar{y}_{2PM} = \omega\bar{y}'_m + (1-\omega)\bar{y}_u$$

y

$$V(\bar{y}_{2PM}) = \omega^2 V(\bar{y}'_m) + (1-\omega)^2 V(\bar{y}_u)$$

Se obtiene el mejor estimador de $\bar{Y}$ en la segunda ocasión utilizando el valor de ω que minimice $V(\bar{y}_{2PM})$

$$\omega_{opt} = \frac{V(\bar{y}_u)}{V(\bar{y}_u)+V(\bar{y}'_m)}$$

Teniendo en cuenta que

$$V(\bar{y}_u) = \frac{S_y^2}{u}$$

y sustituyendo en la expresión de la varianza se tiene que

$$V_{min}(\bar{y}_{2PM}) = \frac{V(\bar{y}'_m)\,V(\bar{y}_u)}{V(\bar{y}'_m)+V(\bar{y}_u)}$$

(3)
$$= \frac{S_y^2}{n}\frac{1+qZ}{1+q^2Z}$$

donde $Z = \Delta\left(2\rho_0+\Delta\frac{1+\rho}{2}\right)$.

Si además

$$\rho = -\rho_0 \quad \text{y} \quad C = C_0$$

obtenemos una expresión así de sencilla para la varianza

$$V_{min}\left(\bar{y}_{2PM}\right) = \frac{S_y^2}{n}\frac{1+q\left(\frac{1-3\rho}{2}\right)}{1+q^2\left(\frac{1-3\rho}{2}\right)}$$

El valor óptimo de u se obtiene minimizando en (4) con respecto a la variación en u, y viene dado por

$$\left(\frac{u}{n}\right)_{opt} = \frac{\sqrt{1+Z}-1}{Z}$$

o lo que es lo mismo, la fracción del apareamiento óptimo vale

$$p_{opt} = \frac{1+Z-\sqrt{1+Z}}{Z}$$

3. COMPARACIÓN DE ESTIMADORES

3.1. Estimador simple y estimador combinado de producto multivariante

Si se considera el estimador usual de la media de la población en la segunda ocasión, $\bar{y}$, que es la media muestral basada sólo en las n unidades muestrales de dicha ocasión, y que no utiliza ninguna información adicional, su varianza toma la siguiente expresión

$$V\left(\bar{y}\right) = \frac{S_y^2}{n}$$

Así, podemos comparar este método de estimación clásica con aquel que emplea, en la fase de estimación, la información auxiliar disponible. Para ello, podemos obtener la ganancia en precisión, G, del estimador combinado, $\bar{y}_{2PM}$, que utiliza un estimador de producto multivariante para la parte apareada de la muestra en la segunda ocasión, sobre el estimador simple, $\bar{y}$, mediante la siguiente expresión

$$G = \frac{V\left(\bar{y}\right) - V\left(\bar{y}_{2PM}\right)}{V\left(\bar{y}_{2PM}\right)} = \frac{-Zp\left(1-p\right)}{1+\left(1-p\right)Z}$$

donde $Z = \Delta\left(2\rho_0 + \Delta\frac{1+\rho}{2}\right)$.

Por definición $p \leq 1$. Si $p = 1$ (apareamiento total) ó $p = 0$ (sin apareamiento), la ganancia vale cero. Para cualquier otro valor de p $(0 < p < 1)$, obtendremos una ganancia

positiva si

$$\frac{2\rho_0}{1+\rho} < -\frac{1}{2}\left(\frac{C}{C_0}\right)$$

Por tanto, se puede concluir que la ganancia en precisión del estimador combinado, $\bar{y}_{2PM}$, sobre el estimador simple, $\bar{y}$, es mayor conforme aumenta ρ_0 en valor absoluto (mayor dependencia entre las variables auxiliares x_1 y x_2 con la variable objeto de estudio y), y disminuye el valor de ρ (menor correlación entre las variables x_1 y x_2).

Si sólo se emplea una variable auxiliar x_1 en la fase de estimación, el estimador combinado que utiliza un estimador producto univariante para la muestra apareada en la segunda ocasión, viene dado por

$$\bar{y}_{2p} = \omega\frac{\bar{x}_1^m}{\bar{x}_1}\bar{y}_m + (1-\omega)\bar{y}_u$$

y mejora en precisión al estimador clásico siempre que

$$\rho < -\frac{1}{2}\left(\frac{C}{C_0}\right)$$

(*Artés, Rueda y Arcos*, 1998).

3.2. Estimador combinado de producto univariante versus multivariante

Se ha estudiado también la precisión del estimador combinado de producto multivariante con aquél que utiliza un estimador producto univariante para la parte apareada de la muestra, a partir de sus varianzas. Según *Artés, Rueda y Arcos*(1998)

$$V_{min}\left(\bar{y}_{2p}\right) = \frac{S_y^2}{n}\frac{1+q\left(2\rho_0+1\right)}{1+q^2\left(2\rho_0+1\right)}$$

Sin embargo, si se utiliza la información auxiliar proporcionada por x_1 y x_2 en la primera ocasión, y se considera un estimador producto multivariante para la parte apareada de la muestra en la segunda ocasión, obtenemos una expresión para la varianza mínima del estimador combinado resultante dada por (4). En este caso, cuando los pesos son uniformes y se cumple la condición (2), obtenemos

$$V_{min}\left(\bar{y}_{2p}\right) - V_{min}\left(\bar{y}_{2PM}\right) \geq 0$$

cuando

$$\frac{1-\rho}{2} > 0$$

(condición análoga a la obtenida por *Singh*(1967) en muestreo simple) es decir, el estimador combinado de producto bivariante es más preciso que el estimador que utiliza

un estimador producto univariante para la parte apareada. (Se puede generalizar a $k \geq 2$ variables auxiliares).

4. ESTUDIO EMPÍRICO

Para evaluar el buen funcionamiento del método propuesto se han utilizado los datos recogidos en una investigación sobre hábitos saludables y nivel de condición física. Dicho estudio se ha llevado a cabo sobre una población de escolares de 4^0 de Educación Secundaria Obligatoria (E.S.O.) en los colegios de Almería capital durante los meses de abril y junio de 1998.

Se ha pretendido desarrollar un plan de muestreo que proporcione estimadores más precisos de las variables estudiadas. Dicho plan se ha basado en el principio del muestreo *sucesivo* de la misma población, y consistió en dos conjuntos de muestras aleatorias independientes: (*i*) una muestra de 135 escolares seleccionados, en la 1^a ocasión (Abril del 98), entre los 2681 escolares que formaban la población, y (*ii*) una segunda muestra de 202 escolares seleccionada, en la 2^a ocasión (Junio del 98), entre los 2546 escolares que no formaron parte de la muestra apareada.

A cada niño de la muestra se le administró un cuestionario sobre hábitos saludables, y se evaluó el nivel de condición física mediante determinados test y medidas antropométricas.

Para el propósito del presente estudio hemos considerado la estimación del *componente endomorfo* (y, una de las múltiples variables implicadas en la investigación) en la 2^a ocasión, tomando como variables auxiliares la *flexión mantenida de brazos* (x_1) y el *volumen máximo de Oxígeno* (x_2) de la 1^a ocasión. El procedimiento de estimación ha consistido en combinar los estimadores de las dos muestras independientes de escolares: $\bar{y}'_m$ y $\bar{y}_u$.

Los datos muestrales sobre el número de escolares y parámetros obtenidos en las dos ocasiones han sido los siguientes:

Primera Ocasión (abril 98): gran muestra $n = 337$

Segunda Ocasión (junio 98): muestra apareada $m = 135$, muestra no apareada $u = 202$

$\hat{C}_0 = 0.42 \qquad \hat{\rho}_{01} = -0.60$

$\hat{C}_1 = 0.27 \qquad \hat{\rho}_{12} = 0.58$

$\hat{C}_2 = 0.27 \qquad \hat{\rho}_{02} = -0.60$

A partir de los datos obtenemos que

$$\hat{V}_{min}\left(\bar{y}_{2PM}\right) = 0.87\frac{s_y^2}{n} < \frac{s_y^2}{n} = \hat{V}\left(\bar{y}\right)$$

lo que supone un 14.53% de ganancia en precisión del estimador propuesto sobre el estimador usual.

Se ha calculado también la fracción del apareamiento óptimo

$$\hat{p}_{opt} = 42.73\%$$

Además, se ha comparado la mejora en precisión del estimador propuesto con otros estimadores indirectos. Los resultados se muestran en la tabla 1.

Tabla 1. Comparación de eficiencias entre estimadores

Estimadores	Variable Auxiliar	Varianza	Precisión sobre $\bar{y}$
1. Directo $\bar{y}$	ninguna	$\frac{s_y^2}{n}$	
2. Producto Univariante $\bar{y}_{2p}$	x_1	$0.94\frac{s_y^2}{n}$	6.38%
3. Razón Bivariante $\bar{y}_{2RM}$	x_1 y x_2	$1.22\frac{s_y^2}{n}$	-18.03%
4. Regresión Bivariante $\bar{y}_{2reg}$	x_1 y x_2	$0.88\frac{s_y^2}{n}$	13.62%
5. Producto Bivariante $\bar{y}_{2PM}$	x_1 y x_2	$0.87\frac{s_y^2}{n}$	14.53%

Como podemos observar, el método de razón no resulta eficiente cuando las variables auxiliares se encuentran negativamente correlacionadas con la variable de interés y, ya que la ganancia en precisión sobre $\bar{y}$ es negativa ($\hat{G} = -18.03\%$). Sin embargo, el estimador combinado basado en un estimador producto bivariante de la parte apareada de la muestra y una media simple de la parte no apareada, $\bar{y}_{2PM}$, es más preciso que el correspondiente estimador que utiliza un estimador de producto univariante para la muestra apareada, $\bar{y}_{2p}$, e incluso mejora en precisión a aquél que utiliza un estimador de regresión para la parte apareada, $\bar{y}_{2reg}$. En la última columna se muestra la ganancia en eficiencia (en %) alcanzada por los distintos estimadores, con respecto a $\bar{y}$.

REFERENCIAS

Artés, E., Rueda, M. y Arcos, A. (1998). «Successive Sampling using a Product Estimate», *Applied Sciences and the Environment, Computational Mechanics Publications*, 85–90.

Casimiro, A. J. (1999). *Comparación, evolución y relación de hábitos saludables y nivel de condición física–salud en escolares, entre final de Educación Primaria (12 años) y final de Educación Secundaria Obligatoria (16 años)*. Tesis Doctoral, Universidad de Granada.

Cochran, W. G. (1977). *Sampling Techniques*, third edition, *John Wiley & Sons*, New York.

Olkin, I. (1958). «Multivariate Ratio Estimation for Finite Populations», *Biometrika*, 43, 154–165.

Rao, P. S. R. S. (1988). «Ratio and regression estimators», *Handbook of Statistics 6. Sampling*. Krishnaiah y Rao (Eds.), North Holland, Amsterdam.

Rao, P. S. R. S. & Mudholkar, G. S. (1967).«Generalized multivariate estimators for the mean of finite populations», *Journal of the American Statistical Association*, 62, 1008–1012.

Sen, A. R. (1972). «Successive Sampling With Two Auxiliary Variables», *Sankhyā: The Indian Journal of Statistics*, B, 371–378.

Sen, A. R., Sellers, S. & Smith, G. E. J. (1975). «The Use of a Ratio Estimate in Successive Sampling», *Biometrics*, 31, 673–683.

Singh, M. P. (1967). «Multivariate Product Method of Estimation for Finite Populations», *Journal of the Indian Society Agricultural Statistics*, 19 (2), 1–10.

Tuteja, R. K. & Bahl, S. (1991). «Multivariate Product Estimators», *Calcutta Statistical Association Bulletin*, 42, 161–164.

ENGLISH SUMMARY

SUCCESSIVE SAMPLING USING A MULTIVARIATE PRODUCT ESTIMATE

EVA M. ARTÉS RODRÍGUEZ
Universidad de Almería*

The problem of estimation of a finite population mean for the current occasion based on the samples selected over two occasions has been considered. For the case when two auxiliary variables are negatively correlated with the main variable, a double–sampling multivariate product estimate from the matched portion of the sample is presented. Expressions for optimum estimator and its variance have been derived. The gain in efficiency of the combined estimate over the direct estimate using no information gathered on the first occasion is computed.
A comparison with the univariate product estimator has been made, giving the specific situations under which either of them may be efficiently used. An empirical study is also included for illustration.

Keywords: Successive sampling, bivariate product estimator, gain in efficiency, matching fraction

AMS Classification (MSC 2000): 62D05

* Departamento de Estadística y Matemática Aplicada. Edificio Científico–Técnico III. 04120 Universidad de Almería. E–mail: eartes@ual.es. Telf.: 950 015172. Fax: 950 015167.

– Received July 1999.
– Accepted February 2001.

1. INTRODUCTION

The theory on successive sampling developed so far aims to gain the optimum estimator by combining two mean estimates: a double sampling indirect estimator for the matched part of the sampling, and a simple estimator for the mean of the unmatched part. It has been shown, in this context, that the combined multivariate ratio estimator improves accuracy over the simple estimator if the relation between auxiliary variables and the principal variable, y, is positive and large. However, when the complementary information is negatively related to the study variable, the ratio method is not that efficient.

In order to cover a wide range of practical situations, this paper focuses on the development of the theory on successive sampling, aiming to build the optimum estimator of the mean at the second occasion, by using a double sampling multivariate product estimator for the matched part of the sampling, and a simple mean based on the unmatched part of the sample on the second occasion. We have used two auxiliary variables, x_1 and x_2, as they are the most frequently applied.

The theory has been applied to provide more accurate estimations of the analysed variables over a study on schoolchildren's health habits and fitness carried out in Almeria schools.

2. DEVELOPMENT OF MULTIVARIATE PRODUCT METHOD OF ESTIMATING THE MEAN ON THE SECOND OCCASION

2.1. Selection of the sample

Suppose that the samples are of size n on both occasions, we use a simple random sampling and the size of the population N is sufficiently great for the factor of correction be ignored.

Let a simple random sample of size n be selected on the first occasion from a universe of size N. When selecting the second sample, we assume that $n - u = m$ of the units of the selected sample on the first occasion are retained for the second occasion (matched sample) and the remaining u units are replaced by a new selection from the universe $N - m$ left after omitting the m units.

Information about both auxiliary variables x_1 and x_2 is available for the first occasion, whose means are denoted $\bar{x}_1$ and $\bar{x}_2$, respectively. Let y be the variable under study on the second occasion, and we suppose that is negatively correlated with x_1 and x_2.

2.2. The multivariate product method of estimation

We construct the optimum estimate of the mean of the population on the second occasion, $\bar{Y}$, by combining two independent estimates: a double sampling multivariate product estimate for the matched portion, $\bar{y}'_m$ and a simple estimate for the unmatched portion, $\bar{y}_u$, with weights ω and $(1-\omega)$ respectively. Thus

$$\bar{y}_{2PM} = \omega\bar{y}'_m + (1-\omega)\bar{y}_u$$

with variance

$$V_{min}(\bar{y}_{2PM}) = \frac{S_y^2}{n}\frac{1+qZ}{1+q^2Z} \tag{4}$$

where $Z = \Delta\left(2\rho_0 + \Delta\frac{1+\rho}{2}\right)$.

The optimum matching fraction is given by

$$p_{opt} = \frac{1+Z-\sqrt{1+Z}}{Z}$$

3. COMPARISON OF ESTIMATORS

We have computed the gain in precision G of the combined multivariate product estimate $\bar{y}_{2PM}$, over the direct estimate $\bar{y}$, which is based exclusively on the n sampling units for the second occasion.

$$G = \frac{V(\bar{y}) - V(\bar{y}_{2PM})}{V(\bar{y}_{2PM})} = \frac{-Zp(1-p)}{1+(1-p)Z}$$

where $Z = \Delta\left(2\rho_0 + \Delta\frac{1+\rho}{2}\right)$.

And we conclude that the gain in precision of the combined estimate, $\bar{y}_{2PM}$, over the direct estimate, $\bar{y}$, increase with increasing the dependence between the auxiliary variables x_1 and x_2 with the variable under study y, and decreasing the correlation between x_1 and x_2.

Also, we have compared the precision of the combined estimator of univariate product versus multivariate. So, if the provided auxiliary information by x_1 and x_2 is utilized on the first occasion, and a double sampling multivariate product estimate from the matched portion of the sample on the second occasion is considered, we obtain more precision than by using an univariate product estimate from the matched portion.

4. EMPIRICAL STUDY

We have used the data collected in a survey on healthy habits and fitness level to assess the optimal operation of the proposed method. This study was carried out over a population of fourteen-year-old schoolchildren in Almeria schools during April and June, 1998. We have intended to develop a sampling scheme that provides us with more accurate estimators of the studied variables. That scheme has been based on the successive sampling principle over the same population, and consisted of two sets of independent random samples: *i*) a selected sample of 135 schoolchildren, at the first occasion (April'98), among the 2681 schoolchildren conforming the population and *ii*) a second sample selected among 202 schoolchildren, at the second occasion (June'98), among the 2546 schoolchildren who did not enter the matched sample.

Every child in the sample was given a questionnaire concerning healthy habits, and the fitness level was assessed by means of some tests and anthropometric measures.

In order to achieve the targets of the study, we have considered the estimation of the endomorphic component ($\bar{y}$, one of the multiple variables which affect the survey) at the second occasion, taking as auxiliary variables the arm maintained flexion ($\bar{x}_1$) and the maximum volume of oxygen ($\bar{x}_2$) from the first occasion. The estimation procedure was performed by combining the estimators for the two independent samples of schoolchildren: $\bar{y}'_m$ and $\bar{y}_u$.

From the sampling data we have obtained a gain in precision of 14.53% of the proposed estimator over the usual estimator. Also, we have calculated the optimum matching fraction $p_{opt} = 42.73\%$. Moreover, we have compared the accuracy of the proposed estimator with other indirect estimators: The ratio method is not efficient when the auxiliary variables are negatively correlated to the principal variable y, as the gain in accuracy over $\bar{y}$ is negative (G = -18.03%); however, the combined estimator based upon a bivariant product estimator for the matched part of the sample $\bar{y}_{2PM}$, is more accurate than the correspondent estimator which makes use of a univariant product estimator for the matched sample $\bar{y}_{2p}$, and it even improves the accuracy of the one which makes use of regression estimator for the matched part, $\bar{y}_{2reg}$. Finally we obtained the efficiency gained from the different estimators, regarding $\bar{y}$.