Nº 9801

AJUSTE DE MUESTRAS CON INFORMACIÓN AUXILIAR

E. Bueno, A. Zarraga y A. Iztueta

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Duque de Wellington, 2
01010 VITORIA-GASTEIZ
Tel.: 945 18 75 00
Fax: 945 18 75 01
E-mail: eustat@eustat.es
www.eustat.es

AJUSTE DE MUESTRAS CON INFORMACIÓN AUXILIAR

Elena BUENO

Becaria de colaboración en EUSTAT

Ainhoa ZARRAGA

Becaria de colaboración en EUSTAT

Anjeles IZTUETA

Responsable de Metodología en EUSTAT

RESUMEN

La fase de ajuste y elevación de muestras tiene por objetivo pasar de la muestra a la población, utilizando para ello toda la información disponible y haciendo compatibles los resultados entre diferentes encuestas o fuentes.

En el cuaderno se describe la metodología utilizada para el ajuste estadístico de muestras, diferenciando dos grupos de metodologías, según la cantidad de información auxiliar disponible:

- Métodos con Máxima Información Auxiliar, utilizados cuando se dispone de la distribución auxiliar multivariante.

- Procedimientos Iterativos, utilizados en el caso de que la información auxiliar disponible sea la univariante.

En la *introducción* se presentan distintos tipos de ajustes, diferenciándolos el tipo de información auxiliar de que se disponga, así como una notación básica utilizada posteriormente en los próximos apartados.

El *segundo apartado* se centra en el estudio de los métodos utilizados con Máxima Información Auxiliar, presentándose primeramente una notación particularizada y más detallada al tipo de situación en el que se utilizan, así como la expresión matemática de dichos estimadores. Seguidamente se realiza un análisis de las propiedades estadísticas de los susodichos estimadores. Para ilustrar el funcionamiento de la metodología empleada y un claro entendimiento de la misma, se presentan unos sencillos ejemplos. Tras una pequeña referencia a los procedimientos informáticos utilizados para implementar estos métodos de ajuste, se mencionan las condiciones óptimas para su aplicación.

En el *tercer apartado* se presentan dos procedimientos iterativos como métodos de ajuste, para los casos en los que la distribución auxiliar que se dispone es la univariante: el *Raking* y una variante del mismo, *RAS*, utilizada en el procedimiento Redre del SPAD. Se presenta primeramente la notación a utilizar y a continuación se realiza una descripción de ambos procedimientos iterativos ilustrando sus etapas para un caso

sencillo. Seguidamente, se aplican ambos métodos a unos ejemplos, también simulados. Finalmente se hace referencia a distintos procedimientos informáticos para la aplicación de ambas metodologías, así como las condiciones óptimas para su aplicación.

En el *cuarto apartado* se presenta una pequeña conclusión y comparación de los dos grupos de métodos de ajuste analizados en el cuaderno.

PALABRAS CLAVE: Ajuste de muestras, Distribución Auxiliar Multivariante, Distribución Auxiliar Univariante, procedimientos iterativos, Raking, pesos, elevadores, estratificación, post-estratificación, variables auxiliares cualitativas y cuantitativa.

Agradecimientos a: Jaime Garrido, Juan José Ortiz y Yolanda Pérez.Indice

Indice

Capítulo

1

Introducción

Introducción y objetivos

En este cuaderno vamos a abordar un tema muy común en la estadística:

Tenemos una población o universo de tamaño N, sobre la que queremos estimar el total de una cierta característica Y. Al ser imposible, en general, encuestar a todos los elementos de una población, lo que se hace es tomar una muestra de tamaño n de dicha población, y se estudia el comportamiento de dicha característica en estos individuos. Pero evidentemente, el total de Y sobre los elementos observados no va a ser el mismo que su total poblacional y para conseguir obtener una estimación de dicho total, considerando las observaciones muestrales obtenidas para Y, se utilizarán unos elevadores. Con esta elevación conseguiremos una primera estimación del total poblacional de la variable Y.

La elevación también es usada para compensar la pérdida de muestra debida a la no-respuesta. Esta última compensación se puede realizar tanto en censos como en encuestas muestrales, pero se ha de hacer siempre que los respondientes no tengan un perfil específico y diferenciado.

En general, además de la *variable objetivo* Y, dispondremos de otras características observadas en la población y en la muestra, denominadas *variables auxiliares*, suponiendo una cierta relación entre la variable objetivo y las auxiliares. Por eso muchas veces denominaremos a la variable Y variable *objetivo o dependiente*. Gracias a esta suposición, modificaremos las elevaciones o pesos iniciales de los individuos muestrales con el fin de ajustar a la distribución poblacional conocida, las variables auxiliares.

Entre los métodos de ajuste y elevación empleados en la muestra, unos utilizan información auxiliar máxima y otros información auxiliar marginal. Analizaremos los métodos utilizados en la actualidad en EUSTAT para ambas situaciones: introduciremos teóricamente los métodos entre todos los posibles, describiremos estos procedimientos de ajuste con la notación adecuada y seguidamente los aplicaremos a un ejemplo sencillo. Se presentarán también distintos software para su aplicación, así como las ventajas y desventajas de cada uno de ellos.

Clasificación de métodos de ajuste

Existen distintos procedimientos para el ajuste, según qué tipo de variables auxiliares, así como la cantidad de información auxiliar de que se dispone. (Ver [1] y [2])

Tipos de Variables Auxiliares

Se pueden diferenciar dos bloques de variables: (Ver [10])

- *Variables Auxiliares Cualitativas*: son variables con un número de modalidades finito. Como ejemplo de este tipo de variables, tenemos el sexo, la relación con el mercado de trabajo, con modalidades como estar ocupado, parado, inactivo,...

- *Variables Auxiliares Cuantitativas*: son variables continuas que toman valores en un intervalo y que, por lo tanto, estos valores pueden ser infinitos. Como ejemplo de este tipo de variables está el total de facturación, ...

Cantidad de Información Disponible

La siguiente clasificación vendrá determinada por la información auxiliar disponible. En todas las encuestas existe una estratificación de la población en celdas o estratos, utilizando para ello el cruce multivariante de variables cualitativas. Se pueden dar dos supuestos:

- cuando las variables auxiliares disponibles son *cualitativas*, la información que se define en los estratos es el número de efectivos de cada uno de estos. En el caso de poseer *Máxima Información Auxiliar* se dispone del número de efectivos poblacionales de cada una de las celdas de la estratificación. En el caso de poseer *No-máxima Información Auxiliar* lo que se dispone es sólo la distribución marginal univariante de las variables auxiliares. (Ver [6] y [7]).

- cuando las variables auxiliares son *cuantitativas*, la información es referente a la distribución de las variables auxiliares a lo largo de los estratos obtenidos. En el caso de *Máxima Información Auxiliar*, la información es el total, media,... poblacionales de las variables auxiliares sobre cada uno de los estratos y en el caso de *No-máxima Información Auxiliar*, la información que se tiene es el total poblacional, media,... de la variable auxiliar en cada una de las modalidades de las variables auxiliares de estratificación univariante.

Métodos de Estimación

Se parte de unos estimadores iniciales de los estadísticos objetivo de estudio, sin considerar la información auxiliar: (ver [15])

- *Método de Horvitz-Thompson*

Se obtienen los estimadores de los estadísticos objetivo de estudio tomando como pesos los inversos de las probabilidades de inclusión en la muestra para cada individuo muestral. Son, por lo tanto, pesos asignados a individuos.

- *Método corregido de Horvitz-Thompson*

  Es utilizado con el fin de corregir la no-respuesta. El peso utilizado es el inverso del producto de la probabilidad de selección y la tasa de no-respuesta, considerada en estratos homogéneos de no-respuesta. En este caso, la ponderación es por celdas, siendo necesaria igual probabilidad de inclusión en la muestra para elementos de una misma celda.

Los métodos de estimación que usan información auxiliar, se pueden agrupar dentro del modelo general de *regresión múltiple*.

MODELO DE REGRESIÓN MÚLTIPLE: supone una relación entre la variable objetivo Y y las variables auxiliares. El modelo de regresión generalmente supuesto entre estas variables es el modelo de regresión lineal. Este estimador de regresión es uno de los procedimientos de estimación más generales, en el que se pueden utilizar varias variables auxiliares y además éstas pueden ser tanto continuas como nominales.

A continuación vamos a mencionar algunos casos particulares del modelo de regresión: (ver [2], [7], [8] y [9])

- El estimador de *estratificación y post-estratificación* o, también denominado, estimador usual de la RAZÓN: se obtiene al tomar las variables auxiliares dicotómicas, asociadas a cada uno de los estratos resultantes de la estratificación o post-estratificación, es decir, coincide con el estimador de regresión adaptado a un modelo de Análisis de la Varianza. Los elevadores resultantes son también por celdas y son el cociente entre el total poblacional y muestral, en efectivos, en cada una de las celdas de la estratificación.

- El estimador *de la Razón*: supone tomar un modelo de regresión lineal en el que el parámetro independiente es nulo, es decir, se toma una recta de regresión que pasa por el origen. El elevador que se obtiene es por celdas y es la razón entre el total poblacional y el total muestral de una variable cuantitativa en dicha celda.

- Los estimadores de *Raking*: es un método que se resuelve de forma iterativa. Nos encontramos con dos tipos de métodos Raking:

  1. La información auxiliar es cualitativa y los datos de que disponemos son los efectivos marginales. Los elevadores de celdas resultantes son los cocientes entre los efectivos marginales poblacionales y los marginales aproximados obtenidos tras las iteraciones. Este es el Raking usual. Como modelo de regresión es un caso particular del estimador usual de la Razón, en el que se toman las variables auxiliares de forma univariante.

  2. La información auxiliar es cuantitativa. Los datos en este caso, son totales marginales de las variables cuantitativas en la estratificación. Como modelo de regresión es un caso particular del estimador de la Razón.

- *MIXTOS*: Existen variaciones de los dos métodos anteriores, métodos que son una combinación de los anteriores. Entre estos, mencionaremos un método denominado *estimador modificado del Raking Ratio*.

Notación básica

Se considera que tenemos una población U y una muestra de la misma S:

U={1,2,...,N}, con k=1,...,N

S={1,2,..,n}, con k=1,...,n

Definimos la variable objetivo Y como un vector Nx1:

$$\mathbf{Y}=\begin{pmatrix} Y(1) \\ \dots \\ Y(k) \\ \dots \\ Y(N) \end{pmatrix}$$

El total poblacional a estimar es:

$$Y=\sum_{k=1}^{N} Y(k)=\sum_{k=1}^{N} y_k \qquad (1)$$

A cada individuo se le asigna un peso inicial

$z_k = 1/p_k$ con k=1,..,n

p_k con k=1,..,n es la probabilidad de selección del individuo muestral k en S, en el caso de un muestreo probabilista.

Llamaremos al peso final:

w_k con k=1,...,n.

Estimador de HORVITZ-THOMPSON del total poblacional de la variable objetivo Y: (ver [4] y [8])

$$\hat{Y}_p = \sum_{k=1}^{n} z_k \bullet y_k \qquad (2)$$

Es *insesgado* respecto del total poblacional de la variable Y, i.e.,

$$E\left(\hat{Y}_p\right) = Y$$

Su *varianza* es la siguiente:

$$Var\left(\hat{Y}_p\right) = \sum_{k,l=1}^{N} \left(p_{kl} - p_k p_l\right) \bullet \frac{y_k}{p_k} \bullet \frac{y_l}{p_l},$$

siendo p_{kl} la probabilidad de selección del elemento k y l simultáneamente.

La estimación de esta varianza es

$$\hat{Var}\left(\hat{Y}_p\right) = \sum_{k,l=1}^{n} \left(p_{kl} - p_k p_l\right) \bullet \frac{y_k}{p_k} \bullet \frac{y_l}{p_l}$$

En el caso particular de que en el muestreo se de igual probabilidad de selección para todos los individuos, tenemos que estas probabilidades de selección son:

$$p_k = \frac{n}{N} \quad \forall k = 1,..,n \qquad \text{y} \qquad p_{kl} = \frac{n}{N} \bullet \frac{n-1}{N-1} \quad \forall k,l = 1,..,n$$

En el muestreo aleatorio simple se cumple esta condición y las expresiones anteriores toman la forma:

$$\hat{Y}_p = \sum_{k=1}^{n} \frac{N}{n} \bullet y_k \tag{3}$$

$$Var\left(\hat{Y}_p\right) = N^2 \left(1 - f\right) * \frac{S^2}{n} \tag{4}$$

$$donde\, S^2 = \frac{\sum_{k=1}^{N} \left(y_k - \bar{Y}\right)}{N-1} \quad \text{y} \quad \bar{Y} = \sum_{k=1}^{N} \frac{y_k}{N}$$

$$\hat{Var}\left(\hat{Y}_p\right) = N^2 \left(1 - f\right) * \frac{s^2}{n} \tag{5}$$

$$donde\, s^2 = \frac{\sum_{k=1}^{n} \left(y_k - \bar{y}\right)}{n-1} \quad \text{y} \quad \bar{y} = \sum_{k=1}^{n} \frac{y_k}{n}$$

La notación respecto a las variables auxiliares la introduciremos ya en las siguientes secciones, particularizando en qué tipo de problema estemos, respecto de dicha información auxiliar.

Capítulo

2

Métodos con Máxima Información Auxiliar

Los métodos que a continuación vamos a analizar se pueden utilizar tan sólo cuando se dispone de MÁXIMA información auxiliar, esto es, conocemos los totales de las variables auxiliares sobre cada una de las celdas de la estratificación.

Notación

Se plantean una serie de restricciones con el fin de que la distribución muestral ponderada sea igual a la distribución poblacional conjunta. A continuación vamos a introducir la notación para representar estas restricciones, diferenciando el caso, respecto del tipo de información auxiliar, en el que estemos.

Variables Auxiliares Cualitativas

La expresión matricial de las restricciones es: (ver [16])

$$\hat{X} \bullet \mathbf{w} = \mathbf{X} \bullet \mathbf{I}, \qquad (6)$$

donde el vector de pesos **w'**=(w_1 ,.., w_k ,.., w_n) y el vector **I** es un vector Nx1 de 1s:

$$\mathbf{I}=\begin{pmatrix} 1 \\ .. \\ 1 \\ .. \\ 1 \end{pmatrix}$$

Las matrices $\hat{X}$ y **X** son las matrices muestral y poblacional de dimensión (m+1)xn y (m+1)xN, respectivamente. Para la definición de las componentes de estas matrices, definimos una serie de variables:

Tenemos L variables auxiliares cualitativas, cada una con $L_1, L_2, ..., L_l, ..., L_L$ modalidades. Se realiza la estratificación de la población mediante el cruce multivariante de estas L variables, obteniendo $L_1 \bullet L_2 \bullet .. \bullet L_l \bullet .. \bullet L_L = m$ estratos.

Definimos las variables identificadoras o dicotómicas de los estratos:

$$X_{h_1..h_l..h_L}(k) = \begin{cases} 1, si\ k \in estrato(h_1,..,h_l,..h_L) \\ 0, en\ caso\ contrario \end{cases}$$

$$\Rightarrow L_1 \bullet .. \bullet L_l \bullet .. \bullet L_L = m$$

siendo m el número de variables dicotómicas.

Tenemos un vector poblacional fila

$X_{h_1..h_l..h_L} = \left(X_{h_1..h_l..h_L}(1),..,X_{h_1..h_l..h_L}(k),..,X_{h_1..h_l..h_L}(N)\right)$ de dimensión (1xN)

y un vector muestral fila

$\hat{X}_{h_1..h_l..h_L} = \left(X_{h_1..h_l..h_L}(1),..,X_{h_1..h_l..h_L}(k),..,X_{h_1..h_l..h_L}(n)\right)$ de dimensión (1xn)

Para asegurar que:

$$w_1 + .. + w_k + ..w_n = N$$

tomamos la variable X_0 idénticamente 1 y se representa mediante:

$X_0 = \left(X_0(1),..,X_0(k),..,X_0(N)\right) = \left(1,..,1,..,1\right)$ vector fila poblacional de dimensión (1xN)

$\hat{X}_0 = \left(X_0(1),..,X_0(k),..,X_0(n)\right) = \left(1,..,1,..,1\right)$ vector fila muestral de dimensión (1xn)

Las matrices **X** y $\hat{X}$ están formadas por estos vectores fila:

$$\mathbf{X}=\begin{pmatrix} X_0 \\ X_1 \\ \cdots \\ X_h \\ \cdots \\ X_m \end{pmatrix} \quad \text{y} \quad \hat{X} = \begin{pmatrix} \hat{X}_0 \\ \hat{X}_1 \\ \cdots \\ \hat{X}_h \\ \cdots \\ \hat{X}_m \end{pmatrix} \qquad [1]$$

Desarrollando la expresión matricial, obtenemos:

$$\begin{cases} \sum_{k=1}^{n} w_k \bullet X_0(k) = \sum_{k=1}^{N} X_0(k) \Leftrightarrow w_1 + .. + w_k + ..w_n = N \\ \sum_{k=1}^{n} w_k \bullet X_{h_1..h_l..hL}(k) = \sum_{k=1}^{N} X_{h_1..h_l..h_L}(k), \forall 1 \le h_1 \le L_1,..,1 \le h_l \le L_l,..,1 \le h_L \le L_L \end{cases}$$

Al ser las variables identificadoras de los estratos, los datos poblacionales toman la forma:

[1] Los indicadores de las variables se han simplificado a un solo índice, sin más que establecer un orden en las variables.

$$\begin{cases} \sum_{k=1}^{N} X_{h1..hl..hL}(k) = N_{h1..hl..hL} \\ \sum_{k=1}^{n} X_{h1..hl...hL}(k) = n_{h1..hl..hL} \end{cases} \quad (7)$$

Resolvemos el sistema, teniendo en cuenta que es un ajuste de celdas y por lo tanto, todos los individuos de un mismo estrato tienen el mismo peso, por lo que el sistema pasa a tener como incógnitas $W_{h1..hl..hL}$: elevadores asignados a las celdas. El sistema es el siguiente:

$$\begin{cases} W_{h1..hl..hL} \bullet n_{h1..hl..hL} = N_{h1..hl..hL} \quad \forall\, 1 \le h_1 \le L_1,..,1 \le h_l \le L_l,..,1 \le h_L \le L_L \\ \sum_{h1..hl..hL} W_{h1..hl..hL} \bullet n_{h1..hl..hL} = N \end{cases} \quad (8)$$

Es un sistema de $L_1 \bullet L_2 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_L = m$ incógnitas y $L_1 \bullet L_2 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_L + 1$ ecuaciones, en el que la última es combinación lineal de las anteriores, por lo que se reduce a un sistema mxm, en el que la matriz asociada es una matriz $A \in (mxm)$ diagonal. Los elementos diagonales de esta matriz son el número de efectivos muestrales en cada celda de la estratificación.

Tras lo anterior, el sistema tiene

Una única solución $\Leftrightarrow n_{h1..hl..hL} \neq 0 \ \forall\, h_1..h_l..h_L$

Bajo estas condiciones, la solución del sistema es:

$$\hat{W}_{h1..hl..hL} = \frac{N_{h1..hl..hL}}{\hat{n}_{h1..hl..hL}}, \quad (9)$$

$$\forall 1 \le h_1 \le L_{1,..},1 \le h_l \le L_{l,..},1 \le h_L \le L_L$$

La estimación resultante tras utilizar estos elevadores es estimación de Estratificación, en la que el muestreo se supone aleatorio estratificado. Esta estimación resulta del ajuste a los datos poblacionales que teníamos y al ajuste en tales condiciones le llamamos ajuste de *ESTRATIFICACIÓN*. Un caso particular de este tipo de ajuste es cuando la estratificación de la población no se realiza antes de la muestra, sino tras la muestra. En este último ajuste la estimación que se obtiene se denomina estimación de Post-estratificación y el método de ajuste método de *POST-ESTRATIFICACIÓN*. En ambos, los elevadores de las celdas son el cociente entre el número de efectivos poblacionales y muestrales para cada una de ellas. (Ver [4], [6] y[14])

Variables Auxiliares Cuantitativas

En este caso, supondremos que tenemos una única variable auxiliar cuantitativa X. La estratificación de la población se realiza mediante unas variables premuestrales: pre-estratificación.

Las restricciones toman la misma forma matricial, con unas nuevas matrices, que se definirán a continuación:

$$\left(\hat{X}^{*}\right)\bullet \mathbf{w} = \left(X^{*}\right)\bullet \mathbf{I}, \qquad (10)$$

donde, como antes, el vector de pesos es **w'**=(w_1 ,.., w_k ,.., w_n) y el vector **I** es un vector Nx1 de 1s:

$$\mathbf{I}=\begin{pmatrix}1\\ ..\\ 1\\ ..\\ 1\end{pmatrix}$$

Para definir estas matrices X^{*} y $\hat{X}^{*}$ definimos las variables auxiliares:

$$X^{*}_{h1..hl..hL}(k) = X_{h1..hl..hL}(k)\bullet X(k),\ \forall 1\le h_1\le L_{1,..,}1\le h_l\le L_{l,..,}1\le h_L\le L_L,$$

con $X_{h1..hl..hL}(k)$ la variable dicotómica identificadora de cada uno de los estratos resultantes y

$X(k)$ es el valor de la variable cuantitativa para cada uno de los elementos poblacionales.

Definimos también:

$$X^{*}_{0}(k) = X_0(k)\bullet X(k) = X(k)$$

Obtenemos los vectores:

$X^{*}_{h1..hl..hL} = \left(X^{*}_{h1..hl..hL}(1),..,X^{*}_{h1..hl..hL}(k),..,X^{*}_{h1..hl..hL}(N)\right)$ vector fila de dimensión (1xN)

$\hat{X}^{*}_{h1..hl..hL} = \left(\hat{X}^{*}_{h1..hl..hL}(1),..,\hat{X}^{*}_{h1..hl..hL}(k),..,\hat{X}^{*}_{h1..hl..hL}(n)\right)$ el vector muestral fila de dimensión (1xn).

Estos vectores son las componentes filas de las matrices X^{*} y $\hat{X}^{*}$. La forma que toman es:

$$X^* = \begin{pmatrix} X_0^* \\ X_1^* \\ \dots \\ X_h^* \\ \dots \\ X_m^* \end{pmatrix} \quad \mathbf{y} \quad \hat{X}^* = \begin{pmatrix} \hat{X}_0^* \\ \hat{X}_1^* \\ \dots \\ \hat{X}_h^* \\ \dots \\ \hat{X}_m^* \end{pmatrix}$$

Desarrollamos la expresión matricial de las restricciones y obtenemos el sistema:

$$\begin{cases} \sum_{k=1}^{n} w_k \bullet X^*_{h_1..h_l..h_L}(k) = \sum_{k=1}^{N} X^*_{h_1..h_l..h_L}(k) \quad \forall 1 \le h_1 \le L_1,..,1 \le h_l \le L_l,..,1 \le h_L \le L_L \\ \sum_{k=1}^{n} w_k \bullet X^*_0(k) = \sum_{k=1}^{N} X^*_0(k) \Leftrightarrow w_1 + .. + w_k + .. + w_n = \sum_{k=1}^{N} X(k) \end{cases}$$

.Es un sistema de $L_1 \bullet L_2 \bullet \dots \bullet L_l \bullet \dots \bullet L_L = m$ incógnitas y $L_1 \bullet L_2 \bullet \dots \bullet L_l \bullet \dots \bullet L_L + 1$ ecuaciones, en el que la última es combinación lineal de las anteriores, por lo que se reduce a un sistema mxm, en el que la matriz asociada es una matriz $A \in (mxm)$ diagonal con elementos diagonales:

$$\sum_{k=1}^{n} X^*_{h_1..h_l..h_L}(k)$$

que serán *no nulos sii*:

la suma de la variable auxiliar X sobre cada estrato es distinta de cero, lo que se puede conseguir pidiendo que la *variable sea >0*

los *efectivos* muestrales en cada estrato son *no nulos*

La primera restricción no influye, ya que normalmente, las variables cuantitativas que se analizan son variables positivas, por ejemplo las económicas: total de facturación, ...

Verificadas estas condiciones de no-singularidad para la matriz del sistema, tenemos que la única solución que se obtiene es:

$$\hat{W}^*_{h_1..h_l..h_L} = \frac{\sum_{k=1}^{N} \hat{X}^*_{h_1..h_l..h_L}(k)}{\sum_{k=1}^{n} \hat{X}^*_{h_1..h_l..h_L}(k)} \qquad (11)$$

Estos $\hat{W}^*_{h_1..h_l..h_L}$ vuelven a ser los elevadores asignados a las celdas de la estratificación, teniendo en cuenta que el ajuste que estamos realizando es por celdas.

El método resultante es el *método de la RAZÓN (RATIO)*, en el que tomamos como factores de elevación de las celdas los cocientes del total poblacional de una variable cuantitativa en dicha celda entre su total muestral.

Hemos visto entonces que en ambos casos, tanto con información auxiliar cualitativa como cuantitativa, el método de ponderación nos da un único elevador para cada estrato, *siempre que* en cada estrato de la estratificación *exista algún elemento muestral*.

Propiedades del estimador de la media

Analicemos las propiedades de los dos estimadores que hemos considerado en el caso de máxima información auxiliar, es decir, para el estimador de *post-estratificación* y el estimador de *la razón*. El estudio lo realizaremos tomando como estadístico la *Media Poblacional* de la variable objetivo.

Variables Auxiliares Cualitativas

El estimador que se obtiene en el muestreo aleatorio estratificado para la media poblacional es:

$$\bar{Y}_{st} = \sum_{h=1}^{m} W_h \bullet \bar{y}_h \text{ , donde } W_h = \frac{N_h}{N}, \quad (12)$$

[2]

siendo $\bar{Y}_h$ la media poblacional sobre el estrato h. Esta estimación se obtiene de forma análoga tanto en el caso de una estratificación pre-muestral como en el de estratificación post-muestral (post-estratificación).

Consideramos en un primer momento que tenemos una *estratificación pre-muestral* y los valores n_h son, por lo tanto, fijos.

Para considerar este estimador necesitamos conocer $W_h = \frac{N_h}{N}$, o, lo que es lo mismo, $N_h \ \ \forall h = 1,..,m$.

Conocidos estos datos, estimamos la media poblacional de Y sobre cada celda:

$\bar{y}_h = \frac{y_h}{n_h}$, donde y_h se considera ahora el total muestral sobre cada celda de la variable objetivo Y.

Sustituimos esta expresión en el estimador y obtenemos la estimación:

[2] Tomamos una numeración de los estratos resultantes del cruce, en la que intervendrá ya un sólo subíndice.

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{m} \frac{N_h}{N} \bullet \frac{y_h}{n_h} \qquad (13)$$

Ésta es la estimación de estratificación de la media poblacional de Y. A partir de ésta se obtiene fácilmente la estimación del total poblacional:

$$\hat{Y}_{st} = \hat{\bar{Y}}_{st} \bullet N$$

Sustituyendo de nuevo en la expresión anterior, obtenemos:

$$\frac{\hat{Y}_{st}}{N} = \sum_{h=1}^{m} \frac{N_h}{N} \bullet \frac{y_h}{n_h} \Leftrightarrow$$

$$\hat{Y}_{st} = \sum_{h=1}^{m} \frac{N_h}{n_h} \bullet y_h \qquad (14)$$

Vamos a estudiar ahora el error que se produce al tomar la estimación de estratificación para la media poblacional:

$$Var(\hat{\bar{Y}}_{st}) = Var(\sum_{h=1}^{m} W_h \bullet \bar{y}_h) = \sum_{h=1}^{m} (W_h)^2 \bullet Var(\bar{y}_h) + 2 \bullet \sum_{h=1}^{m} \sum_{j>h}^{m} Cov(\bar{y}_h, \bar{y}_j)$$

y el error es:

$$\sqrt{Var(\hat{\bar{Y}}_{st})}$$

Al tratarse de un muestreo aleatorio estratificado, tenemos que el término de las covarianzas desaparece, debido a la independencia del muestreo en cada uno de los estratos y la expresión de la varianza toma la forma

$$Var(\hat{\bar{Y}}_{st}) = Var(\sum_{h=1}^{m} W_h \bullet \bar{y}_h) = \sum_{h=1}^{m} (W_h)^2 \bullet Var(\bar{y}_h)$$

El muestreo es *aleatorio* simple en cada uno de los estratos y la varianza en cada uno de ellos es:

$$Var(\bar{y}_h) = \frac{S_h{}^2}{n_h} \bullet \frac{N_h - n_h}{N_h}$$

De esta forma, la expresión final de la varianza es

$$Var(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{m} \frac{W_h{}^2 \bullet S_h{}^2}{n_h} - \frac{\sum_{h=1}^{m} W_h \bullet S_h{}^2}{N}$$

siendo S_h la varianza verdadera de la variable Y sobre el estrato h.

Estimamos la varianza en cada estrato por

$$\hat{V}ar(\bar{y}_h) = \frac{s_h^2}{n_h} \bullet \frac{N_h - n_h}{N_h}$$

con s_h la varianza muestral de Y sobre el estrato h. Y a continuación obtenemos una estimación de la varianza del estimador de estratificación para la media poblacional:

$$\hat{V}ar(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{m} \frac{W_h^2 \bullet s_h^2}{n_h} - \frac{1}{N} \bullet \sum_{h=1}^{m} W_h \bullet s_h^2 \qquad (15)$$

En la expresión de la estimación del error se ve, claramente, que el error de la estimación depende de la varianza de la variable objetivo Y en cada estrato, por lo que si estos *estratos son homogéneos respecto de la variable objetivo*, la *varianza total será pequeña*, siempre que todos los n_h *sean lo suficientemente grandes*. De esta forma se podrá controlar relativamente el error.

El control de la varianza de la variable objetivo sobre las celdas se puede conseguir tomando las variables de estratificación lo más altamente correlacionadas con la objetivo. Los estratos resultantes mediante el cruce multivariante de las variables serán de esta forma homogéneos respecto de la variable objetivo.

De forma análoga consideramos el estimador de *post-estratificación*. En este caso los valores n_h son aleatorios, dependen de la muestra, y los denotamos como $\hat{n}_h$. Una vez tomada la muestra, estos valores $\hat{n}_h$, que exceden todos a cero, son fijos.

El estimador que obtenemos de la media poblacional toma forma análoga al estimador de estratificación:

$$\bar{Y}_W = \sum_{h=1}^{m} W_h \bullet \bar{Y}_h$$

En cada muestra, tomamos la media muestral en cada estrato $\bar{y}_h = \frac{y_h}{\hat{n}_h}$ y obtenemos el estimador de post-estratificación:

$$\hat{\bar{Y}}_W = \sum_{h=1}^{m} W_h \bullet \bar{y}_h = \sum_{h=1}^{m} W_h \bullet \frac{y_h}{\hat{n}_h} = \sum_{h=1}^{m} \frac{N_h}{N} \bullet \frac{y_h}{\hat{n}_h} = \frac{1}{N} \bullet \sum_{h=1}^{m} \frac{N_h}{\hat{n}_h} \bullet y_h$$

Una vez realizada la muestra, como ya hemos dicho, estos $\hat{n}_h$ son fijos, por lo que la estimación de la varianza para este estimador es igual que para el estimador de estratificación y la expresión que tenemos es:

$$Var(\hat{\bar{Y}}_W) = Var(\sum_{h=1}^{m} W_h \bullet \bar{y}_h) = \sum_{h=1}^{m} (W_h)^2 \bullet Var(\bar{y}_h) = \sum_{h=1}^{m} \frac{W_h^2 \bullet S_h^2}{\hat{n}_h} - \frac{1}{N} \bullet \sum_{h=1}^{m} W_h \bullet S_h^2$$

Ahora bien, esta varianza no es fija, desde el momento que los $\hat{n}_h$ no son fijos y debemos hallar por lo tanto un valor promedio de dicha varianza. Este valor promedio se halla sin más que tomar la esperanza:

$$E\left[Var(\hat{\bar{Y}}_W)\right] = E\left[\sum_{h=1}^{m} \frac{W_h{}^2 \bullet S_h{}^2}{\hat{n}_h} - \frac{1}{N} \bullet \sum_{h=1}^{m} W_h \bullet S_h{}^2\right] =$$

$$= \sum_{h=1}^{m} W_h{}^2 \bullet S_h{}^2 \bullet E\left(\frac{1}{\hat{n}_h}\right) - \frac{1}{N} \bullet \sum_{h=1}^{m} W_h \bullet S_h{}^2$$

Ahora bien $E\left(\frac{1}{\hat{n}_h}\right) = \frac{1}{n \bullet W_h} + \frac{1 - W_h}{n^2 \bullet W_h{}^2}$, con lo que la expresión anterior se reduce a:

$$E\left[Var(\hat{\bar{Y}}_W)\right] = \frac{1-f}{n} \sum_{h=1}^{m} W_h \bullet S_h{}^2 + \frac{1}{n^2} \bullet \sum_{h=1}^{m} (1 - W_h) \bullet S_h{}^2$$

El primer término corresponde a la varianza del estimador de estratificación $Var(\bar{Y}_{st})$, siendo el segundo el incremento producido en la varianza al tomar el estimador de post-estratificación. Desarrollando este segundo término obtenemos la expresión:

$$\frac{1}{n^2} \bullet \sum_{h=1}^{m} (1 - W_h) \bullet S_h{}^2 = \frac{1}{n \bullet \bar{n}_h} \bullet \bar{S}_h{}^2 - \frac{1}{n^2} \bullet \sum_{h=1}^{m} W_h \bullet S_h{}^2$$

donde $\bar{n}_h = \frac{n}{m}$ es el número promedio de unidades muestrales por estrato y $\bar{S}_h{}^2$ es el promedio de las $S_h{}^2$. Se obtiene finalmente, que el incremento de la varianza obtenido al tomar el estimador de post-estratificación se mantiene pequeño siempre que $\bar{n}_h$ se mantenga razonablemente grande.

La estimación del promedio de la varianza que se obtiene es:

$$\hat{E}\left[Var(\hat{\bar{Y}}_W)\right] = \frac{1-f}{n} \sum_{h=1}^{m} W_h \bullet s_h{}^2 + \frac{1}{n^2} \bullet \sum_{h=1}^{m} (1 - W_h) \bullet s_h{}^2$$

En este caso de post-estratificación, el control de la varianza de la variable objetivo sobre las celdas se puede conseguir tomando las variables auxiliares cualitativas altamente correlacionadas con la objetivo. Los estratos resultantes mediante el cruce multivariante de las variables auxiliares serán de esta forma homogéneos respecto de la variable objetivo. (Ver [4] y [9]).

Concluimos que el error cometido, tanto con la estimación de estratificación como con la post-estratificación, es directamente proporcional a la varianza de la variable sobre las celdas e inversamente proporcional al tamaño muestral de los estratos.

Variables Auxiliares Cuantitativas

Hay dos tipos de estimaciones de razón en muestreo aleatorio estratificado:

1) Estimación de razón separada

2) Estimación de razón combinada

La que nosotros vamos a tomar es la estimación de razón separada. Este estimador corresponde a un modelo de regresión en el que la pendiente de la recta de regresión no es necesariamente la misma para cada una de las celdas de la estratificación. El estimador de razón combinada, sin embargo, supone la misma pendiente de regresión para todas las celdas. Nosotros consideraremos en este cuaderno el Estimador de la Razón Separada.

El ***Estimador de la Razón Separada*** para la media poblacional es: (ver [4])

$$\overline{Y}_{Rs} = \sum_{h=1}^{m} W_h^* \bullet \overline{Y}_h \qquad (16)$$

siendo $\overline{Y}_h$ la media poblacional sobre el estrato h:

$$\overline{Y}_h = \sum_{\substack{k \in estrato h \\ poblacional}} \frac{y_k}{N_h}$$

y W_h^* es el cociente del total poblacional de la variable X sobre dicho estrato entre su total muestral, también sobre el estrato:

$$W_h^* = \frac{\sum_{k=1}^{N} X_h^*(k)}{\sum_{k=1}^{n} X_h^*(k)},$$

con $X_h^*(k) = X_h(k) \bullet X(k)$

La estimación resultante es:

$$\hat{\overline{Y}}_{Rs} = \sum_{h=1}^{m} W_h^* \bullet \overline{y}_h$$

con $$\overline{y}_h = \sum_{\substack{k \in estrato h \\ muestral}} \frac{y_k}{n_h}$$

Respecto a la varianza de este estimador, tenemos que en el caso de muestreo aleatorio estratificado con tamaños muestrales suficientemente grandes en todos los estratos, la varianza de este estimador es:

$$Var(\hat{\overline{Y}}_{Rs}) = \sum_{h=1}^{m} \frac{N_h^2 \bullet (1 - f_h)}{n_h} \bullet \left(S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h r_h S_{yh} S_{xh} \right) \qquad (17)$$

sin más que, dado que se utiliza el estimador de la razón en cada celda de la estratificación, tomar la varianza de dicho estimador sobre cada una de estas celdas, sobre las que se da un muestreo aleatorio simple:

$$Var\left(\hat{\bar{Y}}_{Rh}\right) = \frac{1-f_h}{n_h} \bullet \left[\frac{\sum_{\substack{k=1 \\ k \in estratoh}}^{N_h} \left(y_k - R_h x_k\right)^2}{N_h - 1} \right] \qquad (18)$$

con R_h la razón entre la media muestral entre Y y X en el estrato h, es decir, la pendiente correspondiente a la recta de regresión entre la variable Y y X para el estrato h:

$$R_h = \frac{\sum_{k=1}^{n_h} y_k}{\sum_{k=1}^{n_h} x_k}$$

Al igual que en el caso de variables auxiliares cualitativas, concluimos que

la varianza del estimador de *la Razón Separado* es *inversamente proporcional* al *tamaño muestral* de las celdas de la estratificación y *directamente proporcional* a la varianza del estimador de la razón sobre cada una de estas celdas.

Ejemplos

En todos los ejemplos consideraremos *muestreos probabilistas* con *igual probabilidad de inclusión* en la muestra para todos los elementos de la misma, es decir, $z_k = N / n \; \forall k = 1,..,n$.

Vamos a considerar los datos almacenados en unas tablas, que en el caso particular en el que las variables auxiliares son cualitativas se llaman *tablas de contingencia*.

Variables Auxiliares Cualitativas

Consideraremos *dos variables objetivo*, correspondientes a los dos tipos de variables objetivo que se pueden dar: *cualitativa y cuantitativa*.

Y_1 es el *número total* de elementos de la población *no-parados*.

Y_2 es los *ingresos totales* en un colectivo.

Las *variables* auxiliares son dos:

- A: *Sexo*. Tiene 2 categorías *a=2*.

- B: *Nivel de institución*: estudios Primarios, Secundarios o Universitarios. Tiene 3 categorías, *b=3*.

Tamaño de la población *N=20* y muestral *n=12*.

Obtenemos axb=6 celdas: 6 variables $X_{hh'}$ con $1 \le h \le 2, 1 \le h' \le 3$ y la variable X_0

Variables auxiliares.

Datos poblacionales. ($N_{hh'}$) *Tabla (1)*

$\sum_{k=1}^{N} X_{hh'}(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	4	4	1	9
Mujer	3	5	3	11
TOTAL	7	9	4	20

Variables auxiliares.

Datos muestrales. ($n_{hh'}$) *Tabla (2)*

$\sum_{k=1}^{n} X_{hh'}(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	1	1	1	3
Mujer	3	3	3	9
TOTAL	4	4	4	12

Variable objetivo Y_1.

Datos muestrales. $\left(Y_{1,hh'}\right)$ *Tabla (3)*

Tabla de nº de *no-parados* muestrales *en cada estrato* de la población.

$\sum_{k=1}^{n} Y_1(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0	1	0	1
Mujer	3	2	0	5
TOTAL	3	3	0	6

Variable objetivo Y_2.

Datos muestrales. $\left(Y_{2,hh'}\right)$ *Tabla (4)*

Ingresos totales por mes sobre los *estratos*.

$\sum_{k=1}^{n} Y_2(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	60000	90000	150000	300000
Mujer	300000	375000	600000	1275000
TOTAL	360000	465000	750000	1575000

Las matrices que obtenemos para las variables *auxiliares* son:

$$\mathbf{X} = \begin{pmatrix} X_{00}' \\ X_{11}' \\ X_{12}' \\ \dots \\ X_{hh}' \\ \dots \\ X_{ab}' \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ X_{11}(1) & X_{11}(2) & X_{11}(N) \\ X_{12}(1) & X_{12}(2) & X_{12}(N) \\ \dots\dots\dots & & \\ X_{hh'}(1) & X_{hh'}(2) & X_{hh'}(N) \\ \dots\dots\dots & & \\ X_{ab}(1) & X_{ab}(2) & X_{ab}(N) \end{pmatrix}$$

que es una matriz 7xN.

En el caso general es una matriz (a*b+1)xN, siendo a*b el nº total de estratos en la estratificación de la población.

La composición la matriz X se obtendrá en el caso de poseer la información almacenada en forma de censo.

La matriz $\hat{X}$ es una estimación de la matriz X, de dimensión 7xn

Aplicando la fórmula (6), obtenemos las restricciones:

$$\begin{pmatrix} 1 & 1 & 1 \\ X_{11}(1) & X_{11}(k) & X_{11}(n) \\ \dots\dots\dots & & \\ X_{hh'}(1) & X_{hh'}(k) & X_{hh'}(n) \\ \dots\dots\dots & & \\ X_{23}(1) & X_{23}(k) & X_{23}(n) \end{pmatrix} * \begin{pmatrix} w_1 \\ \dots \\ w_k \\ \dots \\ w_n \end{pmatrix} = \begin{pmatrix} X_{00}(1) & X_{00}(k) & X_{00}(n) \\ X_{11}(1) & X_{11}(k) & X_{11}(n) \\ \dots\dots\dots & & \\ X_{hh'}(1) & X_{hh'}(k) & X_{hh'}(n) \\ \dots\dots\dots & & \\ X_{23}(1) & X_{23}(k) & X_{23}(n) \end{pmatrix} * \begin{pmatrix} w_1 \\ \dots \\ w_k \\ \dots \\ w_n \end{pmatrix} =$$

$$= \begin{pmatrix} N \\ \sum_{k=1}^{N} X_{11}(k) \\ \dots \\ \sum_{k=1}^{N} X_{hh'}(k) \\ \dots \\ \sum_{k=1}^{N} X_{23}(k) \end{pmatrix} = \begin{pmatrix} 20 \\ 4 \\ 4 \\ 1 \\ 3 \\ 5 \\ 3 \end{pmatrix}$$

Resolvemos el problema de elevación en los estratos para ambos casos (caso de variable objetivo cualitativa, como cuantitativa) al unísono. Su resolución es sencilla, sin más que aplicar las fórmulas (9) a partir de los datos de las tablas (1) y (2).

Los *elevadores* que se obtienen son: $\hat{W}_{hh'} = \frac{N_{hh'}}{\hat{n}_{hh'}}$

	ESTUDIOS		
SEXO	Primarios	Secundarios	Universitarios
Hombre	4=4/1	4=4/1	1=1/1
Mujer	1=3/3	1,667=5/3	1=3/3

La tabla de *no-parados ajustada* que obtenemos es:

Tabla(3)xelevadores

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0	4	0	4
Mujer	3	3	0	6
TOTAL	3	7	0	10

La tabla de *ingresos totales ajustada* que obtenemos es:

Tabla(4)xelevadores

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	240000	360000	150000	750000
Mujer	300000	625125	600000	1525125
TOTAL	540000	985125	750000	2275125

Las estimaciones del total, tanto para la variable cualitativa SER PARADO como para la cuantitativa INGRESOS TOTALES, son:

$\hat{Y}_1$ =10 número de no-parados estimados

$\hat{Y}_2$ =2275125 ingresos totales mensuales estimados

Tenemos entonces que:

- El total de no-parados, estimado, entre los 20 individuos de la población es 10 individuos, y que la media de no-parados en dicha población es, por lo tanto 10/20=0,5.
- Así mismo, los ingresos totales mensuales estimados en la población son de 2275125 y la media estimada es de 2275125/20=113756,25 pesetas.

Variables Auxiliares Cuantitativas

Consideraremos de nuevo dos *variables objetivo*: una *cualitativa* y la otra *cuantitativa*.

Y_1 es el *número total* de elementos de la población *no-parados*.

Y_2 es el *ahorro* mensual.

La variable auxiliar es los *ingresos totales mensuales* X(k)

La estratificación de la *población* la realizamos en función a dos variables:

- A**:** *Vivir en zona Rural o Urbana***.** Tiene 2 categorías *a=2***.**
- B: *Territorios Históricos***:** Bizkaia, Araba o Gipuzkoa. Tiene 3 categorías *b=3***.**

Consideramos tamaño *poblacional N=20* y *muestral n=12.*

Obtenemos axb=6 celdas: 6 variables $X_{hh'}$ con $1 \le h \le 2, 1 \le h' \le 3$ y la variable X_0

Variable auxiliar.

Datos poblacionales. $\left(\sum_{K=1}^{N} X^{*}_{hh'}(K) = X^{*}_{hh'} \right)$ *Tabla (5)*

Ingresos mensuales por *estratos* en la *población*.

$\sum_{k=1}^{N} X^{*}_{hh'}(k)$	TERRITORIO HISTÓRICO			
ZONA	Bizkaia	Araba	Gipuzkoa	TOTAL
Rural	440000	460000	420000	1320000
Urbana	380000	420000	400000	1200000
TOTAL	820000	880000	820000	2520000

Variable auxiliar.

Datos muestrales. $\left(\sum_{K=1}^{n} X^{*}_{hh'}(K) = \hat{X}^{*}_{hh'} \right)$ *Tabla (6)*

Ingresos mensuales por *estratos* en la *muestra*.

$\sum_{k=1}^{n} X^{*}_{hh'}(k)$	PROVINCIA			
ZONA	Bizkaia	Araba	Gipuzkoa	TOTAL
Rural	75000	250000	275000	600000
Urbana	120000	220000	250000	590000
TOTAL	195000	470000	525000	1190000

La notación que tomamos es igual que en los casos anteriores, pero donde antes aparecían las variables $X_{hh'}$, ahora nos aparecen las variables $X^{*}_{hh'}$, que ya están definidas en el apartado 1) de esta sección.

Variable objetivo Y_1.

Datos muestrales. $\left(Y_{1,hh'}\right)$ *Tabla (7)*

Número de no-parados en cada *estrato* de la *muestra*.

$Y_{1,hh'}$	TERRITORIO HISTÓRICO			
ZONA	Bizkaia	Araba	Gipuzkoa	TOTAL
Rural	0	1	0	1
Urbana	3	2	0	5
TOTAL	3	3	0	6

Variable objetivo Y_2.

Datos muestrales. $\left(Y_{2,hh'}\right)$ *Tabla (8)*

Ahorro mensual total sobre los estratos.

$Y_{2,hh'}$	TERRITORIO HISTÓRICO			
ZONA	Bizkaia	Araba	Gipuzkoa	TOTAL
Rural	20000	175000	200000	395000
Urbana	25000	175000	70000	270000
TOTAL	45000	350000	270000	665000

Aplicando la fórmula (10) para las restricciones, obtenemos:

$$\begin{pmatrix} 1 & 1 & 1 \\ X^*_{11}(1) & X^*_{11}(k) & X^*_{11}(n) \\ & \cdots\cdots\cdots & \\ X^*_{hh'}(1) & X^*_{hh'}(k) & X^*_{hh'}(n) \\ & \cdots\cdots\cdots & \\ X^*_{23}(1) & X^*_{23}(k) & X^*_{23}(n) \end{pmatrix} * \begin{pmatrix} w_1 \\ \cdots \\ w_k \\ \cdots \\ w_n \end{pmatrix} = \begin{pmatrix} X^*_{00}(1) & X^*_{00}(k) & X^*_{00}(n) \\ X^*_{11}(1) & X^*_{11}(k) & X^*_{11}(n) \\ & \cdots\cdots\cdots & \\ X^*_{hh'}(1) & X^*_{hh'}(k) & X^*_{hh'}(n) \\ & \cdots\cdots\cdots & \\ X^*_{23}(1) & X^*_{23}(k) & X^*_{23}(n) \end{pmatrix} * \begin{pmatrix} w_1 \\ \cdots \\ w_k \\ \cdots \\ w_n \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{k=1}^{N} X(k) \\ \sum_{k=1}^{N} X^{*}_{11}(k) \\ \dots \\ \sum_{k=1}^{N} X^{*}_{hh'}(k) \\ \dots \\ \sum_{k=1}^{N} X^{*}_{23}(k) \end{pmatrix} = \begin{pmatrix} 2520000 \\ 440000 \\ 460000 \\ 420000 \\ 380000 \\ 420000 \\ 400000 \end{pmatrix}$$

Los *elevadores* que se obtienen son:

$W^{*}_{hh'}$	TERRITORIO HISTÓRICO		
ZONA	Bizkaia	Araba	Gipuzkoa
Rural	5,867= 440000/75000	1,84= 460000/250000	1,527= 420000/275000
Urbana	3,167= 380000/120000	1,909= 420000/220000	1,6= 400000/250000

Con $W^{*}_{hh'} = \frac{X^{*}_{hh'}}{\hat{X}^{*}_{hh'}}$, según ya los habíamos definido en la fórmula (11)

La tabla de *no-parados ajustada* que obtenemos es: *Tabla(7)xelevadores*

$W^{*}_{hh'} \bullet Y_{1,hh'}$	TERRITORIO HISTÓRICO			
ZONA	Bizkaia	Araba	Gipuzkoa	TOTAL
Rural	0	2	0	2
Urbana	9	4	0	13
TOTAL	9	6	0	15

La tabla de *ahorros mensuales totales ajustada* que obtenemos es:

Tabla(8) x elevadores

$W^{*}_{hh'} \bullet Y_{2,hh'}$	TERRITORIO HISTÓRICO			
	Bizkaia	Araba	Gipuzkoa	TOTAL
Rural	117340	322000	305400	744740
Urbana	79175	334075	112000	525250
TOTAL	196515	656075	417400	1269990

Las estimaciones del total, tanto para la variable cualitativa *ser parado* como para la cuantitativa *ahorro*, son:

$\hat{Y}_1$ =15 es el número estimado de no-parados en la población

$\hat{Y}_2$ =1269990 es el ahorro mensual total estimado en la población

Tenemos entonces que:

- El total de no-parados, estimado, entre los 20 individuos de la población es 15 individuos, y que la media de no-parados en dicha población es, por lo tanto 15/20=0,75.

- Así mismo, el ahorro mensual total estimado en la población es de 1269990 pesetas al mes y la media estimada es de 1269990/20=63499,5 pesetas al mes.

Procedimientos informáticos

Para la resolución del ajuste de muestras, cuando tenemos Máxima Información Auxiliar, se pueden utilizar distintos software:

EXCEL

Como hoja de cálculo, EXCEL nos permite el cálculo de elevadores fácilmente. Sumado a esto, es una herramienta de Office bajo WINDOWS, lo que facilita su manejo y acceso. La dificultad en el empleo de esta herramienta para el ajuste es que necesitamos elaborar una macro para el cálculo de los elevadores. Esta macro, debe tener prefijadas las dimensiones de las tablas de datos con las que se trabajará, lo que no permite su uso a bases de datos generales.

SAS

Es un lenguaje de programación para el manejo y tratamiento de base de datos. Nos permite definir fácilmente los elevadores para el ajuste, pero el entorno de trabajo es menos cómodo que en el caso de Excel.

Otros

Además de los mencionados, todos aquellos lenguajes de programación que permitan el manejo de matrices son adecuados para el procedimiento. Entre estos lenguajes tenemos el FORTRAN, PASCAL, C++,... El núcleo del programa será la definición de unos elevadores, resultantes del cociente entre totales poblacionales y muestrales en cada uno de los estratos.

Situaciones óptimas para su aplicación

Considerando todo lo mencionado anteriormente, tenemos que la situación necesaria para su aplicación será:

- El número de celdas de la estratificación no debe ser muy grande, ya que en este caso, la dispersión de la muestra a lo largo de las celdas no será muy densa y, por lo tanto, nos podemos encontrar con un número muy pequeño de elementos en algunos de los estratos: todas aquellas celdas que tengan un tamaño inferior al mínimo de 20 ó 25 darán problemas. Para solucionar estos problemas se plantea como solución el colapso de celdas, que supone una pérdida de información poblacional y por lo tanto tiene sesgo, pero se obtiene una disminución de la varianza, ya que ésta era inversamente proporcional al tamaño muestral de las celdas y de esta forma conseguimos que estos tamaños se mantengan lo suficientemente grandes.

- El método compensa la no-cobertura. En caso de utilizar junto a las probabilidades de inclusión en la muestra de los elementos las tasas de no-respuesta obtendríamos un ajuste que compensaría la no-respuesta y la no-cobertura.

Capítulo

3

Procedimientos Iterativos con Información Auxiliar Cualitativa

Los procedimientos iterativos que vamos a presentar en este capítulo son el RAKING USUAL y una variación de éste utilizada en el procedimiento REDRE. Ambos se utilizan cuando tenemos variables auxiliares cualitativas y tenemos máxima información en las distribuciones univariantes de estas variables auxiliares. La estratificación se realiza con el cruce multivariante de estas variables auxiliares.

Notación

Se plantean unas restricciones con el fin de que la distribución univariante ponderada de la muestra iguale a la poblacional. La representación matricial de estas restricciones sigue siendo la misma:

$$\hat{X} \bullet \mathbf{w} = \mathbf{X} \bullet \mathbf{I}$$

La representación de las matrices auxiliares **X** y $\hat{X}$ la damos a continuación:

Tenemos L variables auxiliares cualitativas, cada una con $L_1, L_2, ..., L_l, ..., L_L$ modalidades. Se realiza la estratificación de la población mediante el cruce multivariante de estas L variables, obteniendo $L_1 \bullet L_2 \bullet .. \bullet L_l \bullet .. \bullet L_L$ estratos. El número total de modalidades es $L_1 + L_2 + .. + L_l + .. + L_L = m$

En este caso, tomamos las variables identificadoras o dicotómicas de las modalidades:

$$X_{h_l}(k) = \begin{cases} 1, \text{si k tiene la modalidad } h_l \\ 0, \text{en caso contrario} \end{cases} \qquad 1 \le h_l \le m$$

con k=1,..,n en la muestra y k=1,..,N en la población.

$\Rightarrow L_1 + .. + L_l + .. + L_L = m$ variables dicotómicas.

Tomamos los vectores fila poblacionales

$X_{h_l} = \left(X_{h_l}(1), .., X_{h_l}(k), .., X_{h_l}(N)\right)$ de dimensión (1xN), con $1 \le h_l \le m$

y los muestrales

$\hat{X}_{h_l} = \left(\hat{X}_{h_l}(1), .., \hat{X}_{h_l}(k), .., \hat{X}_{h_l}(n)\right)$ de dimensión (1xn), con $1 \le h_l \le m$

Para asegurar que:

$$w_1 + .. + w_k + .. w_n = N$$

tomamos la variable X_0 idénticamente 1 y se representa mediante:

$X_0 = (X_0(1),.., X_0(k),.., X_0(N)) = (1,..,1,..,1)$ vector fila poblacional de dimensión (1xN)

$\hat{X}_0 = (X_0(1),.., X_0(k),.., X_0(n)) = (1,..,1,..,1)$ vector fila muestral de dimensión (1xn)

Las matrices **X** y $\hat{X}$ están formadas por estos vectores fila:

$$\mathbf{X}=\begin{pmatrix} X_0 \\ X_1 \\ ... \\ X_h \\ ... \\ X_m \end{pmatrix} \quad \mathbf{y} \quad \hat{X} = \begin{pmatrix} \hat{X}_0 \\ \hat{X}_1 \\ ... \\ \hat{X}_h \\ ... \\ \hat{X}_m \end{pmatrix} \quad \#$$

Es decir, son dos matrices de dimensión (m+1)xN y (m+1)xn

Desarrollando la expresión matricial, obtenemos:

$$\begin{cases} \sum_{k=1}^{n} w_k \bullet X_0(k) = \sum_{k=1}^{N} X_0(k) \Leftrightarrow w_1 + .. + w_k + .. w_n = N \\ \sum_{k=1}^{n} w_k \bullet X_h(k) = \sum_{k=1}^{N} X_h(k), \forall 1 \le h \le m \end{cases}$$

Vamos a simplificar la notación, tomando únicamente dos variables auxiliares con un número de modalidades *a* y *b*. Numeramos las modalidades, de tal forma que si los datos que tenemos son los marginales $N_{h.}$ y $N_{.h'}$ tenemos que:

$$\begin{cases} N_h = N_{h.} \text{ con } 1 \le h \le a \quad \text{y} \\ N_{h'} = N_{.h'} \text{ con } a+1 \le h' \le a+b \end{cases}$$ siendo *a*+*b*=m

$\rightarrow$ Distribución univariante de la 1ª variable

$\rightarrow$ Distribución univariante de la 2ª variable

y de forma análoga, definimos la notación para los datos muestrales de la distribución univariante:

Numeramos las modalidades de 1 a m.

$$\begin{cases} n_h = n_{h\bullet} \text{ con } 1 \le h \le a \quad \text{y} \\ n_{h'} = n_{\bullet h'} \text{ con } a+1 \le h' \le a+b \end{cases}$$

Tomando la representación con variables auxiliares de estos datos, tenemos:

$$\begin{cases} \sum_{k=1}^{N} X_h(k) = N_h \\ \sum_{k=1}^{n} X_h(k) = n_h \end{cases} \text{con} 1 \le h \le \text{m}$$

Resolvemos el sistema

El ajuste es por celdas, teniendo todos los individuos de una misma celda el mismo peso. Las incógnitas pasan a ser entonces, los elevadores de las celdas $W_{hh'}$ con $1 \le h \le a,\ a+1 \le h' \le a+b$. Y el sistema tiene la forma:

$$\begin{cases} \sum_{h=1}^{a} \sum_{h'=a+1}^{a+b} W_{hh'} = N \\ \sum_{h'=a+1}^{a+b} W_{hh'} \bullet n_{hh'} = \sum_{k=1}^{N} X_h(k) = N_h = N_{h\bullet}, \forall 1 \le h \le a \\ \sum_{h=1}^{a} W_{hh'} \bullet n_{hh'} = \sum_{k=1}^{N} X_{h'}(k) = N_{h'} = N_{\bullet h'}, \forall a+1 \le h' \le a+b \end{cases} \quad (18)$$

En general se obtiene un sistema de $L_1 \bullet L_2 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_L$ incógnitas y $L_1 + L_2 + \ldots + L_l + \ldots + L_L + 1 = m + 1$ ecuaciones, en el que la última es combinación lineal de las anteriores, por lo que se reduce a un sistema de $L_1 \bullet L_2 \bullet \ldots \bullet L_l \bullet \ldots \bullet L_L$ incógnitas y $L_1 + L_2 + \ldots + L_l + \ldots + L_L = m$ ecuaciones. En el caso de sólo dos variables auxiliares, tenemos un sistema de *a*b* incógnitas y *a+b* ecuaciones.

En general hay más incógnitas que ecuaciones, por lo que el sistema tendrá infinitos elevadores posibles. Para evitar esto se toma un criterio de minimización para una función, con lo que se obtendrá la unicidad de la solución. La función que se tomará será la función distancia del vector de pesos iniciales al final. Según qué función distancia se tome, se obtendrán distintas soluciones.

Descripción del procedimiento

Vamos a describir a continuación los procedimientos iterativos que se siguen, tanto con el método Raking Usual como con su variación en el Redre.

Antes de describir la fórmula iterativa de cada uno de los dos procesos vamos a presentar la situación, respecto de la información que se dispone, que es común para los dos procesos:

La tabla multivariante con la información poblacional es:

A\B	1	2		h'		b	Total
1	W_{11}	W_{12}		$W_{1h'}$		W_{1b}	$W_{1.} = W_1$
2	W_{21}	W_{22}		$W_{2h'}$		W_{2b}	$W_{2.} = W_2$
h	W_{h1}	W_{h2}		$W_{hh'}$		W_{hb}	$W_{h'.} = W_h$
a	W_{a1}	W_{a2}		$W_{ah'}$		W_{ab}	$W_{a.} = W_a$
Total	$W_{.1} = W_{a+1}$	$W_{.2} = W_{a+2}$		$W_{.h'} = W_{a+h'}$		$W_{.b} = W_{a+b}$	1

De esta tabla sólo conocemos los datos de las distribuciones univariantes y que corresponden a las celdas sombreadas.

Respecto a los datos muestrales, la tabla que se obtiene toma la siguiente forma:

A\B	1	2		h’		b	Total
1	q_{11}	q_{12}		$q_{1h'}$		q_{1b}	$q_{1.} = q_1$
2	q_{21}	q_{22}		$q_{2h'}$		q_{2b}	$q_{2.} = q_2$
h	q_{h1}	q_{h2}		$q_{hh'}$		q_{hb}	$q_{h.} = q_h$
a	q_{a1}	q_{a2}		$q_{ah'}$		q_{ab}	$q_{a.} = q_a$
Total	$q_{.1} = q_{a+1}$	$q_{.2} = q_{a+2}$		$q_{.h'} = q_{a+h'}$		$q_{.b} = q_{a+b}$	1

En este caso se conoce la distribución conjunta, la multivariante.

Hemos reducido el caso general a la situación de únicamente dos variables auxiliares cualitativas y bajo esta suposición vamos a presentar ambos procesos iterativos. Para ello, vamos a ilustrar ambos procesos iterativos aplicándolos a unas tablas sencillas imaginarias:

Simulación

El tamaño de la *población es 20* y la *muestra tiene 10* individuos.

Los datos poblacionales que conocemos son los siguientes $(W_{hh'})$:

A\B	1	2	3	Total
1	W_{11}	W_{12}	W_{13}	$W_{1.} = 0{,}75$
2	W_{21}	W_{22}	W_{23}	$W_{2.} = 0{,}25$
Total	$W_{.1} = 0{,}2$	$W_{.2} = 0{,}4$	$W_{.3} = 0{,}4$	1

La tabla de datos muestrales es $(q_{hh'})$:

A\B	1	2	3	Total
1	$q_{11} = 0{,}1$	$q_{12} = 0{,}2$	$q_{13} = 0{,}2$	$q_{1.} = 0{,}5$
2	$q_{21} = 0{,}2$	$q_{22} = 0{,}1$	$q_{23} = 0{,}2$	$q_{2.} = 0{,}5$
Total	$q_{.1} = 0{,}3$	$q_{.2} = 0{,}3$	$q_{.3} = 0{,}4$	1

Raking Usual

El algoritmo iterativo que define el método RAKING es: (ver [4] y [5])

$$w(k,m) = w(k,m-1) * \frac{W_{h_0}}{q_{h_0,\,m-1}} * \frac{W_{h'_0}}{q'_{h'_0,\,m-1}} \text{ con}$$

$$\begin{cases} q'_{h',0} = \sum_{h=1}^{a} q'_{hh',0} = \sum_{h=1}^{a} q_{hh',0} \bullet \frac{W_{h.}}{q_{h,0}} \\ q'_{h',m} = \sum_{h=1}^{a} q'_{hh',m} = \sum_{h=1}^{a} q_{hh',m} \bullet \frac{W_h}{q_{h,m}}, \text{con m} \geq 1 \end{cases} \quad \text{para } a+1 \leq h' \leq a+b$$

y el elemento k en la celda (h_0, h'_0).

Seguidamente vamos a aplicar la fórmula iterativa al ejemplo simulado. Para ello, presentamos tablas en las que presentaremos los pesos obtenidos, así como la distribución muestral ponderada obtenida al elevarla por estos pesos.

PESOS(0) =20/10:

2	2	2
2	2	2

Tabla muestral ponderada por los *pesos 0* $(q_{hh',0})$**:**

A\B	1	2	3	Total
1	0,2	0,4	0,4	1
2	0,4	0,2	0,4	1
Total	0,6	0,6	0,8	2

El *primer paso de la iteración* es el ajuste de la muestra respecto a la primera característica auxiliar o de estratificación. Los factores de elevación que se obtienen debido al ajuste son:

0,75	0,75	0,75
0,25	0,25	0,25

Y obtenemos la siguiente *tabla muestral en porcentajes ponderada* $(q'_{hh',0})$**:**

A\B	1	2		Total
1	0,15	0,3	0,3	0,75
2	0,1	0,05	0,1	0,25
Total	0,25	0,35	0,4	1

El *segundo paso* de esta *primera iteración* consiste en ajustar la distribución muestral a la segunda variable de estratificación. Los factores de elevación de muestra en este segundo paso de la iteración son:

0,8	1,143	1
0,8	1,143	1

La *tabla muestral en porcentajes* obtenida al *elevar* la muestra por estos nuevos elevadores es $(q_{hh',1})$:

A\B	1	2		Total
1	0,12	0,343	0,3	0,763
2	0,08	0,057	0,1	0,237
Total	0,2	0,4	0,4	1

La tabla muestral se ha *desajustado ahora respecto de la primera variable* de estratificación, por lo que procedemos a la *segunda iteración* del procedimiento:

El *primer paso de la segunda iteración*. Los *factores de elevación* que se obtienen debido al ajuste son:

0,983	0,983	0,983
1,055	1,055	1,055

Y obtenemos la siguiente *tabla muestral en porcentajes ponderada* $(q'_{hh',1})$**:**

A\B	1	2		Total
1	0,118	0,337	0,295	0,75
2	0,084	0,06	0,106	0,25
Total	0,202	0,397	0,401	1

El *segundo paso de la segunda iteración*. Los factores de son:

0,99	1,008	0,998
0,99	1,008	0,998

La *tabla muestral en porcentajes* que se obtiene al *elevar* la muestra por estos nuevos elevadores es $(q_{hh',2})$**:**

A\B	1	2		Total
1	0,117	0,34	0,294	0,751
2	0,083	0,06	0,106	0,249
Total	0,2	0,4	0,4	1

El error producido con estos porcentajes muestrales ajustados respecto de los porcentajes poblacionales conocidos es del orden de 0,0035 y dado que estamos trabajando únicamente con 3 dígitos, esta tolerancia es suficiente para el ajuste, con lo que se da por finalizado dicho ajuste, con tan sólo dos iteraciones, cada una de ellas con sus dos respectivos pasos.

La tabla obtenida es la de *porcentajes muestrales ajustados a la población*. Para obtener la tabla de número de *efectivos poblacionales* estimados en cada celda de la estratificación, basta con multiplicar a tabla por el tamaño de la población N=20. La tabla estimada de efectivos que obtenemos es:

A\B	1	2		Total
1	2	7	6	15
2	2	1	2	5
Total	4	8	8	20

Como se puede apreciar, la distribución muestral se ha desajustado ahora de la primera variable de estratificación tras el ajuste a la segunda, con lo que se vuelve a iterar el procedimiento, iniciando ahora el proceso con una nueva tabla ajustada, la obtenida tras la primera iteración del Raking.

Esta fórmula iterativa surge del siguiente proceso:

- Tomamos la distribución univariante de la primera variable auxiliar. Se ponderan todas las celdas de esa fila h por el factor $\frac{W_{h\bullet}}{q_{h\bullet,0}} = \frac{W_h}{q_{h,0}}$, con el fin de ajustar la distribución marginal auxiliar ponderada de la muestra a la poblacional. La distribución muestral obtenida tras este primer paso se denota por $q'_{hh',0}$

- Se ajusta la distribución muestral obtenida con el fin de ajustar a la distribución marginal de la segunda variable auxiliar. Para ello se toma el factor de elevación $\frac{W_{.h'}}{q'_{.h',0}} = \frac{W_{h'}}{q'_{h',0}}$

Con los pasos anteriores se finaliza una iteración, tras la cual, la distribución muestral queda elevada por $\frac{W_{h.}}{q_{h.,0}} \bullet \frac{W_{.h'}}{q'_{.h',0}} = \frac{W_h}{q_{h,0}} \bullet \frac{W_{h'}}{q'_{h',0}}$

La distribución resultante ajusta exactamente respecto de la distribución univariante para la segunda variable auxiliar, pero debido al segundo paso se ha desajustado ligeramente respecto de la distribución univariante de la primera variable auxiliar.

Para una iteración m cualquiera obtenemos en el ajuste a la primera variable los factores $\frac{W_{h.}}{q_{h.,\,m-1}} = \frac{W_h}{q_{h,\,m-1}}$ y en el segundo paso $\frac{W_{.h'}}{q'_{.h',\,m-1}} = \frac{W_{h'}}{q'_{h',\,m-1}}$, con lo que los pesos finales son:

$$w(k,m) = w(k,m-1) \bullet \frac{W_h}{q_{h',\,m-1}} \bullet \frac{W_{h'}}{q'_{h',\,m-1}} \text{ con k en el estrato } (h,h')$$

El proceso se sigue iterando hasta un *número fijo de iteraciones*, o hasta que se da la *convergencia* a la distribución poblacional univariante.

Este proceso iterativo resulta de tomar el sistema de restricciones y dividirlo en *dos sitemas*:

1. Un primer sistema en el que se plantean las restricciones correspondientes a la *primera variable auxiliar* y en el que se consideran que los pesos resultantes son de la forma:

$$\boxed{w'(k,m-1) = w(k,m-1) * c_h}$$

Resulta entonces un sistema de a incógnitas y a ecuaciones, con matriz asociada diagonal, de elementos diagonales el número de efectivos muestrales obtenidos para cada modalidad de la primera variable auxiliar. Este sistema tiene *única solución* siempre que dichos elementos diagonales sean no nulos. El primer paso de cada iteración resulta entonces de la resolución de este sistema.

2. El segundo paso surge análogamente de tomar los pesos de la forma:

$$\boxed{w(k,m) = w'(k,m-1) * c_{h'}}$$

y se plantea entonces un sistema de b ecuaciones y b restricciones, en el que las incónigtas son $c_{h'}$.

Este sistema tiene como matriz asociada una matriz diagonal con elementos diagonales el número de efectivos muestrales marginales obtenidos para las modalidades de esta

segunda variable auxiliar. Bajo la hipótesis de no nulidad de estos elementos, el sistema tiene una *única solución*, con lo que se obtiene la distribución ponderada tras el segundo paso de la iteración y por lo tanto, al tener únicamente dos variables auxiliares, los pesos finales para la iteración.

En el caso de tener más variables auxiliares, tendríamos más sistemas planteados en cada iteración, tantos como variables.

La solución que se obtiene mediante el procedimiento iterativo RAKING, coincide además con la solución que se obtiene con el método de *Ajuste de Mínimo Cuadrados*: se plantea un sistema con unas restricciones, que corresponden a la consistencia con respecto de la información auxiliar poblacional de la distribución muestral ponderada, y una función objetivo en el que se mide la exactitud de dicha distribución muestral ponderada tomando como distancia una función de la forma $D(x,y)=\frac{(x-y)^2}{x}$. Así, se toma la función distancia como la suma de las distancias de las distribuciones muestrales univariantes con respecto a las poblacionales correspondientes a cada modalidad, obteniendo así la función distancia a minimizar.

Para resolver este problema de minimización de una función objetivo con variables sujetas a una serie de restricciones se emplea el método de los *multiplicadores de Lagrange* y la resolución del mismo mediante dichos multiplicadores nos lleva a la misma solución del Raking, lo que pasa que, mediante este último el proceso es más rápido.

Respecto al Raking, podemos además obtener el ajuste para una celda que sea para nosotros de especial interés de una forma relativamente rápida, sin más que mediante una compresión de las celdas restantes en celdas colindantes a la celda en cuestión.

Como vemos el Raking es un procedimiento iterativo que ajusta la distribución marginal de las modalidades por variables auxiliares.

Resaltar que cuando alguna de las celdas en la estratificación no tiene ningún elemento muestral $\left(n_{hh'}=0\right)$, entonces en el Raking se le asigna un número de efectivos poblacionales aproximado $\tilde{N}_{hh'}=0$ y respectivamente $\tilde{W}_{hh'}=0$.

El Raking converge bajo ciertas condiciones de regularidad. Además se obtiene que en estos casos de convergencia, los estimadores de los $N_{hh'}$, los $\tilde{N}_{hh'}$ son asintóticamente insesgados, de distribución normal y de mínima varianza, es decir, son estimadores BAN (best asymptocally normal estimators).

Redre

Es una variación del método RAKING USUAL. La fórmula iterativa que lo define es:

$$w(k,m) = w(k,m-1) * \left(\sum_{h=1}^{a} \sum_{h'=1}^{b} X_{hh'}(k) * \left(\frac{\left(\sum_{k=1}^{N} X_h(k) \right) / N}{\left(\sum_{k=1}^{n} X_h(k) * w(k,m-1) \right) / n} + \right. \right.$$

$$\left. \left. + \frac{\left(\sum_{k=1}^{N} X_{h'}(k) \right) / N}{\left(\sum_{k=1}^{n} X_{h'}(k) * w(k,m-1) \right) / n} \right) \right) / 2$$

Con:

- $w(k,m)$ peso asignado al elemento k en la iteración m-ésima.

$$w\ k,0 = 1$$

$\left(\sum_{k=1}^{N} {}_h\ k) \right) / N$ designa el porcentaje poblacional de elementos con la modalidad h, siendo esta modalidad una de las de la primera variable auxiliar. Tomando la nomenclatura del número

$$\begin{cases} N_h / N & con\ 1 \le h \le a \\ N_{h'} / N & con\ a+1 \le h' \le a+b \end{cases}$$

Para simplificar aún más la notación denotamos estos porcentajes poblacionales como

$$W_h = \frac{\left(\sum_{k=1}^{N} X_h(k) \right)}{N} = \frac{N_h}{N}\ y\ W_{h'} = \frac{\left(\sum_{k=1}^{N} X_{h'}(k) \right)}{N} = \frac{N_{h'}}{N}$$

- $\left(\sum_{k=1}^{n} X_{h'}(k) * w(k,m-1) \right) / n$ y $\left(\sum_{k=1}^{n} X_h(k) * w(k,m-1) \right) / n$ son los porcentajes muestrales ponderados mediante los pesos obtenidos tras la iteración m. Son porcentajes marginales correspondientes a las modalidades de la primera y la segunda variable auxiliar, respectivamente. Para simplificar las expresiones tomamos la notación:

$$q_{h,\,m-1}=\left(\sum_{k=1}^{n} X_h(k)*w(k,m-1)\right)\Big/ n \quad y$$

$$q_{h',\,m-1}=\left(\sum_{k=1}^{n} X_{h'}(k)*w(k,m-1)\right)\Big/ n$$

En el caso de m=1 estos porcentajes muestrales ponderados son: $\begin{cases}\left(n_h/n\right) con 1\le h\le a \quad y\\ \left(n_{h'}/n\right) con a+1\le h\le a+b\end{cases}$,

Estas expresiones se deducen directamente de tomar $w(k,0)=1$.

- 2: representa el número de variables auxiliares.
- $\sum_{h=1}^{a}\sum_{h'=1}^{b} X_{hh'}(k)$ es un filtro para hallar los factores que intervendrán en el peso obtenido en cada iteración, al ser las variables $X_{hh'}(k)\ con\ 1\le h\le a\ \ y\ 1\le h'\le b$ dicotómicas asociadas a cada celda de la estratificación. Es decir, según la notación que hemos tomado en el problema, $X_{hh'}(k)=X_h(k)*X_{h'}(k)$.

Tras todo lo mencionado la fórmula iterativa para REDRE es:

$$w(k,m)=w(k,m-1)*\left(\sum_{h=1}^{a}\sum_{h'=1}^{b} X_{hh'}(k)*\left(\frac{\frac{W_h}{q_{h,\,m-1}}+\frac{W_{h'}}{q_{h',\,m-1}}}{2}\right)\right)$$

Generalizamos la fórmula iterativa anterior, al caso general de más variables auxiliares y obtenemos:

$$w(k,m)=w(k,m-1)*\left(\frac{suma\left(\frac{W_h}{q_{h,\,m-1}}\right)}{n^{o}\,\mathrm{var}\,iables\ auxiliares}\right)$$ siendo la suma en todas las modalidades en las que está el elemento k de la muestra, para todas las variables auxiliares.

Estos pesos se multiplican luego por $\frac{N}{n}$ en EUSTAT, con el fin de que $\sum_{k=1}^{n} w(k,m) = N$ Así, esto es equivalente a tomar como pesos iniciales $w(k,0) = \frac{N}{n}$, es decir, corresponde a tomar como pesos iniciales los inversos de la probabilidad de inclusión en la muestra para cada individuo, es decir, corresponde a tomar el estimador inicial de H-T en los muestreos probabilistas con igual probabilidad de inclusión $z_k = \frac{n}{N}$ para todo individuo muestral.

Es un ajuste iterativo por celdas, en el que a todos los individuos de una misma celda se les asigna el mismo peso.

Vemos cómo actúa el REDRE aplicado a nuestro ejemplo de simulación:

Primera iteración

Elevadores

=(0,75/0,5+0,2/0,3)/2 =**1,083**	=(0,75/0,5+0,4/0,3)/2 =**1,417**	=(0,75/0,5+0,4/0,4)/2 =**1,25**
=(0,25/0,5+0,2/0,3)/2 =**0,583**	=(0,25/0,5+0,4/0,3)/2 =**0,917**	=(0,25/0,5+0,4/0,4)/2 =**0,75**

*Tabla muestral en porcentajes ajustada a la población***:**

A\B	1	2	3	Total
1	0,108	0,283	0,25	0,641
2	0,117	0,092	0,15	0,359
Total	0,225	0,375	0,4	1

Segunda iteración

Elevadores

=(0,75/0,641+0,2/0,225)/2 =**1,029**	=(0,75/0,641+0,4/0,375)/2 =**1,118**	=(0,75/0,641+0,4/0,4)/2 =**1,085**
=(0,25/0,359+0,2/0,225)/2 =**0,793**	=(0,25/0,359+0,4/0,375)/2 =**0,882**	=(0,25/0,359+0,4/0,4)/2 =**0,848**

Tabla muestral en *porcentajes ajustada a la población***:**

A\B	1	2	3	TOTAL
1	0,111	0,316	0,271	0,698
2	0,093	0,081	0,127	0,301
TOTAL	0,204	0,397	0,398	0,999

Tercera iteración

Elevadores

=(0,75/0,698+0,2/0,204)/2 =**1,027**	=(0,75/0,698+0,4/0,397)/2 =**1,041**	=(0,75/0,698+0,4/0,398)/2 =**1,04**
=(0,25/0,301+0,2/0,204)/2 =**0,905**	=(0,25/0,301+0,4/0,397)/2 =**0,919**	=(0,25/0,301+0,4/0,398)/2 =**0,918**

Tabla muestral en porcentajes ajustada a la población

A\B	1	2	3	TOTAL
1	0,114	0,329	0,282	0,725
2	0,084	0,074	0,117	0,275
TOTAL	0,198	0,403	0,399	1

Cuarta iteración

Elevadores

=(0,75/0,725+0,2/0,198)/2 =**1,022**	=(0,75/0,725+0,4/0,403)/2 =**1,014**	=(0,75/0,725+0,4/0,399)/2 =**1,018**
=(0,25/0,275+0,2/0,198)/2 =**0,96**	=(0,25/0,275+0,4/0,403)/2 =**0,951**	=(0,25/0,275+0,4/0,399)/2 =**0,956**

Tabla muestral en porcentajes ajustada a la población

A\B	1	2	3	TOTAL
1	0,117	0,334	0,287	0,738
2	0,081	0,07	0,112	0,263
TOTAL	0,198	0,404	0,399	1,001

Quinta iteración

Elevadores

=(0,75/0,738+0,2/0,198)/2 =**1,013**	=(0,75/0,738+0,4/0,404)/2 =**1,003**	=(0,75/0,738+0,4/0,399)/2 =**1,009**
=(0,25/0,263+0,2/0,198)/2 =**0,98**	=(0,25/0,263+0,4/0,404)/2 =**0,97**	=(0,25/0,263+0,4/0,399)/2 =**0,977**

Tabla muestral en porcentajes ajustada a la población

A\B	1	2	3	TOTAL
1	0,119	0,335	0,29	0,744
2	0,079	0,068	0,109	0,256
TOTAL	0,198	0,403	0,399	1

Sexta iteración

Elevadores

=(0,75/0,744+0,2/0,198)/2 =**1,009**	=(0,75/0,744+0,4/0,403)/2 =**1**	=(0,75/0,744+0,4/0,399)/2 =**1,005**
=(0,25/0,256+0,2/0,198)/2 =**0,993**	=(0,25/0,256+0,4/0,403)/2 =**0,985**	=(0,25/0,256+0,4/0,399)/2 =**0,99**

Tabla muestral en porcentajes ajustada a la población

A\B	1	2	3	TOTAL
1	0,12	0,335	0,291	0,746
2	0,078	0,067	0,108	0,253
TOTAL	0,198	0,402	0,399	0,999

Séptima iteración

Elevadores

=(0,75/0,746+0,2/0,198)/2 =**1,008**	=(0,75/0,746+0,4/0,402)/2 =**1**	=(0,75/0,746+0,4/0,399)/2 =**1,004**
=(0,25/0,253+0,2/0,198)/2 =**0,999**	=(0,25/0,253+0,4/0,402)/2 =**0,992**	=(0,25/0,253+0,4/0,398)/2 =**0,995**

Tabla muestral en porcentajes ajustada a la población

A\B	1	2	3	TOTAL
1	0,121	0,335	0,292	0,748
2	0,078	0,066	0,107	0,251
TOTAL	0,199	0,401	0,399	0,999

La distribución ha quedado ya ajustada con una *tolerancia de 0,002*, tras siete iteraciones.

La tabla con el número de efectivos estimados tras el ajuste a la distribución univariante de las variables de estratificación es:

A\B	1	2	3	TOTAL
1	2	7	6	15
2	2	1	2	5
TOTAL	4	8	8	20

La tabla estimada para nuestro ejemplo simulado obtenida con ambos métodos es la misma.

Este *método de ajuste*, el *RAS,* lo que plantea es evitar tantas modificaciones de los pesos en cada iteración, por lo que toma el factor de modificación de los pesos como una media de los factores resultantes de los ajustes a las modalidades a las que pertenece cada elemento. Así, lo que hace es plantear un único sistema a solucionar, en el que obtenemos $a+b$ incógnitas y ecuaciones. Los pesos resultantes toman entonces la forma

$$w(k,m) = w(k,m-1) * \sum_{h,h'} \left(X_h(k) * c_h\right) * \left(X_{h'}(k) * c_{h'}\right)$$

$$\text{con} 1 \le h \le a \text{ y } a+1 \le h' \le a+b+1$$

Los factores c_h y $c_{h'}$ se hallan de resolver el sistema de restricciones planteado y el peso final para cada elemento resulta de elevar el peso inicial por la media de los factores c_h y $c_{h'}$

Ejemplos

Consideraremos dos variables objetivos, correspondientes a los dos tipos de variables objetivo que se pueden dar: cualitativa y cuantitativa.

Y_1 es el *Número total* de elementos de la población *no-parados*.

Y_2 es la *facturación total por mes* de la población.

Las *variables* auxiliares son dos:

- A: *Sexo.* Tiene 2 categorías *a=2*.
- B: *Nivel de institución*: estudios Primarios, Secundarios o Universitarios. Tiene 3 categorías, *b=3*.

Tamaño de la *población N=20 y muestral n=12*.

Consideramos siempre un *muestreo probabilista*, con *igual probabilidad de inclusión* para todos los elementos de la muestra 12/20.

Variables auxiliares

Datos poblacionales ($N_{hh'}$) *Tabla (1)*

$\sum_{k=1}^{N} X_{hh'}(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	?	?	?	9
Mujer	?	?	?	11
TOTAL	7	9	4	20

Datos muestrales ($n_{hh'}$) *Tabla (2)*

$\sum_{k=1}^{n} X_{hh'}(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	1	1	1	3
Mujer	3	3	3	9
TOTAL	4	4	4	12

Variable objetivo Y_1.

Datos muestrales. $\left(Y_{1,hh'}\right)$ *Tabla (3)*

Tabla de nº de *no-parados* muestrales en cada *estrato* de la población.

$\sum_{k=1}^{n} Y_1(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0	1	0	1
Mujer	3	2	0	5
TOTAL	3	3	0	6

Variable objetivo Y_2.

Datos muestrales. $\left(Y_{2,hh'}\right)$ *Tabla (4)*

Totales de facturación por mes sobre los *estratos*.

$\sum_{k=1}^{n} Y_2(k)$	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	75000	250000	275000	600000
Mujer	120000	220000	250000	590000
TOTAL	195000	470000	525000	1190000

Presentamos estos datos para las variables auxiliares en forma de porcentajes:

Variables auxiliares. Datos poblacionales. ($W_{hh'}$)

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	?	?	?	0,45
Mujer	?	?	?	0,55
TOTAL	0,35	0,45	0,2	1

Variables auxiliares. Datos muestrales. ($q_{hh'}$)

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0,0833	0,0833	0,0833	0,25
Mujer	0,25	0,25	0,25	0,75
TOTAL	0,333	0,333	0,333	1

La resolución de ambos problemas es la misma, es decir, el ajuste que vamos a obtener con la variable objetivo cualitativa o cuantitativa es el mismo. Resolvamos entonces por Raking Usual y seguidamente lo haremos por Redre.

Raking Usual

Vamos a resolver el problema con el método Raking, usando hojas de cálculo de Excel.

Comenzamos las iteraciones, recordando que cada iteración se compone de 2 pasos, al haber únicamente dos variables de ajuste:

1. Se ajusta la distribución respecto de la distribución poblacional marginal de las modalidades de la primera variable.

2. En la segunda fase, con esta nueva distribución muestral ponderada se ajustará la distribución a las modalidades de la segunda característica, finalizando así la iteración.

Primeramente tomamos *la tabla de porcentajes muestrales* $(q_{hh'})$:

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0,0833	0,0833	0,0833	0,25
Mujer	0,25	0,25	0,25	0,75
TOTAL	0,333	0,333	0,333	1

El primer peso que se asigna a los individuos es el de *Horvitz-Thompson*, es decir, N/n. La tabla que se obtiene al ponderar por estos pesos se denomina *tabla expandida* $(q_{hh',0})$**:**

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0,138	0,138	0,138	0,415
Mujer	0,417	0,417	0,417	1,25
TOTAL	0,555	0,555	0,555	1,665

Primer paso de la *primera iteración*: ajuste de la distribución a la distribución univariante de la variable *estudios*

	ESTUDIOS			$(q_{hh',0})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0,138	0,138	0,138	0,415
Mujer	0,417	0,417	0,417	1,25
TOTAL	0,555	0,555	0,555	1,665
	0,35	0,45	0,2	
elevadores	0,631	0,811	0,36	

Ponderamos la tabla muestral con los elevadores hallados y se procede al *segundo paso:* al ajuste a la distribución univariante de la variable *sexo*

	ESTUDIOS					
SEXO	Primarios	Secundarios	Universitarios	TOTAL		elevadores
Hombre	0,087	0,112	0,05	0,249	0,45	1,807
Mujer	0,263	0,338	0,15	0,751	0,55	0,732
TOTAL	0,35	0,45	0,2	1		

Tabla final

	ESTUDIOS			$(q''_{hh',0} = q_{hh',1})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0,157	0,202	0,09	0,449
Mujer	0,193	0,247	0,11	0,55
TOTAL	0,35	0,449	0,2	0,999

Tras una única iteración la distribución univariante de las variables Estudios y Sexo ha quedado ajustada.

La tabla de contingencia muestral, ya ajustada que se obtiene es:

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	2	2	1	5
Mujer	2	3	1	6
TOTAL	4	5	2	11

Así mismo, la tabla de contingencia poblacional ajustada que obtenemos es:

	ESTUDIOS			
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	3	4	2	9
Mujer	4	5	2	11
TOTAL	7	9	4	20

Se obtiene la siguiente *tabla de pesos para cada individuo muestral*:

	ESTUDIOS		
SEXO	Primarios	Secundarios	Universitarios
Hombre	3	4	2
Mujer	1,33333333	1,66666667	0,66666667

Las tablas con los totales estimados sobre las celdas para cada variable objetivo son las siguientes:

	ESTUDIOS			$(\tilde{Y}_{1,hh'})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0	4	0	4
Mujer	4	3	0	7
TOTAL	4	7	0	11

	ESTUDIOS			$(\tilde{Y}_{2,hh'})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	225000	1000000	550000	1775000
Mujer	160000	366666,67	166666,67	693333,34
TOTAL	385000	1366666,67	716666,67	2468333,34

Obtenemos por lo tanto que:

$\hat{Y}_1$ =11 no-parados

$\hat{Y}_2$ =2468333,34 facturación total estimada

Así mismo, obtenemos que las media estimada de no-parados es 0,55 y de forma análoga, la media de la facturación estimada es 123416,66.

Hemos trabajado con 3 decimales, por lo que consideramos que la tolerancia es de 0,005. Con esta tolerancia, vemos que se consigue la convergencia del método tras esta primera iteración.

Redre

Dada la base de datos para los elementos de la muestra, junto distribuciones univariantes poblacionales, se operan estos datos mediante el procedimiento REDRE del paquete SPADN. Las tablas resultantes para las variables auxiliares son:

Tabla muestral ajustada:

	ESTUDIOS			$(\tilde{n}_{hh'})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	2	3	1	6
Mujer	2	3	1	6
TOTAL	4	6	2	12

Tabla poblacional ajustada = tabla anterior * 20/12

	ESTUDIOS			$(\tilde{N}_{hh'})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	4	5	2	11
Mujer	3	4	2	9
TOTAL	7	9	4	20

Tabla de pesos para cada elemento de la muestra según a qué estrato pertenece es:

	ESTUDIOS		
SEXO	Primarios	Secundarios	Universitarios
Hombre	4	5	2
Mujer	1	1,333	0,667

Con esto hemos obtenido ya la distribución muestral ponderada ajustada a la distribución poblacional marginal, que no es exacta, dado que hemos empleado una tolerancia de 0.001 y con un número máximo de iteraciones de 10.

Tabla ajustada a la población para la variable ser *no-parado*:

	ESTUDIOS			$(\tilde{Y}_{1,\,hh'})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	0	5	0	5
Mujer	3	3	0	6
TOTAL	3	8	0	11

Tabla de totales para la variable *total de facturación* ajustada a la población:

	ESTUDIOS			$(\tilde{Y}_{2,\,hh'})$
SEXO	Primarios	Secundarios	Universitarios	TOTAL
Hombre	300000	1250000	550000	2100000
Mujer	120000	293260	166750	580010
TOTAL	420000	1543260	716750	2680010

Las estimaciones resultantes para las dos variables objetivo que tenemos son:

$\hat{Y}_1$ =11 no-parados

$\hat{Y}_2$ =2680010 facturación total estimada

Así mismo, obtenemos que la media estimada de no-parados es 0,55 y de forma análoga, la media estimada de la facturación es 134000,5.

Procedimientos informáticos

Raking Usual

Los software informáticos conocidos para la aplicación del Raking son:

CALMAR, del INSEE, *Institut Nataional de la Statistique et des Etudes Economiques*. Es una macro programada en S.A.S., que aplica el Raking tanto para variables *cualitativas* como *cuantitativas*. Permite por lo tanto el ajuste de muestras a distintos niveles de individuos simultáneamente.

BASCULA, del NCBS, *Netherlands Central Bureau of Statistics*, software que aplica el Raking usual y por lo tanto únicamente para variables *cualitativas*.

Programa elaborado de manera experimental en FORTRAN en el EUSTAT, *Euskal Estatistika Erakundea*, cuya finalidad es el ajuste de muestras tanto para variables *cualitativas* y *cuantitativas*, así como la *actualización de datos* y el *ajuste a distintos niveles de individuos*.

Redre

Para la aplicación de la variación del RAKING estudiada en este capítulo existe un procedimiento informático incluido dentro del programa SPADN, el REDRE. SPADN es un paquete informático bajo WINDOWS y uno de los procedimientos informáticos que incluye es el REDRE. Éste permite el ajuste de muestras, pero se limita a variables *cualitativas*.

Situaciones óptimas para su aplicación

Se requiere el conocimiento de la *distribución poblacional univariante* de las variables auxiliares.

Estos procedimientos iterativos evitan el problema de celdas pequeñas que surgen cuando hay estratificaciones con muchas celdas. En estos casos, el número de efectivos en muchas de las celdas es próximo a cero, con lo que el elevador correspondiente es muy grande y sumado a esto, a pequeñas variaciones de la estratificación, y por lo tanto, en el número de elementos muestrales en cada una de las

celdas, la *variación del elevador* que a estas celdas se asigna es muy *grande*, es decir, el *estimador es inestable*. Para evitar esta inestabilidad, se evita la estratificación multivariante y se considera únicamente la distribución univariante de las variables auxiliares, realizándose el ajuste mediante alguno de los procedimientos iterativos en este capítulo mencionados. De esta forma, el estimador resultante es estable respecto de la estratificación muestral: el problema de celdas pequeñas desaparece.

Capítulo

4

Conclusiones y Propuestas

En este cuaderno hemos abordado el ajuste estadístico de muestras, considerando dos grandes bloques de estimadores:

Métodos utilizados con Máxima Información Auxiliar. Se obtienen elevadores celda a celda. Estos métodos los hemos considerado y analizado en dos situaciones: *estratificación* y *post-estratificación*. Cuando la información auxiliar es *cualitativa* los elevadores resultan del cociente entre el número de efectivos poblacionales por el número de efectivos muestrales de cada celda. En el caso de información auxiliar *cuantitativa*, se ha presentado el *Método de la Razón* (*Ratio*), en el que los elevadores resultan del cociente entre los totales poblacionales y muestrales de la variable auxiliar en las celdas obtenidas.

A la hora de utilizar estos métodos de ajuste a menudo se presenta el problema de celdas vacías o muy pequeñas: el tamaño muestral de las celdas se establece en un mínimo de 20 ó 25. Para evitar esto, se considera el *colapso de celdas* para todas aquellas que no superen un tamaño mínimo.

Procedimientos Iterativos, utilizados cuando se dispone únicamente de *Distribuciones Auxiliares Univariantes*. En este cuaderno hemos considerado este tipo de ajuste cuando las variables auxiliares son cualitativas, únicamente. Como métodos de ajuste se han presentado el Raking Usual y una variación de éste. Respecto a software informáticos el CALMAR, BASCULA y un programa en FORTRAN elaborado de forma experimental en EUSTAT, para la aplicación del Raking, y el REDRE, dentro del SPADN, para la aplicación de la variación del Raking.

Otra opción para situaciones en las que tenemos celdas muy pequeñas y celdas muy grandes al unísono es el *MÉTODO MIXTO*: (ver [13])

Se toman primero aquellas celdas donde el número de elementos muestrales es *muy grande* y se usa en éstas el método de *estratificación* o *post-estratificación*. Tras el ajuste, estas celdas se extraen de la tabla.

A continuación se aplica una *variación del RAKING* para las *celdas* restantes, el *Raking Acotado*: este método de ajuste acota la variación de los elevadores sobre ciertas celdas respecto de los elevadores iniciales. Su procedimiento es el siguiente:

- Para aquellas celdas en las que el número de elementos muestrales sea pequeño se acota la estimación del elevador correspondiente $\tilde{W}_{hh'}$ tal que no difiera demasiado del elevador inicial y después de obtener esta estimación se toma una estimación del total de elementos poblacionales en

dicha celda, sin más que hacer $\tilde{N}_{hh'} = \tilde{W}_{hh'} \bullet n_{hh'}$. Las celdas ajustadas se extraen de las tablas.

- Entonces se ajustan las observaciones restantes por RAKING, es decir, se ajustan a las distribuciones univariantes.

Respecto a este *método mixto*, hacer referencia a ciertas cuestiones:

La arbitrariedad a la hora de definir *celdas grandes* y un factor de ponderación o elevación que *varía demasiado*.

Una vez de aplicar en el método para las celdas grandes la estimación de la Razón, ¿por qué no se puede realizar el ajuste celda a celda con colapso para las celdas restantes?

¿*Converge* el *Raking Acotado*? Una condición que es seguro que hay que imponer es que el número de veces a aplicar el procedimiento no puede ser demasiado grande.

Todas estas cuestiones están todavía en fase de estudio y mejora. Por esta causa, no se conoce todavía ningún software para la aplicación de este método.

Como hemos visto, aún disponiendo de la distribución auxiliar multivariante, se plantean múltiples opciones y variaciones para el ajuste.

Capítulo

5

Bibliografía

[1] Alexander, C.H.

A Class of Methods for Using Person Controls in Household Weighting

Survey Methodology. December 1987 Vol.13 Nº2

[2] Bethlehem, J.G.& Keller, W.J.

Linear Weighting of Sample Survey Data

Journal of Official Statistics. 1987 Vol.3 Nº2

[3] Bishop, Y. M. M. & Fienberg, S.E.

Incomplete Two-Dimensional Contingency Tables

Biometrics. March 1969 pp. 119-128

[4] Cochran W. G.

Técnicas de muestreo. Segunda Edición 1981. Ed. Continental

[5] Copeland, K. R. , Pertzmeier, F.K. & Hoy, C.E.

An Alternative Method of Controlling Current Population Survey Estimates to Population Counts

Survey Methodology. December 1987 Vol.13 Nº2

[6] Deming, W.E. & Stephan, F.F.

On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known

Annals of Mathematical Statistics. 1940, Nº11 pp. 427-444

[7] Dupont, F.

Alternative Adjustments Where There Are Several Levels of Auxiliary Information

Survey Methodology. December 1995, Vol.21 Nº2 pp. 125-135

[8] Fuller, W.A., Loughin, M.M.& Baker, H.D.

Regression weighting in the Presence of Nonresponse

Survey Methodology. Junio 1994 Vol.20 Nº1

[9] Kalton, G.

Compensating for Missing Survey Data

Research Report Series. University of Michigan

[10] Ku, H. H. & Kullback, S.

Loglinear Models in Contingency Table Analysis

The American Statiscian. November 1974 Vol.28 Nº4

[11] Little & Rubin

Adjustment Cells

Statistical Analysis with Missing Data. Ed. Wiley 1987

[12] Niyorvenga, T.

Nonparametric Estimation of Response Probabilities and Sampling Theory

Survey Methodology. Vol.20 Nº2

[13] Oh, H.L. & Scheuren, F.

Modified Raking Ratio Estimation

Survey Methodology. December 1987, Vol.13 Nº2 pp. 209-219

[14] Rao, P.S.R.S.

Ratio Estimation with Subsampling the Nonrespondents

Survey Methodology. 1986 Vol.12 Nº2.

[15] Särndal, C.E.

A regression Approach to Estimation in the Presence of Nonresponse

Survey Methodology. 1986 Vol.12 Nº2

[16] Sitter, R. R. & Skinner, D.J.

Multi-way Stratification by Linear Programming

Survey Methodology. Junio 1994 Vol.20 Nº1

[17] Stephan, F.F.

An Iterative Method of Adjusting Sample Frequency Tables when Expected Marginal Totals are Known

Annals of Mathematical Statistics. Vol.13, pp. 166-178