USO DE DOBLE MARCO EN MUESTREO

Por Richard E. Lund[1]

Centro de Estadística y Cálculo, Chapingo, Méx.

Sinopsis

El uso de dos marcos es necesario en varias ocasiones para cubrir adecuadamente el universo de interés. Cuando los dos marcos tienen duplicaciones se requiere la eliminación de las duplicaciones o el uso de estimadores especiales. La distribución óptima de la muestra sobre los dos marcos es un problema importante cuando varía el costo relativo de muestreo entre los marcos considerados. Este artículo propone estimadores adecuados para el uso de dos marcos así como sus varianzas. Los estimadores que se proponen representan una ganancia en eficiencia sobre los sugeridos por Hartley (1962). La solución del problema de la distribución óptima, se presenta esquemáticamente para casos especiales y algebraicamente para casos generales.

Summary

The use of two sampling frames is often needed to provide adequate coverage of the universe of interest. When the two frames overlap, either the duplicate elements must be removed or special estimators are required. Optimum allocation of the sample among the two frames under varying costs is also a problem. This article suggests estimators suitable for use with two sampling frames. Variances are provided. The estimators are more efficient than those suggested by Hartley (1962). The solution of optimum allocation is presented schematically for some special cases and algebraically for the general cases.

Introducción

En el trabajo de investigación generalmente existe interés en un conjunto definido de unidades, denominado universo en el lenguaje técnico. El propósito del muestreo estadístico es el hacer inferencia sobre características del universo medio de una muestra.

Simbólicamente emplearemos Yi para referirnos al valor de la característica bajo estudio correspondiente a la i-ésima unidad. En esta forma se puede definir un conjunto de valores, asociados con el universo, de tamaño N como:

$$Y_1, Y_2, \ldots\ldots Y_N$$

Este conjunto se denomina población. Lo importante entonces es estimar una función de los valores en la población, por ejemplo, el total (ΣY_i). El método de muestreo para hacer tales estimaciones, consiste en elegir al azar una muestra de la totalidad de las unidades del universo. Se mide la característica de interés en las unidades muestrales para construir una estimación de un valor en la población con base en una función de los valores muestrales. Por ejemplo, para obtener una estimación del total en la población (ΣY_i), se puede emplear la media aritmética representada por y multiplicada por el número de unidades en el universo. Así, emplearemos la función $\hat{Y} = N\overline{y}$, como un estimador del total: (ΣY_i).

1 319 N. 25th Avenue. Bozeman, Montana, U.S.A.

El concepto de muestreo es sencillo, no así su aplicación. La teoría presupone la existencia de un sistema mecánico que permita elegir las unidades muestrales de la totalidad de las unidades del universo con probabilidades conocidas. Así por ejemplo, se supone la disponibilidad de una lista de nombres y direcciones de personas que forman un grupo definido que constituye el universo. Estas listas o cualquier otro mecanismo para identificar las unidades individuales del universo se conocen como marcos.

Como ejemplos de marcos utilizados en la práctica tenemos: cuando el universo a investigar son comercios, el marco se puede construir mediante registros de licencias, lista de miembros de una organización profesional, archivos de impuestos pagados, etc. Cuando nuestro universo son agricultores, entonces emplearemos como marco: listas del último censo, archivos de impuestos pagados, registros de personas que vendieron cultivos especiales, etc. En ocasiones se usan muestras basadas en mapas usando áreas geográficas como unidades muestrales. Sin considerar más ejemplos de marcos, es obvio que los marcos individuales sugeridos tienen muchas deficiencias en la práctica. En particular cuando la inclusión en el marco de todos los elementos del universo de interés es muy deficiente, se usan dos o más marcos o listas, eliminando las duplicaciones de las mismas. con objeto de formar un solo marco que sea más adecuado para representar el universo definido. El único inconveniente de este procedimiento es que la eliminación de las duplicaciones de las listas puede ser muy costosa. Por supuesto, existe la alternativa de usar los marcos sin eliminar las duplicaciones y, suponiendo que el grado de duplicación no pueda omitirse, existirá la posibilidad de que esta alternativa dé resultados absurdos con relación a la investigación. El propósito de este artículo es proponer estimadores satisfactorios que controlen los problemas generados con el uso del doble marco.

Para empezar, ejemplifiquemos la manera como se obtienen resultados absurdos de una investigación cuando no se emplean los estimadores adecuados. Supongamos que el universo está compuesto de miembros de dos organizaciones. Se dispone de listas de los miembros de la organización A que incluye 2,000 personas, mientras que la organización B contiene 1,000 personas. La incógnita es que 500 personas son miembros de ambas organizaciones. pero el investigador trata de estimar la doble membresía con una muestra al azar de los 3,000 nombres o elementos. Dentro de una muestra de 300 personas, suponer que se encontrarán 105 que eran miembros de ambas organizaciones. El valor 105 es solamente 5% mayor del valor esperado. En consecuencia, se estimó que era:

$$\frac{105}{300}\ (3{,}000) = 1{,}050$$

miembros de ambas organizaciones. Según este resultado, existen más miembros comunes a ambas organizaciones que los miembros existentes en la organización B.

La manera correcta de obtener una muestra aleatoria de las dos listas, es dividir entre dos todos los datos relacionados con elementos duplicados, es decir, elementos que se encuentran en los dos marcos. El dividir entre dos solamente los que están duplicados en la muestra, no es suficiente; la multiplicación de la media

de los datos transformados, por el número de elementos en ambas listas nos da un estimador insesgado del total de la población. Un estimador de la media de la población, se obtiene dividiendo la estimación total entre el número de componentes del universo; así se encuentra el número de elementos únicos en ambas listas. Por ejemplo, la estimación correcta del número de miembros de ambas organizaciones, usando los datos arriba mencionados es:

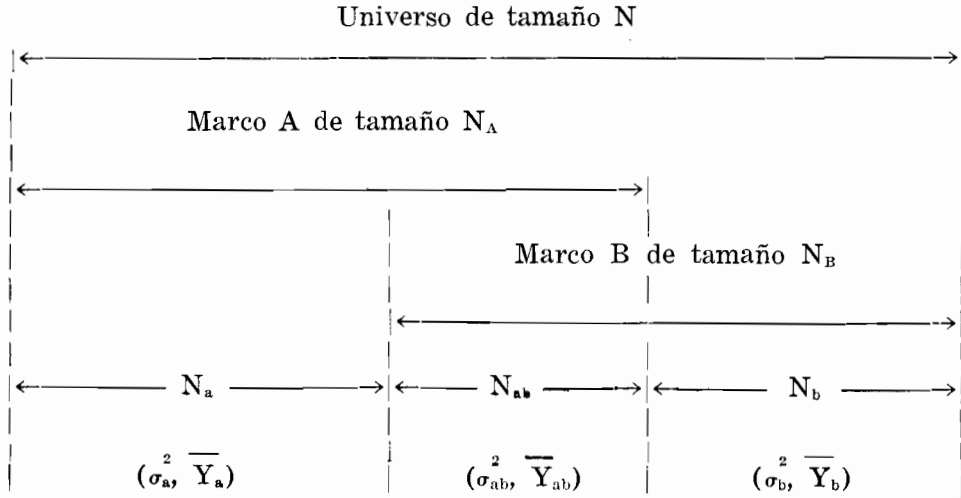
$$\frac{105/2}{300}\ (3{,}000) = 525$$

que es una buena estimación del valor verdadero.

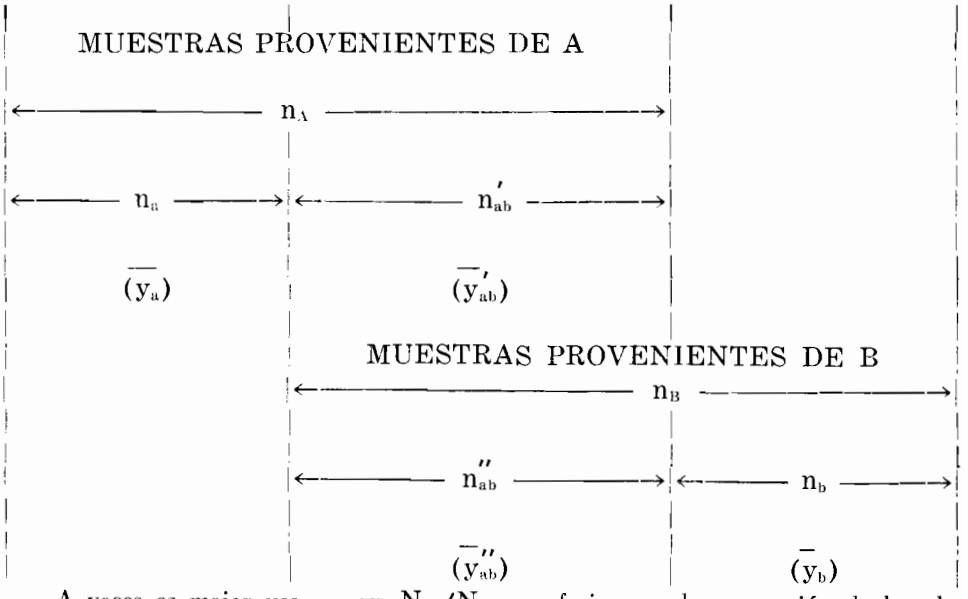
Nomenclatura y notación en caso de dos marcos

El ejemplo anterior trata el caso de una sola muestra al azar de las dos listas como un marco único. En la práctica es más factible elegir una muestra separada de cada lista, o sea de cada marco. Los costos generados con el uso separado de los marcos puede variar significativamente. También, es posible que las varianzas difieran en los diferentes segmentos de los marcos; así, el desarrollo presentado en este artículo tratará el problema de un muestreo estratificado con tres estratos: la parte no duplicada en el marco *A* (a), la parte duplicada y común en los marcos *A* y *B* (ab) y la parte no duplicada en *B* (b). Por lo tanto, el problema principal consiste en encontrar estimadores insesgados, sus varianzas y al mismo tiempo determinar cómo distribuir la muestra entre los dos marcos.

Supongamos la existencia de dos marcos A y B, de tamaños N_A y N_B respectivamente, que cubren completamente el universo de tamaño N. Los tamaños de los tres estratos son: N_a, N_{ab} y N_b; respectivamente se usan los símbolos σ^2_a, σ^2_{ab} y σ^2_a para las varianzas y $\overline{Y}_a$, $\overline{Y}_{ab}$ y $\overline{Y}_b$ para representar las medias de poblaciones. De esta manera, se tiene esquemáticamente:



Del marco A se elige al azar una muestra de tamaño n_A y del marco B, una de tamaño n_B. Debe entenderse que estas muestras son obtenidas en forma aleatoria de los dos estratos definidos para cada marco. Los símbolos n_a y n'_{ab} se refieren a la división de n_A mientras que n''_{ab} y n_b se refieren a las divisiones de n_B. Definiendo en forma similar las medias muestrales, como $\bar{y}_a$, $\bar{y}'_{ab}$, $\bar{y}''_{ab}$ y $\bar{y}_b$, respectivamente. tenemos:



A veces es mejor usar $\alpha = N_{ab}/N_A$ y referirnos a la proporción de los elementos del marco A que están duplicados en B así como $\beta = N_{ab}/N_B$ de igual manera. Los costos muestrales por unidad, cuando usamos A. es igual a c_A. El costo relacionado a B es c_B.

Con esta base, la discusión se puede referir a dos casos principales: uno en el cual, N_a, N_{ab} y N_b son conocidos y un segundo caso en el cual ninguno de estos valores es conocido.

Casos en que N_a, N_{ab} y N_b *son conocidos*

Hartley (1962) sugirió el uso de coeficientes ponderados, p y (1-p), $0 < p < 1$, para crear dos estratos del estrato (ab). Un estrato (ab'), es representativo de los valores originales ponderados por p y el otro estrato (ab"), de los valores originales ponderados por (1-p). Es claro que para cualquier valor de p, la suma de los cuatro estratos (a, ab', ab", b), es igual a la suma de los tres estratos originales (a, ab, b). Por lo tanto, es posible utilizar la teoría de muestreo estratificado.

Se sabe que el estimador del total de la población es:

$$\hat{Y} = N_a\bar{y}_a + N_{ab}p\bar{y}'_{ab} + N_{ab}(1-p)\bar{y}''_{ab} + N_b\bar{y}_b \qquad (1)$$

donde en adición a las definiciones anteriores, $p\bar{y}'_{ab}$ y $(1\text{-}p)\,\bar{y}''_{ab}$ son medios de los estratos ab' y ab". Como se puede ver, se hace una estratificación posterior entre a y ab' y entre ab" y b. Se conoce bien que la varianza en el caso de estratificación posterior es aproximadamente igual a la varianza en el caso de distribución proporcional. De esta manera, excluyendo los multiplicadores finitos, tenemos:

$$\text{Var}\ (\hat{\bar{Y}}) = \frac{N_A}{n_A}\left\{(1-\alpha)\ \sigma_a^2 + \alpha\, p^2\, \sigma_{ab}^2\right\} + \frac{N_B}{n_B}\left\{(1-\beta)\ \sigma_b^2 + \beta\,(1\text{-}p)^2\ \sigma_{ab}^2\right\} \qquad (2)$$

Así, minimizando la expresión (2) con respecto a la variable p resulta en:

$$p = \frac{\alpha n_A}{\alpha n_A + \beta n_B} \qquad (3)$$

De esta manera, la varianza del estimador $\hat{\bar{Y}}$, se reduce al nivel mínimo cuando la ponderación p es igual a la razón de la esperanza de n'_{ab} con respecto a la esperanza de $(n'_{ab} + n''_{ab})$. Debe hacerse notar que n'_{ab} y n''_{ab} se definieron como los tamaños muestrales en la parte duplicada de los marcos A y B respectivamente y son variables aleatorias.

La solución encontrada para p es la solución que aparece en la literatura estadística. Las investigaciones del autor indican que es ventajoso basar p en los valores n'_{ab} y n''_{ab} obtenidos en la muestra; es decir,

$$p = n'_{ab}/(n'_{ab} + n''_{ab}) \qquad (4)$$

Esta última expresión es la solución que se obtiene al minimizar la varianza de $\hat{\bar{Y}}$, condicionada por los tamaños de muestra n'_{ab} y n''_{ab}, con respecto a p. Se puede verificar que la varianza calculada, sin considerar la aproximación presentada en la expresión (2), es igual o menor usando la expresión (4) en contraste al uso de la expresión (3); sin embargo, la reducción en la varianza no es importante, excepto para el caso de muestras muy pequeñas. El hecho es que (2) puede servir también como una buena aproximación de la varianza cuando se usa el valor de p indicado en la expresión (4).

La substitución de (4) en (1) y (2), da lugar a las expresiones de $\hat{\bar{Y}}$ y varianza $\hat{\bar{Y}}$ que el autor recomienda. Estas son:

$$\hat{\bar{Y}} = N_a\bar{y}_a + N_{ab}\bar{y}_{ab} + N_b\bar{y}_b \qquad (5)$$

donde:

$$\bar{y}_{ab} = \frac{n'_{ab}\bar{y}'_{ab} + n''_{ab}\bar{y}''_{ab}}{n'_{ab} + n''_{ab}}$$

y:

$$\text{Var } (\bar{Y}) = \frac{N_A^2}{n_A} \sigma_a^2 + \frac{N_A N_B \alpha \beta}{\alpha n_A + \beta n_B} \sigma_{ab}^2 + \frac{N_B^2}{n_B} \sigma_b^2 \qquad (6)$$

La media $\bar{y}_{ab}$ es simplemente la media sobre todos los elementos seleccionados en el área de duplicación. Así tenemos que, otra ventaja del estimador $\hat{Y}$, es su simplicidad.

El investigador tiene opción para determinar los tamaños de las muestras de cada marco (n_A y n_B); sin embargo, los valores óptimos, para reducir al grado mínimo la varianza aproximada (6), se determinan sujetos a la función lineal de costo: Costos totales $= n_A c_A + n_B c_B$. (7)

La solución general se expresa en términos de la razón (r), de los tamaños, es decir, $r = n_A/n_B$. La técnica iterativa para encontrar r está dada por

$$r_1 = \sqrt{\frac{c_B}{c_A}} \left(\frac{\beta}{\alpha}\right) \qquad (8)$$

y

$$r_j^2 + 1 \qquad \frac{c_B}{c_A} \left(\frac{\beta}{\alpha}\right)^2 \left\{ \frac{(r_i + \frac{\beta}{\alpha})^2 (1-\alpha) \sigma_a^2 + r_i^2 \alpha \sigma_{ab}^2}{(r_i + \frac{\beta}{\alpha})^2 (1-\beta) \sigma_b^2 + (\frac{\beta}{\alpha})^2 \beta \alpha_{ab}^2} \right\} \qquad (9)$$

El procedimiento es empezar con r_1 y substituir este valor en la expresión (9) para obtener r_2. Luego, substituir r_2 en la expresión (9) para obtener r_3 y así sucesivamente. La práctica indica que son pocas las iteraciones que es necesario hacer para obtener un valor final más o menos constante.

La solución óptima del tamaño relativo (n_A/n_B), fue encontrada por el autor para un rango amplio de valores de los parámetros. La varianza del estimador de la solución o distribución óptima, se comparó con las varianzas de los casos en que las distribuciones difieren en un 10% de la óptima. Las diferencias no fueron importantes, ya que los valores de las diferencias, en términos de varianza fueron aproximadamente el 1%. Es decir, la sensibilidad del estimador a desviaciones en la distribución óptima es pequeña.

Se puede considerar un caso especial. Muchas veces se tiene disponible un marco costoso completo y un marco de bajo costo pero incompleto. Definiendo A como el marco completo, es claro que $N_b = 0$ y $\beta = 1$. Las soluciones de las distribuciones óptimas de n_A y n_B se pueden presentar en forma simple. La figura 1, presenta las soluciones para cuatro niveles de costo y tres razones de varianza.

En este caso especial, cuando se usa la varianza en ambos marcos de manera óptima, se puede comparar el caso en que toda la muestra se elige del marco completo. La comparación se presenta en la Figura 2; la disponibilidad económica es igual en ambos casos. Es claro que la ventaja de usar ambos marcos es importante cuando los costos relativos difieren significativamente.

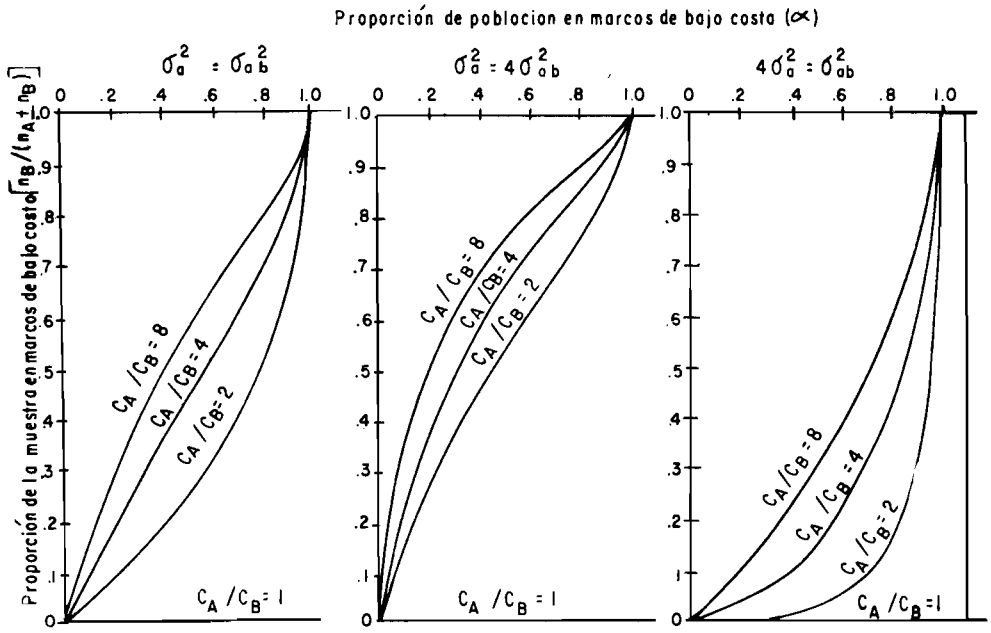


Fig. 1. Distribución óptima de la Muestra entre los dos Marcos para costos relativos diferentes y el Caso Especial de Cubrimiento Completo mediante un Marco Costoso: $\beta = 1$ y marco A más costoso que marco B ($C_A/C_B \geq 1$).

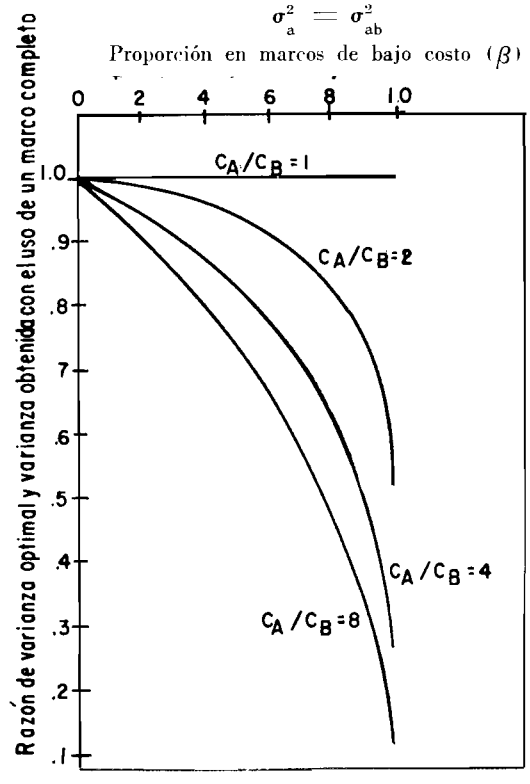


Fig. 2 Reducción en Varianza por el Uso de la Distribución óptima en lugar del uso de un solo Marco Completo para Varios Costos Relativos y el Caso Especial de Cubrimiento Completo mediante un Marco Costoso: $\beta = 1$ y marco A más costoso que marco B $C_A/C_B \geq 1$).

Caso en que N_a, N_{ab} y N_b *no son conocidos*

Cuando N_a, N_{ab} y N_b no son conocidos, es necesario estimarlos. Los estimadores insesgados de N_a y N_b son:

$N_A\,(n_a/n_A)$ y $N_B\,(n_b/n_B)$ respectivamente. De esta manera, dos estimadores insesgados de Nab son: $N_A\,(n'_{ab}/n_A)$ y $N_B\,(n'_{ab}/n_B)$

Usando estos estimadores y ponderando los dos valores de Nab por p y (1-p), obtenemos un estimador insesgado del total de la población.

$$\hat{Y} = \frac{N_A}{n_A}\,n_a\bar{y}_a + \left\{\frac{N_A}{n_A}\,n'_{ab}p + \frac{N_B}{n_B}\,n''_{ab}(1-p)\right\}\bar{y}_{ab} + \frac{N_B}{n_B}\,n_b\bar{y}_b \quad (10)$$

donde

$$\bar{y}_{ab} = \frac{n'_{ab}\,\bar{y}'_{ab} + n''_{ab}\,\bar{y}''_{ab}}{n'_{ab} + n''_{ab}}$$

La varianza del estimador es:

$$\mathrm{Var}\,(\hat{Y}) = \frac{N_A^2(1-\alpha)}{n_A}\,\sigma_a^2 + \frac{N_A N_B\,\alpha\beta}{(n_A+n_B)}\,\sigma_{ab}^2 + \frac{N_B^2(1-\beta)}{n_B}\,\sigma_b^2 +$$

$$\frac{N_A^2\,(1-\alpha)\,\alpha}{n_A}\left\{\bar{Y}_a - p\bar{Y}_{ab}\right\}^2 + \frac{N_B^2\,(1-\beta)}{n_B}\left\{\bar{Y}_b - (1-p)\,\bar{Y}_{ab}\right\}^2 \quad (11)$$

En esta última expresión, existe una aproximación que no es importante aun para muestras pequeñas.

Se puede ver que el desconocimiento de los tamaños de los estratos produce un aumento en la varianza. El aumento de la varianza es importante a menos que la sobreposición de los marcos sea despreciable o prácticamente completa.

Es claro que para cualquier valor de p, independientemente de n'_{ab} y n''_{ab}, el estimador (10) es insesgado. Sin embargo, la varianza del estimador varía de acuerdo a los valores seleccionados de p. El valor de p que reduce la varianza al mínimo es:

$$p = \frac{\dfrac{N_A(1-\alpha)}{n_A}\,\bar{Y}_a + \dfrac{N_B\,(1-\beta)}{n_B}\,(\bar{Y}_{ab} - \bar{Y}_b)}{\left\{\dfrac{N_A(1-\alpha)}{n_A} + \dfrac{N_B\,(1-\beta)}{n_B}\right\}\bar{Y}_{ab}}$$

En esta expresión desde luego los valores α, β, $\bar{Y}_a$, $\bar{Y}_{ab}$ y $\bar{Y}_b$ no están determinados. Por lo tanto, en la práctica, es necesario hacer una estimación de p utilizando los datos derivados de la muestra como sigue:

$$p = \frac{\frac{n''_{ab}\, n_a}{n_A}\, \overline{y}_a + \frac{n'_{ab}\, n_b}{n_B}\, (\overline{y}_b - \overline{y}_{ab})}{(\frac{n_{ab}\, n_a}{n_A} + \frac{n_{ab}\, n_b}{n_B})\, \overline{y}_{ab}} \qquad (13)$$

Esta estimación no es más que una función de n_{ab} y n''_{ab}, por lo que produce en (10) un sesgo pequeño, en el sentido matemático, pero sin consecuencias en la práctica.

Se puede minimizar la varianza (11) con respecto a n_A y n_B sujeto a la función de costos (7) y encontrar la distribución de la muestra entre los dos marcos. La solución se obtiene, como antes, mediante un proceso iterativo considerando:

$$r_1 = \sqrt{\frac{c_B}{c_A}\, (\frac{\beta}{\alpha})} \qquad (14)$$

$$r^2_{i+1} = \frac{c_B}{c_A}\, (\frac{\beta}{\alpha})^2$$

$$\frac{(1-\alpha)\, \sigma^2_a + \frac{r_i^2\, \alpha\, \sigma^2_{ab}}{r_i + (\frac{\beta}{\alpha})^2} + \frac{r_i^2\, \alpha\, (1-\alpha)\, (\overline{Y}_a + \overline{Y}_b - \overline{Y}_{ab})^2}{\left\{ r_i + \frac{\beta}{\alpha} - \left(\frac{1-\alpha}{1-\beta}\right) \right\}^2}}{(1-\beta)\, \sigma^2_b + \frac{(\frac{\beta}{\alpha})^2 \beta\, \sigma^2_{ab}}{(r_i + \frac{\beta}{\alpha})^2} + \frac{\left\{ \frac{\beta\,(1-\alpha)}{\alpha\,(1-\beta)} \right\}^2 \beta\,(1-\beta)\, (\overline{Y}_a + \overline{Y}_b - \overline{Y}_{ab})^2}{\left\{ r_i + \frac{\beta}{\alpha} \left(\frac{1-\alpha}{1-\beta}\right) \right\}^2}} \qquad (15)$$

Como antes, la distribución se expresa en términos de $r = n_A/n_B$. La práctica indica que se necesitan muy pocas iteraciones en la mayoría de los casos.

Es claro que se necesita una aproximación de la proporción de elementos duplicados y de los parámetros de la población para determinar la distribución óptima. Una investigación numérica indicó que la sensibilidad de la varianza de $\hat{Y}$, con respecto a desviaciones de la distribución óptima, es pequeña. Lo mismo ocurre para desviaciones del valor óptimo de p. La Figura 3 presenta las distribuciones óptimas para los casos especiales de

$$\alpha = \beta,\ \sigma^2_a = \sigma^2_{ab} = \sigma^2_b \text{ y } \overline{Y}_a = \overline{Y}_{ab} = \overline{Y}_b$$

para tres valores de los coeficientes de variación.

El estimador (10) no es el que se presenta en la literatura, sugerido por Hartley (1962). Las investigaciones realizadas por el estimador (11) es más eficiente significativamente en varios casos prácticos que el común descrito por Hartley. Además (11), es más fácil de usar.

Ejemplos

Supongamos que se va a realizar una encuesta para medir el consumo diario de leche en una ciudad grande en México. La unidad de muestreo seleccionada es la familia. Entre los varios marcos posibles se decidió usar los siguientes:

1. Directorio telefónico.
2. Registro de propiedades existentes en la Secretaría de Hacienda.

Las investigaciones preliminares indicaron que el segundo marco es suficientemente completo para usarse satisfactoriamente. Sin embargo, el uso del primero tiene la ventaja en costos reducidos por la razón de que se pueden obtener los datos por teléfono. Se supone que es más factible obtener todos los datos del segundo marco pues se realizan entrevistas personales. También, las investigaciones iniciales sugirieron que el directorio de teléfonos cubre el 40% de las familias y el costo relativo de entrevistas personales es 8 veces el costo de entrevistas telefónicas. Se supone que las varianzas son iguales en ambas partes del universo. Usando estos datos se puede ver en la Figura 1 que la distribución óptima de la muestra es asignada aproximadamente el 50% de la muestra a cada marco. Si el valor de costo relativo es 2, la asignación óptima es de 20% en el primer marco.

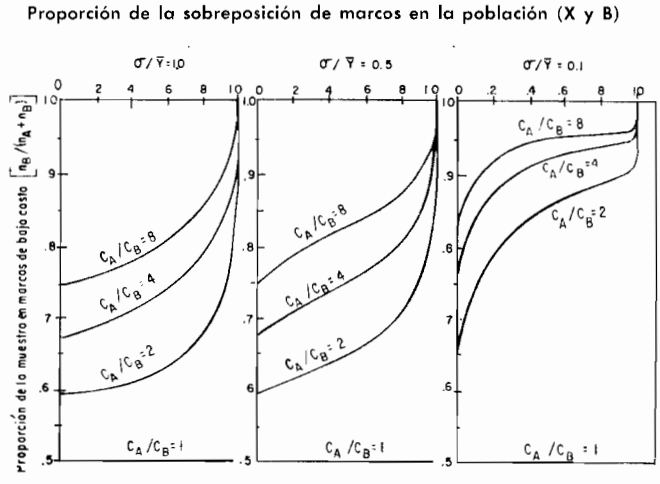


Fig. 3 Distribución óptima de la muestra entre los dos Marcos para Costos Relativos diferentes y el Caso Especial de Marcos de Tamaños Iguales, Varianzas Iguales, Medias Iguales aproximadamente y Tamaños (Na, Nab, Nb) no conocidos: Na, Nab y Nb no conocidas, $\sigma^2_a = \sigma^2_{ab} = \sigma^2_b$, $\overline{Y}_a = \overline{Y}_{ab} = \overline{Y}_a$ y $\propto = \beta$ y marco A más costoso que marco B ($C_A/C_B \geq 1$).

Como un ejemplo de condiciones diferentes, supongamos que una encuesta semejante se plantea para una ciudad de los Estados Unidos. El investigador decide que el uso conjunto de dos marcos semejantes a los descritos anteriormente es adecuado, pero en este caso el directorio de teléfonos cubre alrededor del 90% del universo de interés y también la lista basada en archivos de impuestos cubre una proporción igual. Es decir, cada marco cubre aproximadamente el 90% del universo en total

mientras que ambos proveen un cubrimiento completo del universo. Los tamaños N_a, N_{ab} y N_b no son conocidos. Suponiendo que el uso de directorio telefónico tiene una ventaja en costos del 50% sobre el segundo marco, que las varianzas y las medias son iguales aproximadamente en todos los sectores y el coeficiente de variación tiene un valor de 0.5, la distribución óptima se muestra en la Figura 3, donde se indica que el 75% de la muestra debe ser asignada al primer marco (directorio telefónico).

Es claro que muchas veces la situación no está de acuerdo con las situaciones descritas en las Figuras 1 ó 3. Así, es necesario determinar la distribución óptima mediante las expresiones (8) o (14).

Bibliografía

1. Cochran, R. S. "Multiple frame survey". En: American Statistical Association, Proceedings of the Social Statistics Section, 1964.
2. Hartley, H. O. "Multiple frame surveys". En: American Statistical Association, Proceedings of the Social Statistics Section, 1962.
3. Kish, Lesle. Survey Sampling. New York. John Wiley and Sons, Inc., 1967.