UNIVERSIDAD DE GRANADA
FACULTAD DE CIENCIAS

APORTACIONES A LA TEORIA DE ESTIMADORES DE RAZON

MARIA DEL MAR RUEDA GARCIA

TESIS DOCTORAL

Sección de Matemáticas

APORTACIONES A LA TEORIA DE ESTIMADORES DE RAZON

Memoria que para optar al grado de Doctor en Ciencias, sección Matemáticas, presenta María del Mar Rueda García.

Vº Bº
Director de Tesis:

Prof. Dr. D. Andrés Gonzalez Carmona

Vº Bº
Director de Tesis:

Prof. Dr. D. Mariano Ruiz Espejo

DEPARTAMENTO DE ESTADISTICA E INVESTIGACION OPERATIVA

FACULTAD DE CIENCIAS

UNIVERSIDAD DE GRANADA

Deseo expresar mi sincero agradecimiento a los directores de este trabajo así como al Departamento de Estadística e Investigación Operativa, por el apoyo y ayuda recibidos.

Indice

Introducción.

El muestreo de una población está dirigido a obtener información acerca de alguna o algunas características de ésta a partir, no de la población entera sino de una parte de ella llamada muestra, debido a la imposibilidad de obtener información de todas las unidades poblacionales, ya sea por falta de recursos económicos, porque la población es excesivamente grande, o simplemente por conveniencia. Apoyándonos en las unidades muestrales debemos proponer estimadores sobre dichas características de forma que proporcionen la información más eficiente.

Los métodos de muestreo más conocidos y utilizados consideran estimadores que utilizan sólo los valores observados de la característica en estudio. Sin embargo es frecuente que la variable objeto de estudio, y, esté altamente relacionada con una característica auxiliar, x, cuyos datos están disponibles o son muy fáciles de obtener para todos los elementos de la población. En esta situación es muy útil considerar métodos de estimación que utilizan información auxiliar o suplementaria, relativa a una variable o característica correlacionada con la que es objeto de estudio, para modificar la forma de los estimadores directos o expandidos (los usuales cuando no existe dicha información suplementaria) consiguiendo estimadores más precisos que los calculados a partir de la muestra.

Como información suplementaria pueden utilizarse observaciones obtenidas con muestras grandes pero no probabilísticas; probabilísticas pero de tamaño excesivamente pequeño; o bien, observaciones obtenidas con muestras relativas a la población en estudio en fechas anteriores, o en último caso, estimaciones relativas a otra población diferente pero relacionada con la que se estudia.

Entre estos métodos, llamados métodos de estimación indirecta, hay dos especialmente importantes: el método de estimación de razón, y el de estimación de regresión.

La literatura de muestro de poblaciones finitas es abundante en ejemplos en los cuales estos métodos son utilizados para estimar medias y totales poblacionales. Al respecto, *Cochran* (1978) hace referencia a que ya en el año 1802, el procedimiento de *Laplace* para estimar la población de Francia utiliza un estimador de tipo razón. En la literatura estadística moderna estos procedimientos se han considerado en los últimos 50 años.

Es conocido que para un tamaño de muestra grande, el error cuadrático medio del estimador de razón es más pequeño que la varianza del estimador de expansión simple (si el coeficiente de correlación entre las variables es positivo y alto), pero mayor que el error cuadrático medio del estimador de regresión. No obstante el método de regresión es bastante complejo computacionalmente, especialmente en el caso de diseños de muestreo multietápicos. Según *Yates* (1960),"*El método de estimación de razón es más sencillo computacionalmente, pero el método de regresión es en ciertas circunstancias más acurado... Cuando la variable auxiliar x representa el tamaño de la unidad, la recta de regresión pasa por el orígen....*". En este último caso el método de razón resulta un estimador óptimo. La superioridad del estimador de razón frente al estimador de expansión simple y su simplicidad respecto al de regresión son dos de las razones por las que se ha dado una importancia especial al estudio del método de estimación de razón. Así los trabajos sobre este método son los más abundantes y muchos de los resultados obtenidos han sido posteriormente extendidos a otros métodos, de ahí que se ha elegido el estimador de razón como centro de nuestro estudio.

En el **capítulo 1**, después de introducir el estimador de razón como un tipo particular de estimadores indirectos, se procede al estudio del estimador de razón usual en el muestreo aleatorio simple. El estudio del estimador de razón usual es bastante complejo pues está definido como cociente de estimadores; así este estimador es en general sesgado, en contraposición con la mayor parte de los estimadores utilizados en la teoría clásica de muestreo en poblaciones finitas. Además, si bien se tiene el valor verdadero de este sesgo, en la práctica no se puede calcular en una muestra concreta, y se utiliza así una aproximación de este sesgo que sí permite su estimación.

El hecho de que el estimador sea sesgado, lleva a medir su precisión por su error cuadrático medio, en vez de por su varianza, como ocurre en la mayor parte de estimadores que usa la teoría de muestras, con la mayor complejidad que ello conlleva. Aquí se presentará la expresión verdadera de este error cuadrático medio, que no es funcional y las aproximaciones obtenidas que son

válidas para tamaños de muestra grandes.

Después de estudiadas las características generales del estimador de razón, se analizan los distintos intentos que han habido de mejorar este estimador. Estos intentos han venido por dos caminos: el primero es la formulación de nuevos estimadores cuyo sesgo sea menor que el del estimador de razón clásico y el segundo formulando estimadores cuyo error cuadrático medio sea más pequeño.

En el **capítulo 2** se estudia el método de estimación de razón bajo otros tipos de muestreo muy utilizados: el muestreo estratificado y el muestreo con probabilidades desiguales.

Se comienza con el estudio en el muestreo estratificado y se consideran en primer lugar dos estimadores ya conocidos: el estimador separado y el estimador combinado.

A continuación se define un posible estimador de la media en el muestreo estratificado. Este estimador que hemos llamado estimador de pesos óptimos, pondera cada media estratal por un peso que minimiza la varianza total del estimador. A partir de esta idea, proponemos un estimador teórico de razón bajo un esquema de muestreo estratificado, que utiliza los estimadores de pesos óptimos de las medias de las variables principal y auxiliar.

Existen situaciones en las que el diseñador de la encuesta posee alguna información acerca de las unidades poblacionales que le lleva a dar más importancia a algunas de ellas, asignandoles probabilidades de acuerdo a su importancia. Así, proponemos un estimador tipo razón, definido como el cociente de dos estimadores de *Hansen* y *Hurwitz*, bajo un esquema de muestreo con reemplazo y probabilidades desiguales. Se estudia su sesgo y error cuadrático medio, estableciendo las condiciones bajo las cuales es preferible al estimador de *Hansen* y *Hurwitz*, que no utiliza información auxiliar. También se propone una forma de definir las probabilidades de selección de cada unidad, de forma que haga insesgado al estimador.

Por último se aborda en el **capítulo 3** el problema de mejorar los estimadores mediante el uso de más de una variable auxiliar. En algunas ocasiones, el diseñador de la encuesta dispone de información acerca de varias variables auxiliares que están muy correlacionadas con la que es objeto de estudio. El primero que abordó el problema fué *Olkin* (1958), quien definió un estimador de razón multivariante bajo un esquema de muestreo aleatorio simple. Aquí se analiza también este problema bajo otros tipos de muestreo. Así se definen dos estimadores multivariantes en muestreo estratificado, y un estimador

multivariante en muestreo con probabilidades desiguales.

Posteriormente se proponen dos nuevas formas de utilizar la información suplementaria que porporciona más de una variable auxiliar.

La primera es la construcción de un nuevo estimador que llamamos "condensado" y que utiliza una cierta variable auxiliar que es combinación lineal de las variables auxiliares, para construir el estimador de razón. Esta variable auxiliar, "condensada", se determinará mediante dos procedimientos distintos:

1. maximizando la covarianza entre la variable condensada y la variable principal

2. minimizando el error cuadrático medio del estimador de razón obtenido.

dando lugar a dos nuevos estimadores, uno de los cuales tiene la propiedad de tener un error cuadrático medio igual al error cuadrático medio del estimador multivariante de *Olkin*.

Si además se exige que la línea de regresión de la variable de estudio, y, sobre la variable condensada pase por el origen, los dos estimadores obtenidos son insesgados y coinciden.

Todos los estimadores multivariantes considerados dependen de ciertos valores poblacionales desconocidos. El problema se solventa sustituyendo estos valores por los respectivos valores muestrales.

El último estimador multivariante que proponemos no tiene este inconveniente pues puede calcularse siempre al no depender de ningún valor desconocido. Este estimador, que llamamos iterado, se construye iterando el estimador de razón, utilizando cada vez una variable auxiliar y presentará grandes ventajas respecto a los antes considerados.

Capítulo 1

El estimador de razón en el muestreo aleatorio simple.

§1.1 Notación.

Llamamos:

y_i *al valor de la variable en estudio para la unidad i-ésima de la población.*

x_i *al valor de la variable auxiliar para la misma unidad.*

Y *al total poblacional de la variable* y.

X *al total poblacional de la variable* x.

N *al tamaño de la población.*

n *al tamaño de la muestra.*

$\overline{y}$ *a la media aritmética de los valores* y_i *en la muestra.*

$\overline{x}$ *a la media aritmética de los valores* x_i *en la muestra.*

$\overline{Y}$ *a la media aritmética de los valores* y_i.

$\overline{X}$ *a la media aritmética de los valores* x_i.

§1.2 Definición de los estimadores indirectos.

Supongamos que queremos estimar el total Y. Se define el estimador directo o expandido de la forma:

Definición 1.2.1 *Estimador directo o expandido del total*

$$\widehat{Y} = N\overline{y}$$

Además del estimador directo o expandido, caso $b_0 = 0$ en la expresión general

$$\widehat{Y}_G = \widehat{Y} + b_0\left(X - \widehat{X}\right) \tag{1.1}$$

en la que b_0 puede interpretarse como un coeficiente de corrección para mejorar el estimador, se definen los siguientes estimadores, como casos particulares de la expresión 1.1:

Definición 1.2.2 *Estimador de razón o por cociente del total*

$$\widehat{Y}_R = \widehat{Y} + \frac{\widehat{Y}}{\widehat{X}}\left(X - \widehat{X}\right) = \frac{\widehat{Y}}{\widehat{X}}X = N\overline{y}\frac{\overline{X}}{\overline{x}}$$

para $b_0 = \dfrac{\widehat{Y}}{\widehat{X}}$

Definición 1.2.3 *Estimador por diferencia del total*

$$\widehat{Y}_D = \widehat{Y} + X - \widehat{X}$$

para $b_0 = 1$.

Definición 1.2.4 *Estimador de regresión del total*

$$\widehat{Y}_{rg} = \widehat{Y} + b\left(X - \widehat{X}\right)$$

para $b_0 = b =$ *coeficiente de regresión de* y *sobre* x

Si interesa estimar la media, se obtienen las fórmulas correspondientes sustituyendo $\widehat{Y}$ y $\widehat{X}$ por los estimadores de las medias

$$\widehat{\overline{X}} = \overline{x} \; ; \quad \widehat{\overline{Y}} = \overline{y}$$

obteniéndose así:

Definición 1.2.5 *Estimador de razón o por cociente de la media*

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}} \overline{X}$$

Definición 1.2.6 *Estimador por diferencia de la media*

$$\widehat{\overline{Y}}_D = \overline{y} + \left(\overline{X} - \overline{x}\right)$$

Definición 1.2.7 *Estimador de regresión de la media*

$$\widehat{\overline{Y}}_{rg} = \overline{y} + b\left(\overline{X} - \overline{x}\right)$$

siendo b el coeficiente de regresión de y sobre x en la muestra.

§1.3 El estimador de razón.

El método de estimación de razón trata de mejorar la precisión del estimador simple, utilizando información sobre una variable auxiliar x, relacionada con la variable de estudio y.

En la práctica x suele ser el valor de y en una ocasión anterior en la que se hizo un censo completo.

Con frecuencia queremos estimar una razón entre dos variables, y no un total o una media. Definimos así:

Definición 1.3.8 *Estimador de la razón*

$$\widehat{R} = \frac{\overline{y}}{\overline{x}} = \frac{\widehat{Y}}{\widehat{X}}$$

Para el estimador de la razón no es necesario conocer el total X.

Vamos a comenzar el estudio de estas estimaciones en el caso de que las extracciones de las unidades sean sin reemplazo y con probabilidades iguales.

1.3.1 Sesgo.

Obviamente, el estimador de razón es consistente, por ser una función continua de estimadores consistentes, pero por ser cociente de dos estimadores, en general no es insesgado. Damos a continuación las expresiones exactas del sesgo, mediante las siguientes proposiciones:

Proposición 1.3.9 *El sesgo de los estimadores de razón de* R, $\overline{Y}$ *e* Y *viene dado por las expresiones:*

$$sesgo\left(\hat{R}\right) = -\frac{\mathrm{Cov}\left(\hat{R},\overline{x}\right)}{\overline{X}}$$

$$sesgo\left(\hat{\overline{Y}}_R\right) = -\mathrm{Cov}\left(\hat{R},\overline{x}\right)$$

$$sesgo\left(\hat{Y}_R\right) = -N\,\mathrm{Cov}\left(\hat{R},\overline{x}\right)$$

Demostración.-

Comenzamos con el estudio del estimador $\hat{R}$, siendo inmediata la extensión al caso de $\hat{\overline{Y}}_R$ e $\hat{Y}_R$.

Puesto que $\overline{y}$ y $\overline{x}$ son estimadores insesgados de $\overline{Y}$ y $\overline{X}$, respectivamente, podemos escribir:

$$R = \frac{\overline{Y}}{\overline{X}} = \frac{E\left(\overline{y}\right)}{E\left(\overline{x}\right)} = \frac{E\left(\hat{R}\overline{x}\right)}{E\left(\overline{x}\right)}$$

y por tanto

$$sesgo\left(\hat{R}\right) = E\left(\hat{R}\right) - R = E\left(\hat{R}\right) - \frac{E\left(\hat{R}\overline{x}\right)}{E\left(\overline{x}\right)} =$$

$$= \frac{E\left(\hat{R}\right)E\left(\overline{x}\right) - E\left(\hat{R}\overline{x}\right)}{E\left(\overline{x}\right)} = -\frac{\mathrm{Cov}\left(\hat{R},\overline{x}\right)}{\overline{X}}$$

Como $\hat{\overline{Y}}_R = \hat{R}\overline{X}$ e $\hat{Y}_R = N\hat{\overline{Y}}_R$, se tiene

$$sesgo\left(\hat{\overline{Y}}_R\right) = E\left(\hat{\overline{Y}}_R\right) - \overline{Y} = \overline{X}E\left(\hat{R}\right) - R\overline{X} = \overline{X}\,sesgo\left(\hat{R}\right) = -\mathrm{Cov}\left(\hat{R},\overline{x}\right)$$

$$sesgo\left(\widehat{Y}_R\right) = N\,sesgo\left(\widehat{\overline{Y}}_R\right) = -N\,\mathrm{Cov}\left(\widehat{R},\overline{x}\right)$$

Este resultado debido a *Hartley* y *Ross* (1954) nos permite deducir una cota superior para el estimador de razón.

Corolario 1.3.10 *Una cota superior del cociente entre el sesgo y la desviación típica de los estimadores de razón* $\widehat{R}$, $\widehat{\overline{Y}}_R$ *e* $\widehat{Y}_R$ *viene dada por el coeficiente de variación de* $\overline{x}$

Demostración.-
Puesto que

$$\left|\mathrm{Cov}\left(\widehat{R},\overline{x}\right)\right| \leq \sigma_{\widehat{R}}\sigma_{\overline{x}}$$

(donde por σ_a denotamos la desviación típica de la variable a) en virtud de la proposición 1.3.9, se tiene:

$$\left|sesgo\left(\widehat{R}\right)\right| \leq \frac{\sigma_{\widehat{R}}\sigma_{\overline{x}}}{\overline{X}} = \sigma_{\widehat{R}}C_{\overline{x}}$$

siendo $C_{\overline{x}}$ el coeficiente de variación de $\overline{x}$ y por tanto

$$\frac{\left|sesgo\left(\widehat{R}\right)\right|}{\sigma_{\widehat{R}}} \leq C_{\overline{x}}$$

Entonces si el coeficiente de variación de $\overline{x}$ es pequeño, es decir si no hay demasiada distorsión de la muestra para la variable auxiliar, el sesgo del estimador es despreciable en comparación con su desviación típica. (Investigaciones empíricas debidas a *Kish, Namboodiri* y *Pillai* (1962), prueban que $\dfrac{\left|sesgo\left(\widehat{R}\right)\right|}{\sigma_{\widehat{R}}}$ es pequeño a menos que el tamaño muestral sea muy pequeño).
El mismo límite se aplica al sesgo de los estimadores del total y media.
Puesto que

$$sesgo\left(\widehat{\overline{Y}}_R\right) = \overline{X}\,sesgo\left(\widehat{R}\right)\ ;\ \ sesgo\left(\widehat{Y}_R\right) = X\,sesgo\left(\widehat{R}\right)$$

$$\sigma_{\widehat{Y}_R} = \sigma_{\widehat{R}}X\ ;\ \ \sigma_{\widehat{\overline{Y}}_R} = \sigma_{\widehat{R}}\overline{X}$$

es inmediato que

$$\frac{\left|sesgo\left(\widehat{\overline{Y}}_R\right)\right|}{\sigma_{\widehat{\overline{Y}}_R}} = \frac{\left|sesgo\left(\widehat{Y}_R\right)\right|}{\sigma_{\widehat{Y}_R}} \leq C_{\overline{x}}$$

Corolario 1.3.11 *En un muestreo aleatorio simple, el estimador de razón es insesgado si*

$$\mathrm{Cov}\left(\widehat{R}, \overline{x}\right) = 0$$

Demostración.-
Inmediata sin más que considerar el resultado de la proposición 1.3.9.

Proposición 1.3.12 *El sesgo del estimador de razón puede expresarse, alternativamente, de la forma siguiente:*

$$sesgo\left(\widehat{\overline{Y}}_R\right) = \sum_{i=1}^{N}\left(T_i^{(1)} - \frac{1}{N}\right) y_i$$

donde

$$T_i^{(1)} = \frac{\overline{X}}{n\binom{N}{n}} \sum_{\substack{s \in S \\ i \in s}} \frac{1}{\overline{x}_s}$$

donde s denota una muestra de tamaño n y S el espacio muestral.

Demostración.-

$$E\left(\widehat{\overline{Y}}_R\right) = \frac{\overline{X}}{\binom{N}{n}} \sum_{s \in S} \frac{\overline{y}_s}{\overline{x}_s} = \frac{\overline{X}}{n\binom{N}{n}} \sum_{s \in S} \frac{\sum_{i \in s} y_i}{\overline{x}_s} =$$

$$= \frac{\overline{X}}{n\binom{N}{n}} \sum_{i=1}^{N} y_i \sum_{\substack{s \in S \\ i \in s}} \frac{1}{\overline{x}_s} = \sum_{i=1}^{N} T_i^{(1)} y_i \qquad (1.2)$$

Entonces el sesgo del estimador viene dado por

$$sesgo\left(\widehat{\overline{Y}}_R\right) = \sum_{i=1}^{N} T_i^{(1)} y_i - \overline{Y} = \sum_{i=1}^{N}\left(T_i^{(1)} - \frac{1}{N}\right) y_i$$

Las expresiones del sesgo de los otros dos estimadores son:

$$sesgo\left(\widehat{Y}_R\right) = N\sum_{i=1}^{N}\left(T_i^{(1)} - \frac{1}{N}\right)y_i$$

$$sesgo\left(\widehat{R}\right) = \frac{1}{\overline{X}}\sum_{i=1}^{N}\left(T_i^{(1)} - \frac{1}{N}\right)y_i$$

Estas expresiones obtenidas por *Rao* (1967) son complejas puesto que los coeficientes $T_i^{(1)}$ envuelven sumatorias sobre todos las $\binom{N}{n}$ muestras posibles de tamaño n, por lo que las expresiones tienen un interés limitado.

Por ello son más útiles expresiones aproximadas del sesgo, que damos a continuación, con el grado de aproximación deseado.

Proposición 1.3.13 *En un muestreo aleatorio simple, primeras aproximaciones de los sesgos de los estimadores de razón* $\widehat{R}$, $\widehat{Y}_R$ *e* $\widehat{\overline{Y}}_R$ *son:*

$$sesgo_1\left(\widehat{R}\right) = \frac{1-f}{n}R\left(C_x^2 - \rho C_x C_y\right)$$

$$sesgo_1\left(\widehat{Y}_R\right) = \frac{1-f}{n}Y\left(C_x^2 - \rho C_x C_y\right)$$

$$sesgo_1\left(\widehat{\overline{Y}}_R\right) = \frac{1-f}{n}\overline{Y}\left(C_x^2 - \rho C_x C_y\right)$$

Demostración.-
Escribimos $\widehat{R}$ de la forma:

$$\widehat{R} = \frac{\overline{y}}{\overline{x}} = \frac{\overline{Y}}{\overline{X}}\frac{\frac{\overline{y}}{\overline{Y}}}{\frac{\overline{x}}{\overline{X}}} = R\frac{1+e_1}{1+e_2}$$

donde

$$e_1 = \frac{\overline{y}-\overline{Y}}{\overline{y}} \; ; \quad e_2 = \frac{\overline{x}-\overline{X}}{\overline{x}}$$

Así

$$\hat{R} - R = R\left[(1+e_1)(1+e_2)^{-1} - 1\right] \tag{1.3}$$

desarrollando $(1+e_2)^{-1}$ en serie de potencias de e_2, se obtiene:

$$\hat{R} - R = R\left[(1+e_1)\left(1 - e_2 + e_2^2 + \cdots\right) - 1\right] \tag{1.4}$$

Este desarrollo es válido si $|e_2| < 1$ para todas las $\binom{N}{n}$ posibles muestras de tamaño n. Esta condición no es difícil de satisfacer excepto en algunas ocasiones especiales. *David & Sukhatme* (1974), presentan una justificación del uso de estas aproximaciones para el sesgo y la varianza.

Si cortamos el desarrollo reteniendo sólo los términos de orden inferior o igual a dos en e_1 y e_2, se obtiene:

$$\hat{R} - R \simeq R\left(e_1 - e_2 - e_1 e_2 + e_2^2\right)$$

y tomando esperanzas, se tiene:

$$sesgo\left(\hat{R}\right) = E\left(\hat{R}\right) - R \simeq R\left[E(e_1) - E(e_2) - E(e_1 e_2) + E\left(e_2^2\right)\right]$$

Ahora bien, $E(e_1) = E(e_2) = 0$ por ser $\overline{y}$ y $\overline{x}$ insesgados y

$$E\left(e_2^2\right) = \frac{1}{\overline{X}^2} E\left(\overline{x} - \overline{X}\right)^2 = \frac{V(\overline{x})}{\overline{X}^2} = \frac{1-f}{n\overline{X}^2} S_x^2$$

$$E(e_1 e_2) = \frac{1}{\overline{X}\,\overline{Y}} E\left[\left(\overline{y} - \overline{Y}\right)\left(\overline{x} - \overline{X}\right)\right] = \frac{1}{\overline{X}\,\overline{Y}} \mathrm{Cov}(\overline{y}, \overline{x}) = \frac{1-f}{n\overline{X}\overline{Y}} \rho S_y S_x$$

donde S_x^2 y S_y^2 son las casivarianzas poblacionales de las variables x e y, respectivamente.

Por tanto el término principal del sesgo viene dado por:

$$sesgo_1\left(\hat{R}\right) = R\left(\frac{1-f}{n\overline{X}^2} S_x^2 - \frac{1-f}{n\overline{X}\,\overline{Y}} \rho S_y S_x\right) = \frac{1-f}{n\overline{X}^2}\left(RS_x^2 - \rho S_x S_y\right)$$

o bien en función de los coeficientes de variación:

$$sesgo_1\left(\widehat{R}\right) = \frac{1-f}{n} R\left(C_x^2 - \rho C_x C_y\right)$$

donde $C_x = \frac{S_x}{\overline{X}}$ y $C_y = \frac{S_y}{\overline{Y}}$.

Entonces es fácil deducir que:

$$sesgo_1\left(\widehat{Y}_R\right) = \frac{1-f}{n} Y\left(C_x^2 - \rho C_x C_y\right)$$

$$sesgo_1\left(\widehat{\overline{Y}}_R\right) = \frac{1-f}{n} \overline{Y}\left(C_x^2 - \rho C_x C_y\right)$$

Si partimos de la expresión 1.4 y consideramos los términos de orden inferior o igual a 4, podemos obtener una aproximación mejor del sesgo de $\widehat{R}$.

Proposición 1.3.14 *Una segunda aproximación del sesgo del estimador $\widehat{R}$ viene dada por la expresión:*

$$sesgo_2\left(\widehat{R}\right) = \frac{N-n}{N-1}\,\frac{R}{n}\left[\left(\frac{\mu_{02}}{\mu_{01}^2} - \frac{\mu_{11}}{\mu_{10}\mu_{01}}\right) + \right.$$

$$\frac{(N-n)(N-2n)}{(N-1)(N-2)}\,\frac{1}{n^2}\left(\frac{\mu_{12}}{\mu_{10}\mu_{01}^2} - \frac{\mu_{03}}{\mu_{01}^3}\right) +$$

$$+\frac{(N-n)(N^2-6Nn+N+6n^2)}{(N-1)(N-2)(N-3)}\,\frac{1}{n^3}\left(\frac{\mu_{04}}{\mu_{01}^4} - \frac{\mu_{13}}{\mu_{10}\mu_{01}^3}\right) +$$

$$\left. + \frac{N(N-n)(N-n-1)3(n-1)}{(N-1)(N-2)(N-3)}\,\frac{1}{n^3}\left(\frac{\mu_{02}^2}{\mu_{01}^4} - \frac{\mu_{11}\mu_{02}}{\mu_{10}\mu_{01}^3}\right)\right]$$

donde

$$\mu_{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^{\alpha}\left(x_i - \overline{Y}\right)^{\beta}$$

Además

$$sesgo_2\left(\widehat{Y}_R\right) = X\, sesgo_2\left(\widehat{R}\right)$$

$$sesgo_2\left(\widehat{\overline{Y}}_R\right) = \overline{X}\, sesgo_2\left(\widehat{R}\right)$$

Demostración.-

Partiendo del desarrollo 1.4, cortamos en los términos de orden 4 en e_1 y e_2 y al tomar esperanzas nos queda:

$$sesgo_2\left(\widehat{R}\right) \simeq R\left[E\left(e_2^2\right) - E\left(e_1 e_2\right) + E\left(e_2^2 e_1\right) - E\left(e_2^3\right) + E\left(e_2^4\right) - E\left(e_2^3 e_1\right)\right].$$

Las esperanzas de estos productos fueron tabuladas por *Sukhatme* (1964). Usando sus resultados se obtiene la expresión dada en la proposición.

Para dar una idea de esta magnitud, supongamos que N es grande y además la población sigue una distribución normal bivariante, entonces:

$$\mu_{12} = \mu_{21} = 0 \; ; \quad \mu_{03} = \mu_{30} = 0 \; ; \quad \mu_{04} = 3\mu_{02}^2 \; ; \quad \mu_{13} = 3\mu_{11}\mu_{02}$$

por lo que tenemos:

$$sesgo_2\left(\widehat{R}\right) \simeq \frac{1}{n}\left(\frac{\mu_{02}}{\mu_{01}^2} - \frac{\mu_{11}}{\mu_{10}\mu_{01}}\right) + \frac{3}{n^2}\left(\frac{\mu_{02}}{\mu_{01}^2} - \frac{\mu_{11}}{\mu_{10}\mu_{01}}\right)\frac{\mu_{02}}{\mu_{01}^2} =$$

$$= \frac{R}{n}\left(C_x^2 - \rho C_y C_x\right)\left(1 + \frac{3}{n}C_x^2\right)$$

En esta segunda aproximación el sesgo relativo del estimador de razón en poblaciones grandes puede aproximarse así:

$$sesgo_2\left(\widehat{R}\right) \simeq sesgo_1\left(\widehat{R}\right)\left(1 + \frac{3}{n}C_x^2\right) \tag{1.5}$$

donde hemos notado con $sesgo_1\left(\widehat{R}\right)$ a la primera aproximación del sesgo dada por la proposición 1.3.13.

La ecuación 1.5 muestra que la contribución de los términos de grado 3 al sesgo relativo del estimador de razón es $\frac{3}{n}C_x^2$ veces el valor de éste en la primera aproximación. A menos que n sea pequeño,la contribución puede considerarse despreciable. *Ayachit* (1953) obtuvo el valor de la contribución del sesgo para sucesivas aproximaciones por medio de muestreo experimental en una amplia gama de poblaciones comúnmente utilizadas en encuestas y comprobó que la contribución de los términos de orden superior es despreciable. Para n grande (30 o mayor) incluso el término más bajo no tiene consecuencias.

Teorema 1.3.15 *Si la regresión de y sobre x es lineal y pasa por el origen, en una población finita, los estimadores* $\hat{Y}_R$, $\hat{\overline{Y}}_R$ *y* $\hat{R}$ *no tienen sesgo en muestras aleatorias simples de tamaño n.*

Demostración.-
Escribimos:

$$y_i = \beta x_i + e_i \quad i = 1, \ldots, N$$

Si es cierta la hipótesis del teorema, se tiene que

$$\overline{Y} = \beta \overline{X} \text{ y } R = \beta$$

y por otra parte la suma de los e_i para las unidades que tienen el mismo valor de x es cero.

Entonces

$$E\left(\hat{R}\right) = \sum_{s \in S} \hat{R}_s P(s) = \frac{1}{\binom{N}{n}} \sum_{(2)} \sum_{(1)} \frac{\overline{y}_s}{\overline{x}_s}$$

donde la sumatoria de todas las muestras se realiza en dos etapas: primero se suman las muestras que contienen el mismo grupo de valores de x (por tanto $\overline{x}_s$ permanece constante en estas muestras) y segundo se suman los grupos de muestras que contienen valores distintos de x.

Así

$$E\left(\hat{R}\right) = \frac{1}{\binom{N}{n}} \sum_{(2)} \sum_{(1)} \left(\beta + \frac{\overline{e}_s}{\overline{x}_s}\right) =$$

$$= \frac{1}{\binom{N}{n}} \sum_{(2)} \sum_{(1)} \beta + \frac{1}{\binom{N}{n}} \sum_{(2)} \sum_{(1)} \frac{\overline{e}_s}{\overline{x}_s} =$$

$$= \beta + \frac{1}{\binom{N}{n}} \sum_{(2)} \frac{1}{\overline{x}_s} \sum_{(1)} \overline{e}_s =$$

Pero $\sum_{(1)} \overline{e}_s = 0$ puesto que todas las muestras de esta sumatoria tienen las mismos valores de x y por tanto se concluye

$$E\left(\hat{R}\right) = \beta = R$$

1.3.2 Error cuadrático medio.

Como se ha comprobado en el apartado anterior, el estimador de razón es en general sesgado, por lo que su precisión se medirá por su error cuadrático medio, que procedemos a estudiar.

Este estudio presenta bastantes problemas puesto que tanto x como y varían de muestra a muestra. La expresión exacta fue obtenida por *Rao* (1967) y viene dada por la siguiente proposición.

Proposición 1.3.16 *En un muestreo aleatorio simple, el estimador de razón* $\widehat{\overline{Y}}_R$ *tiene un error cuadrático medio dado por la expresión:*

$$ECM\left(\widehat{\overline{Y}}_R\right) = \sum_{i=1}^{N} T_i^{(2)} y_i^2 + \sum_{i \neq j}^{N} T_{ij}^{(2)} y_i y_j - 2\overline{Y} \sum_{i=1}^{N} T_i^{(1)} y_i + \overline{Y}^2$$

donde

$$T_i^{(2)} = \frac{\overline{X}^2}{n^2 \binom{N}{n}} \sum_{\substack{s \in S \\ i \in s}} \frac{1}{\overline{x}_s^2}$$

$$T_{ij}^{(2)} = \frac{\overline{X}^2}{n^2 \binom{N}{n}} \sum_{\substack{s \in S \\ i,j \in s}} \frac{1}{\overline{x}_s^2}$$

Demostración.-

$$ECM\left(\widehat{\overline{Y}}_R\right) = E\left(\widehat{\overline{Y}}_R - \overline{Y}\right) = E\left(\widehat{\overline{Y}}_R^2\right) - 2\overline{Y}E\left(\widehat{\overline{Y}}_R\right) + \overline{Y}^2 \qquad (1.6)$$

Ahora bien

$$E\left(\widehat{\overline{Y}}_R^2\right) = \overline{X}^2 E\left(\hat{R}^2\right) = \frac{\overline{X}^2}{\binom{N}{n}} \sum_{s \in S} \frac{\overline{y}_s^2}{\overline{x}_s^2} = \frac{\overline{X}^2}{n^2 \binom{N}{n}} \sum_{s \in S} \frac{\left(\sum_{i \in s} y_i\right)^2}{\overline{x}_s^2} =$$

$$= \frac{\overline{X}^2}{n^2\binom{N}{n}} \left\{ \sum_{i=1}^{N} y_i^2 \sum_{\substack{s \in S \\ i \in s}} \frac{1}{\overline{x}_s^2} + \sum_{i \neq j} y_i y_j \sum_{\substack{s \in S \\ i,j \in S}} \frac{1}{\overline{x}_s^2} \right\} = \sum_{i=1}^{N} T_i^{(2)} y_i^2 + \sum_{i \neq j}^{N} T_{ij}^{(2)} y_i y_j$$

Además, como demostramos en 1.2, $E\left(\widehat{\overline{Y}}_R\right) = \sum_{i=1}^{N} T_i^{(1)} y_i$ y por tanto sustituyendo en 1.6 obtenemos la expresión deseada:

$$ECM\left(\widehat{\overline{Y}}_R\right) = \sum_{i=1}^{N} T_i^{(2)} y_i^2 + \sum_{i \neq j}^{N} T_{ij}^{(2)} y_i y_j - 2\overline{Y} \sum_{i=1}^{N} T_i^{(1)} y_i + \overline{Y}^2$$

La expresión del error cuadrático medio de los estimadores $\widehat{R}$ e $\widehat{Y}_R$ se obtiene de forma inmediata:

$$ECM\left(\widehat{R}\right) = \frac{1}{\overline{X}^2} \sum_{i=1}^{N} T_i^{(2)} y_i^2 + \sum_{i \neq j}^{N} T_{ij}^{(2)} y_i y_j - 2\overline{Y} \sum_{i=1}^{N} T_i^{(1)} y_i + \overline{Y}^2$$

$$ECM\left(\widehat{Y}_R\right) = N^2 \sum_{i=1}^{N} T_i^{(2)} y_i^2 + \sum_{i \neq j}^{N} T_{ij}^{(2)} y_i y_j - 2\overline{Y} \sum_{i=1}^{N} T_i^{(1)} y_i + \overline{Y}^2$$

Estas expresiones, al igual que ocurre con la expresión exacta del sesgo, tienen el inconveniente de depender de los coeficientes $T_i^{(1)}$, $T_i^{(2)}$ y $T_{ij}^{(2)}$, que envuelven sumatorias sobre todas las $\binom{N}{n}$ muestras posibles.

Sin embargo se tienen aproximaciones del error cuadrático medio, válidas para muestras suficientemente grandes, que son más sencillas de calcular y pueden utilizarse en un problema real.

Proposición 1.3.17 *Una primera aproximación del error cuadrático medio de los estimadores de razón $\widehat{R}$, $\widehat{\overline{Y}}_R$ e $\widehat{Y}_R$ vienen dadas por las expresiones:*

$$ECM_1\left(\widehat{R}\right) = \frac{1-f}{n} R^2 \left(C_y^2 + C_x^2 - 2\rho C_y C_x\right)$$

$$ECM_1\left(\widehat{\overline{Y}}_R\right) = \frac{1-f}{n} \overline{Y}^2 \left(C_y^2 + C_x^2 - 2\rho C_y C_x\right)$$

$$ECM_1\left(\widehat{Y}_R\right) = \frac{1-f}{n} Y^2 \left(C_y^2 + C_x^2 - 2\rho C_y C_x\right)$$

Demostración.-
Como vimos en 1.3

$$\widehat{R} - R = R\left[(1+e_1)(1-e_2)^{-1} - 1\right] =$$
$$= R\left[(1+e_1)\left(1 - e_2 + e_2^2 - e_2^3 + \cdots\right) - 1\right] =$$
$$= R\left(-e_2 + e_2^2 - e_2^3 + e_1 - e_1e_2 + e_1e_2^2 - \cdots\right)$$

elevando al cuadrado

$$\left(\widehat{R} - R\right)^2 = R^2\left(-e_2 + e_2^2 - e_2^3 + e_1 - e_1e_2 + e_1e_2^2 - \cdots\right)^2 \qquad (1.7)$$

Reteniendo sólo los términos de potencias menores o iguales a 2 en e_1 y e_2 y tomando esperanzas obtenemos:

$$E\left(\widehat{R} - R\right)^2 = R^2E\left(e_1^2 + e_2^2 - e_1e_2\right) = R^2\left[E\left(e_1^2\right) + E\left(e_2^2\right) - 2E\left(e_1e_2\right)\right] =$$

$$= R^2\left[\frac{V(\overline{y})}{\overline{Y}^2} + \frac{V(\overline{x})}{\overline{X}^2} - 2\frac{\mathrm{Cov}(\overline{x},\overline{y})}{\overline{X}\,\overline{Y}}\right] = R^2\left[\frac{1-f}{n}\left(\frac{S_y^2}{\overline{Y}^2} + \frac{S_x^2}{\overline{X}^2} - 2\frac{\rho S_xS_y}{\overline{X}\,\overline{Y}}\right)\right]$$

Entonces

$$ECM_1\left(\widehat{R}\right) = \frac{1-f}{n}R^2\left(C_y^2 + C_x^2 - 2\rho C_yC_x\right)$$

y así

$$ECM_1\left(\widehat{\overline{Y}}_R\right) = ECM_1\left(\widehat{R}\right)\overline{X}^2 = \frac{1-f}{n}\overline{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_yC_x\right)$$

$$ECM_1\left(\widehat{Y}_R\right) = ECM_1\left(\widehat{R}\right)X^2 = \frac{1-f}{n}Y^2\left(C_y^2 + C_x^2 - 2\rho C_yC_x\right)$$

Corolario 1.3.18 *Una primera aproximación de los coeficientes de variación de los estimadores* $\widehat{R}$, $\widehat{\overline{Y}}_R$ *e* $\widehat{Y}_R$ *viene dada por:*

$$CV_1 = \left[\frac{1-f}{n}\left(C_y^2 + C_x^2 - 2C_{yx}\right)\right]^{\frac{1}{2}}$$

Demostración.-

$$CV_1\left(\hat{R}\right) = \sqrt{\frac{ECM_1\left(\hat{R}\right)}{\hat{R}^2}} = \left[\frac{1-f}{n}\left(C_y^2 + C_x^2 - 2C_{yx}\right)\right]^{\frac{1}{2}}$$

siendo el mismo para los otros dos estimadores.

Corolario 1.3.19 *Una expresión alternativa de la primera aproximación del error cuadrático medio del estimador de razón viene dada por*

$$ECM_1\left(\hat{R}\right) = \frac{1-f}{n\overline{X}^2}\frac{\sum_{i=1}^{N}\left(y_i - Rx_i\right)^2}{N-1}$$

Demostración.-

$$ECM_1\left(\hat{R}\right) = \frac{1-f}{n}R^2\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right) =$$

$$= \frac{1-f}{n}\left(R^2\frac{S_y^2}{\overline{Y}^2} + R^2\frac{S_x^2}{\overline{X}^2} - 2R\rho\frac{S_y S_x}{\overline{X}^2}\right) =$$

$$= \frac{1-f}{n}\frac{1}{\overline{X}^2}\left(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x\right) =$$

$$= \frac{1-f}{n\overline{X}^2}\frac{1}{N-1}\left(\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 + R^2\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2 - \right.$$

$$\left. -2R\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right)\right) =$$

$$= \frac{1-f}{n\overline{X}^2}\frac{1}{N-1}\sum_{i=1}^{N}\left(\left(y_i - \overline{Y}\right) - R\left(x_i - \overline{X}\right)\right)^2 = \frac{1-f}{n\overline{X}^2}\frac{\sum_{i=1}^{N}\left(y_i - Rx_i\right)^2}{N-1}$$

Las fórmulas obtenidas del error cuadrático medio son buenas aproximaciones si las muestras son de tamaño grande. Una regla aproximada de lo

grande que ha de ser el tamaño de muestra, n, para que sean aceptables estas aproximaciones es que n sea lo sufientemente grande para que $C_{\overline{x}} < 0.05$ (Una justificación teórica de este resultado puede verse en *Hansen, Hurwitz & Madow* (1953), Vol. 2).

Corolario 1.3.20 *En un muestreo aleatorio simple, el sesgo del estimador de razón decrece más rápidamente que el error estándar y con un tamaño muestral moderadamente grande, el sesgo de $\widehat{R}$ se vuelve despreciable en relación a su error estándar.*

Demostración.-

$$sesgo_1\left(\widehat{R}\right) = \frac{1-f}{n} R\left(C_x^2 - \rho C_x C_y\right)$$

$$ECM_1\left(\widehat{R}\right) = \frac{1-f}{n} R^2\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right)$$

Entonces

$$\frac{sesgo_1^2\left(\widehat{R}\right)}{ECM_1\left(\widehat{R}\right)} = \frac{1-f}{n} \frac{\left(C_x^2 - \rho C_x C_y\right)^2}{\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right)}$$

$$\frac{sesgo}{\sigma_{\widehat{R}}} = \sqrt{\frac{1-f}{n}} \sqrt{\frac{\left(C_x^2 - \rho C_x C_y\right)^2}{\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right)}} \tag{1.8}$$

que decrece a medida que n aumenta.

La expresión 1.8 muestra que la razón del sesgo al error es de orden $O\left(n^{-\frac{1}{2}}\right)$ y por lo tanto el sesgo es prácticamente despreciable en relación con el error estándar para tamaños de muestra moderadamente grandes.

Proposición 1.3.21 *Una segunda aproximación del error cuadrático medio del estimador de razón $\widehat{R}$ viene dada por:*

$$ECM_2\left(\widehat{R}\right) = R^2 \frac{N-n}{N-1} \frac{1}{n} K$$

donde notamos K por:

$$K = \left(\frac{\mu_{20}}{\mu_{10}^2} + \frac{\mu_{02}}{\mu_{01}^2} - 2\frac{\mu_{11}}{\mu_{01}\mu_{10}}\right) +$$

$$+\frac{2(N-n)(N-2n)}{(N-1)(N-2)n^2}\left(2\frac{\mu_{12}}{\mu_{10}\mu_{01}^2}-\frac{\mu_{21}}{\mu_{10}^2\mu_{01}}-\frac{\mu_{03}}{\mu_{01}^3}\right)+$$

$$+\frac{3(N-n)(N^2+N-6nN+6n^2)}{(N-1)(N-2)(N-3)n^3}\left(\frac{\mu_{22}}{\mu_{10}^2\mu_{01}^2}-2\frac{\mu_{13}}{\mu_{10}\mu_{01}^3}+\frac{\mu_{04}}{\mu_{01}^4}\right)+$$

$$+\frac{3(n-1)N(N-n)(N-n-1)}{(N-1)(N-2)(N-3)n^3}\left(\frac{\mu_{20}\mu_{02}+2\mu_{11}^2}{\mu_{10}^2\mu_{01}^2}-\frac{6\mu_{11}\mu_{02}}{\mu_{10}\mu_{01}^3}+\frac{3\mu_{02}^2}{\mu_{01}^4}\right)$$

Demostración.-

Consideramos le expresión 1.7 y cortamos el desarrollo quedándonos con los términos en e_1 y e_2 de grado inferior o igual a cuatro. Al tomar esperanzas se tiene:

$$ECM\left(\widehat{R}\right)=R^2E\left(e_2^2+e_1^2-2e_1e_2+e_2^4+e_1^2e_2^2-\right.$$

$$\left.-2e_2^3+2e_2^4+2e_1e_2^2-2e_1e_2^3+2e_2^2e_1-2e_1e_2^3-2e_1e_2^3+2e_1^2e_2^2-2e_1^2e_2\right)=$$

$$=R^2E\left(e_1^2+e_2^2-2e_1e_2-2e_2^3+3e_2^4+3e_1^2e_2^2+4e_1e_2^2-6e_1e_2^3-2e_1^2e_2\right)$$

$$ECM_2\left(\widehat{R}\right)=R^2\left(E\left(e_1^2\right)+E\left(e_2^2\right)-2E\left(e_1e_2\right)-2E\left(e_2^3\right)+\right.$$

$$\left.+2E\left(e_2^4\right)+3E\left(e_1^2e_2^2\right)+4E\left(e_1e_2^2\right)-6E\left(e_1e_2^3\right)-2E\left(e_1^2e_2\right)\right)$$

Ahora bien, se tiene:

$$E\left(e_1^2\right)=\frac{N-n}{N-1}\frac{1}{n}\left(\frac{\mu_{20}}{\mu_{10}^2}\right)\ ;\ E\left(e_2^2\right)=\frac{N-n}{N-1}\frac{1}{n}\left(\frac{\mu_{02}}{\mu_{01}^2}\right)$$

$$E\left(e_1e_2\right)=\frac{N-n}{N-1}\frac{1}{n}\left(\frac{\mu_{11}}{\mu_{01}\mu_{10}}\right)\ ;\ E\left(e_2^3\right)=\frac{(N-n)(N-2n)}{(N-1)(N-2)n^2}\left(\frac{\mu_{03}}{\mu_{01}^3}\right)$$

$$E\left(e_1 e_2^2\right) = \frac{(N-n)(N-2n)}{(N-1)(N-2)n^2}\left(\frac{\mu_{12}}{\mu_{10}\mu_{01}^2}\right)$$

$$E\left(e_1^2 e_2\right) = \frac{(N-n)(N-2n)}{(N-1)(N-2)n^2}\left(\frac{\mu_{21}}{\mu_{10}^2\mu_{01}}\right)$$

$$\begin{aligned} E\left(e_2^4\right) &= \frac{N-n}{n^3\mu_{01}^4}\left[\frac{N^2+N-6nN+6n^2}{(N-1)(N-2)(N-3)}\mu_{04} + \right. \\ &+ \left. \frac{3(n-1)(N-n-1)}{(N-1)(N-2)(N-3)}N\mu_{02}^2\right] \end{aligned}$$

$$\begin{aligned} E\left(e_1 e_2^3\right) &= \frac{N-n}{n^3\mu_{10}\mu_{01}^3}\left[\frac{N^2+N-6nN+6n^2}{(N-1)(N-2)(N-3)}\mu_{13} + \right. \\ &+ \left. \frac{3(n-1)(N-n-1)}{(N-1)(N-2)(N-3)}N\mu_{11}\mu_{02}\right] \end{aligned}$$

$$\begin{aligned} E\left(e_1^2 e_2^2\right) &= \frac{N-n}{n^3\mu_{10}^2\mu_{01}^2}\left[\frac{N^2+N-6nN+6n^2}{(N-1)(N-2)(N-3)}\mu_{22} + \right. \\ &+ \left. \frac{(n-1)(N-n-1)}{(N-1)(N-2)(N-3)}N\left(\mu_{20}\mu_{02}+2\mu_{11}^2\right)\right] \end{aligned}$$

y sustituyendo estas expresiones se obtiene la fórmula deseada.

Es inmediato que

$$ECM_2\left(\widehat{\overline{Y}}_R\right) = \overline{Y}^2\frac{N-n}{N-1}\frac{1}{n}K$$

$$ECM_2\left(\widehat{Y}_R\right) = Y^2\frac{N-n}{N-1}\frac{1}{n}K$$

Precisión de las aproximaciones del sesgo y el error cuadrático medio.

Hemos obtenido las aproximaciones tanto del sesgo como del error cuadrático medio de los estimadores $\hat{R}$, $\widehat{\overline{Y}}_R$ e $\hat{Y}_R$. Es importante determinar la precisión de estas aproximaciones, que haremos a continuación, y daremos una cota para la diferencia entre las aproximaciones y los valores verdaderos. Para ello utilizaremos el siguiente lema propuesto por *David & Sukhatme* (1974).

Este lema utiliza el concepto de "superpoblación". La idea de las superpoblaciones se basa en considerar que los valores de la población, y_i, $i = 1, \ldots, N$, en vez de ser constantes, son valores obtenidos de una muestra extraída a partir de una "superpoblación" de N variables aleatorias independientes "y_i", $i = 1, \ldots, N$ que tienen una distribución de probabilidad conjunta ξ.

En este trabajo se seguirá el procedimiento clásico caracterizado por el uso de diseños muestrales como base para la inferencia acerca de los parámetros poblacionales. Ocasionalmente utilizaremos los modelos de superpoblaciones. Estos modelos datan de *Cochran*, (1946), y fueron desarrollados principalmente por *Richard Royall* y su escuela. Hoy día existe una gran controversia sobre la utilización de estos modelos.

Lema 1.3.22 *Sea una sucesión de poblaciones $\{S_n\}$ de tamaños $\{N_n\}$ extraídas mediante muestreo aleatorio simple de una superpoblación infinita $S = \{U_i,\ V_i\}$. Sea $\{N_n\}$ estrictamente creciente y tal que*

$$\lim_{n\to\infty} t_n = \lim_{n\to\infty} \frac{n}{N_n} = t \quad 0 < t < 1 \ y\ 0 < t_n < 1\ \forall n$$

Suponiendo que para algunos enteros positivos fijados r y s, los momentos de orden $r+s$ e inferiores de S_n permanecen acotados. Sean $\overline{u}$ y $\overline{v}$ las medias muestrales basadas en una muestra de tamaño n extraída de S_n mediante muestreo aleatorio simple. Entonces si $\overline{U}$ y $\overline{V}$ son las medias poblacionales para S_n, se tiene

$$E\left[\left(\overline{u}-\overline{U}\right)^r\left(\overline{v}-\overline{V}\right)^s\right] = \begin{cases} O\left(n^{-\frac{r+s}{2}}\right) & si\ r+s\ es\ par \\ O\left(n^{-\frac{r+s+1}{2}}\right) & si\ r+s\ es\ impar \end{cases}$$

Teorema 1.3.23 *Las aproximaciones del sesgo y del error cuadrático medio del estimador $\hat{R}$ dadas por $sesgo_1\left(\hat{R}\right)$ y $ECM_1\left(\hat{R}\right)$ son de orden $O\left(n^{-2}\right)$ y las de $sesgo_2\left(\hat{R}\right)$ y $ECM_2\left(\hat{R}\right)$ son de orden $O\left(n^{-3}\right)$.*

Demostración.-
Para la determinación de las aproximaciones, partimos de la expresión:

$$\hat{R} - R = R\left[(1+e_1)(1+e_2)^{-1} - 1\right] = R\left[(e_1 - e_2)\left(1 - e_2 + e_2^2 + \cdots\right)\right] \quad (1.9)$$

Veamos en primer lugar las aproximaciones de sesgo.
La expresión exacta del sesgo es:

$$sesgo\left(\hat{R}\right) = B\left(\hat{R}\right) = R\overline{X}E\left[\frac{e_1 - e_2}{\overline{x}}\right] = \overline{Y}E\left[\frac{e_1 - e_2}{\overline{x}}\right]$$

Consideramos la aproximación de orden k del sesgo:

$$B_k\left(\hat{R}\right) = R\sum_{i=0}^{2k-1} E\left[(-1)^i e_2^i (e_1 - e_2)\right] \quad (1.10)$$

obtenida a partir de la expresión 1.9 cortando en el término $2k$ y tomando esperanzas (para $k=1$ y $k=2$ ya estudiadas con $sesgo_1\left(\hat{R}\right)$ y $sesgo_2\left(\hat{R}\right)$).

Puesto que los términos de la sumatoria son de la forma $E\left(e_2^{i+1}\right)$ y $E\left(e_2^i e_1\right)$ podemos aplicar el lema 1.3.22 según el cual $B_k\left(\hat{R}\right)$ contiene los términos de e_1 y e_2 hasta los de orden $O\left(n^{-k}\right)$.
Por otra parte:

$$B\left(\hat{R}\right) - B_k\left(\hat{R}\right) = E\left[\overline{Y}\frac{e_1 - e_2}{\overline{x}}\right] - R\sum_{i=0}^{2k-1} E\left[(-1)^i e_2^i (e_1 - e_2)\right] =$$

$$= \overline{Y}E\left[\frac{e_1 - e_2}{\overline{x}} - \sum_{i=0}^{2k-1}(-1)^i \frac{e_2^i}{\overline{x}}(e_1 - e_2)\frac{\overline{x}}{\overline{X}}\right] =$$

$$= \overline{Y}E\left[\frac{e_1 - e_2}{\overline{x}} - \sum_{i=0}^{2k-1}(-1)^i \frac{e_2^i}{\overline{x}}(e_1 - e_2)(e_2 + 1)\right] =$$

$$= \overline{Y}E\left[\frac{e_1 - e_2}{\overline{x}} - \sum_{i=0}^{2k-1}(-1)^i \frac{e_2^{i+1}}{\overline{x}}(e_1 - e_2) - \sum_{i=0}^{2k-1}(-1)^i \frac{e_2^i}{\overline{x}}(e_1 - e_2)\right] =$$

$$= \overline{Y}E\left[\frac{e_1 - e_2}{\overline{x}} - \sum_{i=1}^{2k}(-1)^{i-1} \frac{e_2^i}{\overline{x}}(e_1 - e_2) - \sum_{i=0}^{2k-1}(-1)^i \frac{e_2^i}{\overline{x}}(e_1 - e_2)\right] =$$

$$= \overline{Y} E\left[\frac{e_2^{2k}}{\overline{x}}\left(e_1 - e_2\right)\right]$$

Entonces si la variable x es positiva y x_0 es una cota inferior positiva de $\overline{x}$, se tiene:

$$\left|B\left(\hat{R}\right) - B_k\left(\hat{R}\right)\right| \leq \left|\frac{\overline{Y}}{x_0} E\left(e_1 e_2^{2k} - e_2^{2k+1}\right)\right| \tag{1.11}$$

que según el lema 1.3.22 es de orden $O\left(n^{-(k+1)}\right)$.

Así, para una elección apropiada de n y k, el término $\left|B\left(\hat{R}\right) - B_k\left(\hat{R}\right)\right|$ puede hacerse muy pequeño en comparación con $B_k\left(\hat{R}\right)$ y éste será una buena aproximación de $B\left(\hat{R}\right)$.

En segundo lugar, veamos las aproximaciones del error cuadrático medio.

La expresión exacta del error cuadrático medio es:

$$ECM\left(\hat{R}\right) = \overline{Y}^2 E\left[\frac{\left(e_1 - e_2\right)^2}{\overline{x}^2}\right]$$

y la aproximación *k-ésima*:

$$ECM_k\left(\hat{R}\right) = R^2 E\left[\left(e_1 - e_2\right)^2 \sum_{i=0}^{2k-2} (-1)^i e_2^i (i+1)\right] \tag{1.12}$$

Entonces:

$$ECM\left(\hat{R}\right) - ECM_k\left(\hat{R}\right) =$$

$$= \overline{Y}^2 E\left[\frac{\left(e_1 - e_2\right)^2}{\overline{x}^2} - \frac{\left(e_1 - e_2\right)^2}{\overline{x}^2} \sum_{i=0}^{2k-2} (-1)^i e_2^i (i+1) \frac{\overline{x}^2}{\overline{X}^2}\right] =$$

$$= \overline{Y}^2 E\left[\frac{\left(e_1 - e_2\right)^2}{\overline{x}^2}\left(1 - \sum_{i=0}^{2k-2} (-1)^i e_2^i (i+1)\left(1 + e_2\right)^2\right)\right] =$$

$$= \overline{Y}^2 E\left[\frac{\left(e_1 - e_2\right)^2}{\overline{x}^2}\left(1 - \sum_{i=0}^{2k-2} (-1)^i (i+1) e_2^{i+2} -\right.\right.$$

$$- \sum_{i=0}^{2k-2} (-1)^i (i+1) e_2^i - 2 \sum_{i=0}^{2k-2} (-1)^i (i+1) e_2^{i+1} \Big) \Big]$$

y simplificando los sumatorios

$$ECM\left(\widehat{R}\right) - ECM_k\left(\widehat{R}\right) = -\overline{Y}^2 E\left[\frac{(e_1 - e_2)^2}{\overline{x}^2}\left(2ke_2^{2k-1} + (2k-1)e_2^{2k}\right)\right]$$

Así:

$$\left|ECM\left(\widehat{R}\right) - ECM_k\left(\widehat{R}\right)\right| \leq \left|\frac{\overline{Y}^2}{x_0^2} E\left[(e_1 - e_2)^2\left(2ke_2^{2k-1} + (2k-1)e_2^{2k}\right)\right]\right|$$

que es de orden $O\left(n^{-(k+1)}\right)$.

Del mismo modo que en el caso del sesgo, para una elección apropiada de n y k, $ECM_k\left(\widehat{R}\right)$ es una buena aproximación de $ECM\left(\widehat{R}\right)$.

Estos resultados en el caso $k = 1$ y $k = 2$ prueban el teorema 1.3.23.

Corolario 1.3.24 *Cotas para la diferencia entre los valores verdaderos y sus primeras aproximaciones del sesgo y del error cuadrático medio vienen dadas por:*

$$\left|B\left(\widehat{R}\right) - B_1\left(\widehat{R}\right)\right| \leq \left|\frac{\overline{Y}}{x_0}\frac{(N-n)(N-2n)}{(N-1)(N-2)n^2}\left(\frac{\mu_{12}}{\mu_{10}\mu_{01}^2} - \frac{\mu_{03}}{\mu_{01}^3}\right)\right|$$

y

$$\left|ECM\left(\widehat{R}\right) - ECM_1\left(\widehat{R}\right)\right| \leq$$

$$\leq \frac{\overline{Y}^2}{x_0^2}\left|\frac{2(N-n)(N-2n)}{(N-1)(N-2)n^2}\left(\frac{\mu_{03}}{\mu_{01}^3} - 2\frac{\mu_{12}}{\mu_{10}\mu_{01}^2} + \frac{\mu_{21}}{\mu_{10}^2\mu_{01}}\right) +\right.$$

$$+\frac{(n-1)N(N-n)(N-n-1)}{(N-1)(N-2)(N-3)n^3}\left(3\frac{\mu_{02}}{\mu_{01}^4} - 6\frac{\mu_{02}}{\mu_{01}^2}\frac{\mu_{11}}{\mu_{10}\mu_{01}} + \frac{\mu_{02}}{\mu_{01}^2}\frac{\mu_{20}}{\mu_{10}^2} + 2\frac{\mu_{11}^2}{\mu_{10}^2\mu_{01}^2}\right) +$$

$$\left.+\frac{(N-n)(N^2+N-6nN+6n^2)}{(N-1)(N-2)(N-3)n^3}\left(\frac{\mu_{04}}{\mu_{01}^4} - 2\frac{\mu_{13}}{\mu_{10}\mu_{01}^3} + \frac{\mu_{22}}{\mu_{10}^2\mu_{01}^2}\right)\right|$$

Demostración.-
Según la expresión 1.11 para $k = 1$:

$$\left|B\left(\hat{R}\right) - B_1\left(\hat{R}\right)\right| \leq \left|\frac{\overline{Y}}{x_0} E\left(e_1 e_2^2 - e_2^3\right)\right|$$

Sustituyendo los valores de estas esperanzas, se tiene:

$$\left|B\left(\hat{R}\right) - B_1\left(\hat{R}\right)\right| \leq$$

$$\leq \left|\frac{\overline{Y}}{x_0}\left(\frac{(N-n)(N-2n)}{(N-1)(N-2)n^2}\frac{\mu_{12}}{\mu_{10}\mu_{01}^2} - \frac{(N-n)(N-2n)}{(N-1)(N-2)n^2}\frac{\mu_{03}}{\mu_{01}^3}\right)\right| =$$

$$= \left|\frac{\overline{Y}}{x_0}\frac{(N-n)(N-2n)}{(N-1)(N-2)n^2}\left(\frac{\mu_{12}}{\mu_{10}\mu_{01}^2} - \frac{\mu_{03}}{\mu_{01}^3}\right)\right|$$

En cuanto al error cuadrático medio, se tiene:

$$\left|ECM\left(\hat{R}\right) - ECM_1\left(\hat{R}\right)\right| \leq \frac{\overline{Y}^2}{x_0^2}\left|E\left[\left(e_1 - e_2\right)^2\left(2e_2 + e_2^2\right)\right]\right|$$

Entonces

$$\left|ECM\left(\hat{R}\right) - ECM_1\left(\hat{R}\right)\right| \leq \frac{\overline{Y}^2}{x_0^2}\left|E\left(2e_2e_1^2 + 2e_2^3 - 4e_1e_2^2 + e_1^2e_2^2 + e_2^4 - 2e_1e_2^3\right)\right|$$

Sustituyendo los valores de las esperanzas de estos productos y simplificando se obtiene el resultado deseado.

1.3.3 Comparación del estimador de razón con el de expansión simple.

El estimador de Y que se obtiene sin utilizar la información suplementaria es $N\overline{y}$, donde $\overline{y}$ es la media muestral por unidad (en el muestreo aleatorio simple) o una media ponderada por unidad (en el muestreo aleatorio estratificado).

Estimadores de esta clase se conocen con el nombre de estimadores obtenidos por expansión simple.

A continuación vamos a comparar el error cuadrático medio del estimador de razón y la varianza del estimador de expansión simple en muestreo aleatorio simple.

Teorema 1.3.25 *Si el tamaño muestral n es suficientemente grande para que los términos de orden* $O(n^{-2})$ *puedan ser ignorados, el estimador de razón* $\widehat{\overline{Y}}_R$ *es más eficiente que el estimador* $\overline{y}$ *si*

$$\rho \frac{C_y}{C_x} > \frac{1}{2}$$

Demostración.-

Como ya sabemos en el muestreo aleatorio simple el estimador del total $\widehat{Y} = N\overline{y}$ tiene varianza:

$$V\left(\widehat{Y}\right) = N^2 \frac{1-f}{n} S_y^2$$

Para el estimador de la razón hemos visto que una aproximación de orden $O(n^{-2})$ es:

$$ECM\left(\widehat{Y}_R\right) = N^2 \frac{1-f}{n}\left(S_y^2 + R^2 S_x^2 - 2\rho R S_x S_y\right)$$

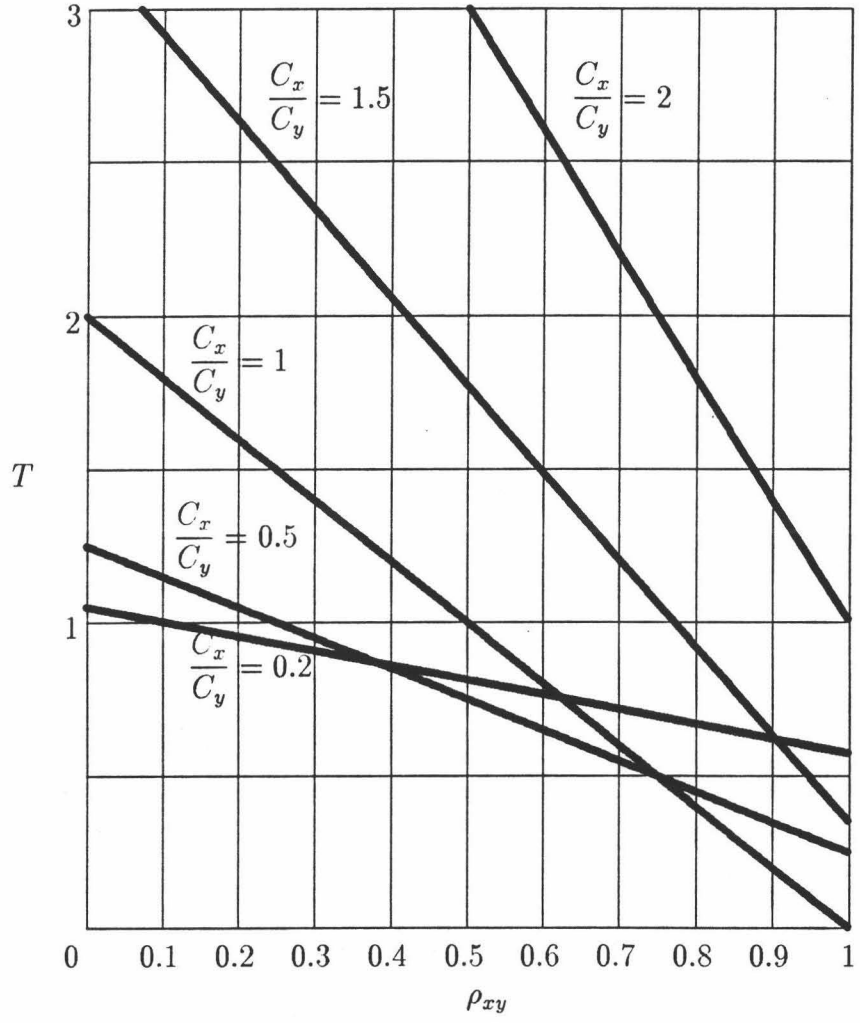
La eficiencia relativa de un estimador comparado con otro es el cociente de sus varianzas, así:

$$Eficiencia = \frac{S_y^2}{S_y^2 + R^2 S_x^2 - 2\rho R S_x S_y} = \frac{1}{1 + \frac{C_x^2}{C_y^2} - 2\rho \frac{C_x}{C_y}}$$

Por tanto para muestras grandes, el estimador de razón es más eficiente si el denominador es menor que 1, es decir

$$\frac{C_x^2}{C_y^2} < 2\rho \frac{C_x}{C_y} \; ; \quad \rho \frac{C_y}{C_x} > \frac{1}{2}$$

La figura de la página siguiente muestra, para valores diferentes de $\frac{C_x}{C_y}$, los valores de $T = 1 + \frac{C_x^2}{C_y^2} - 2\rho \frac{C_x}{C_y}$ con respecto a los valores que toma ρ.



A partir de esta figura es claro que cuando $\frac{C_x}{C_y} = 1$, el estimador de la razón tendrá una varianza menor que el estimador expandido si la correlación entre x_i e y_i es mayor que 0.5 y tendrá varianza mayor si es menor que 0.5. Más aún, el estimador de razón tiene varianza sustancialmente menor si la correlación es alta, y en este caso su varianza es cero si la correlación es perfecta.

Si $\frac{C_x}{C_y} > 2$ se perdería precisión si se utiliza el estimador de la razón en lugar del estimador expandido, incluso con correlación perfecta.

Con valores menores de éste cociente, habrá una ganancia utilizando el el estimador de razón siempre que la correlación sea lo suficientemente alta.

Si $\frac{C_x}{C_y}$ es pequeño, obtenemos ganancia utilizando el estimador de razón incluso con muy baja correlación.

Pero si $\frac{C_x}{C_y}$ es pequeño, las ganancias no son muy importantes incluso con alta correlación. Como ejemplo, fijémosnos en $\frac{C_x}{C_y} = 0.2$.

Si $\rho = 0$ se pierde precisión, pero ésta perdida es pequeña; sin embargo, si $\rho = 1$, hay ganancia, pero sólo reducimos la varianza del estimador expandido en la tercera parte.

Por tanto, si $\frac{C_x}{C_y}$ es pequeño se puede tomar cualquiera de los dos estimadores aunque si la correlación es alta sería conveniente adoptar el estimador de la razón y si es excesivamente baja el estimador expandido.

Todo esto, como ya hemos dicho, tiene el inconveniente de que la muestra ha de ser suficientemente grande. En muestras más pequeñas el método de la razón no se puede comparar tan bien como en el caso anterior ya que la fórmula aproximada de la varianza es normalmente una subestimación.

1.3.4 Intervalos de confianza.

La distribución exacta del estimador de razón no puede expresarse de una forma sencilla. Para construir intervalos de confianza se pueden utilizar dos procedimientos, dependiendo del tamaño de la muestra:

1. Para muestras grandes (según se verá en apartado siguiente) la distribución del cociente $\frac{\hat{R}}{R}$ puede considerarse normal, con error estándar dado por la expresión:

$$K = \frac{1}{\sqrt{n}} \left(C_x^2 - 2\rho C_x C_y + C_y^2 \right)^{\frac{1}{2}}$$

y por tanto

$$P\left(\frac{\hat{R}}{R} - t_{\alpha,\infty}K \leq 1 \leq \frac{\hat{R}}{R} + t_{\alpha,\infty}K\right) = 1 - \alpha$$

obteniéndose los siguientes límites de confianza para R:

$$\left(\frac{\hat{R}}{1 - t_{\alpha,\infty}K}, \frac{\hat{R}}{1 + t_{\alpha,\infty}K}\right)$$

siendo $t_{\alpha,\infty}$ el valor de la variable N(0,1) (o de una t de Student con $n-1$ grados de libertad) que deja por encima de él un área de $\frac{\alpha}{2}$.

2. Cuando las muestras son pequeñas es aconsejable utilizar el siguiente método debido a *Fieller* (1932):

   Sea (x_i, y_i) una distribución normal bivariante. Si consideramos la variable:

$$u_i = y_i - Rx_i$$

está distribuida normalmente con media cero y varianza:

$$V = S_y^2 - 2RS_{xy} + R^2S_x^2$$

Entonces se tiene que la variable

$$\overline{u} = \overline{y} - R\overline{x}$$

se distribuye también según una normal con media cero y varianza $\frac{V}{n}$.

Un estimador insesgado de V se obtiene a partir de

$$V = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \overline{u})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((y_i - Rx_i) - (\overline{y} - R\overline{x}))^2 =$$

$$= s_y^2 - 2Rs_{xy} + R^2s_x^2$$

Los límites de confianza para R con un coeficiente de confianza de $(1-\alpha)$ están determinados por las dos raíces de una forma cuadrática en R dada por:

$$t_{\alpha,n-1} = \frac{\sqrt{n}\,(\overline{y} - R\overline{x})}{\left(s_y^2 - 2Rs_{xy} + R^2 s_x^2\right)^{\frac{1}{2}}} \tag{1.13}$$

Reemplazando s_{xy} por $rs_x s_y$, donde r es el coeficiente de correlación muestral y usando que $c_x = \frac{s_x}{\overline{x}}$ y $c_y = \frac{s_y}{\overline{y}}$, la solución de la ecuación 1.13 da los siguientes valores para los límites de confianza de R:

$$\frac{\hat{R}}{1 - t_{\alpha,n-1}\frac{c_x^2}{n}}\left[\left(1 - \frac{t_{\alpha,n-1}^2}{n} r c_x c_y\right) \pm \right.$$

$$\left. \pm \frac{t_{\alpha,n-1}}{\sqrt{n}}\left(\left(c_x^2 - 2rc_x c_y + c_y^2\right) - \frac{t_{\alpha,n-1}^2}{n} c_x^2 c_y^2 \left(1 - r^2\right)\right)^{\frac{1}{2}}\right]$$

1.3.5 Distribución asintótica del estimador de razón.

En este apartado vamos a especificar las condiciones bajo las cuales el Teorema Central del Límite para poblaciones finitas es cierto para el estimador de razón bajo muestreo aleatorio simple.

Siguiendo la formulación usual del Teorema Central del Límite para poblaciones finitas, vamos a considerar nuestra población finita U como una sucesión de poblaciones $\{U_\nu\}$ donde n_ν y N_ν tienden a infinito de forma que $N_\nu - n_\nu$ también tienda a infinito cuando ν lo haga.

Para estudiar la distribución asintótica de $\overline{y}_R$, necesitamos utilizar los siguientes teoremas:

Teorema 1.3.26 *(Hájek, 1960) Suponiendo que $n_\nu \to \infty$ y $N_\nu - n_\nu \to \infty$ para $\nu \to \infty$, entonces para un muestreo aleatorio simple*

$$\sqrt{n_\nu}\left(\overline{y}_\nu - \overline{Y}_\nu\right) \xrightarrow{L} N(0,1) \quad \textit{cuando} \quad \nu \to \infty$$

sí y sólo si $\left\{Y_{\nu_j}\right\}$ satisface la condición de Lindeberg–Hájek:

$$\lim_{\nu\to\infty} \sum_{T_\nu(\delta)} \frac{\left(Y_{\nu_j} - \overline{Y}_\nu\right)^2}{(N_\nu - 1)\, S_{\nu y}^2} = 0\ , \quad \forall \delta > 0 \tag{1.14}$$

donde $T_\nu(\delta)$ es el conjunto de unidades de U_ν para las cuales

$$\frac{\left|Y_{\nu_j} - \overline{Y}_\nu\right|}{\sqrt{1-f_\nu}S_{\nu y}} > \delta\sqrt{n_\nu}$$

Teorema 1.3.27 *Supongamos $\{Y_{\nu_j}\}$ satisfaciendo la condición:*

$$(1-f_\nu)\frac{S^2_{\nu y}}{n_\nu} \to 0 \quad cuando \quad \nu \to \infty \tag{1.15}$$

Entonces, bajo un muestreo aleatorio simple

1. $E\left|\overline{y}_\nu - \overline{Y}_\nu\right| \to 0.$

2. $\overline{y}_\nu - \overline{Y}_\nu \xrightarrow{P} 0$ *si* $\nu \to \infty.$

Consideramos ahora el estimador de razón. Llamamos

$$R_{\nu_j} = Y_{\nu_j} - b_\nu X_{\nu_j}$$

donde $b_\nu = \dfrac{\overline{Y}_\nu}{\overline{X}_\nu}$.

Los valores R_{ν_j} tienen media poblacional $\overline{R}_\nu = 0$ y varianza

$$S^2_{\nu R} = \sum_{j=1}^{N_\nu} \frac{R^2_{\nu_j}}{N_\nu - 1}$$

Entonces:

Teorema 1.3.28 *Sea $\hat{\overline{Y}}_{\nu R} = \overline{X}_\nu \dfrac{\overline{y}_\nu}{\overline{x}_\nu}$. Bajo un muestreo aleatorio simple, se tiene:*

$$\frac{\sqrt{n_\nu}\left(\hat{\overline{Y}}_{\nu R} - \overline{Y}_\nu\right)}{\sqrt{1-f_\nu}S_{\nu R}} \xrightarrow{L} N(0,1) \text{ si } \nu \to \infty$$

siempre que $\{R_{\nu_j}\}$ satisfaga la condición de Lindeberg-Hájek, (1.14) y $\{\dfrac{X_{\nu_j}}{\overline{X}_\nu}\}$ satisfaga la condición 1.15.

Demostración.-
Podemos escribir:

$$\hat{\overline{Y}}_{\nu R} - \overline{Y}_{\nu} = (\overline{y}_{\nu} - b_{\nu}\overline{x}_{\nu})\frac{\overline{X}_{\nu}}{\overline{x}_{\nu}} = \overline{r}_{\nu}\frac{\overline{X}_{\nu}}{\overline{x}_{\nu}}$$

donde

$$\overline{r}_{\nu} = \sum_{i \in S} \frac{R_{\nu_i}}{n_{\nu}}$$

Puesto que $\{R_{\nu_j}\}$ satisface 1.14 y $\overline{R}_{\nu} = 0$, se sigue del teorema 1.3.26 que

$$\frac{\sqrt{n_{\nu}}\,\overline{r}_{\nu}}{\sqrt{1-f_{\nu}}S_{\nu R}} \xrightarrow{L} N(0,1) \quad \text{cuando} \quad \nu \to \infty$$

Por otra parte, puesto que $\dfrac{X_{\nu_j}}{\overline{X}_{\nu}}$ satisface la condición 1.15, se tiene:

$$\left|\frac{\overline{x}_{\nu}}{\overline{X}_{\nu}} - \frac{\overline{X}_{\nu}}{\overline{X}_{\nu}}\right| \xrightarrow{P} 0 \quad \text{cuando} \quad \nu \to \infty$$

según el teorema 1.3.27, por lo que $\dfrac{\overline{x}_{\nu}}{\overline{X}_{\nu}} \xrightarrow{P} 1$, y así, según las propiedades de la convergencia:

$$\frac{\sqrt{n_{\nu}}\left(\hat{\overline{Y}}_{\nu R} - \overline{Y}_{\nu}\right)}{\sqrt{1-f_{\nu}}S_{\nu R}} \xrightarrow{L} N(0,1) \text{ cuando } \nu \to \infty$$

Los resultados asintóticos para poblaciones finitas son bastante artificiales puesto que sólo hay realmente una población. Sin embargo, estos resultados son útiles para determinar las condiciones bajo las cuales se puede trabajar con las aproximaciones normales.

Scott y *Wu* (1981), determinan las condiciones para que el teorema 1.3.28 sea también cierto al sustituir $S_{\nu R}$ por el valor muestral $s_{\nu R}$.

§1.4 Estimadores de tipo razón.

Una vez comprobada la eficiencia (bajo ciertas condiciones) del estimador de razón frente al de expansión simple, diversos autores han considerado el problema de construir otros estimadores utilizando los cocientes de las variables

principal y auxiliar, que "mejoren" el estimador de razón clásico. Estos intentos han ido encaminados bien a disminuir el sesgo o bien a disminuir el error cuadrático medio. Aquí se van a exponer algunos de los resultados más interesantes.

1.4.1 Disminución del sesgo: Estimadores insesgados y cuasi–insesgados.

Como hemos visto en la proposición 1.3.9, el estimador de razón en el muestreo aleatorio simple es en general sesgado. Este sesgo es despreciable si el tamaño de la muestra es grande, pero puede ser considerable su efecto si el tamaño muestral es muy pequeño.

De ahí que en los últimos años está tomando un gran interés el desarrollo de estimadores del tipo razón que sean insesgados o sujetos a un sesgo inferior al estimador de razón ordinario.

Estos estimadores suelen utilizarse en muestreos con muchos estratos y muestras pequeñas en cada estrato, si el estimador de razón separado no parece apropiado para ellas.

Aquí se van a considerar algunos de los más importantes.

El estimador de *Hartley* y *Ross*.

El estimador $\overline{y}$ es insesgado de $\overline{Y}$ en muestreo aleatorio simple, y hemos visto que puede ser mejorado usando la información auxiliar que se posee de una variable x correlacionada con y.

El cociente $\dfrac{\overline{X}}{\overline{x}}$ mide cómo de representativa es la muestra obtenida. Si es menor que uno, la muestra revela un predominio de unidades cuyo valor de x es superior a la media. Si es mayor que uno ocurre lo contrario. Entonces se usa dicho coeficiente para mejorar el estimador insesgado $\overline{y}$ dando lugar al estimador de razón $\overline{y}_R$.

Se puede retener la idea que subyace en el estimador de razón en orden a estudiar otros estimadores por medio del grado de representación de cada unidad de la muestra considerada individualmente. Se puede llegar así a un estimador *media de razones*:

$$\overline{r} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i} = \frac{1}{n}\sum_{i=1}^{n} r_i \; ; \quad r_i = \frac{y_i}{x_i}$$

$$\overline{y}_{MR} = \overline{r}\ \overline{X}$$

Este estimador media de razones es sesgado, pero a partir de él se pueden construir otros estimadores insesgados que se conocen con el nombre de *estimadores insesgados del tipo razón.*

Un estimador, debido a *Hartley* y *Ross* (1954) se obtiene partiendo de la media $\overline{r}$ de las razones $\frac{y}{x}$ y el estimador $\overline{r}$ corregido su sesgo es:

$$\overline{r}' = \overline{r} + \frac{n(N-1)}{(n-1)N\overline{X}}(\overline{y} - \overline{r}\ \overline{x})$$

El correspondiente estimador insesgado de $\hat{Y}$ es:

$$\overline{r}'X = \overline{r}X + \frac{n(N-1)}{(n-1)}(\overline{y} - \overline{r}\ \overline{x})$$

y el de la media $\overline{Y}$:

$$\overline{y}_{HR} = \overline{r}'\ \overline{X} = \overline{r}\ \overline{X} + \frac{n(N-1)}{(n-1)N}(\overline{y} - \overline{r}\ \overline{x})$$

La varianza exacta de este estimador fué obtenida por *Robson* en 1957. Nosotros vamos a considerar una aproximación de ella válida para muestras grandes:

$$V(\overline{r}'X) = \frac{N^2}{n}\sum_{i=1}^{N}\frac{\left(y_i - \overline{Y} - \overline{r}_N\left(x_i - \overline{X}\right)\right)^2}{N-1}$$

donde $\overline{r}_N$ es la media poblacional de r_i.

Comparándola con la varianza del estimador de razón usual:

$$V\left(\hat{Y}_R\right) = \frac{N^2}{n}\sum_{i=1}^{N}\frac{(y_i - Rx_i)^2}{N-1}$$

Entonces $\overline{r}'X$ es más preciso que $\hat{Y}_R$ en muestras grandes si la línea $\overline{Y} + \overline{r}_N\left(x_i - \overline{X}\right)$ se ajusta a los valores de y_i mejor que la recta Rx_i, o lo que es equivalente, si el coeficiente de regresión β está más próximo a $\overline{r}_N$ que a R. Este resultado fué obtenido por *Goodman* y *Hartley* (1958).

Las condiciones bajo las cuales el estimador insesgado de *Hartley* y *Ross* es más eficiente que el estimador de razón usual $\overline{y}_R$ son bastante restrictivas y difíciles de realizar en la práctica.

El estimador de *Ruiz* y *Santos*.

Ruiz y *Santos* (1989) proponen un estimador insesgado del tipo razón alternativo al de *Hartley* y *Ross*, obtenido a partir de una expresión alternativa del estimador media de razones.

Como hemos visto anteriormente una expresión del sesgo de $\overline{r}$ viene dada por la fórmula:

$$sesgo_1(\overline{r}) = -\frac{1}{\overline{X}N}\sum_{i=1}^{N} r_i\left(x_i - \overline{X}\right) = -\frac{\mathrm{Cov}(r_i, x_i)}{\overline{X}}$$

por lo que un estimador insesgado de este sesgo vale:

$$\widehat{sesgo}_1(\overline{r}) = -\frac{\overline{y}}{\overline{X}} + \overline{r}$$

Otra expresión del sesgo se puede obtener de la siguiente forma:

$$sesgo_2(\overline{r}) = E(\overline{r}) - R = E(\overline{r}) - \frac{E(\overline{y})}{E(\overline{x})} =$$

$$= \frac{-\mathrm{Cov}(\overline{r}, \overline{x}) + E(\overline{r}\,\overline{x} - \overline{y})}{\overline{X}} \tag{1.16}$$

Un estimador insesgado del $sesgo_2(\overline{r})$ viene dado por:

$$\widehat{sesgo}_2(\overline{r}) = \frac{1}{\overline{X}}\left[-\widehat{\mathrm{Cov}}(\overline{r}, \overline{x}) + \hat{E}(\overline{r}\,\overline{x} - \overline{y})\right] \tag{1.17}$$

con

$$\hat{E}(\overline{r}\,\overline{x} - \overline{y}) = \overline{r}\,\overline{x} - \overline{y} = \frac{-\overline{y}_{HR} + \overline{X}\overline{r}}{(N-1)n} N(n-1) \tag{1.18}$$

y un estimador insesgado de $\mathrm{Cov}(\overline{r}, \overline{x})$ en muestreo aleatorio simple es:

$$\widehat{\mathrm{Cov}}(\overline{r}, \overline{x}) = \frac{N-n}{(N-1)n}\left(\overline{y} - \overline{X}\,\overline{r}\right) \tag{1.19}$$

Entonces sustituyendo 1.18 y 1.19 en 1.17:

$$\widehat{sesgo}_3(\overline{r}) = \frac{1}{\overline{X}}\left[\overline{r}\,\overline{x} + \overline{r}\,\overline{X}\frac{N-n}{(N-1)n} - \overline{y}\frac{N(n+1)-2n}{(N-1)n}\right]$$

que es un estimador insesgado análogo al deducido por *Hartley* y *Ross*, y que viene expresado en función de los mismos estadísticos.

Entonces un estimador alternativo al de *Hartley* y *Ross* para $\overline{Y}$ es:

$$\overline{y}_{RE} = \overline{X}\left(\overline{r} - \widehat{sesgo}_3(\overline{r})\right) = \frac{N(n-1)}{(N-1)n}\overline{r}\,\overline{X} + \frac{N(n+1)-2n}{(N-1)n}\overline{y} - \overline{r}\,\overline{x}$$

que es insesgado.

Este estimador es similar al de *Hartley* y *Ross* cuando n es grande, pero sensiblemente distinto cuando el tamaño de la muestra es pequeño.

Puesto que estos dos estimadores son insesgados de $\overline{Y}$, cualquier combinación lineal de ellos será también un estimador insesgado. En efecto:

$$E\left(\lambda\overline{y}_{HR} + (1-\lambda)\,\overline{y}_{RE}\right) = \overline{Y}$$

En particular si $\lambda = -\dfrac{N(n-1)}{N-n}$ entonces $\lambda\overline{y}_{HR} + (1-\lambda)\,\overline{y}_{RE} = \overline{y}$.

Además:

$$E\left(\overline{r}\,\overline{x}\right) = \frac{N(n-1)}{(N-1)n}\overline{X}E\left(\overline{r}\right) + \frac{N-n}{(N-1)n}\overline{Y}$$

y entonces:

$$z = (N-n)\overline{y} + (n-1)N\overline{X}\,\overline{r} - n(N-1)\overline{r}\,\overline{x}$$

es un estimador insesgado de cero, puesto que:

$$\overline{y}_{HR} = \overline{y} + \frac{z}{N(N-1)}\;;\quad \overline{y}_{RE} = \overline{y} + \frac{z}{n(N-1)}$$

son insesgados de $\overline{Y}$.

Entonces un nuevo estimador insesgado de $\overline{Y}$ es:

$$\overline{y}_{RS} = \overline{y} - \frac{z}{N-n} = \frac{n(N-1)}{N-n}\overline{r}\,\overline{x} - \frac{(n-1)N}{N-n}\overline{X}\,\overline{r}$$

Este estimador es mejor desde el punto de vista computacional puesto que no hace falta el cálculo de $\overline{y}$, necesario para el cálculo de los otros dos estimadores.

El estimador de *Tin*.

Como hemos visto anteriormente las condiciones bajo las cuales el estimador insesgado $\overline{y}_{HR}$ es preferible a $\overline{y}_R$ no son fácilmente realizables en la práctica. Por ello se han propuesto otros estimadores de tipo razón que aún sin ser insesgados reducen considerablemente el sesgo de $\overline{y}_R$, pero son más precisos que $\overline{y}_{HR}$.

Hemos visto anteriormente que $\overline{y}_R$ es sesgado y su esperanza viene dada por la expresión:

$$E(\overline{y}_R) = \overline{Y}\left[1 + \frac{1-f}{n}\left(C_x^2 - \rho C_x C_y\right) + O\left(\frac{1}{n^2}\right)\right]$$

En un intento de reducir el sesgo de $\overline{y}_R$, *Tin* propone el estimador:

$$\overline{y}_T = \overline{y}_R\left[1 - \frac{1-f}{n}\left(\frac{s_x^2}{\overline{x}^2} - \frac{s_{xy}}{\overline{x}\,\overline{y}}\right)\right]$$

Teorema 1.4.29 *El sesgo del estimador de Tin es de orden* $O(n^{-2})$.

Demostración.-

$$E(\overline{y}_T) = E\left[\overline{y}_R\left[1 - \frac{1-f}{n}\left(\frac{s_x^2}{\overline{x}^2} - \frac{s_{xy}}{\overline{x}\,\overline{y}}\right)\right]\right] =$$

$$= E[\overline{y}_R] - E\left[\overline{y}_R \frac{1-f}{n}\left(\frac{s_x^2}{\overline{x}^2} - \frac{s_{xy}}{\overline{x}\,\overline{y}}\right)\right] =$$

$$= \overline{Y}\left[1 + \frac{1-f}{n}\left(\frac{S_x^2}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}\,\overline{Y}}\right) + O\left(\frac{1}{n^2}\right)\right] - E\left[\frac{1-f}{n}\overline{X}\left(\frac{s_x^2\overline{y}}{\overline{x}^3} - \frac{s_{xy}}{\overline{x}^2}\right)\right] =$$

$$= \overline{Y} + \frac{1-f}{n}E\left[\overline{X}\left(-\frac{s_x^2\overline{y}}{\overline{x}^3} + \frac{s_{xy}}{\overline{x}^2}\right) + \frac{S_x^2\overline{Y}}{\overline{X}^2} - \frac{S_{xy}}{\overline{X}}\right] + O\left(n^{-2}\right) =$$

$$= \overline{Y} + \frac{1-f}{n}\overline{X}E\left[\left(-\frac{s_x^2\overline{y}}{\overline{x}^3} + \frac{S_x^2\overline{Y}}{\overline{X}^3}\right) + \left(\frac{s_{xy}}{\overline{x}^2} - \frac{S_{xy}}{\overline{X}^2}\right)\right] + O\left(n^{-2}\right) =$$

$$= \overline{Y} + \frac{1-f}{n}\overline{X}E\left[\left(\frac{s_{xy}}{\overline{x}^2} - \frac{S_{xy}}{\overline{X}^2}\right) - \left(\frac{s_x^2\overline{y}}{\overline{x}^3} - \frac{S_x^2\overline{Y}}{\overline{X}^3}\right)\right] + O\left(n^{-2}\right)$$

Si llamamos

$$e_s = \frac{s_{xy} - S_{xy}}{S_{xy}}$$

tenemos:

$$E\left(\frac{s_{xy}}{\overline{x}^2} - \frac{S_{xy}}{\overline{X}^2}\right) = E\left(\frac{S_{xy}(1+e_s)}{\overline{X}^2(1+e_2)^2} - \frac{S_{xy}}{\overline{X}^2}\right) =$$

$$= \frac{S_{xy}}{\overline{X}^2}E\left(\frac{1+e_s-(1+e_2)^2}{(1+e_2)^2}\right) = \frac{S_{xy}}{\overline{X}^2}E\left(\frac{e_s-2e_2-e_2^2}{(1+e_2)^2}\right)$$

Así:

$$E\left(\frac{s_{xy}}{\overline{x}^2} - \frac{S_{xy}}{\overline{X}^2}\right) \simeq \frac{S_{xy}}{\overline{X}^2}E\left[\left(e_s-2e_2-e_2^2\right)(1-2e_2) - 2e_2^3\right] = A$$

Entonces, si x_0 es una cota inferior positiva de $\overline{x}$, tenemos:

$$\left|E\left(\frac{s_{xy}}{\overline{x}^2} - \frac{S_{xy}}{\overline{X}^2}\right) - A\right| = \left|S_{xy}\left\{E\left[\frac{1}{\overline{x}^2}\left(3e_2^2e_s - 4e_2^3 - 3e_2^4 + 2e_2^3e_s\right)\right]\right\}\right| \leq$$

$$\leq \frac{1}{x_0^2}\left|S_{xy}E\left[3e_2^2e_s - 4e_2^3 - 3e_2^4 + 2e_2^3e_s\right]\right|$$

Según el lema 1.3.22:

$$\left|E\left(\frac{s_{xy}}{\overline{x}^2} - \frac{S_{xy}}{\overline{X}^2}\right) - A\right| = O\left(n^{-2}\right)$$

y como

$$A = \frac{S_{xy}}{\overline{X}^2}E\left[e_s - 2e_se_2 - 2e_2 + 3e_2^2\right]$$

$$E(e_se_2) = \frac{N-n}{n(N-2)}\frac{\mu_{21}}{\mu_{11}\overline{X}} \; ; \quad E\left(e_2^2\right) = \frac{N-n}{n(N-1)}\frac{\mu_{20}}{\overline{X}^2}$$

A es orden $O(n^{-1})$ y también lo será

$$E\left(\frac{s_{xy}}{\overline{x}^2} - \frac{S_{xy}}{\overline{X}^2}\right)$$

De forma análoga se comprueba que

$$E\left(\frac{s_x^2\overline{y}}{\overline{x}^3} - \frac{S_x^2\overline{Y}}{\overline{X}^3}\right) = O\left(n^{-1}\right)$$

con lo que

$$E\left(\overline{y}_T\right) = \overline{Y} + O\left(n^{-2}\right)$$

como se quería probar.

Puesto que el sesgo de $\overline{y}_T$ es de orden $O(n^{-2})$, el estimador recibe el nombre de *estimador cuasi-insesgado* de $\overline{Y}$.

Otros estimadores insesgados o con sesgo inferior al del estimador $\overline{y}_R$ son los propuestos por *Mickey* (1959), *Beale* (1962), *Quenouille* (1965) y otros.

Estimador de razón bajo el esquema de muestreo de *Midzuno*.

Otra forma de reducir el sesgo del estimador clásico de razón es asignar un esquema de muestreo adecuado. Aquí se va a considerar el esquema de muestreo propuesto por *Midzuno* (1952), el cual va a tener unas propiedades óptimas para nuestro propósito, puesto que el estimador de razón va a ser insesgado.

El esquema de muestreo es el siguiente: se selecciona la primera unidad con probabilidad proporcional al valor de la variable auxiliar, mientras que el resto de las $n-1$ unidades que componen la muestra son elegidas con probabilidades iguales y sin reemplazo.

Bajo este esquema, la probabilidad de seleccionar una muestra específica $(u_1, u_2, \ldots, u_n)$ viene dada por:

$$P\left(u_1, u_2, \ldots, u_n\right) = \sum_{i=1}^{n} \frac{x_i}{\sum_{i=1}^{n} x_i} \frac{1}{\binom{N-1}{n-1}} = \frac{\overline{x}}{\overline{X}} \frac{1}{\binom{N}{n}}$$

Consideremos el estimador de razón:

$$\overline{y}_R = \frac{\overline{y}}{\overline{x}}\overline{X}$$

y vamos a estudiar su comportamiento bajo este esquema.

Teorema 1.4.30 *El estimador de razón bajo el esquema de selección de Midzuno es insesgado y su varianza viene dada por la expresión:*

$$V(\overline{y}_R) = \frac{\overline{X}}{\binom{N}{n}} \sum_{s \in S} \frac{\overline{y}_s^2}{\overline{x}_s^2} - \overline{Y}^2$$

donde $\sum_{s \in S}$ *denota la suma sobre todas las posibles muestras de tamaño* n.

Demostración.-
En primer lugar el estimador es insesgado puesto que

$$E(\overline{y}_R) = \overline{X} E\left(\frac{\overline{y}}{\overline{x}}\right) = \overline{X} \sum_{s \in S} \frac{\overline{y}_s}{\overline{x}_s} \frac{\overline{x}_s}{\overline{X}\binom{N}{n}}$$

y

$$\sum_{s \in S} \frac{\overline{y}_s}{\binom{N}{n}} = \overline{Y}$$

En cuanto a la varianza:

$$V(\overline{y}_R) = E\left(\overline{y}_R^2\right) - \overline{Y}^2 = \sum_{s \in S} \left(\frac{\overline{X}\overline{y}_s}{\overline{x}_s}\right)^2 \frac{\overline{x}_s}{\overline{X}} \frac{1}{\binom{N}{n}} - \overline{Y}^2$$

y simplificando se obtiene el resultado.

A partir de aquí se obtiene el siguiente corolario:

Corolario 1.4.31 *Un estimador insesgado de la varianza del estimador de razón viene dada por la expresión:*

$$\hat{V}(\overline{y}_R) = \overline{y}_R^2 - \frac{\overline{X}}{\overline{x}} \left[\overline{y}^2 - \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2\right]$$

Para la estimación del total, Y, se obtienen, sin dificultad, los siguientes resultados:

$$\widehat{Y}_R = \frac{\overline{y}}{\overline{x}} X$$

$$V\left(\widehat{y}_R\right) = N^2 \frac{\overline{X}}{\binom{N}{n}} \sum_{s \in S} \frac{\overline{y}_s^2}{\overline{x}_s^2} - \overline{Y}^2$$

$$\widehat{V}\left(\widehat{Y}_R\right) = \widehat{Y}_R^2 - N^2 \frac{\overline{X}}{\overline{x}} \left[\overline{y}^2 - \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2\right]$$

1.4.2 Disminución del error cuadrático medio.

En los últimos años diversos autores han sugerido distintos métodos de mejorar el estimador clásico de razón, en el sentido de disminuir el error cuadrático medio. La mayor parte de estos intentos o bien necesitan el conocimiento de ciertos valores de la población en general desconocidos, o bien el aumento de precisión no es lo suficientemente grande como para compensar la mayor complejidad de los métodos. Vamos a considerar en este tema algunos de dichos métodos.

Método repetido de sustitución.

Srivastava (1967) propone un método de reformar el estimador de razón mediante el método de exponenciación. Para ello parte del estimador de razón para el total:

$$\widehat{Y}_R = \widehat{Y} \frac{X}{\widehat{\overline{X}}}$$

Este estimador es mejor que el usual $\widehat{Y}$ en un esquema de muestreo aleatorio simple siempre que $\rho > \dfrac{C_x}{2C_y}$. Entonces si sustituimos $\widehat{Y}_R = \widehat{Y}_R^{(1)}$ por $\widehat{Y} = N\overline{y}$ en la expresión del estimador de razón obtenemos:

$$\widehat{Y}_R^{(2)} = \widehat{Y}_R \frac{X}{\widehat{\overline{X}}} = \widehat{Y} \left(\frac{X}{\widehat{\overline{X}}}\right)^2$$

Este nuevo estimador $\widehat{Y}_R^{(2)}$ puede utilizarse en lugar de $\widehat{Y}$, dando lugar a un nuevo estimador:

$$\hat{Y}_R^{(3)} = \hat{Y}_R^{(2)} \frac{X}{\widehat{X}} = \hat{Y} \left(\frac{X}{\widehat{X}} \right)^3$$

Repitiendo el proceso α veces (donde α es un entero), podemos llegar al estimador:

$$\hat{Y}_R^{(\alpha)} = \hat{Y}_R^{(\alpha-1)} \frac{X}{\widehat{X}} = \hat{Y} \left(\frac{X}{\widehat{X}} \right)^\alpha$$

que se conoce con el nombre de *estimador de razón por sustitución repetida.*

En primer lugar se va a calcular su varianza, para posteriormente determinar el número óptimo de iteraciones que minimiza dicha varianza.

Si llamamos :

$$e_1 = \frac{\left(\hat{Y} - Y\right)}{Y} \; ; \quad e_2 = \frac{\left(\widehat{X} - X\right)}{X}$$

podemos escribir:

$$\hat{Y}_R^{(\alpha)} = Y \frac{(1 + e_1)}{(1 + e_2)^\alpha} \simeq (1 + e_1)(1 - \alpha e_2) = Y(1 - \alpha e_2 + e_1 - \alpha e_1 e_2)$$

Entonces:

$$\left(\hat{Y}_R^{(\alpha)} - Y\right)^2 = Y^2 (1 - \alpha e_2 - \alpha e_1 e_2)^2$$

Desarrollamos el término de la derecha quedándonos sólo con los términos de orden inferior o igual a dos, y al tomar esperanzas queda:

$$V\left(\hat{Y}_R^{(\alpha)}\right) = Y^2 E\left(e_1^2 + \alpha^2 e_2^2 - 2\alpha e_1 e_2\right) =$$

$$= Y^2 \left[\frac{V\left(\hat{Y}\right)}{Y^2} + \alpha^2 \frac{V\left(\widehat{X}\right)}{X^2} - 2\alpha \frac{\mathrm{Cov}\left(\hat{Y}, \widehat{X}\right)}{Y\,X} \right] =$$

$$= V\left(\hat{Y}\right) + \alpha^2 R^2 V\left(\widehat{X}\right) - 2\alpha R\,\mathrm{Cov}\left(\hat{Y}, \widehat{X}\right)$$

A partir de esta expresión podemos determinar el número de iteraciones óptimo en el sentido de que produzca menor varianza en el estimador.

Si derivamos la expresión anterior respecto a α e igualamos a cero, obtenemos un valor para α igual a:

$$\alpha = \frac{\text{Cov}\left(\widehat{Y}, \widehat{X}\right)}{V\left(\widehat{X}\right) R} = \frac{\beta}{R}$$

siendo β el coeficiente de regresión entre y y x poblacional.

Entonces el número óptimo de iteraciones :

$$\alpha_{\acute{o}pt} = \left[\frac{\beta}{R}\right] \quad \left(\text{o } \left[\frac{\beta}{R}\right] + 1 \text{ si } \left[\frac{\beta}{R}\right] < 1\ \right)$$

puesto que α debe de ser entero.

Este valor de α es válido para muestras grandes ya que hemos partido de una aproximación de la varianza de orden $O(n^{-2})$.

La varianza del estimador para $\alpha_{\acute{o}pt}$ se deduce de inmediato y vale:

$$V\left(\widehat{Y}_R^{(\alpha)}\right) = V\left(\widehat{Y}\right)\left(1 - \rho^2\right)$$

que coincide con el valor mínimo de la varianza del estimador de regresión para una muestra de tamaño grande.

Sin embargo, en muchas situaciones prácticas $\frac{\beta}{R} = \rho$ ya que $C_x = C_y$ ($\frac{\beta}{R} = \rho\frac{C_y}{C_x}$). Incluso cuando C_y no está próximo a C_x suele ocurrir que $\frac{\beta}{R}$ no está demasiado alejado de 1. En general $\alpha_{\acute{o}pt}$ está próximo a 1, y no hay mejora sustancial usando más de una iteración.

T.J. Rao (1991) compara los sesgos y el error cuadrático medio para los estimadores $\widehat{Y}_R^{(\alpha)}$ con $\alpha = 1, 2$ y $\alpha_{\acute{o}pt}$ en diversas poblaciones naturales concluyendo que la elección de $\alpha = 1$ que tiene una justificación práctica, es mejor que la de $\alpha_{\acute{o}pt}$ que es el óptimo teórico, y que no hay una ganancia sustancial de precisión si se reitera más de una vez el proceso.

Este método de sustitución repetida puede utilizarse también para "mejorar" los estimadores de diferencia y de regresión.

El estimador de razón generalizado.

Menéndez y *Ferrales* (1989) proponen un estimador de tipo razón al que llaman *estimador de razón generalizado*, siguiendo la idea del *estimador producto generalizado* dado por *Ray* y *Singh* (1981).

Los nuevos estimadores de la razón, total y media poblacional se definen de la forma:

$$\hat{R}_G = \frac{\overline{y}}{\overline{X} - k\left(\overline{x} - \overline{X}\right)} \; ; \; \hat{Y}_G = \hat{R}_G X \; ; \; \hat{\overline{Y}}_G = \hat{R}_G \overline{X}$$

donde k es un parámetro a determinar.

Sesgo del estimador.

Este estimador es sesgado como demuestra el siguiente teorema.

Teorema 1.4.32 *El estimador de razón generalizado es sesgado y su sesgo viene dado por la expresión:*

$$sesgo\left(\hat{R}_G\right) = -\frac{\mathrm{Cov}\left(\hat{R}_G, X - k\left(\overline{x} - \overline{X}\right)\right)}{\overline{X}}$$

Demostración.-

$$sesgo\left(\hat{R}_G\right) = E\left(\hat{R}_G - R\right) = E\left(\hat{R}_G\right) - \frac{E\left(\overline{y}\right)}{E\left(\overline{x}\right)} =$$

$$= \frac{E\left(\hat{R}_G\right) E\left(\overline{x}\right) - E\left(\hat{R}_G\left(\overline{X} - k\left(\overline{x} - \overline{X}\right)\right)\right)}{\overline{X}} =$$

$$= \frac{E\left(\hat{R}_G\right) E\left(\overline{X} - k\left(\overline{x} - \overline{X}\right)\right) - E\left(\hat{R}_G\left(\overline{X} - k\left(\overline{x} - \overline{X}\right)\right)\right)}{\overline{X}} =$$

$$= -\frac{\mathrm{Cov}\left(\hat{R}_G, \overline{X} - k\left(\overline{x} - \overline{X}\right)\right)}{\overline{X}}$$

Sin embargo este resultado no es útil en la práctica. Vamos a obtener una aproximación del sesgo que sea más fácil de calcular en un problema real.

Teorema 1.4.33 *Una aproximación del sesgo de orden* $O(n^{-2})$ *viene dada por la expresión:*

$$sesgo\left(\hat{R}_G\right) \simeq R\frac{1-f}{n}k\left(kC_x^2 + C_{xy}\right)$$

Demostración.-
Consideramos las variables

$$e_1 = \frac{\overline{y} - \overline{Y}}{\overline{Y}} \; ; \quad e_2 = \frac{\overline{x} - \overline{X}}{\overline{X}}$$

podemos expresar $\widehat{R}_G$ de la forma:

$$\widehat{R}_G = R\,(1 + e_1)\,(1 - ke_2)^{-1}$$

Desarrollando en serie el término $(1 - ke_2)^{-1}$ y despreciando los términos en e_1 y e_2 de orden superior a dos tendremos la aproximación:

$$\widehat{R}_G \simeq R\left(1 + ke_2 + e_1 + ke_1e_2 + k^2e_2^2\right)$$

Así:

$$sesgo\left(\widehat{R}_G\right) \simeq R\left[kE\left(e_2\right) + E\left(e_1\right) + kE\left(e_1e_2\right) + k^2E\left(e_2^2\right)\right] =$$

$$= R\left[k^2\frac{1-f}{n\overline{X}^2}S_x^2 + k\frac{1-f}{\overline{X}\,\overline{Y}}\frac{S_xS_y}{n}\right] = R\frac{1-f}{n}k\left[kC_x^2 + C_{xy}\right]$$

Esta expresión es nula si $k = 0$ o $k = -\dfrac{C_{xy}}{C_x^2} = -\rho\dfrac{C_y}{C_x}$. Como $k = 0$ carece de sentido, el sesgo desaparece si:

$$k = k_0 = -\rho\frac{C_y}{C_x}$$

Veámos a continuación bajo qué condiciones es posible la determinación de este valor k_0 que hace el estimador insesgado.

El desarrollo anterior es válido para:

$$|k|\,\frac{\left|\overline{x} - \overline{X}\right|}{\left|\overline{X}\right|} < 1$$

por tanto el valor k_0 tendrá sentido si

$$|k_0| < \frac{\left|\overline{X}\right|}{\left|\overline{x} - \overline{X}\right|} \Leftrightarrow |-\rho| < \frac{\left|\overline{X}\right|}{\left|\overline{x} - \overline{X}\right|}\frac{|C_x|}{|C_y|}$$

Para n suficientemente grande $\left|\overline{X}\right| > \left|\overline{x} - \overline{X}\right|$, entonces si x e y son variables tales que $|C_x| \geq \left|C_y\right|$ se cumple que el término de la derecha es mayor que uno, y como $|-\rho| \leq 1$ entonces será posible la determinación del valor óptimo de k.

Error cuadrático medio.

Teorema 1.4.34 *Una aproximación del error cuadrático medio del estimador de razón generalizado viene dado por la expresión:*

$$ECM\left(\hat{R}_G\right) \simeq R^2 \frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + 2k\rho\frac{S_x S_y}{\overline{X}\,\overline{Y}} + k^2\frac{S_x^2}{\overline{X}^2}\right]$$

Demostración.-

Partiendo de la aproximación:

$$\hat{R}_G = R\left(1 + ke_2 + e_1 + ke_1e_2 + k^2e_2^2\right)$$

y procediendo como en el caso anterior de la determinación del sesgo del estimador, obtenemos la aproximación deseada.

Corolario 1.4.35 *El error cuadrático medio toma su valor mínimo para $k = k_0$, y este mínimo vale:*

$$ECM_0\left(\hat{R}_G\right) = \frac{1-f}{n\overline{X}^2}S_y^2\left(1-\rho^2\right)$$

Demostración.-

Derivando respecto a k en la expresión del $ECM\left(\hat{R}_G\right)$ e igualando a cero, obtenemos que el valor $k_0 = -\rho\dfrac{C_y}{C_x}$ es el mínimo puesto que la segunda derivada es positiva.

Para este valor, se tiene:

$$ECM_0\left(\hat{R}_G\right) = R^2\frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} - 2\rho^2C_y^2 + \rho^2C_y^2\right] = \frac{1-f}{n\overline{X}^2}S_y^2\left(1-\rho^2\right)$$

Por tanto si elegimos el valor k_0 tenemos un estimador de razón que es insesgado y que produce menor error entre la clase de estimadores de razón generalizados.

Comparación con el estimador de razón usual.

El estimador de razón usual, $\widehat{R}$ es sesgado y su error cuadrático medio vale:

$$ECM\left(\widehat{R}\right) = \frac{1-f}{n}R^2\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right)$$

entonces:

$$ECM\left(\widehat{R}\right) - ECM\left(\widehat{R}_G\right) = \frac{1-f}{n}R^2\left[C_x^2\left(1-k^2\right) - \rho C_y C_x\left(2\rho + 2k\right)\right]$$

y el estimador de razón generalizado es más eficaz que el estimador de razón usual si:

$$\begin{array}{ll} -1 < k < 1 + 2k_0 & \text{para} \quad 1 + 2k_0 > -1 \\ 1 + 2k_0 < k < -1 & \text{para} \quad 1 + 2k_0 < -1 \end{array}$$

El valor k_0 está dentro de alguno de estos intervalos, y para este valor la diferencia entre los errores cuadráticos medios de ámbos estimadores es mínima y vale:

$$ECM\left(\widehat{R}\right) - ECM\left(\widehat{R}_G\right) = \frac{1-f}{n}R^2C_x^2\left(1+k_0\right)^2$$

Comparación con el estimador de expansión simple.

La media muestral $\overline{y}$ es insesgado de $\overline{Y}$ y $V\left(\overline{y}\right) = \dfrac{1-f}{n}S_y^2$, por tanto:

$$V\left(\overline{y}\right) - ECM\left(\widehat{R}_G\right) = -\frac{1-f}{n}\overline{Y}^2\left(2k\rho C_y C_x + k^2C_x^2\right)$$

y el estimador de razón generalizado es más eficaz que el estimador de expansión simple si:

$$\begin{array}{ll} k \in (0, 2k_0) & \text{para} \quad k_0 > 0 \\ k \in (2k_0, 0) & \text{para} \quad k_0 < 0 \end{array}$$

Además en $k = k_0$ la diferencia entre los errores cuadráticos medios de los estimadores tiene un mínimo, y este mínimo vale:

$$V\left(\overline{y}\right) - ECM\left(\widehat{R}_G\right) = \frac{1-f}{n}\rho^2S_y^2$$

El estimador de *Prasad*.

Prasad (1986) propone un nuevo estimador tipo razón que es insesgado bajo muestreo aleatorio simple. Este autor considera el siguiente estimador de $\overline{Y}$:

$$\widehat{\overline{Y}}_a = \overline{y} - a\frac{\overline{x}}{\overline{X}} + a$$

donde a es una constante cualquiera.

En primer lugar es evidente que el estimador es insesgado, por lo que habría que considerar la varianza como medida de precisión:

$$V\left(\widehat{\overline{Y}}_a\right) = V\left(\overline{y} - a\frac{\overline{x}}{\overline{X}}\right) = V(\overline{y}) + \frac{a^2}{\overline{X}^2}V(\overline{x}) - 2\frac{a}{\overline{X}}\operatorname{Cov}(\overline{y},\overline{x}) =$$

$$= \frac{1-f}{n}\left[S_y^2 + \frac{a^2}{\overline{X}^2}S_x^2 - 2\frac{a}{\overline{X}}\rho S_x S_y\right]$$

Comparando la varianza de $\widehat{\overline{Y}}_a$ con la del estimador de expansión simple se tiene que $\widehat{\overline{Y}}_a$ es más eficiente que $\overline{y}$ si:

$$V\left(\widehat{\overline{Y}}_a\right) < V(\overline{y}) = \frac{1-f}{n}S_y^2$$

o equivalentemente, si:

$$\rho > \frac{a}{2\overline{X}}\left(\frac{S_x}{S_y}\right)$$

Entonces, si S_x^2 es aproximadamente igual a S_y^2 podemos elegir una constante a suficientemente pequeña de forma que $\widehat{\overline{Y}}_a$ sea más eficiente siempre que $\overline{y}$, mientras que $\widehat{\overline{Y}}_R$ será preferible a $\overline{y}$ sólo si $\rho > \frac{1}{2}$.

Si se compara $\widehat{\overline{Y}}_a$ con $\widehat{\overline{Y}}_R$ para muestras grandes, se tendrá que el primero es más eficiente si

$$V\left(\widehat{\overline{Y}}_a\right) < ECM\left(\widehat{\overline{Y}}_R\right)$$

es decir

$$\rho > \frac{1}{2}\left(\frac{a+\overline{Y}}{\overline{X}}\right)\left(\frac{S_x}{S_y}\right) \quad \text{si} \quad 0 < \overline{Y} < a$$
$$\rho < \frac{1}{2}\left(\frac{a+\overline{Y}}{\overline{X}}\right)\left(\frac{S_x}{S_y}\right) \quad \text{si} \quad \overline{Y} > a$$

y son igual de precisos si $\overline{Y} = a$

Entonces, si S_x^2 y S_y^2 son aproximadamente iguales, tomando a adecuadamente, las desigualdades anteriores pueden ser siempre ciertas, por lo que el estimador $\widehat{\overline{Y}}_a$ será mejor que $\widehat{\overline{Y}}_R$.

Familia biparamétrica de estimadores tipo razón.

Ray y *Sahai* (1980) proponen la siguiente familia biparamétrica de estimadores tipo razón:

$$\overline{y}_{RK\theta} = \overline{y}\frac{K\overline{X} + \theta\overline{x}}{\overline{x} + (K + \theta - 1)\overline{X}}$$

donde $0 \leq \theta < 1$ y K es un entero no negativo.

Este estimador se reduce al estimador de razón $\overline{y}_R$ para $\theta = 0$ y $K = 1$, y al estimador $\overline{y}$ para $\theta = 1$ y $\forall K$.

Si consideramos como en casos anteriores las variables:

$$e_1 = \frac{\left(\overline{y} - \overline{Y}\right)}{\overline{Y}} \; ; \quad e_2 = \frac{\left(\overline{x} - \overline{X}\right)}{\overline{X}}$$

y desarrollamos $\overline{y}_{RK\theta}$ en potencias de e_2 quedándonos sólo con los términos de orden inferior o igual a dos en e_1 y e_2, llegamos a las siguientes aproximaciones para el sesgo y error cuadrático medio:

$$sesgo\left(\overline{y}_{RK\theta}\right) = \frac{\overline{Y}}{n}(1-f)\,Q\left\{(K+\theta)^{-1} - C\right\}C_x^2$$

$$ECM\left(\overline{y}_{RK\theta}\right) = \frac{\overline{Y}^2}{n}(1-f)\left\{C_y^2 + Q\,(Q-2C)\,C_x^2\right\}$$

donde

$$Q = \frac{1-\theta}{K+\theta} \; ; \quad C = \rho\frac{C_y}{C_x}$$

Entonces el error cuadrático medio del estimador $\overline{y}_{RK\theta}$, $ECM\left(\overline{y}_{RK\theta}\right)$, decrece con K si $C < Q$ y el rango de Q para diferentes valores de K es $0 < Q < \frac{1}{K}$, $k = 0, 1, \ldots$.

Si se tiene información acerca del rango de C, se puede utilizar para determinar un estimador tipo razón que sea más preciso que el estimador de razón usual o el estimador de expansión simple, en las situaciones en que cada uno de ellos es más eficiente.

Capítulo 2

El estimador de razón en otros tipos de muestreo.

§2.1 El estimador de razón en el muestreo estratificado.

2.1.1 Introducción.

Como es sabido la precisión de los estimadores de la media o el total en un muestreo aleatorio simple depende no sólo del tamaño muestral y de la fracción de muestreo, sino también de la variabilidad o heterogeneidad entre las unidades de la población. Además de incrementar el tamaño muestral, una vía de incrementar la precisión de estos estimadores es dividir la población en diversos grupos, de manera que cada uno de ellos sea más homogéneo que la población entera, mediante el procedimiento de estratificación.

La estratificación es una técnica común. Hay muchas razones para realizarla, entre las que destacan:

1. Si se desea información con cierta precisión en algunas subdivisiones de la población es aconsejable tratar cada subdivisión como una población por sí sola.

2. Por conveniencias de tipo administrativo.

3. Los problemas de muestreo pueden diferir marcadamente en diferentes partes de la población, haciendo aconsejable elegir un esquema de muestreo diferente en cada estrato.

4. La estratificación puede dar lugar a una ganancia sustancial en precisión. Es posible dividir una población heterogénea en subpoblaciones, cada una de las cuales es internamente homogénea. Dentro de cada estrato se puede obtener un estimador muy preciso de cualquier parámetro del estrato para una muestra pequeña y combinar estos estimadores en un estimador preciso de toda la población.

En este capítulo vamos a considerar un esquema de muestreo estratificado y a estudiar estimadores de tipo razón.

En primer lugar estudiamos los dos estimadores conocidos en la Teoría de Muestras y que reciben el nombre de estimador de razón separado y estimador de razón combinado.

El muestreo estratificado utiliza el estimador de la media

$$\overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{y}_h$$

donde $\overline{y}_h$ es la media muestral en cada estrato y L el número de estratos. Es decir, es la suma de las medias de cada estrato ponderadas por el peso del estrato.

Nosotros vamos a proponer en la segunda parte del capítulo un posible estimador de la media, que pondera cada media del estrato según unos pesos que hacen mínima la varianza del estimador.

Siguiendo esta idea de ponderar los estratos de forma que se minimize la varianza, vamos a construir otros dos estimadores de tipo razón.

2.1.2 El estimador de razón separado.

Sea N_h el número de unidades en el estrato h y sea n_h el tamaño de la muestra seleccionada de él. Entonces:

$$\sum_{h=1}^{L} N_h = N \quad \text{y} \quad \sum_{h=1}^{L} n_h = n$$

Sean:

Y_h *el total de la variable* y *en el estrato* h.

X_h *el total de la variable* x *en el estrato* h.

$\overline{Y}_h$ *la media de la variable y en el estrato h.*

$\overline{X}_h$ *la media de la variable x en el estrato h.*

$\overline{y}_h$ *la media muestral de la variable y en el estrato h.*

$\overline{x}_h$ *la media muestral de la variable x en el estrato h.*

Una forma sencilla de obtener un estimador de razón de Y es obtener un estimador de razón del total de cada estrato y sumar estos totales.

Definición 2.1.1 *Se define el estimador de razón separado del total como*

$$\hat{Y}_{Rs} = \sum_{h=1}^{L} \frac{\overline{y}_h}{\overline{x}_h} X_h$$

En este estimador no se supone que la razón verdadera permanezca constante en cada estrato. Por otra parte se requiere del conocimiento de los totales X_h en cada estrato.

Del mismo modo se obtiene el estimador de razón separado para $\overline{Y}$.

Puesto que:

$$\overline{Y} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{Y}_h$$

definimos:

Definición 2.1.2 *Se llama estimador separado de la media al estimador*

$$\hat{\overline{Y}}_{Rs} = \sum_{h=1}^{L} \frac{N_h}{N} \hat{\overline{Y}}_{R_h}$$

siendo $\hat{\overline{Y}}_{R_h}$ el estimador de razón de la media del estrato h.

$$\hat{\overline{Y}}_{R_h} = \frac{\overline{y}_h}{\overline{x}_h} \overline{X}_h$$

por lo que $N\hat{\overline{Y}}_{Rs} = \hat{Y}_{Rs}$.

Propiedades.

El estimador de razón es consistente, pues cuando $n = N$, $\hat{\overline{Y}}_{Rs} = \overline{Y}$.

Sesgo.

Evidentemente el estimador de razón separado es sesgado pues se obtiene como suma de los estimadores de razón en cada estrato, los cuales son sesgados.

Para investigar este sesgo vamos a determinar en primer lugar una cota suya.

$$sesgo\left(\widehat{\overline{Y}}_{Rs}\right) = \sum_{h=1}^{L} W_h E\left(\widehat{\overline{Y}}_{R_h} - \overline{Y}_h\right) =$$

$$= \sum_{h=1}^{L} W_h \left(E\left(\widehat{R}_h\right)\overline{X}_h - \overline{Y}_h\right) =$$

$$= \sum_{h=1}^{L} W_h \left(E\left(\widehat{R}_h\right) E\left(\overline{x}_h\right) - E\left(\widehat{R}_h \overline{x}_h\right)\right) =$$

$$= -\sum_{h=1}^{L} W_h \operatorname{Cov}\left(\widehat{R}_h, \overline{x}_h\right)$$

Entonces

$$\left|sesgo\left(\widehat{\overline{Y}}_{Rs}\right)\right| \leq \sum_{h=1}^{L} W_h \sigma_{\widehat{R}_h} \sigma_{\overline{x}_h} = \sum_{h=1}^{L} W_h \sigma_{\widehat{\overline{Y}}_{R_h}} CV\left(\overline{x}_h\right)$$

Si $CV\left(\overline{x}_h\right) \leq C_0 \ \ \forall h$, tenemos:

$$\left|sesgo\left(\widehat{\overline{Y}}_{Rs}\right)\right| \leq C_0 \sum_{h=1}^{L} W_h \sigma_{\widehat{\overline{Y}}_{R_h}}$$

Ahora bien

$$\sum_{h=1}^{L} W_h^2 V\left(\widehat{\overline{Y}}_{R_h}\right) \geq \frac{1}{L}\left(\sum_{h=1}^{L} W_h \sigma_{\widehat{\overline{Y}}_{R_h}}\right)^2$$

y por tanto

$$\frac{\left|sesgo\left(\widehat{\overline{Y}}_{Rs}\right)\right|}{\sigma_{\widehat{\overline{Y}}_{RS}}} \leq C_0 \sqrt{L}$$

Por tanto, si el número de estratos es grande, el sesgo del estimador separado puede ser importante. A continuación vamos a determinar una expresión aproximada del sesgo.

Proposición 2.1.3 *Una aproximación del sesgo del estimador de razón separado de la media viene dado por la expresión:*

$$sesgo\left(\widehat{\overline{Y}}_{Rs}\right) \simeq \sum_{h=1}^{L} \frac{N_h}{N}\overline{Y}_h \left(\frac{N_h - n_h}{N_h n_h}\right)\left(C_{xh}^2 - \rho_h C_{xh} C_{yh}\right)$$

Demostración.-
En una primera aproximación:

$$E\left(\widehat{\overline{Y}}_{Rs}\right) = \sum_{h=1}^{L} \frac{N_h}{N} E\left(\widehat{\overline{Y}}_{R_h}\right) \simeq \sum_{h=1}^{L} \frac{N_h}{N}\overline{Y}_h \left(1 + \frac{N_h - n_h}{N_h n_h}\right)\left(C_{xh}^2 - \rho_h C_{xh} C_{yh}\right)$$

siendo $C_{xh} = \dfrac{S_{xh}}{\overline{X}_h}$ y $C_{yh} = \dfrac{S_{yh}}{\overline{Y}_h}$, de donde se sigue que $\widehat{\overline{Y}}_{Rs}$ es un estimador sesgado de la media poblacional y su sesgo viene dado por la expresión anterior.

Error cuadrático medio.

Proposición 2.1.4 *Si los tamaños de muestra en cada estrato son grandes, una aproximación del error cuadrático medio del estimador de razón separado,* $\widehat{\overline{Y}}_{Rs}$ *viene dado por la expresión:*

$$ECM\left(\widehat{\overline{Y}}_{Rs}\right) \simeq \sum_{h=1}^{L} N_h^2 \frac{1 - f_h}{n_h}\left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{xh} S_{yh}\right]$$

Demostración.-
Calculemos una aproximación del error cuadrático medio. Usando la aproximación del error cuadrático medio del estimador de razón de la media en cada estrato y teniendo en cuenta que los términos de productos cruzados desaparacen pues el muestreo es independiente en los diferentes estratos, obtenemos:

$$ECM\left(\widehat{\overline{Y}}_{Rs}\right) = \sum_{h=1}^{L} \frac{N_h^2}{N^2} ECM\left(\widehat{\overline{Y}}_{R_h}\right) \simeq$$

$$\simeq \sum_{h=1}^{L} \frac{N_h^2}{N^2} \frac{1-f_h}{n_h} \left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{xh} S_{yh} \right]$$

Esta fórmula sólo es válida si la muestra en cada estrato es suficientemente grande como para que la fórmula aproximada de la varianza sea aplicable en cada estrato.

Sin embargo, en la práctica no siempre ocurre esto. Para solventar esta dificultad *Hansen, Hurwitz y Gurney* (1946) sugirieron otro estimador de razón, que se conoce con el nombre de estimador de razón combinado.

2.1.3 El estimador de razón combinado.

A partir de los datos de la muestra se calcula:

$$\widehat{Y}_{st} = \sum_{h=1}^{L} N_h \overline{y}_h \; ; \quad \widehat{X}_{st} = \sum_{h=1}^{L} N_h \overline{x}_h$$

estimadores de los totales de la población Y y X respectivamente, obtenidos de una muestra estratificada.

Definición 2.1.5 *El estimador combinado se define por:*

$$\widehat{R}_C = \frac{\widehat{Y}_{st}}{\widehat{X}_{st}} \; ; \quad \widehat{Y}_{R_C} = \frac{\widehat{Y}_{st}}{\widehat{X}_{st}} X \; ; \quad \widehat{\overline{Y}}_{R_C} = \frac{\widehat{Y}_{st}}{\widehat{X}_{st}} \overline{X}$$

según sea de la razón, del total o de la media.

El estimador combinado, a diferencia del separado no necesita del conocimiento de las X_h en cada estrato, sino sólo del total X.

Sesgo.

Proposición 2.1.6 *Una aproximación del sesgo del estimador de razón combinado viene dada por la expresión:*

$$sesgo\left(\widehat{R}_C\right) \simeq \sum_{h=1}^{L} \frac{1-f_h}{n_h \overline{X}^2} \left(R S_{xh}^2 - \rho_h S_{xh} S_{yh} \right)$$

Demostración.-
Siguiendo el mismo proceso que en el muestreo aleatorio simple, se obtiene:

$$\widehat{R}_C - R = \frac{\overline{y}_{st}}{\overline{x}_{st}} - R \simeq \frac{\overline{y}_{st} - R\overline{x}_{st}}{\overline{X}} \left(1 - \frac{\overline{x}_{st} - \overline{X}}{\overline{X}}\right)$$

si

$$\left|\frac{\overline{x}_{st} - \overline{X}}{\overline{X}}\right| < 1$$

Tomando esperanzas:

$$E\left(\widehat{R}_C - R\right) = E\left(-\frac{\overline{y}_{st} - R\overline{x}_{st}}{\overline{X}} \frac{\overline{x}_{st} - \overline{X}}{\overline{X}}\right)$$

y

$$E\left(\overline{y}_{st}\left(\overline{x}_{st} - \overline{X}\right)\right) = \sum_{h=1}^{L} \frac{1 - f_h}{n_h} \rho_h S_{xh} S_{yh}$$

$$E\left(\overline{x}_{st}\left(\overline{x}_{st} - \overline{X}\right)\right) = \sum_{h=1}^{L} \frac{1 - f_h}{n_h} \rho_h S_{xh}^2$$

Entonces:

$$sesgo\left(\widehat{R}_C\right) = \sum_{h=1}^{L} \frac{1 - f_h}{n_h \overline{X}^2} \left(RS_{xh}^2 - \rho_h S_{xh} S_{yh}\right)$$

Error cuadrático medio.

Proposición 2.1.7 *Una aproximación del error cuadrático medio del estimador de razón combinado del total viene dada por la expresión:*

$$ECM\left(\widehat{Y}_{R_C}\right) \simeq \sum_{i=1}^{N_h} \frac{N_h^2 (1 - f_h)}{n_h} \left(S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{xh} S_{yh}\right)$$

Demostración.-
Calculemos el error cuadrático medio:

$$\widehat{R}_c - R = \frac{\overline{y}_{st} - R\overline{x}_{st}}{\overline{x}_{st}} \simeq \frac{\overline{y}_{st} - R\overline{x}_{st}}{\overline{X}}$$

Ahora bien, $\overline{y}_{st} - R\overline{x}_{st}$ es la media de la muestra de la variable $d_{h_i} = y_{h_i} - Rx_{h_i}$ cuya media poblacional es $\overline{D} = \overline{Y} - R\overline{X} = 0$ y por tanto:

$$E\left(\overline{y}_{st} - R\overline{x}_{st}\right)^2 = \sum_{i=1}^{N} \frac{S_{hd}^2}{n_h}\left(1 - f_h\right) =$$

$$= \sum_{i=1}^{N} \frac{1 - f_h}{n_h} \frac{\sum_{i=1}^{N_h} \left(d_{h_i} - \overline{D}_h\right)^2}{N_h - 1}$$

por lo que:

$$ECM\left(\widehat{Y}_{R_C}\right) = \frac{1 - f_h}{n_h} \frac{\sum_{i=1}^{N_h} \left(y_{h_i} - Rx_{h_i}\right)^2}{N_h - 1} =$$

$$= \sum_{i=1}^{N_h} \frac{N_h^2\left(1 - f_h\right)}{n_h}\left(S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{xh} S_{yh}\right)$$

2.1.4 Comparación de estimadores separado y combinado.

Es interesante comparar las expresiones de la varianza muestral para los estimadores separado y combinado, las cuales tienen la misma forma excepto en que la razón simple R es sustituida por R_h.

Podemos expresar la diferencia entre los errores cuadráticos medios de la siguiente forma:

$$ECM\left(\widehat{Y}_{R_C}\right) - ECM\left(\widehat{Y}_{R_S}\right) =$$

$$= \sum_{i=1}^{N_h} \frac{N_h^2\left(1 - f_h\right)}{n_h}\left(\left(R^2 - R_h^2\right) S_{xh}^2 - 2\left(R - R_h\right)\rho_h S_{xh} S_{yh}\right) =$$

$$= \sum_{i=1}^{N_h} \frac{N_h^2\left(1 - f_h\right)}{n_h}\left(\left(R - R_h\right)^2 S_{xh}^2 - 2\left(R - R_h\right)\left(\rho_h S_{xh} S_{yh} - R_h S_{xh}^2\right)\right)$$

De aquí se sigue que la diferencia entre los errores depende por una parte de la magnitud de la variación entre las razones de los estratos, y por otra del valor de $\rho_h S_{xh} S_{yh} - R_h S_{xh}^2$.

En las situaciones en las cuales es apropiado el estimador de razón, este término es usualmente pequeño (desaparece si dentro de cada estrato la relación entre y_{h_i} y x_{h_i} es una línea recta a través del orígen)

Entonces a menos que sea R_h constante de estrato a estrato, es probable que sea más preciso el uso del estimador separado.

Esta discusión supone, sin embargo, que la muestra en cada estrato es suficientemente grande, de tal manera que la fórmula aproximada para la varianza del estimador separado es válida. Si se tiene una pequeña muestra en cada estrato, es recomendable el estimador combinado. Un estudio más detallado de la comparación entre el estimador separado y combinado puede encontrarse en *Rao* y *Ramachandran* (1974).

Por último vamos a considerar la estimación de los errores cuadráticos medios. Estos no presentan ningún problema.

$$\widehat{ECM}\left(\hat{Y}_{Rs}\right) = \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h}\left(s_{yh}^2 + \hat{R}_h^2 s_{xh}^2 - 2\hat{R}_h s_{yxh}\right)$$

$$\widehat{ECM}\left(\hat{Y}_{Rc}\right) = \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h}\left(s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R} s_{yxh}\right)$$

siendo $\hat{R}_h = \dfrac{\overline{y}_h}{\overline{x}_h}$ y s_{yh}^2, s_{xh}^2, s_{xyh} las varianzas y covarianzas muestrales en el estrato.

2.1.5 Afijación usando el estimador de razón.

Vamos a determinar la afijación óptima de la muestra entre los diferentes estratos cuando se utiliza el muestreo por razón.

Supongamos una función de coste C, que es función de los tamaños muestrales n_h, y queremos determinar éstos de forma que el coste sea como máximo un cierto valor C_0, y la precisión del estimador sea máxima.

Para ello construimos la función:

$$\phi = ECM\left(\hat{Y}_{R'}\right) + \mu\left(C - C_0\right) = \sum_{h=1}^{L} \frac{N_h^2\left(1-f_h\right)}{n_h} S_{yh}'^2 + \mu\left(C - C_0\right)$$

donde

μ constante

$S'^2_{yh} = S^2_{yh} + R^2_h S^2_{xh} - 2R_h\rho_h S_{xh}S_{yh}$ para el estimador de razón separado y $S'^2_{yh} = S^2_{yh} + R^2 S^2_{xh} - 2R\rho_h S_{xh}S_{yh}$ en el caso del estimador combinado.

Esta función es la misma que se construye para el estimador $\widehat{Y}_{st}$, salvo que S'^2_{yh} sustituye a S^2_{yh}, por tanto los resultados obtenidos para el estimador de razón separado y combinado son los mismos que para el muestreo estratificado aleatorio, sin más que sustituir S^2_{yh} por el valor S'^2_{yh} apropiado.

Así para las funciones de coste más sencillas:

1. $C = cn$ (c coste por unidad de muestra)

2. $C = \sum_{h=1}^{L} c_h n_h$ (c_h coste por unidad del estrato h)

obtenemos los siguientes resultados:

1. $n_h \sim N_h S'_{yh}$

2. $n_h \sim N_h \dfrac{S'_{yh}}{\sqrt{c_h}}$

2.1.6 Estimador estratificado con pesos óptimos.

Como ya es sabido, el muestreo estratificado es una vía tradicional de estimar la media de una población finita y presenta ventajas, en ciertos casos, debido a su mayor precisión con respecto a algunos estimadores clásicos como la media muestral.

Se han dado algunos pasos para mejorar la precisión del estimador estratificado clásico, como son: el problema de afijación muestral de *Tschuprow-Neyman*, la optimización de construción de estratos estudiado por *Dalenius*, etc.

Aquí vamos a introducir el problema de optimización de pesos usados en el estimador tradicional:

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h$$

donde $W_h = \dfrac{N_h}{N}$.

Definición del estimador.

Con objeto de mejorar la precisión del muestreo estratificado, vamos a introducir un nuevo estimador:

Definición 2.1.8 *Se define el estimador estratificado de pesos óptimos como*

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h^* \overline{y}_h$$

de forma que sea insesgado y cuyos pesos sean tales que minimizen la varianza del estimador.

Proposición 2.1.9 *Los pesos óptimos que minimizan la varianza del estimador tradicional*

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h$$

son los dados por

$$W_h^* = \frac{\overline{Y}\,\overline{Y}_h}{V(\overline{y}_h)\,S_{(2)}}$$

donde

$$S_{(2)} = \sum_{h=1}^{L} \frac{\overline{Y}_h^2}{V(\overline{y}_h)}$$

Demostración.-
Queremos minimizar:

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\overline{y}_h)$$

(siempre que la elección de la muestra en cada estrato sea independiente), sujeto a la restricción:

$$\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h$$

Derivando en la expresión

$$L^* = \sum_{h=1}^{L} W_h^2 V(\overline{y}_h) - \lambda \left(\sum_{h=1}^{L} W_h \overline{Y}_h - \overline{Y} \right)$$

respecto a W_h, se obtiene:

$$\frac{\partial L^*}{\partial W_h} = 2W_h V(\overline{y}_h) - \lambda \overline{Y}_h = 0$$

de donde $W_h^* = \dfrac{\lambda \overline{Y}_h}{2V(\overline{y}_h)}$ y como

$$\overline{Y} = \sum_{h=1}^{L} W_h^* \overline{Y}_h = \frac{\lambda}{2} \sum_{h=1}^{L} \frac{\overline{Y}_h^2}{V(\overline{y}_h)}$$

resulta $\lambda = \dfrac{2\overline{Y}}{S_{(2)}}$ donde

$$S_{(i)} = \sum_{h=1}^{L} \frac{\overline{Y}_h^i}{V(\overline{y}_h)} \ , \quad i = 0, 1, 2$$

Entonces los pesos óptimos que minimizan $V(\overline{y}_{st})$ son:

$$W_h^* = \frac{\overline{Y}\,\overline{Y}_h}{V(\overline{y}_h)\, S_{(2)}}$$

donde $\overline{y}_{st}$ es insesgado para tales pesos.

Varianza del estimador.

Proposición 2.1.10 *La varianza óptima viene dada por la expresión:*

$$V_{opt}^*(\overline{y}_{st}) = \frac{\overline{Y}^2}{S_{(2)}}$$

Demostración.-

$$V_{opt}^*(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^{*2} V(\overline{y}_h) = \sum_{h=1}^{L} \left(\frac{\overline{Y}\,\overline{Y}_h}{V(\overline{y}_h)\, S_{(2)}} \right)^2 V(\overline{y}_h) =$$

$$= \sum_{h=1}^{L} \frac{\overline{Y}^2 \overline{Y}_h^2}{V(\overline{y}_h) S_{(2)}^2} = \frac{\overline{Y}^2}{S_{(2)}^2} \sum_{h=1}^{L} \frac{\overline{Y}_h^2}{V(\overline{y}_h)} = \frac{\overline{Y}^2}{S_{(2)}}$$

Este procedimiento es simple y válido en el caso $L \geq 2$ para cualquier estratificación, cualquier afijación y cualquier estimador $\overline{y}_h$ insesgado de $\overline{Y}_h$ que se use.

Corolario 2.1.11 *En muestreo estratificado aleatorio, la varianza óptima viene dada por la expresión:*

$$V_{opt}^{*}\ (\overline{y}_{st}\ ) = \frac{\overline{Y}^2}{\sum_{h=1}^{L} \frac{\overline{Y}_h^2}{(1-f_h)\frac{S_h^2}{n_h}}}$$

Demostración.-
Para el caso del muestreo estratificado aleatorio, se tiene:

$$\overline{y}_h = \sum_{i=1}^{n_h} \frac{y_{i_h}}{n_h}\ ;\quad V(\overline{y}_h) = (1-f_h)\frac{S_h^2}{n_h}$$

y por tanto

$$V_{opt}^{*}\ (\overline{y}_{st}\ ) = \frac{\overline{Y}^2}{\sum_{h=1}^{L} \frac{\overline{Y}_h^2}{(1-f_h)\frac{S_h^2}{n_h}}}$$

El uso de estos pesos requiere el conocimiento de $\overline{Y}_h$ y $V(\overline{y}_h)$ para $h = 1, \ldots, L$, pero es suficiente sustituir estos valores por sus valores muestrales para tener un conocimiento aproximado de estos pesos.

Hay que notar que en este caso óptimo la suma $\sum_{h=1}^{L} W_h^*$ no ha de ser necesariamente 1. Se puede incluir si se quiere una restricción nueva:

$$\sum_{h=1}^{L} W_h = 1$$

Tendremos así una nueva función a minimizar:

$$L = \sum_{h=1}^{L} W_h^2 V(\overline{y}_h) - \lambda_1 \left(\sum_{h=1}^{L} W_h \overline{Y}_h - \overline{Y} \right) - \lambda_2 \left(\sum_{h=1}^{L} W_h - 1 \right)$$

cuya solución es

$$W_h = \frac{\lambda_1 \overline{Y}_h + \lambda_2}{2V(\overline{y}_h)}$$

donde

$$\lambda_1 = \frac{2 - \lambda_2 S_{(0)}}{S_{(1)}} \; ; \quad \lambda_2 = \frac{2\left(\overline{Y} S_{(1)} - S_{(2)}\right)}{S_{(1)}^2 - S_{(0)} S_{(2)}}$$

Con estos dos contrastes nuevos hay una ganancia en precisión válida para $L \geq 3$, pero menor que la dada en el caso anterior.

2.1.7 Estimador de razón de pesos óptimos.

Siguiendo la idea del estimador combinado de *Hansen*, *Hurwitz* y *Gurney* (1946), vamos a definir un nuevo estimador de razón que utiliza en vez de los estimadores estratificados usuales los estimadores con pesos óptimos.

La idea es la siguiente:

Sea

$$\overline{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \overline{y}_h$$

donde

$$W_h^{*} = \frac{\overline{Y}\,\overline{Y}_h}{V(\overline{y}_h)\, S_{(2)y}}$$

con $S_{(i)y} = \sum_{h=1}^{L} \frac{\overline{Y}_h^i}{V(\overline{y}_h)}$ y

$$\overline{x}_{st}^{*} = \sum_{h=1}^{L} Z_h^{*} \overline{x}_h$$

donde

$$Z_h^{*} = \frac{\overline{X}\,\overline{X}_h}{V(\overline{x}_h)\, S_{(2)x}}$$

con $S_{(i)x} = \sum_{h=1}^{L} \frac{\overline{X}_h^i}{V(\overline{x}_h)}$.

Definición 2.1.12 *Definimos el estimador de pesos óptimos de la razón:*

$$\widehat{R}_C^* = \frac{\overline{y}_{st}^*}{\overline{x}_{st}^*}$$

del total:

$$\widehat{Y}_C^* = \widehat{R}_C^* X$$

y de la media:

$$\widehat{\overline{Y}}_C^* = \widehat{R}_C^* \overline{X} = \frac{\widehat{Y}_C^*}{N}$$

En primer lugar veamos que forma adopta este estimador:

$$\widehat{R}_C^* = \frac{\overline{y}_{st}^*}{\overline{x}_{st}^*} = \frac{\displaystyle\sum_{h=1}^{L} \left[\frac{\overline{Y}\,\overline{Y}_h}{V(\overline{y}_h) \displaystyle\sum_{h=1}^{L} \frac{\overline{Y}_h^2}{V(\overline{y}_h)}} \right] \overline{y}_h}{\displaystyle\sum_{h=1}^{L} \left[\frac{\overline{X}\,\overline{X}_h}{V(\overline{x}_h) \displaystyle\sum_{h=1}^{L} \frac{\overline{X}_h^2}{V(\overline{x}_h)}} \right] \overline{x}_h} = \frac{\overline{Y}}{\overline{X}} \frac{S_{(2)x}}{S_{(2)y}} \frac{\displaystyle\sum_{h=1}^{L} \frac{\overline{y}_h \overline{Y}_h}{V(\overline{y}_h)}}{\displaystyle\sum_{h=1}^{L} \frac{\overline{x}_h \overline{X}_h}{V(\overline{x}_h)}}$$

Sesgo.

Proposición 2.1.13 *El sesgo de los estimadores de razón de pesos óptimos de la razón, de la media y del total, vienen dados por las expresiones:*

$$sesgo\left(\widehat{R}_C^*\right) = -\frac{R}{S_{(2)y} S_{(2)x}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h}{V(\overline{y}_h)\, V(\overline{x}_h)} \operatorname{Cov}(\overline{x}_h, \overline{y}_h)$$

$$sesgo\left(\widehat{\overline{Y}}_C^*\right) = \overline{X}\, sesgo\left(\widehat{R}_C^*\right)$$

$$sesgo\left(\widehat{Y}_C^*\right) = X\, sesgo\left(\widehat{R}_C^*\right)$$

Demostración.-
Para calcular el sesgo utilizaremos las siguientes variables:

$$e_1^* = \frac{\overline{y}_{st}^* - \overline{Y}}{\overline{Y}} \quad e_2^* = \frac{\overline{x}_{st}^* - \overline{X}}{\overline{X}}$$

Puesto que $\overline{y}_{st}^*$ y $\overline{x}_{st}^*$ son insesgados, $E(e_1^*) = 0$ y $E(e_2^*) = 0$. Además:

$$E\left(e_1^{*\,2}\right) = \frac{1}{\overline{Y}^2} V\left(\overline{y}_{st}^*\right)$$

$$E\left(e_2^{*\,2}\right) = \frac{1}{\overline{X}^2} V\left(\overline{x}_{st}^*\right)$$

$$E(e_1^* e_2^*) = \frac{1}{\overline{Y}\,\overline{X}} \operatorname{Cov}\left(\overline{y}_{st}^*,\ \overline{x}_{st}^*\right)$$

El estimador $\hat{R}_C^*$ se puede expresar en función de e_1^* y e_2^* en la forma

$$\hat{R}_C^* = R\frac{1+e_1^*}{1+e_2^*} \simeq R(1+e_1^*)(1-e_2^*) \tag{2.1}$$

si se aproxima $(1+e_2^*)^{-1}$ por sus dos primeros términos de su desarrollo en serie de Taylor.

Así

$$E\left(\hat{R}_C^*\right) - R = E[R(1+e_1^*)(1-e_2^*) - R] = RE[(1+e_1^*)(1-e_2^*) - 1] =$$

$$= R[E(e_1^*) - E(e_2^*) - E(e_1^* e_2^*) = -RE(e_1^* e_2^*)] =$$

$$= \frac{-R}{\overline{Y}\,\overline{X}} \operatorname{Cov}\left(\overline{y}_{st}^*, \overline{x}_{st}^*\right) = \frac{-\operatorname{Cov}\left(\overline{y}_{st}^*, \overline{x}_{st}^*\right)}{\overline{X}^2}$$

Calculemos pues $\operatorname{Cov}\left(\overline{y}_{st}^*, \overline{x}_{st}^*\right)$:

$$\operatorname{Cov}\left(\overline{y}_{st}^*, \overline{x}_{st}^*\right) = E\left[\left(\overline{x}_{st}^* - \overline{X}\right)\left(\overline{y}_{st}^* - \overline{Y}\right)\right] =$$

$$= E\left[\left(\sum_{h=1}^{L} Z_h^* \overline{x}_h - \overline{X}\right)\left(\sum_{h=1}^{L} W_h^* \overline{y}_h - \overline{Y}\right)\right] =$$

$$= E\left[\left(\sum_{h=1}^{L} Z_h^* \left(\overline{x}_h - \overline{X}_h\right)\right)\left(\sum_{h=1}^{L} W_h^* \left(\overline{y}_h - \overline{Y}_h\right)\right)\right] =$$

$$= E\left[\sum_{h=1}^{L} Z_h^* W_h^* \left(\overline{x}_h - \overline{X}_h\right)\left(\overline{y}_h - \overline{Y}_h\right) + \sum_{h\neq h'} Z_h^* W_{h'}^* \left(\overline{x}_h - \overline{X}_h\right)\left(\overline{y}_{h'} - \overline{Y}_{h'}\right)\right] =$$

$$= \sum_{h=1}^{L} Z_h^* W_h^* E\left[\left(\overline{x}_h - \overline{X}_h\right)\left(\overline{y}_h - \overline{Y}_h\right)\right] +$$

$$+ \sum_{h\neq h'} Z_h^* W_{h'}^* E\left[\left(\overline{x}_h - \overline{X}_h\right)\left(\overline{y}_{h'} - \overline{Y}_{h'}\right)\right] =$$

$$= \sum_{h=1}^{L} Z_h^* W_h^* \operatorname{Cov}\left(\overline{x}_h, \overline{y}_h\right)$$

Puesto que $\operatorname{Cov}\left(\overline{x}_h, \overline{y}_{h'}\right) = 0$ por ser independientes las muestras en cada estrato, se tiene:

$$sesgo\left(\widehat{R}_C^*\right) = -\frac{\sum_{h=1}^{L} Z_h^* W_h^* \operatorname{Cov}\left(\overline{x}_h, \overline{y}_h\right)}{\overline{X}^2} =$$

$$= -\sum_{h=1}^{L} \frac{\overline{Y}\,\overline{Y}_h \overline{X}\,\overline{X}_h}{\overline{X}^2 V\left(\overline{y}_h\right) V\left(\overline{x}_h\right) S_{(2)y} S_{(2)x}} \operatorname{Cov}\left(\overline{x}_h, \overline{y}_h\right) =$$

$$= -\frac{R}{S_{(2)y} S_{(2)x}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h}{V\left(\overline{y}_h\right) V\left(\overline{x}_h\right)} \operatorname{Cov}\left(\overline{x}_h, \overline{y}_h\right)$$

$$sesgo\left(\widehat{Y}_C^*\right) = X\, sesgo\left(\widehat{R}_C^*\right) = -\frac{Y}{S_{(2)y} S_{(2)x}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h}{V\left(\overline{y}_h\right) V\left(\overline{x}_h\right)} \operatorname{Cov}\left(\overline{x}_h, \overline{y}_h\right)$$

$$sesgo\left(\widehat{\overline{Y}}_C^*\right) = \overline{X}\, sesgo\left(\widehat{R}_C^*\right) = -\frac{\overline{Y}}{S_{(2)y} S_{(2)x}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h}{V\left(\overline{y}_h\right) V\left(\overline{x}_h\right)} \operatorname{Cov}\left(\overline{x}_h, \overline{y}_h\right)$$

Corolario 2.1.14 *En muestreo estratificado aleatorio el sesgo del estimador de razón de pesos óptimos de la razón tiene una expresión aproximada dada por la expresión:*

$$sesgo\left(\widehat{R}_C^*\right) \simeq \frac{-R}{S_{(2)y}S_{(2)x}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h \rho_h}{\frac{(1-f_h)}{n_h} S_{yh} S_{xh}}$$

Demostración.-
En el caso del muestreo estratificado aleatorio, se tiene:

$$V(y_h) = (1-f_h)\frac{S_{yh}^2}{n_h} \; ; \quad V(x_h) = (1-f_h)\frac{S_{xh}^2}{n_h}$$

$$\mathrm{Cov}\left(\overline{x}_h, \overline{y}_h\right) = \frac{1-f_h}{n_h}\rho_h S_{xh} S_{yh}$$

Entonces:

$$sesgo\left(\widehat{R}_C^*\right) \simeq \frac{-R}{\sum_{h=1}^{L} \frac{\overline{Y}_h^2 n_h}{(1-f_h)S_{yh}^2} \sum_{h=1}^{L} \frac{\overline{X}_h^2 n_h}{(1-f_h)S_{xh}^2}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h}{V(\overline{y}_h)V(\overline{x}_h)} \mathrm{Cov}\left(\overline{x}_h, \overline{y}_h\right) =$$

$$= \frac{-R}{S_{(2)y}S_{(2)x}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h}{\frac{(1-f_h)^2}{n_h^2} S_{yh}^2 S_{xh}^2} \left(\frac{1-f_h}{n_h}\right) \rho_h S_{xh} S_{yh} =$$

$$= \frac{-R}{S_{(2)y}S_{(2)x}} \sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h \rho_h}{\frac{(1-f_h)}{n_h} S_{yh} S_{xh}}$$

Error cuadrático medio.

Proposición 2.1.15 *Una aproximación del error cuadrático medio del estimador de razón combinado de pesos óptimos viene dada por la expresión:*

$$ECM\left(\widehat{R}_C^*\right) = R^2\left[\frac{1}{S_{(2)y}} + \frac{1}{S_{(2)x}} - \frac{2}{S_{(2)x}S_{(2)y}}\left(\sum_{h=1}^{L} \frac{\overline{Y}_h \overline{X}_h}{V(\overline{y}_h)V(\overline{x}_h)} \mathrm{Cov}\left(\overline{y}_h, \overline{x}_h\right)\right)\right]$$

Demostración.-
Partimos de la aproximación 2.1:

$$\hat{R}_C^* = R\frac{1+e_1^*}{1+e_2^*} \simeq R\,(1+e_1^*)\,(1-e_2^*)$$

y con ella:

$$ECM\left(\hat{R}_C^*\right) = E\left(\hat{R}_C^* - R\right)^2 = E\left[R\,(1+e_1^*)\,(1-e_2^*) - R\right]^2 =$$
$$= R^2 E\left[(1+e_1^*)\,(1-e_2^*) - 1\right]^2 = R^2\left[e_1^* - e_2^* - e_1^* e_2^*\right]^2$$

Si nos quedamos sólo con los términos de grado inferior o igual a dos en e_1^* y e_1^*, se obtiene:

$$ECM\left(\hat{R}_C^*\right) \simeq R^2 E\left[e_1^{*\,2} + e_2^{*\,2} - 2e_1^* e_2^*\right] =$$
$$= R^2\left[E\left(e_1^{*\,2}\right) + E\left(e_2^{*\,2}\right) - 2E\left(e_1^* e_2^*\right)\right] \qquad (2.2)$$

Sustituyendo ahora los valores:

$$E\left(e_1^{*\,2}\right) = \frac{V\left(\overline{y}_{st}^*\right)}{\overline{Y}^2}\;;\quad E\left(e_2^{*\,2}\right) = \frac{V\left(\overline{x}_{st}^*\right)}{\overline{X}^2}$$

$$E\left(e_1^* e_2^*\right) = \frac{\mathrm{Cov}\left(\overline{y}_{st}^*, \overline{x}_{st}^*\right)}{\overline{Y}\;\overline{X}}$$

en la expresión 2.2, se obtiene:

$$ECM\left(\hat{R}_C^*\right) = R^2\left[\frac{1}{\overline{Y}^2}\frac{\overline{Y}^2}{S_{(2)y}} + \frac{1}{\overline{X}^2}\frac{\overline{X}^2}{S_{(2)x}} - 2\frac{1}{\overline{X}\;\overline{Y}}\sum_{h=1}^{L} W_h^* Z_h^*\,\mathrm{Cov}\left(\overline{y}_h, \overline{x}_h\right)\right] =$$

$$= R^2\left[\frac{1}{S_{(2)y}} + \frac{1}{S_{(2)x}} - 2\frac{1}{\overline{X}\;\overline{Y}}\sum_{h=1}^{L}\frac{\overline{Y}\;\overline{Y}_h}{V\left(\overline{y}_h\right)S_{(2)y}}\frac{\overline{X}\;\overline{X}_h}{V\left(\overline{x}_h\right)S_{(2)x}}\,\mathrm{Cov}\left(\overline{y}_h, \overline{x}_h\right)\right] =$$

$$= R^2\left[\frac{1}{S_{(2)y}} + \frac{1}{S_{(2)x}} - \frac{2}{S_{(2)x}S_{(2)y}}\left(\sum_{h=1}^{L}\frac{\overline{Y}_h\overline{X}_h}{V\left(\overline{y}_h\right)V\left(\overline{x}_h\right)}\,\mathrm{Cov}\left(\overline{y}_h, \overline{x}_h\right)\right)\right]$$

como queríamos demostrar.

Corolario 2.1.16 *Unas aproximaciones de los errores cuadráticos medios de los estimadores* $\widehat{Y}_C^*$ *e* $\widehat{\overline{Y}}_C^*$ *vienen dadas por las siguientes expresiones:*

$$ECM\left(\widehat{Y}_C^*\right) = Y^2\left[\frac{1}{S_{(2)y}} + \frac{1}{S_{(2)x}} - \frac{2}{S_{(2)x}S_{(2)y}}\left(\sum_{h=1}^{L}\frac{\overline{Y}_h\overline{X}_h}{V\left(\overline{y}_h\right)V\left(\overline{x}_h\right)}\operatorname{Cov}\left(\overline{y}_h,\overline{x}_h\right)\right)\right]$$

$$ECM\left(\widehat{\overline{Y}}_C^*\right) = \overline{Y}^2\left[\frac{1}{S_{(2)y}} + \frac{1}{S_{(2)x}} - \frac{2}{S_{(2)x}S_{(2)y}}\left(\sum_{h=1}^{L}\frac{\overline{Y}_h\overline{X}_h}{V\left(\overline{y}_h\right)V\left(\overline{x}_h\right)}\operatorname{Cov}\left(\overline{y}_h,\overline{x}_h\right)\right)\right]$$

Demostración.-
Es inmediata a partir de la relación entre los estimadores.

Corolario 2.1.17 *En muestreo estratificado aleatorio el error cuadrático medio del estimador de razón combinado de pesos óptimos viene dado por la expresión:*

$$ECM\left(\widehat{R}_C^*\right) = R^2\left[\frac{1}{S_{(2)y}} + \frac{1}{S_{(2)x}} - \frac{2}{S_{(2)x}S_{(2)y}}\left(\sum_{h=1}^{L}\frac{\overline{Y}_h\overline{X}_h}{\dfrac{1-f_h}{n_h}S_{yh}S_{xh}}\rho_h\right)\right]$$

donde

$$S_{(2)y} = \sum_{h=1}^{L}\frac{\overline{Y}_h^2 n_h}{\left(1-f_h\right)S_{yh}^2}$$

$$S_{(2)x} = \sum_{h=1}^{L}\frac{\overline{X}_h^2 n_h}{\left(1-f_h\right)S_{xh}^2}$$

Demostración.-
En el muestreo estratificado aleatorio:

$$V\left(y_h\right) = \left(1-f_h\right)\frac{S_{yh}^2}{n_h}\ ;\quad V\left(x_h\right) = \left(1-f_h\right)\frac{S_{xh}^2}{n_h}$$

$$\operatorname{Cov}\left(\overline{y}_h,\overline{x}_h\right) = \frac{1-f_h}{n}\rho_h S_{xh}S_{yh}$$

y por la proposición 2.1.15, se concluye sin más que sustituir.

§2.2 El estimador de razón en el muestreo con probabilidades desiguales.

2.2.1 Introducción.

El estudio del estimador de razón realizado en el capítulo 1 está centrado en suponer que la muestra se elige mediante muestreo aleatorio simple.

Las extracciones con probabilidades iguales estan basadas en la idea de que si no se conoce nada acerca de las variables a estudiar, o si no se puede utilizar dicha información, lo mejor es asignar probabilidades iguales a las unidades de la población. Sin embargo lo usual es que el diseñador de la encuesta posea información acerca de las unidades poblacionales que le lleve a dar más peso a algunas de ellas, asignándoles probabilidades según su importancia.

En este apartado vamos a estudiar la estimación de razón bajo un esquema de muestreo con probabilidades desiguales y con reemplazo. Se propone un estimador basado en el cociente de estimadores de *Hansen* y *Hurwitz* (1943) para la media de las variables principal y auxiliar, y se estudian las propiedades de dicho estimador.

2.2.2 Definición del estimador.

Consideremos una población de N unidades de la que se selecciona una muestra de n unidades con probabilidades desiguales y con reemplazo.

Sea p_i la probabilidad de seleccionar la unidad u_i en cada extracción ($i = 1, \ldots, N$). Es conocido en la teoría de muestras que los estimadores de *Hansen* y *Hurwitz*

$$\overline{y}_{HH} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{np_i}$$

$$\overline{x}_{HH} = \frac{1}{N}\sum_{i=1}^{n}\frac{x_i}{np_i}$$

son insesgados de $\overline{Y}$ y $\overline{X}$, respectivamente.

Definición 2.2.18 *En un muestreo con probabilidades desiguales, llamamos estimadores de la razón, del total y de la media a*

$$\widehat{R}_D = \frac{\overline{y}_{HH}}{\overline{x}_{HH}} \; ; \quad \widehat{Y}_{RD} = \widehat{R}_D X \; ; \quad \widehat{\overline{Y}}_{RD} = \widehat{R}_D \overline{X}$$

Propiedades de los estimadores.

Los estimadores $\widehat{R}_D$, $\widehat{Y}_{RD}$ y $\widehat{\overline{Y}}_{RD}$ no son consistentes, lo cual no es extraño puesto que el estimador de *Hansen* y *Hurwitz* tampoco lo es.

Sesgo.

Estos estimadores son también sesgados, pero bajo ciertas condiciones proporcionan estimaciones más precisas que los estimadores de expansión simple.

Vamos a obtener, en primer lugar, una cota para el sesgo relativo del estimador. Para ello definimos las variables:

$$z_i = \frac{y_i}{Np_i} \; ; \quad v_i = \frac{x_i}{Np_i}$$

Entonces $\overline{z} = \overline{y}_{HH}$ y $\overline{v} = \overline{x}_{HH}$ y por tanto el estimador de razón se puede escribir de la forma:

$$\widehat{R}_D = \frac{\overline{z}}{\overline{v}}$$

Proposición 2.2.19 *En muestreo con probabilidades desiguales y con reemplazo, una cota para el sesgo relativo del estimador de razón viene dada por la expresión:*

$$\frac{\left| sesgo\left(\widehat{\overline{Y}}_{RD}\right)\right|}{\sigma_{\widehat{\overline{Y}}_{RD}}} \leq \frac{C_v}{\sqrt{n}}$$

donde C_v es el coeficiente de variación de la variable v.

Demostración.-
Puesto que

$$sesgo\left(\widehat{\overline{Y}}_{RD}\right) = E\left(\widehat{R}_D\right)\overline{X} - \overline{Y} = E\left(\widehat{R}_D\right)E\left(\overline{v}\right) - E\left(\widehat{R}_D\overline{v}\right) =$$

$$= \text{Cov}\left(\widehat{R}_D, \overline{v}\right) \leq \sigma_{\widehat{R}_D}\sigma_{\overline{v}}$$

Entonces

$$\frac{sesgo\left(\widehat{\overline{Y}}_{RD}\right)}{\sigma_{\widehat{\overline{Y}}_{RD}}} \leq C_{\overline{v}}$$

donde $C_{\overline{v}} = \dfrac{\sigma_{\overline{v}}}{E(\overline{v})}$ y como $\sigma^2_{\overline{v}} = \dfrac{\sigma^2_v}{n}$, se obtiene el resultado.

El paso siguiente será determinar, como en el caso del muestreo aleatorio simple, una aproximación del sesgo.

Proposición 2.2.20 *Una aproximación del sesgo del estimador* $\widehat{\overline{Y}}_{RD}$ *viene dada por la expresión*

$$sesgo\left(\widehat{\overline{Y}}_{RD}\right) \simeq \frac{\overline{Y}}{n}\left(C_v^2 - \rho_{vz}C_vC_z\right)$$

Considerando las variables

$$o_1 = \frac{\overline{z} - \overline{Y}}{\overline{Y}} \; ; \quad o_2 = \frac{\overline{v} - \overline{X}}{\overline{X}}$$

podemos expresar el estimador de la forma:

$$\widehat{\overline{Y}}_{RD} = \overline{Y}(1 + o_1)(1 + o_2)^{-1}$$

Desarrollando $(1 + o_2)^{-1}$ en serie de Taylor y quedándonos sólo con los términos de orden inferior o igual a dos en o_1 y o_2, tenemos:

$$\widehat{\overline{Y}}_{RD} - \overline{Y} \simeq \overline{Y}\left(o_1 - o_2 + o_2^2 - o_1o_2\right)$$

y por tanto

$$sesgo\left(\widehat{\overline{Y}}_{RD}\right) \simeq \overline{Y}\left(E(o_1) - E(o_2) + E\left(o_2^2\right) - E(o_1o_2)\right)$$

Ahora bien:

$$E(o_1) = E(o_2) = 0$$

$$E\left(o_2^2\right) = \frac{1}{\overline{X}^2}\frac{1}{n}\sum_{i=1}^{N} p_i\left(v_i - \overline{X}\right)^2$$

$$E\left(o_1 o_2\right) = \frac{1}{n\overline{X}\,\overline{Y}}\sum_{i=1}^{N} p_i\left(z_i - \overline{Y}\right)\left(v_i - \overline{X}\right)$$

y sin más que sustituir se obtiene el resultado deseado.

Corolario 2.2.21 *En una muestra concreta se puede obtener una estimación del sesgo del estimador* $\widehat{\overline{Y}}_{RD}$ *mediante*

$$\widehat{sesgo}\left(\widehat{\overline{Y}}_{RD}\right) = \frac{\overline{y}_{HH}}{n}\left(\frac{s_v^2}{\overline{X}^2} - \frac{rs_z s_v}{\overline{X}\overline{y}_{HH}}\right)$$

donde

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \overline{y}_{HH}\right)^2 \;;\quad s_v^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(v_i - \overline{x}_{HH}\right)^2$$

$$rs_z s_v = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \overline{y}_{HH}\right)\left(v_i - \overline{x}_{HH}\right)$$

Error cuadrático medio.

Puesto que el estimador es segado, su precisión se medirá por su error cuadrático medio que procedemos a estudiar.

Proposición 2.2.22 *Una aproximación del error cuadrático medio del estimador* $\widehat{\overline{Y}}_{RD}$ *viene dada por la expresión:*

$$ECM\left(\widehat{\overline{Y}}_{RD}\right) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(z_i - Rv_i\right)^2$$

Demostración.-
Partiendo de la expresión:

$$\widehat{\overline{Y}}_{RD} = \overline{Y}\left(1 + o_1\right)\left(1 + o_2\right)^{-1}$$

desarrollando $(1+o_2)^{-1}$ en serie de potencias, elevando al cuadrado y reteniendo los términos de orden inferior o igual a dos en o_1 y o_2, obtenemos:

$$\left(\widehat{\overline{Y}}_{RD}-\overline{Y}\right)^2 \simeq \overline{Y^2}\left(o_1^2+o_2^2-2o_1o_2\right)$$

y así

$$ECM\left(\widehat{\overline{Y}}_{RD}\right) \simeq \overline{Y}^2\left(E\left(o_1^2\right)+E\left(o_2^2\right)-2E\left(o_1o_2\right)\right) =$$

$$= \overline{Y}^2\left(\frac{\sigma_z^2}{n\overline{Y}^2}+\frac{\sigma_v^2}{n\overline{X}^2}-2\frac{\rho\sigma_z\sigma_v}{n\overline{XY}}\right)$$

Si sustituimos σ_z^2, σ_v^2 y $\rho\sigma_z\sigma_v$ por sus respectivos valores, el error cuadrático medio toma la expresión:

$$ECM\left(\widehat{\overline{Y}}_{RD}\right) \simeq \frac{1}{n}\sum_{i=1}^{N} p_i\left(z_i-Rv_i\right)^2$$

Corolario 2.2.23 *En una muestra concreta, una estimación del error cuadrático medio del estimador* $\widehat{\overline{Y}}_{RD}$ *viene dada por*

$$\widehat{ECM}\left(\widehat{\overline{Y}}_{RD}\right) = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(z_i-\hat{R}v_i\right)^2\right]$$

Demostración.-
Puesto que

$$\sum_{i=1}^{N} p_i\left(z_i-Rv_i\right)^2$$

puede considerarse como la varianza de la variable $u_i = z_i - Rv_i$, un estimador de esa varianza es

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(z_i-\hat{R}v_i\right)^2$$

y sustituyendolo se obtiene un estimador para $ECM\left(\widehat{\overline{Y}}_{RD}\right)$.

Precisión de la aproximación del sesgo y del error cuadrático medio.

Consideremos el sesgo exacto del estimador $\widehat{\overline{Y}}_{RD}$:

$$sesgo\left(\widehat{\overline{Y}}_{RD}\right) = \overline{Y}\,\overline{X}E\left[\frac{(o_1 - o_2)}{\overline{v}}\right]$$

y consideremos la aproximación de este sesgo dada por:

$$sesgo_k\left(\widehat{\overline{Y}}_{RD}\right) = \overline{Y}E\left\{\sum_{i=0}^{2k-1}(-1)^i o_2^i\,(o_1 - o_2)\right\}$$

para k entero y positivo.

Para $k = 1$ obtenemos la expresión antes considerada:

$$sesgo_1\left(\widehat{\overline{Y}}_{RD}\right) = \frac{\overline{Y}}{n}\left(C_v^2 - \rho_{vz}C_vC_z\right)$$

Siguiendo el mismo desarrollo que en el capítulo 1 se demuestra que

$$\left|sesgo\left(\widehat{\overline{Y}}_{RD}\right) - sesgo_k\left(\widehat{\overline{Y}}_{RD}\right)\right| \leq \left|\frac{\overline{Y}\,\overline{X}}{v_0}E\left(o_1o_2^{2k} - o_2^{2k+1}\right)\right|$$

siendo v_0 una cota inferior de $\overline{v}$.

Ahora bien

$$E\left(o_1o_2^{2k}\right) = \frac{1}{\overline{Y}}\frac{1}{\overline{X}^{2k}}E\left[\left(\overline{z} - \overline{Y}\right)\left(\overline{v} - \overline{X}\right)\right]^{2k}$$

y

$$E\left(o_2^{2k+1}\right) = \frac{1}{\overline{X}^{2k+1}}E\left[\left(\overline{v} - \overline{X}\right)\right]^{2k+1}$$

$E\left[\left(\overline{z} - \overline{Y}\right)\left(\overline{v} - \overline{X}\right)\right]^{2k}$ es de orden $O\left(n^{-(k+1)}\right)$ y $E\left[\left(\overline{v} - \overline{X}\right)\right]^{2k+1}$ es de orden $O\left(n^{-(k+1)}\right)$ como demostraron *Hansen, Hurwitz* y *Madow* (1953) para un muestreo con reemplazo en poblaciones finitas y por tanto

$$\left|sesgo\left(\widehat{\overline{Y}}_{RD}\right) - sesgo_k\left(\widehat{\overline{Y}}_{RD}\right)\right| = O\left(n^{-(k+1)}\right) \qquad (2.3)$$

y así, eligiendo n y k apropiadamente $\left|sesgo\left(\widehat{\overline{Y}}_{RD}\right) - sesgo_k\left(\widehat{\overline{Y}}_{RD}\right)\right|$ se puede tomar muy pequeño en comparación con $sesgo_k\left(\widehat{\overline{Y}}_{RD}\right)$, por lo que esta será una buena aproximación para $sesgo\left(\widehat{\overline{Y}}_{RD}\right)$.

Proposición 2.2.24 *La expresión*

$$sesgo_1\left(\widehat{\overline{Y}}_{RD}\right) \simeq \frac{\overline{Y}}{n}\left(C_v^2 - \rho_{vz}C_vC_z\right)$$

es una aproximación de sesgo $\left(\widehat{\overline{Y}}_{RD}\right)$ *de orden* $O(n^{-2})$.

Demostración.-

Inmediata sin más que considerar la expresión 2.3 para $k = 1$.

Del mismo modo podemos estudiar la precisión de la aproximación obtenida para el error cuadrático medio del estimador.

Escribiendo este error en función de las variables o_1 y o_2 adopta la forma:

$$ECM\left(\widehat{\overline{Y}}_{RD}\right) = E\left(\widehat{\overline{Y}}_{RD} - \overline{Y}\right)^2 = \overline{X}^2\overline{Y}^2E\left(\frac{(o_1 - o_2)^2}{\overline{v}^2}\right)$$

Dada la aproximación:

$$ECM_k\left(\widehat{\overline{Y}}_{RD}\right) = \overline{Y}^2E\left[(o_1 - o_2)^2\sum_{i=0}^{2k-2}(-1)^i o_2^i(i+1)\right]$$

operando, llegamos a

$$ECM\left(\widehat{\overline{Y}}_{RD}\right) - ECM_k\left(\widehat{\overline{Y}}_{RD}\right) =$$

$$= -\overline{Y}^2\overline{X}^2E\left[\frac{(o_1 - o_2)^2}{\overline{v}^2}\left(2ko_2^{2k-1} + (2k-1)o_2^{2k}\right)\right]$$

y por tanto

$$\left|ECM\left(\widehat{\overline{Y}}_{RD}\right) - ECM_k\left(\widehat{\overline{Y}}_{RD}\right)\right| \leq$$

$$\leq \frac{\overline{Y}^2\overline{X}^2}{v_0^2} E\left[(o_1 - o_2)^2 \left(2ko_2^{2k-1} + (2k-1)o_2^{2k}\right)\right]$$

siendo v_0 una cota inferior de $\overline{v}$.

Ahora bien, $E\left(o_1^2 o_2^{2k-1}\right)$, $E\left(o_2^{2k+1}\right)$ y $E\left(o_1 o_2^{2k}\right)$ son de orden $O\left(n^{-(k+1)}\right)$ y por tanto

$$\left|ECM\left(\widehat{\overline{Y}}_{RD}\right) - ECM_k\left(\widehat{\overline{Y}}_{RD}\right)\right| = O\left(n^{-(k+1)}\right)$$

Para el caso particular de $k = 1$, obtenemos la siguiente proposición:

Proposición 2.2.25 *La expresión*

$$ECM_1\left(\widehat{\overline{Y}}_{RD}\right) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(z_i - Rv_i\right)^2$$

es una aproximación del valor $ECM\left(\widehat{\overline{Y}}_{RD}\right)$ *de orden* $O(n^{-2})$.

2.2.3 Condiciones bajo las cuales el estimador de razón es insesgado.

Vamos a considerar en este apartado el caso en que las probabilidades de selección de cada unidad sean proporcionales a la variable auxiliar:

$$p_i = \frac{x_i}{X}$$

Para este caso particular, el estimador aquí propuesto coincide con el estimador usado por *Barnett* (1982):

$$\widehat{\overline{Y}}_{RD} = \widehat{R}_D\overline{X} = \frac{\overline{y}_{HH}}{\overline{x}_{HH}}\overline{X} = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{Np_i}}{\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{Np_i}}\overline{X} = \frac{\overline{X}}{n}\sum_{i=1}^{n}\frac{y_i}{x_i} = \overline{X}\overline{r}$$

donde $r_i = \dfrac{y_i}{x_i}$ y $\overline{r} = \displaystyle\sum_{i=1}^{n}\frac{r_i}{n}$.

Proposición 2.2.26 *Si las probabilidades se eligen proporcionalmente a la variable auxiliar, el estimador de razón* $\widehat{\overline{Y}}_{RD}$ *es insesgado y una expresión aproximada de su varianza es:*

$$V\left(\widehat{\overline{Y}}_{RD}\right) \simeq \frac{\overline{X}^2}{n}\sum_{i=1}^{N} p_i\left(r_i - R\right)^2$$

Demostración.-
$\widehat{\overline{Y}}_{RD}$ es insesgado puesto que $\widehat{\overline{Y}}_{RD} = \overline{X}\overline{r}$ y $E\left(\overline{r}\right) = E\left(r_i\right) = R$.
Por otra parte, se tiene $z_i = r_i\overline{X}$, $v_i = \overline{X}$ y por tanto

$$V\left(\widehat{\overline{Y}}_{RD}\right) \simeq \frac{1}{n}\sum_{i=1}^{N} p_i\left(z_i - Rv_i\right)^2 = \frac{\overline{X}^2}{n}\sum_{i=1}^{N} p_i\left(r_i - R\right)^2$$

2.2.4 Comparación con el estimador de expansión simple.

Vamos a comparar la precisión del estimador de razón, $\widehat{\overline{Y}}_{RD}$, con la del estimador de expansión simple, $\overline{y}_{HH}$, bajo este esquema de muestreo con reemplazo y probabilidades desiguales.

Proposición 2.2.27 *El estimador* $\widehat{\overline{Y}}_{RD}$ *es más eficiente que el estimador de expansión simple,* $\overline{y}_{HH}$ *sí y sólo si*

$$\rho\frac{C_z}{C_v} > \frac{1}{2}$$

Demostración.-
Si no se utiliza información auxiliar, el estimador $\overline{y}_{HH}$ tiene por varianza

$$V\left(\overline{y}_{HH}\right) = \frac{1}{n}\sigma_z^2$$

Además, el error cuadrático medio del estimador $\widehat{\overline{Y}}_{RD}$ adopta la forma:

$$ECM\left(\widehat{\overline{Y}}_{RD}\right) = \frac{1}{n}\left(\sigma_z^2 + R^2\sigma_v^2 - 2R\rho\sigma_z\sigma_v\right)$$

Por tanto, el método de estimación de razón será más eficaz que el de expansión simple si:

$$R^2\sigma_z^2 - 2R\rho\sigma_z\sigma_v < 0$$

es decir

$$\frac{\overline{Y}}{\overline{X}}\frac{\sigma_v}{\sigma_z} < 2\rho \quad \Leftrightarrow \quad \rho\frac{C_z}{C_v} > \frac{1}{2}$$

Aquí hemos considerado dos tipos importantes de muestreo, aunque el método de estimación de razón proporciona buenos resultados en otros tipos de diseños muestrales.

Cuando la media, $\overline{X}$, es desconocida puede ser estimada por $\overline{x}_1$, obtenida de una muestra aleatoria simple inicial de tamaño n_1. Una submuestra de tamaño n de esta muestra inicial proporciona las medias $(\overline{x}, \overline{y})$ y permite la construcción del estimador de razón. La aplicación del estimador de razón en el muestreo bifásico ha sido considerada por *Rao* (1975 a, 1975 b, 1981) y por *Bose* (1943).

El estimador de razón en el muestreo polietápico ha sido considerado por *Williams* (1961, 1962) y por *Rao* (1964).

Swain (1964) y *Singh* (1966) consideran la estimación de razón en el muestreo sistemático.

Capítulo 3

Extensión del método de estimación de razón al caso de varias variables auxiliares.

En los dos primeros capítulos hemos estudiado cómo utilizar la información auxiliar que proporciona una variable positivamente correlacionada con la variable objeto de estudio, y hemos estudiado distintos estimadores de razón que proporcionan, bajo ciertas condiciones, ganancias en precisión respecto a los estimadores de expansión simple.

Algunas veces el organizador de una encuesta dispone de información acerca de varias variables auxiliares y en tales circunstancias es importante saber utilizar toda la información disponible, tanto en la fase de diseño como en la de estimación.

Una primera posibilidad consiste en determinar la variable que tiene mayor correlación con la variable principal y construir a partir de ella el estimador de razón, despreciando el resto de las variables. Sin embargo, es obvio que este procedimiento implica un importante desperdicio de recursos.

Olkin (1958) propuso un estimador de razón en muestreo aleatorio simple que utiliza varias variables auxiliares. La idea de este trabajo es construir con cada variable auxiliar un estimador de razón y hacer una combinación lineal de ellos. En este capítulo comenzamos con el estudio de este estimador y después, siguiendo su idea, construiremos otros estimadores de razón para muestras estratificadas y muestras con probabilidades desiguales.

Posteriormente se proponen estimadores que hemos llamado condensados

y que responden a una nueva idea: condensar la información de todas las variables auxiliares en una nueva variable auxiliar y a partir de ella construir un estimador de razón. Según se determine la variable "condensada" se obtendrán estimadores distintos.

Por último se propone un nuevo estimador, que hemos llamado estimador iterado de razón, que se basa en un procedimiento iterativo en el cual el estimador de $\overline{Y}$ obtenido en el paso $i-1$ se utiliza, en vez del estimador de expansión simple, para construir el estimador de razón en el paso i. Este nuevo estimador presentará grandes ventajas en cuanto a su puesta en práctica, respecto a los otros considerados.

§3.1 El estimador de razón multivariante en el muestreo aleatorio simple.

Este método de estimación en el muestreo aleatorio simple fue propuesto por *Olkin* (1958).

Representamos por y el caracter o variable que constituye el objeto de estudio, y por $x_1, x_2, \ldots, x_k$ las variables auxiliares que suponemos correlacionadas con la primera.

Si queremos estimar el total poblacional Y, y conocemos los totales poblacionales $X_1, X_2, \ldots, X_k$, podemos utilizar además del estimador directo o expandido:

$$\widehat{Y} = N\overline{y} \tag{3.1}$$

otros estimadores:

$$\widehat{Y}_{R_i} = \frac{\overline{y}}{\overline{x}_i} X_i \quad i = 1, \ldots, k \tag{3.2}$$

donde:

$\overline{y}$ es la media muestral de y.

$\overline{x}_i$ es la media muestral de x_i, $i = 1, \ldots, k$.

Estos estimadores $\widehat{Y}_{R_i}$ son los estimadores de razón univariantes del total Y, que sabemos proporcionan mejores estimaciones que el estimador directo, siempre que la razón entre las variables permanezca aproximadamente constante. Vamos a estudiar a continuación un estimador que combine estos estimadores, de forma que utilice simultáneamente la información auxiliar que proporcionan las variables $x_1, \ldots, x_k$.

Definición 3.1.1 *Se llama estimador de razón multivariante del total Y, al estimador:*

$$\widehat{Y}_R^{1,2,\ldots,k} = w_1\widehat{Y}_{R_1} + w_2\widehat{Y}_{R_2} + \cdots + w_k\widehat{Y}_{R_k} \tag{3.3}$$

siendo $w_1, w_2, \ldots, w_k$ pesos definidos de forma que $\sum_{i=1}^{k} w_i = 1$.

Si interesa estimar la media poblacional $\overline{Y}$ se puede definir el estimador:

Definición 3.1.2 *Se llama estimador de razón multivariante de la media $\overline{Y}$, al estimador:*

$$\widehat{\overline{Y}}_R^{1,2,\ldots,k} = N^{-1}\widehat{Y}_R^{1,2,\ldots,k} \tag{3.4}$$

Vamos a pasar a continuación a estudiar las principales propiedades del estimador así definido. El estudio se hace, en primer lugar, para el caso bivariante. La extensión al caso de más variables se estudia después.

3.1.1 El estimador de razón bivariante.

Se llama estimador de razón bivariante del total Y, al siguiente estimador:

$$\widehat{Y}_{R_{12}} = w_1\widehat{Y}_{R_1} + w_2\widehat{Y}_{R_2} \tag{3.5}$$

siendo w_1, w_2 pesos definidos de forma que $\sum_{i=1}^{2} w_i = 1$.

Si interesa estimar la media poblacional $\overline{Y}$ se utiliza el estimador

$$\widehat{\overline{Y}}_{R_{12}} = N^{-1}\widehat{Y}_{R_{12}} \tag{3.6}$$

Estos estimadores son consistentes pues son funciones lineales de estimadores consistentes. Sin embargo, no son insesgados, como vamos a comprobar a continuación.

Sesgo.

Comenzamos estudiando el caso del estimador del total. A partir de él, es muy fácil calcular el sesgo del estimador de la media.

De la expresión 3.5 se obtiene:

$$sesgo\left(\widehat{Y}_{R_{12}}\right) = w_1 sesgo\left(\widehat{Y}_{R_1}\right) + w_2 sesgo\left(\widehat{Y}_{R_2}\right)$$

Puesto que la muestra ha sido elegida por muestreo aleatorio simple, tenemos:

$$sesgo\left(\widehat{Y}_{R_i}\right) = -N\,\mathrm{Cov}\left(\widehat{R}_i, \overline{x}_i\right) \quad i = 1, 2$$

con $\widehat{R}_i = \dfrac{\overline{y}}{\overline{x}_i}$, por lo que

$$sesgo\left(\widehat{Y}_{R_{12}}\right) = -N\left\{w_1\,\mathrm{Cov}\left(\widehat{R}_1, \overline{x}_1\right) + w_2\,\mathrm{Cov}\left(\widehat{R}_2, \overline{x}_2\right)\right\} \tag{3.7}$$

Se obtiene así la siguiente proposición:

Proposición 3.1.3 *Una cota para el sesgo de los estimadores de razón* $\widehat{Y}_{R_{12}}$ *e* $\widehat{\overline{Y}}_{R_{12}}$ *es la siguiente:*

$$\mid sesgo\left(\widehat{Y}_{R_{12}}\right) \mid \leq N\,máx\left(\sigma_{\widehat{R}_1}\sigma_{\overline{x}_1}, \sigma_{\widehat{R}_2}\sigma_{\overline{x}_2}\right)$$

$$\mid sesgo\left(\widehat{\overline{Y}}_{R_{12}}\right) \mid \leq máx\left(\sigma_{\widehat{R}_1}\sigma_{\overline{x}_1}, \sigma_{\widehat{R}_2}\sigma_{\overline{x}_2}\right)$$

Demostración.-
Es inmediata a partir de la expresión 3.7 y teniendo en cuenta que

$$\mid \rho_i \mid \leq 1 \Rightarrow \mid \mathrm{Cov}\left(\widehat{R}_i, \overline{x}_i\right) \mid \leq \sigma_{\widehat{R}_i}\sigma_{\overline{x}_i} \quad \forall i$$

Tenemos así una cota para el sesgo del estimador del total, a partir de la cual se puede obtener fácilmente una cota para el estimador de la media poblacional. En efecto, puesto que $\widehat{Y}_{R_{12}} = N\widehat{\overline{Y}}_{R_{12}}$, se tiene que

$$sesgo\left(\widehat{Y}_{R_{12}}\right) = N\,sesgo\left(\widehat{\overline{Y}}_{R_{12}}\right)$$

por lo que una cota para $sesgo\left(\widehat{\overline{Y}}_{R_{12}}\right)$ será:

$$\mid sesgo\left(\widehat{\overline{Y}}_{R_{12}}\right) \mid \leq máx\left(\sigma_{\widehat{R}_1}\sigma_{\overline{x}_1}, \sigma_{\widehat{R}_2}\sigma_{\overline{x}_2}\right)$$

Pero estas cotas no suelen ser útiles pues dependen de las varianzas de $\widehat{R}_1$ y $\widehat{R}_2$ en general desconocidas.

Hay veces que es más útil obtener una aproximación del sesgo, lo que procedemos a hacer a continuación.

Proposición 3.1.4 *Unas aproximaciones de orden $O(n^{-2})$ del sesgo de los estimadores de $\widehat{Y}_R$ e $\widehat{\overline{Y}}_R$ vienen dadas por las expresiones siguientes:*

$$sesgo\left(\widehat{Y}_{R_{12}}\right) \simeq N\frac{1-f}{n}\left\{\frac{w_1}{\overline{X}_1}\left(R_1 S_{x_1}^2 - \rho_1 S_y S_{x_1}\right) + \frac{w_2}{\overline{X}_2}\left(R_2 S_{x_2}^2 - \rho_2 S_y S_{x_2}\right)\right\}$$

$$sesgo\left(\widehat{\overline{Y}}_{R_{12}}\right) \simeq \frac{1-f}{n}\left\{\frac{w_1}{\overline{X}_1}\left(R_1 S_{x_1}^2 - \rho_1 S_y S_{x_1}\right) + \frac{w_2}{\overline{X}_2}\left(R_2 S_{x_2}^2 - \rho_2 S_y S_{x_2}\right)\right\}$$

donde:

f es la fracción de muestreo
ρ_i es el coeficiente de correlación entre x_i e y, $i = 1,2$
$S_{x_i}^2$ la cuasivarianza de x_i, $i = 1,2$.

Demostración.-
Partimos de la igualdad:

$$sesgo\left(\widehat{Y}_{R_{12}}\right) = w_1 sesgo\left(\widehat{Y}_{R_1}\right) + w_2 sesgo\left(\widehat{Y}_{R_2}\right) \tag{3.8}$$

En el capítulo 1 (proposición 1.3.13) demostramos que una aproximación de orden $O(n^{-2})$ del sesgo del estimador univariante es:

$$sesgo\left(\widehat{Y}_{R_i}\right) \simeq \frac{N}{\overline{X}}\left(\frac{1-f}{n}\right)\left(R_i S_{x_i}^2 - \rho_i S_y S_{x_i}\right) \ , \quad i = 1,2 \tag{3.9}$$

por lo que sustituyendo en la expresión 3.8 se obtiene el resultado deseado.

El sesgo del estimador de la media se obtiene inmediatamente a partir del sesgo del estimador del total, considerando la igualdad:

$$sesgo\left(\widehat{\overline{Y}}_{R_{12}}\right) = \frac{sesgo\left(\widehat{Y}_{R_{12}}\right)}{N} \tag{3.10}$$

De aquí se sigue que los estimadores son sesgados. Como se observa en la fórmula 3.10, el sesgo es de orden $\frac{1}{n}$. Como el error estándar de $\widehat{Y}_{R_{12}}$, como se comprobará más adelante , es de orden $O\left(n^{-\frac{1}{2}}\right)$, la razón del sesgo al error es también del orden $O\left(n^{-\frac{1}{2}}\right)$, y se vuelve despreciable cuando n aumenta.

En la práctica ocurre que el sesgo no tiene importancia aún cuando el tamaño de la muestra es moderado, siguiendo un razonamiento similar al dado en el capítulo 1.

Por último se considera el caso de que el estimador de razón no tiene sesgo.

Proposición 3.1.5 *Si las curvas de regresión de y sobre x_i son líneas rectas que pasan por el origen, entonces el estimador de razón bivariante no tiene sesgo.*

Demostración.-

Si la regresión de y sobre x_i es lineal y pasa por el origen para $i = 1, 2$, entonces $\widehat{Y}_{R_1}$ e $\widehat{Y}_{R_2}$ son insesgados de Y, como ya se demostró en el teorema 1.3.15, por lo que $\widehat{Y}_{R_{12}}$ también lo será. Del mismo modo $\widehat{\overline{Y}}_{R_{12}}$ es insesgado de $\overline{Y}$.

Error cuadrático medio.

Proposición 3.1.6 *Una aproximación de orden $O(n^{-2})$ del error cuadrático medio del estimador $\widehat{Y}_{R_{12}}$ viene dada por:*

$$ECM\left(\widehat{Y}_{R_{12}}\right) = \frac{1-f}{n}Y^2\left\{w_1^2\left(C_y^2 + C_{x_1}^2 - 2C_{yx_1}\right) + \right.$$

$$+w_2^2\left(C_y^2 + C_{x_2}^2 - 2C_{yx_2}\right) +$$

$$\left.+2w_1w_2\left(C_y - C_{yx_2} - C_{yx_1} + C_{x_1x_2}\right)\right\}$$

Demostración.-

Partimos del resultado:

$$ECM\left(\widehat{Y}_{R_{12}}\right) = w_1^2 ECM\left(\widehat{Y}_{R_1}\right) +$$

$$+ w_2^2 ECM\left(\widehat{Y}_{R_2}\right) + 2w_1w_2E\left[\left(\widehat{Y}_{R_1} - Y\right)\left(\widehat{Y}_{R_2} - Y\right)\right] \qquad (3.11)$$

Para un muestreo aleatorio simple se tiene que:

$$ECM\left(\widehat{Y}_{R_i}\right) = \frac{1-f}{n}Y^2\left(C_y^2 + C_{x_i}^2 - 2C_{yx_i}\right) \quad i = 1, 2$$

donde $C_y = \frac{S_y}{\overline{Y}}$, $C_{x_i} = \frac{S_{x_i}}{\overline{X}_i}$ y $C_{yx_i} = \frac{S_{yx_i}}{\overline{X}_i \overline{Y}}$. Además

$$E\left[\left(\hat{Y}_{R_1} - Y\right)\left(\hat{Y}_{R_2} - Y\right)\right] =$$

$$= \overline{Y}^2 E\left[\left((1+e_1)(1+e_2)^{-1} - 1\right)\left((1+e_1)(1+e_2')^{-1} - 1\right)\right]$$

donde

$$e_2 = \frac{\overline{x}_1 - \overline{X}_1}{\overline{X}_1} \; ; \quad e_2' = \frac{\overline{x}_2 - \overline{X}_2}{\overline{X}_2}$$

Si desarrollamos $(1+e_2)^{-1}$ y $(1+e_2')^{-1}$ en serie de Taylor y retenemos sólo los términos de orden inferior o igual a dos en e_1, e_2 y e_2', se llega a que

$$E\left[\left(\hat{Y}_{R_1} - Y\right)\left(\hat{Y}_{R_2} - Y\right)\right] \simeq \frac{1-f}{n} Y^2 \left(C_y^2 - \rho_2 C_y C_{x_2} - \rho_1 C_y C_{x_1} + \rho_{12} C_{x_1} C_{x_2}\right) \tag{3.12}$$

donde

ρ_i es el coeficiente de correlación entre y y x_i, $i = 1, 2$ y

ρ_{12} el coeficiente de correlación entre x_1 y x_2.

Se puede demostrar que esta aproximación es de orden $O(n^{-2})$. Sustituyendo en 3.11 los valores de estas esperanzas, se obtiene:

$$ECM\left(\hat{Y}_{R_{12}}\right) = \frac{1-f}{n} Y^2 \left\{w_1^2 \left(C_y^2 + C_{x_1}^2 - 2C_{yx_1}\right) + \right.$$
$$+ w_2^2 \left(C_y^2 + C_{x_2}^2 - 2C_{yx_2}\right) +$$
$$\left. + 2w_1 w_2 \left(C_y - C_{yx_2} - C_{yx_1} + C_{x_1 x_2}\right)\right\}$$

Análogamente:

$$ECM\left(\hat{\overline{Y}}_{R_{12}}\right) = \frac{1-f}{n} \overline{Y}^2 \left\{w_1^2 \left(C_y^2 + C_{x_1}^2 - 2C_{yx_1}\right) + \right.$$
$$\left. + w_2^2 \left(C_y^2 + C_{x_2}^2 - 2C_{yx_2}\right) + 2w_1 w_2 \left(C_y - C_{yx_2} - C_{yx_1} + C_{x_1 x_2}\right)\right\}$$

A continuación vamos a calcular los pesos óptimos en el sentido de que produzcan el menor error cuadrático medio posible.

Si tomamos $w_2 = 1 - w_1$, derivamos $ECM\left(\widehat{Y}_{R_{12}}\right)$ respecto a w_1 e igualamos a cero, obtenemos el siguiente valor para w_1:

$$w_1 = \frac{ECM\left(\widehat{Y}_{R_2}\right) - E\left[\left(\widehat{Y}_{R_1} - Y\right)\left(\widehat{Y}_{R_2} - Y\right)\right]}{ECM\left(\widehat{Y}_{R_1}\right) + ECM\left(\widehat{Y}_{R_2}\right) - 2E\left[\left(\widehat{Y}_{R_1} - Y\right)\left(\widehat{Y}_{R_2} - Y\right)\right]}$$

Estos valores obtenidos para w_1 y w_2 dependen de $E\left(\widehat{Y}_{R_1}\right)$, $E\left(\widehat{Y}_{R_2}\right)$ y $E\left[\left(\widehat{Y}_{R_1} - Y\right)\left(\widehat{Y}_{R_2} - Y\right)\right]$ que son desconocidos. En la práctica se aproximan utilizando los valores muestrales en lugar de los poblacionales.

En el caso particular en que $\rho_1 = \rho_2 = \rho_{yx}$ y $C_{x_1} = C_{x_2} = C_x$, $\Rightarrow w_1 = w_2 = \frac{1}{2}$ y el error cuadrático medio es:

$$ECM\left(\widehat{Y}_{R_{12}}\right) = \frac{1-f}{n} Y^2 \left(C_y^2 - 2\rho_{yx} C_y C_x + \frac{C_x^2}{2}\left(1 + \rho_{12}\right)\right)$$

Comparación con el caso de estimadores univariantes y estimadores de expansión simple.

Si no se tiene información auxiliar, el estimador del total es $\widehat{Y} = N\overline{y}$ que se conoce con el nombre de estimador de expansión simple cuya varianza vale

$$V\left(\widehat{Y}\right) = \frac{1-f}{n} Y^2 C_y^2$$

Entonces el estimador de razón bivariante será más eficiente si

$$\frac{ECM\left(\widehat{Y}_{R_{12}}\right)}{V\left(\widehat{Y}\right)} < 1 \Leftrightarrow \frac{\rho_{yx}}{1 + \rho_{12}} \frac{C_y}{C_x} > \frac{1}{4}$$

En el caso en que $\rho_{12} = 1$ se obtiene la condición $\rho_{yx}\frac{C_y}{C_x} > \frac{1}{2}$, que es la que se obtiene en el caso de una sóla variable auxiliar, y en este, se tiene:

$$ECM\left(\widehat{Y}_R\right) = \frac{1-f}{n} Y^2 \left(C_y^2 + C_x^2 - 2C_{xy}\right)$$

por tanto el estimador bivariante será más eficaz que el univariante si:

$$\frac{ECM\left(\widehat{Y}_{R_{12}}\right)}{ECM\left(\widehat{Y}_R\right)} < 1 \Leftrightarrow C_x^2\left(\rho_{12} - 1\right) < 0$$

condición que siempre ocurre excepto en el caso $\rho_{12} = 1$.

3.1.2 Caso de $k > 2$ variables auxiliares.

Definición del estimador.

Consideremos ahora que hay información disponible acerca de x_1,x_2, ...,x_k con $k > 2$ variables auxiliares. Vamos a estudiar las propiedades del estimador:

$$\widehat{Y}_R^{1,2,\ldots,k} = \sum_{i=1}^{k} w_i \widehat{Y}_{R_i} \tag{3.13}$$

Sesgo.

Los resultados siguientes acerca del sesgo de los estimadores, se deducen inmediatamente a partir de los resultados obtenidos para $k = 2$.

Proposición 3.1.7 *Una cota del sesgo de los estimadores* $\widehat{Y}_R^{1,2,\ldots,k}$ *e* $\widehat{\overline{Y}}_R^{1,2,\ldots,k}$ *viene dada por las expresiones siguientes:*

$$\mid sesgo\left(\widehat{Y}_R^{1,2,\ldots,k}\right) \mid \leq N \max_i \left(\sigma_{\widehat{R}_i} \cdot \sigma_{\overline{x}_i}\right)$$

$$\mid sesgo\left(\widehat{\overline{Y}}_R^{1,2,\ldots,k}\right) \mid \leq \max_i \left(\sigma_{\widehat{R}_i} \cdot \sigma_{\overline{x}_i}\right)$$

Proposición 3.1.8 *Una aproximación del sesgo de los estimadores* $\widehat{Y}_R^{1,2,\ldots,k}$ *e* $\widehat{\overline{Y}}_R^{1,2,\ldots,k}$ *viene dada por las expresiones siguientes:*

$$sesgo\left(\widehat{Y}_R^{1,2,\ldots,k}\right) = N\frac{1-f}{n}\sum_{i=1}^{k}\left(R_i S_{x_i}^2 - \rho_i S_y S_{x_i}\right)\frac{w_i}{\overline{X}_i}$$

$$sesgo\left(\widehat{\overline{Y}}_R^{1,2,\ldots,k}\right) = \frac{1-f}{n}\sum_{i=1}^{k}\left(R_i S_{x_i}^2 - \rho_i S_y S_{x_i}\right)\frac{w_i}{\overline{X}_i}$$

donde ρ_i *es el coeficiente de correlación entre* x_i *e* y, *y* $S_{x_i}^2$ *la cuasivarianza poblacional de* x_i, $i = 1, 2, \ldots, k$.

Teorema 3.1.9 *Si la regresión de y sobre x_i $(i = 1, \ldots, k)$ son líneas rectas que pasan por el origen los estimadores $\widehat{Y}_R^{1,2,\ldots,k}$ e $\widehat{\overline{Y}}_R^{1,2,\ldots,k}$ son insesgados.*

Demostración.-

La demostración es inmediata pues en estas condiciones todos los $\widehat{Y}_{R_i}$ e $\widehat{\overline{Y}}_{R_i}$ son insesgados de sus respectivos parámetros y así $\widehat{Y}_R^{1,2,\ldots,k}$ e $\widehat{\overline{Y}}_R^{1,2,\ldots,k}$ también lo son.

Error cuadrático medio.

En el estudio del error cuadrático medio es donde hay diferencias notables respecto al caso $k = 2$. El problema es hallar el error cuadrático medio del estimador:

$$ECM\left(\widehat{Y}_R^{1,2,\ldots,k}\right) = ECM\left(\sum_{i=1}^{k} w_i \widehat{Y}_{R_i}\right) =$$

$$= E\left[\sum_{i=1}^{k} w_i \widehat{Y}_{R_i} - Y\right]^2 = E\left[\sum_{i=1}^{k} w_i \left(\widehat{Y}_{R_i} - Y\right)\right]^2$$

$$= E\left[\sum_{i=1}^{k} w_i^2 \left(\widehat{Y}_{R_i} - Y\right)^2 + \sum_{i \neq j} w_i w_j \left(\widehat{Y}_{R_i} - Y\right)\left(\widehat{Y}_{R_j} - Y\right)\right] =$$

$$= \sum_{i=1}^{k} w_i^2 ECM\left(\widehat{Y}_{R_i}\right) + \sum_{i \neq j} w_i w_j E\left[\left(\widehat{Y}_{R_i} - Y\right)\left(\widehat{Y}_{R_j} - Y\right)\right]$$

Ahora bien, sabemos que:

$$ECM\left(\widehat{Y}_{R_i}\right) \simeq \frac{1-f}{n} Y^2 \left(C_y^2 + C_{x_i}^2 - 2C_{yx_i}\right) \tag{3.14}$$

$$C_y = \frac{S_y}{\overline{Y}};\ C_{x_i} = \frac{S_{x_i}}{\overline{X}_i};\ C_{yx_i} = \frac{S_{yx_i}}{\overline{X}_i \overline{Y}}$$

$$E\left[\left(\widehat{Y}_{R_i} - Y\right)\left(\widehat{Y}_{R_j} - Y\right)\right] =$$

$$= \frac{1-f}{n} Y^2 \left(C_y^2 - \rho_j C_y C_{x_j} - \rho_i C_y C_{x_i} + \rho_{ij} C_{x_i} C_{x_j}\right)$$

Entonces si llamamos $W = (w_1, \ldots, w_k)'$, $A = \left(a_{ij}\right)_{k\times k}$ donde

$$a_{ij} = C_y^2 - \rho_j C_y C_{x_j} - \rho_i C_y C_{x_i} + \rho_{ij} C_{x_i} C_{x_j}$$

podemos expresar matricialmente el ECM de la forma siguiente:

$$ECM\left(\hat{Y}_R^{1,2,\ldots,k}\right) = \frac{1-f}{n} Y^2 \sum_{i,j=1}^{k} w_i w_j a_{ij} = \frac{1-f}{n} Y^2 W' A W \tag{3.15}$$

y se obtiene la siguiente proposición:

Proposición 3.1.10 *Una aproximación del error cuadrático medio de los estimadores de razón de $\hat{Y}_R$ e $\hat{\overline{Y}}_R$ viene dada por la expresión:*

$$ECM\left(\hat{Y}_R\right) = \frac{1-f}{n} Y^2 W' A W$$

$$ECM\left(\hat{\overline{Y}}_R\right) = \frac{1-f}{n} \overline{Y}^2 W' A W$$

con $W = (w_1, \ldots, w_k)'$ y $A = \left(a_{ij}\right)_{k\times k}$ donde

$$a_{ij} = C_y^2 - \rho_j C_y C_{x_j} - \rho_i C_y C_{x_i} + \rho_{ij} C_{x_i} C_{x_j}$$

A partir de este resultado podemos calcular los pesos w_i que hacen mínimo este ECM. Para ello necesitamos los siguientes resultados sobre matrices:

Desigualdad extendida de Cauchy–Schwartz.

Sean $b_{(k\times 1)}$ y $d_{(k\times 1)}$ vectores y $B_{(k\times k)}$ definida positiva. Entonces

$$(b'd)^2 \leq (b'Bb)\left(d'B^{-1}d\right) \tag{3.16}$$

y la igualdad es cierta si y sólo si $b = cB^{-1}d$ para alguna constante c.

Proposición 3.1.11 *Si A es una matriz de orden $(p \times p)$ definida positiva y B es de orden $(q \times p)$ de rango r, entonces BAB' es definida positiva si $r = q$ y semidefinida positiva si $r < q$.*

Teorema 3.1.12 *Los pesos óptimos que minimizan el error cuadrático medio del estimador* $\widehat{Y}_R^{1,2,\ldots,k}$ *vienen dados por*

$$\widehat{W} = \frac{A^{-1}e}{e'A^{-1}e}$$

donde $e_{(k\times 1)} = (1,\ldots,1)'$ *y el valor mínimo del ECM es:*

$$ECM\left(\widehat{Y}_R^{1,2,\ldots,k}\right) = \frac{1-f}{n}Y^2\frac{1}{e'A^{-1}e}$$

Demostración.-
En primer lugar comprobemos que la matriz A es definida positiva:
Llamando $e_{(k\times 1)} = (1,\ldots,1)'$, I a la matriz identidad de orden $(k\times k)$, y

$$L_{(k\times(k+1))} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

y C la matriz de coeficientes de variación de orden $(k+1)\times(k+1)$, que asumimos definida positiva (C es como mínimo semidefinida positiva puesto que es la matriz de covarianzas, salvo una constante, del vector $(e_1, o_1, o_2, \ldots, o_k)'$ donde $o_i = \dfrac{\overline{x}_i - \overline{X}_i}{\overline{X}_i}$):

$$C = \begin{bmatrix} C_y^2 & \rho_1 C_y C_{x_1} & \rho_2 C_y C_{x_2} & \cdots & \rho_k C_y C_{x_k} \\ \rho_1 C_y C_{x_1} & C_{x_1}^2 & \rho_{12} C_{x_1} C_{x_2} & \cdots & \rho_{1k} C_{x_1} C_{x_k} \\ \rho_2 C_y C_{x_2} & \rho_{12} C_{x_1} C_{x_2} & C_{x_2}^2 & \cdots & \rho_{2k} C_{x_2} C_{x_k} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_i C_y C_{x_i} & \rho_{1i} C_{x_1} C_{x_i} & \rho_{2i} C_{x_2} C_{x_i} & \cdots & \rho_{ik} C_{x_i} C_{x_k} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_k C_y C_{x_k} & \rho_{1k} C_{x_1} C_{x_k} & \rho_{2k} C_{x_2} C_{x_k} & \cdots & C_{x_k}^2 \end{bmatrix}$$

se obtiene que la matriz A se puede expresar de la forma $A = LCL'$, luego A es definida positiva, por la proposición 3.1.11, al ser L de rango k. Estamos pues en condiciones de aplicar la desigualdad de Cauchy-Schwartz extendida.

Consideremos A como la matriz B, el vector W como b, y el vector e como d, obteniendo:

$$(W'e)^2 \leq (W'AW)\left(e'A^{-1}e\right)$$

y al ser $e \neq 0$ y A definida positiva, $e'Ae > 0$. Dividiendo ambos términos obtenemos:

$$\frac{(W'e)^2}{e'A^{-1}e} \leq W'AW$$

Como $W'e = \sum_{i=1}^{k} w_i = 1$, nos queda la desigualdad:

$$\left(e'A^{-1}e\right)^{-1} \leq W'AW$$

o lo que es lo mismo:

$$\frac{1-f}{n}Y^2\left(e'A^{-1}e\right)^{-1} \leq ECM\left(\widehat{Y}_R^{1,2,\ldots,k}\right) \tag{3.17}$$

Por tanto el valor mínimo es:

$$ECM\left(\widehat{Y}_R^{1,2,\ldots,k}\right) = \frac{1-f}{n}Y^2\left(e'A^{-1}e\right)^{-1} \tag{3.18}$$

La desigualdad se convierte en igualdad para $W = CA^{-1}e$, para alguna constante C, que calculamos a continuación:

$$\left(e'A^{-1}e\right)^{-1} = \left(CA^{-1}e\right)'A\left(CA^{-1}e\right) =$$

$$= C^2\left(A^{-1}e\right)'AA^{-1}e = C^2\left(A^{-1}e\right)'e = C^2e'A^{-1}e$$

y despejando:

$$C = \left(e'A^{-1}e\right)^{-1} \tag{3.19}$$

Es decir, la matriz de pesos óptimos viene dada por:

$$W = \frac{A^{-1}e}{e'A^{-1}e} \tag{3.20}$$

Los pesos son uniformes si y sólo si las sumas de las columnas de A son iguales.

Comparación con el caso univariante.

Vamos a comparar la precisión del estimador multivariante que usa las variables auxiliares $x_1, x_2, \ldots, x_p$ con la del estimador que utiliza las variables auxiliares $x_1, x_2, \ldots, x_q$ con $q > p$. Para el caso particular de $p = 1$, obtendremos la comparación con el caso univariante.

Teorema 3.1.13 *En un muestreo aleatorio simple, el error cuadrático medio del estimador* $\hat{Y}_R^{1,2,\ldots,q}$ *es menor que el error cuadrático medio del estimador* $\hat{Y}_R^{1,2,\ldots,p}$ *si* $q > p$ *y los pesos son elegidos de forma que minimicen el error cuadrático medio.*

Demostración.-
Según el teorema 3.1.12:

$$ECM\left(\hat{Y}_R^{1,2,\ldots,p}\right) = \frac{1-f}{n} Y^2 \frac{1}{e_p' A_p^{-1} e_p}$$

$$ECM\left(\hat{Y}_R^{1,2,\ldots,q}\right) = \frac{1-f}{n} Y^2 \frac{1}{e_q' A_q^{-1} e_q}$$

Expresando la matriz A_q en función de la matriz A_p, tenemos:

$$A_q = \begin{pmatrix} A_p & B \\ C & D \end{pmatrix}$$

con

$B_{p\times(q-p)}$, $C = B'$ y $D_{(q-p)\times(q-p)}$.

Ahora bien, aplicando las propiedades de las matrices por cajas y puesto que tanto A_p como D son matrices no singulares, obtenemos:

$$A_q^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

donde

$B_{11} = A_p^{-1} + FH^{-1}G$
$B_{12} = FH^{-1}$
$B_{21} = H^{-1}G$
$B_{22} = H^{-1}$

siendo

$A_pF = -B$, $GA_p = -C$ y $H = D - CA_p^{-1}B$.
Vamos a comprobar que

$$e_q'A_q^{-1}e_q \geq e_p'A_p^{-1}e_p$$

Para ello calculamos

$$e_q'A_q^{-1}e_q - e_p'A_p^{-1}e_p =$$

$$= e_q'\begin{pmatrix} A_p^{-1} + A_p^{-1}BH^{-1}B'A_p^{-1} & -A_p^{-1}BH^{-1} \\ -H^{-1}B'A_p^{-1} & H^{-1} \end{pmatrix}e_q - e_q'\begin{pmatrix} A_p^{-1} & 0 \\ 0 & 0 \end{pmatrix}e_q =$$

$$= e_q'\begin{pmatrix} A_p^{-1}BH^{-1}B'A_p^{-1} & -A_p^{-1}BH^{-1} \\ -H^{-1}B'A_p^{-1} & H^{-1} \end{pmatrix}e_q =$$

$$= e_q'\begin{pmatrix} A_p^{-1}B \\ -I \end{pmatrix}H^{-1}\begin{pmatrix} B'A_p^{-1} & -I \end{pmatrix}e_q \geq 0$$

puesto que H^{-1} es definida positiva al serlo la matriz A_q^{-1}. Entonces

$$e_q'A_q^{-1}e_q \geq e_p'A_p^{-1}e_p$$

y

$$ECM\left(\widehat{Y}_R^{1,2,\ldots,q}\right) \leq ECM\left(\widehat{Y}_R^{1,2,\ldots,p}\right)$$

Corolario 3.1.14 *El estimador de razón multivariante es al menos tan preciso como el estimador univariante.*

§3.2 El estimador de razón multivariante en el muestreo estratificado.

Como ya hemos visto, en el muestreo estratificado existen diversas formas de obtener un estimador de razón. Entre ellas las más conocidas son la estimación separada y la estimación combinada. A partir de los estimadores así obtenidos se pueden construir otros estimadores que utilicen más de una variable auxiliar y que llamaremos estimadores de razón multivariantes separado y combinado.

3.2.1 El estimador de razón bivariante separado.

Como en el caso del muestreo aleatorio simple, estudiaremos primero por su interés el caso $k = 2$, para después dar su generalización al caso de más de dos variables.

Supongamos que queremos estudiar una variable y y tenemos dos variables auxiliares x_1 y x_2.

Sea una población de N elementos dividida en L estratos.

Denotamos por:

N_h *al número de unidades de la población en el estrato* h.

n_h *al número de unidades de la muestra que pertenecen al estrato* h.

Y_h *al total poblacional de la variable* y *en el estrato* h.

X_{1h} *al total poblacional de la variable* x_1 *en el estrato* h.

X_{2h} *al total poblacional de la variable* x_2 *en el estrato* h.

$\overline{Y}_h$ *a la media poblacional de la variable* y *en el estrato* h.

$\overline{X}_{1h}$ *a la media poblacional de la variable* x_1 *en el estrato* h.

$\overline{X}_{2h}$ *a la media poblacional de la variable* x_2 *en el estrato* h.

$\overline{y}_h$ *a la media muestral de la variable* y *en el estrato* h.

$\overline{x}_{1h}$ *a la media muestral de la variable* x_1 *en el estrato* h.

$\overline{x}_{2h}$ *a la media muestral de la variable* x_2 *en el estrato* h.

Una forma sencilla de obtener un estimador de razón de Y es obtener un estimador de razón bivariante del total de cada estrato y sumar estos totales:

$$\widehat{Y}_{R_{12}h} = w_1\widehat{Y}_{R_1h} + w_2\widehat{Y}_{R_2h} \; ; \quad h = 1, \ldots, L \; ; \quad w_1 + w_2 = 1$$

donde $\widehat{Y}_{R_ih}$ es el estimador de razón de Y en el estrato h, utilizando como variable auxiliar x_i, $(i = 1, 2)$. Entonces:

Definición 3.2.15 *Definimos el estimador de razón separado bivariante del total como:*

$$\widehat{Y}_{R_{12}S} = \sum_{h=1}^{L} \widehat{Y}_{R_{12}h}$$

Es inmediato por definición que este estimador es una suma ponderada de los estimadores de razón separados univariantes:

$$\widehat{Y}_{R_{12}S} = w_1 \widehat{Y}_{R_1 S} + w_2 \widehat{Y}_{R_2 S}$$

donde

$$\widehat{Y}_{R_i S} = \sum_{h=1}^{L} \frac{\overline{y}_h}{\overline{x}_{ih}} X_{ih} \; ; \quad (i = 1, 2)$$

El estimador así definido no presupone que las razones verdaderas $\frac{y}{x_1}$ e $\frac{y}{x_2}$ permanecen constantes en cada estrato, por tanto el estimador funcionará bien aún cuando las razones varíen de estrato a estrato.

Sin embargo este estimador tiene el inconveniente de que para su uso se necesita el conocimiento de los totales poblacionales X_{1h} y X_{2h} $\forall h$.

La definición del estimador para la media es inmediato:

Definición 3.2.16 *Definimos el estimador de razón separado bivariante de la media como:*

$$\widehat{\overline{Y}}_{R_{12}S} = w_1 \widehat{\overline{Y}}_{R_1 S} + w_2 \widehat{\overline{Y}}_{R_2 S}$$

por lo que $N\widehat{\overline{Y}}_{R_{12}S} = \widehat{Y}_{R_{12}S}$.

A partir de ahora consideraremos un muestreo estratificado aleatorio, es decir, en cada estrato se elige una muestra aleatoria simple.

Sesgo.

Proposición 3.2.17 *El estimador de razón es sesgado y una aproximación de orden $O(n^{-2})$ del sesgo viene dada por la expresión:*

$$sesgo\left(\widehat{Y}_{R_{12}S}\right) = \sum_{h=1}^{L} N_h \frac{1 - f_h}{n_h} \left\{ \frac{w_1}{\overline{X}_{1h}} \left(R_{1h} S_{x_1 h}^2 - \rho_{1h} S_{yh} S_{x_1 h} \right) + \right.$$

$$\left. + \frac{w_2}{\overline{X}_{2h}} \left(R_{2h} S_{x_2 h}^2 - \rho_{2h} S_{yh} S_{x_2 h} \right) \right\}$$

Demostración.-

$$sesgo\left(\widehat{Y}_{R_{12}S}\right) = \sum_{h=1}^{L} sesgo\left(\widehat{Y}_{R_{12}h}\right) \tag{3.21}$$

Ahora bien, dentro de cada estrato y según la proposición 3.1.4, el sesgo viene dado por:

$$sesgo\left(\widehat{Y}_{R_{12}h}\right) = N_h \frac{1-f_h}{n_h} \sum_{i=1}^{2} \left(R_{ih} S^2_{x_i h} - \rho_{ih} S_{yh} S_{x_i h}\right) \frac{w_i}{\overline{X}_i}$$

donde:

ρ_{ih} es el coeficiente de correlación entre x_i e y en el estrato h.

$S^2_{x_i h}$ es la cuasivarianza de x_i en el estrato h.

$R_{ih} = \dfrac{Y_h}{X_{ih}}$ $(i = 1, 2)$.

Por tanto, sustituyendo en la expresión 3.21 se obtiene el resultado deseado.

Inmediatamente se obtiene que

$$sesgo\left(\widehat{\overline{Y}}_{R_{12}S}\right) = \frac{1}{N} \sum_{h=1}^{L} N_h \frac{1-f_h}{n_h}$$

$$\left(\sum_{i=1}^{2} \frac{w_i}{\overline{X}_{ih}} \left(R_{ih} S^2_{x_i h} - \rho_{ih} S_{yh} S_{x_i h}\right)\right)$$

Por tanto $\widehat{Y}_{R_{12}S}$ e $\widehat{\overline{Y}}_{R_{12}S}$ son sesgados de sus respectivos parámetros.

La proposición siguiente nos da las condiciones bajo las cuales estos estimadores son insesgados.

Proposición 3.2.18 *Si la regresión de y sobre x_1 y la regresión de y sobre x_2 dentro de cada estrato son rectas que pasan por el origen, el estimador de razón separado bivariante es insesgado.*

Demostración.-

Si dentro de cada estrato la regresión de y sobre x_i es lineal y pasa por el origen entonces $\widehat{Y}_{R_i h}$ es insesgado de Y_h, y por tanto $\widehat{Y}_{R_i S} = \sum_{h=1}^{L} \widehat{Y}_{R_i h}$ es insesgado de $Y = \sum_{h=1}^{L} Y_h$.

En consecuencia

$$E\left(\widehat{Y}_{R_{12}S}\right) = w_1 E\left(\widehat{Y}_{R_1S}\right) + w_2 E\left(\widehat{Y}_{R_2S}\right) = (w_1 + w_2)\, Y = Y$$

y del mismo modo $\widehat{\overline{Y}}_{R_{12}S}$ es insesgado de $\overline{Y}$.

Error cuadrático medio.

Vamos a dar una aproximación del error cuadrático medio de los estimadores $\widehat{Y}_{R_{12}S}$ e $\widehat{\overline{Y}}_{R_{12}S}$ válida para muestras grandes en cada estrato.

Proposición 3.2.19 *Una aproximacion de orden* $O(n^{-2})$ *del error cuadrático medio de los estimadores* $\widehat{Y}_{R_{12}S}$ *e* $\widehat{\overline{Y}}_{R_{12}S}$ *viene dada por las expresiones:*

$$ECM\left(\widehat{Y}_{R_{12}S}\right) \simeq \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(w_1^2\left(C_{yh}^2 + C_{x_1h}^2 - 2C_{yx_1h}\right) + \right.$$

$$+ w_2^2\left(C_{yh}^2 + C_{x_2h}^2 - 2C_{yx_2h}\right) +$$

$$\left. + 2w_1w_2\left(C_{yh}^2 - \rho_{2h}C_{yh}C_{x_2h} - \rho_{1h}C_{yh}C_{x_1h} + \rho_{x_1x_2h}C_{x_1h}C_{x_2h}\right)\right)$$

$$ECM\left(\widehat{\overline{Y}}_{R_{12}S}\right) \simeq \frac{1}{N^2}\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(w_1^2\left(C_{yh}^2 + C_{x_1h}^2 - 2C_{yx_1h}\right) + \right.$$

$$+ w_2^2\left(C_{yh}^2 + C_{x_2h}^2 - 2C_{yx_2h}\right) +$$

$$\left. + 2w_1w_2\left(C_{yh}^2 - \rho_{2h}C_{yh}C_{x_2h} - \rho_{1h}C_{yh}C_{x_1h} + \rho_{x_1x_2h}C_{x_1h}C_{x_2h}\right)\right)$$

Demostración.-
Partimos de la igualdad:

$$ECM\left(\widehat{Y}_{R_{12}S}\right) = w_1^2 ECM\left(\widehat{Y}_{R_1S}\right) + w_2^2 ECM\left(\widehat{Y}_{R_2S}\right) +$$

$$+ 2w_1w_2 E\left[\left(\widehat{Y}_{R_1S} - Y\right)\left(\widehat{Y}_{R_2S} - Y\right)\right] \qquad (3.22)$$

Ahora bien, el error cuadrático medio del estimador univariante en el muestreo estratificado tiene la expresión:

$$ECM\left(\hat{Y}_{R_iS}\right) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{yh}^2 + C_{x_ih}^2 - 2C_{yx_ih}\right)$$

según demostramos en el capítulo 2, donde

$$C_{yh} = \frac{S_{yh}}{\overline{Y}_h} \; ; \quad C_{x_ih} = \frac{S_{x_ih}}{\overline{X}_{ih}} \; ; \quad C_{yx_ih} = \frac{S_{yx_ih}}{\overline{X}_{ih}\overline{Y}_{ih}} \quad (i = 1,2)$$

Sólo queda calcular $E\left[\left(\hat{Y}_{R_1S} - Y\right)\left(\hat{Y}_{R_2S} - Y\right)\right]$:

$$E\left[\left(\hat{Y}_{R_1S} - Y\right)\left(\hat{Y}_{R_2S} - Y\right)\right] =$$

$$= E\left[\left(\sum_{h=1}^{L} \hat{Y}_{R_1h} - Y\right)\left(\sum_{h=1}^{L} \hat{Y}_{R_2h} - Y\right)\right] =$$

$$= E\left[\sum_{h=1}^{L} \left(\hat{Y}_{R_1h} - Y_h\right) \sum_{h=1}^{L} \left(\hat{Y}_{R_2h} - Y_h\right)\right] =$$

$$= \sum_{h=1}^{L} E\left[\left(\hat{Y}_{R_1h} - Y_h\right)\left(\hat{Y}_{R_2h} - Y_h\right)\right] +$$

$$+ \sum_{h' \neq h} E\left[\left(\hat{Y}_{R_1h} - Y_h\right)\left(\hat{Y}_{R_2h'} - Y_{h'}\right)\right]$$

Pero la elección de la muestra en cada estrato es independiente y por tanto

$$E\left[\left(\hat{Y}_{R_1h} - Y_h\right)\left(\hat{Y}_{R_2h'} - Y_{h'}\right)\right] = 0$$

Además, según la expresión 3.12, como dentro de cada estrato

$$E\left[\left(\hat{Y}_{R_1h} - Y_h\right)\left(\hat{Y}_{R_2h} - Y_h\right)\right] =$$

$$= \frac{1-f_h}{n_h} Y_h^2 \left(C_{yh}^2 - \rho_{2h} C_{yh} C_{x_2h} - \rho_{1h} C_{yh} C_{x_1h} + \rho_{x_1x_2h} C_{x_1h} C_{x_2h}\right)$$

sustituyendo los valores de $ECM\left(\hat{Y}_{R_iS}\right)$ y

$$E\left[\left(\hat{Y}_{R_1S} - Y\right)\left(\hat{Y}_{R_2S} - Y\right)\right]$$

en la expresión 3.22, se obtiene el resultado deseado.

La fórmula para la aproximación del error cuadrático medio del estimador de la media se obtiene sin más que considerar la igualdad:

$$ECM\left(\widehat{\bar{Y}}_{R_{12}S}\right) = \frac{1}{N^2} ECM\left(\widehat{Y}_{R_{12}S}\right)$$

Pero estas fórmulas sólo son válidas si la muestra en cada estrato es suficientemente grande como para que la fórmula aproximada del error cuadrático medio sea aplicable en cada estrato. Este problema no lo va a tener el estimador combinado que se estudia después.

Por último, vamos a determinar los pesos óptimos en el sentido de que produzcan menor error cuadrático medio del estimador de razón.

Proposición 3.2.20 *Los pesos que minimizan el* $ECM\left(\widehat{Y}_{R_{12}S}\right)$ *vienen dados por las expresiones siguientes:*

$$w_1 = \frac{\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{x_2h}^2 - C_{yx_2h} + C_{yx_1h} - C_{x_1x_2h}\right)}{\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{x_2h}^2 + C_{x_1h}^2 - 2C_{x_1x_2h}\right)}$$

$$w_2 = 1 - w_1$$

Demostración.-

Por la expresión 3.22, se tiene:

$$ECM\left(\widehat{Y}_{R_{12}S}\right) = w_1^2 ECM\left(\widehat{Y}_{R_1S}\right) + w_2^2 ECM\left(\widehat{Y}_{R_2S}\right) +$$

$$+2w_1w_2E\left[\left(\widehat{Y}_{R_1S} - Y\right)\left(\widehat{Y}_{R_2S} - Y\right)\right]$$

tomando $w_2 = 1 - w_1$, derivando respecto a w_1 e igualando a cero, se obtiene:

$$w_1 = \frac{ECM\left(\widehat{Y}_{R_2S}\right) - E\left[\left(\widehat{Y}_{R_1S} - Y\right)\left(\widehat{Y}_{R_2S} - Y\right)\right]}{ECM\left(\widehat{Y}_{R_1S}\right) + ECM\left(\widehat{Y}_{R_2S}\right) - 2E\left[\left(\widehat{Y}_{R_1S} - Y\right)\left(\widehat{Y}_{R_2S} - Y\right)\right]}$$

Sustituyendo los valores de $ECM\left(\widehat{Y}_{R_1S}\right)$, $ECM\left(\widehat{Y}_{R_2S}\right)$ y

$$E\left[\left(\widehat{Y}_{R_1S} - Y\right)\left(\widehat{Y}_{R_2S} - Y\right)\right]$$

calculados anteriormente, se obtiene:

$$w_1 = \frac{\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{x_2h}^2 - C_{yx_2h} + C_{yx_1h} - C_{x_1x_2h}\right)}{\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{x_2h}^2 + C_{x_1h}^2 - 2C_{x_1x_2h}\right)}$$

En el caso particular en que $C_{x_1h} = C_{x_2h}$ y $\rho_{1h} = \rho_{2h}$ para todo h, se obtiene

$$w_1 = w_2 = \frac{1}{2}$$

Comparación con el estimador separado univariante.

Hemos visto que:

$$ECM\left(\widehat{Y}_{R_iS}\right) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{yh}^2 + C_{x_ih}^2 - 2C_{yx_ih}\right) \quad i = 1,2$$

y según la proposición 3.2.19

$$ECM\left(\widehat{Y}_{R_{12}S}\right) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(w_1^2 \left(C_{yh}^2 + C_{x_1h}^2 - 2C_{yx_1h}\right) + \right.$$

$$+ w_2^2 \left(C_{yh}^2 + C_{x_2h}^2 - 2C_{yx_2h}\right)$$

$$\left. + 2w_1w_2 \left(C_{yh}^2 - \rho_{2h}C_{yh}C_{x_2h} - \rho_{1h}C_{yh}C_{x_1h} + \rho_{x_1x_2h}C_{x_1h}C_{x_2h}\right)\right)$$

En el caso en que $C_{x_1h} = C_{x_2h}$ y $\rho_{1h} = \rho_{2h}, \forall h \Rightarrow w_1 = w_2 = \frac{1}{2}$ y

$$ECM\left(\widehat{Y}_{R_{12}S}\right) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{yh}^2 + \frac{1}{2}C_{x_1h}^2 - 2C_{yx_1h} + C_{x_1x_2h}\right)$$

Por tanto el estimador de razón separado bivariante será más eficaz que el univariante si y sólo si

$$ECM\left(\widehat{Y}_{R_{12}S}\right) < ECM\left(\widehat{Y}_{R_iS}\right) \quad i = 1,2$$

es decir, si y sólo si

$$\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{yh}^2 + \frac{1}{2} C_{x_1 h}^2 - 2C_{y x_1 h} + C_{x_1 x_2 h} \right) <$$

$$< \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{yh}^2 + C_{x_i h}^2 - 2C_{y x_i h} \right)$$

cosa que es cierta, pues dentro de cada estrato

$$ECM\left(\widehat{Y}_{R_{12h}}\right) < ECM\left(\widehat{Y}_{R_{ih}}\right) \quad i = 1,2$$

y por tanto

$$\sum_{h=1}^{L} ECM\left(\widehat{Y}_{R_{12h}}\right) < \sum_{h=1}^{L} ECM\left(\widehat{Y}_{R_{ih}}\right)$$

excepto en el caso en que $\rho_{x_1 x_2 h} = 1 \ \forall h$, para el que se tendría la igualdad (ver la comparación entre el estimador bivariante y univariante en el muestreo aleatorio simple)

En conclusión, el estimador separado bivariante será mejor (en el sentido de producir menor error cuadrático medio) que el estimador separado univariante excepto si en todos los estratos el coeficiente de correlación entre las variables auxiliares sea 1, en cuyo caso será igual.

3.2.2 El estimador separado con $k > 2$ variables auxiliares.

Sean $x_1, x_2, \ldots, x_k$ $(k > 2)$ las variables auxiliares de las que disponemos información.

Definición del estimador.

Consideremos ahora el estimador:

$$\widehat{Y}_{R_S}^{1,2,\ldots,k} = \sum_{h=1}^{L} \widehat{Y}_{R_h} \tag{3.23}$$

donde $\widehat{Y}_{R_h} = \sum_{i=1}^{k} w_i \widehat{Y}_{R_{ih}}$, $\widehat{Y}_{R_{ih}} = \frac{\overline{y}_h}{\overline{x}_{ih}} X_{ih}$ y $w_1, w_2, \ldots, w_k$ son pesos tales que $\sum_{i}^{k} w_i = 1$.

Este estimador se puede escribir de forma alternativa:

$$\widehat{Y}_{R_S}^{1,2,\ldots,k} = \sum_{i=1}^{k} w_i \widehat{Y}_{R_i S} \tag{3.24}$$

donde

$$\widehat{Y}_{R_i S} = \sum_{i=1}^{L} \frac{\overline{y}_h}{\overline{x}_{ih}} X_{ih}$$

La definición de estimador de la media se obtiene de inmediato:

$$\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k} = \sum_{i=1}^{k} w_i \widehat{\overline{Y}}_{R_i S} = \frac{\widehat{Y}_{R_S}^{1,2,\ldots,k}}{N} \tag{3.25}$$

Sesgo.

Proposición 3.2.21 *El estimador de razón separado multivariante es sesgado y una aproximación de este sesgo viene dada por la expresión:*

$$sesgo\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right) \simeq \sum_{h=1}^{L} N_h \frac{1-f_h}{n_h} \left\{ \sum_{i=1}^{k} \frac{w_i}{\overline{X}_{ih}} \left(R_{ih} S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right\}$$

Demostración.-

Puesto que $\widehat{Y}_{R_S}^{1,2,\ldots,k} = \sum_{h=1}^{L} \widehat{Y}_{R_h}$, el sesgo del estimador será la suma de los sesgos de los estimadores de razón multivariantes en cada estrato y en estos el muestreo es aleatorio simple, por tanto:

$$sesgo\left(\widehat{Y}_{R_h}\right) \simeq N_h \frac{1-f_h}{n_h} \sum_{i=1}^{k} \frac{w_i}{\overline{X}_{ih}} \left(R_{ih} S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right)$$

donde

f_h es la fracción de muestreo en el estrato h,

ρ_{ih} el coeficiente de correlación entre x_i e y en el estrato h,

$S_{x_i h}$ la cuasivarianza de x_i en el estrato h y $R_{ih} = \dfrac{Y_h}{X_{ih}}$, para $i = 1, \ldots, k$.

Entonces se obtiene el resultado deseado sin más que sumar estos sesgos.

Corolario 3.2.22 *El sesgo del estimador de razón separado multivariante de la media* $\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}$ *viene dado por la expresión:*

$$sesgo\left(\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}\right) \simeq \frac{1}{N}\sum_{h=1}^{L} N_h \frac{1-f_h}{n_h}$$

$$\left\{\sum_{i=1}^{k} \frac{w_i}{\overline{X}_{ih}}\left(R_{ih}S_{x_ih}^2 - \rho_{ih}S_{yh}S_{x_ih}\right)\right\}$$

Demostración.-
Inmediata sin más que considerar que

$$sesgo\left(\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}\right) = \frac{1}{N} sesgo\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right)$$

Condiciones bajo las cuales el estimador de razón separado multivariante es insesgado.

Son una extensión inmediata del caso $k = 2$:

Proposición 3.2.23 *Si la regresión de y sobre* x_i $(i = 1, \ldots, k)$ *dentro de cada estrato es una línea recta que pasa por el origen, los estimadores de razón multivariante separados* $\widehat{Y}_{R_S}^{1,2,\ldots,k}$ *e* $\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}$ *son insesgados.*

Error cuadrático medio.

Vamos a obtener una expresión del error cuadrático medio del estimador $\widehat{Y}_{R_S}^{1,2,\ldots,k}$, para posteriormente calcular los pesos $w_1, \ldots, w_k$ que hacen mínimo $ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right)$.

Proposición 3.2.24 *Expresiones aproximadas del error cuadrático medio de los estimadores* $\widehat{Y}_{R_S}^{1,2,\ldots,k}$ *e* $\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}$ *son:*

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left\{\sum_{i=1}^{k} w_i^2\left(C_{yh}^2 + C_{x_ih}^2 - 2C_{yx_ih}\right) + \right.$$

$$\left. + \sum_{i\neq j} w_i w_j \left(C_{yh}^2 - \rho_{jh}C_{yh}C_{x_jh} - \rho_{ih}C_{yh}C_{x_ih} + \rho_{x_ix_jh}C_{x_ih}C_{x_jh}\right)\right\} \qquad (3.26)$$

$$ECM\left(\widehat{\bar{Y}}_{R_S}^{1,2,\ldots,k}\right) = \frac{1}{N}\sum_{h=1}^{L}\frac{1-f_h}{n_h}Y_h^2\left\{\sum_{i=1}^{k}w_i^2\left(C_{yh}^2 + C_{x_ih}^2 - 2C_{yx_ih}\right) + \right.$$

$$\left. + \sum_{i\neq j} w_i w_j \left(C_{yh}^2 - \rho_{jh}C_{yh}C_{x_jh} - \rho_{ih}C_{yh}C_{x_ih} + \rho_{x_ix_jh}C_{x_ih}C_{x_jh}\right)\right\}$$

Demostración.-

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right) = E\left[\sum_{i=1}^{k} w_i\widehat{Y}_{R_iS} - Y\right]^2 =$$

$$= E\left[\sum_{i=1}^{k}\left(w_i\left(\widehat{Y}_{R_iS} - Y\right)\right)\right]^2 =$$

$$= E\left[\sum_{i=1}^{k} w_i^2\left(\widehat{Y}_{R_iS} - Y\right)^2 + \sum_{i\neq j} w_i w_j\left(\widehat{Y}_{R_iS} - Y\right)\left(\widehat{Y}_{R_jS} - Y\right)\right] =$$

$$= \sum_{i=1}^{k} w_i^2 ECM\left(\widehat{Y}_{R_iS}\right) + \sum_{i\neq j} w_i w_j E\left[\left(\widehat{Y}_{R_iS} - Y\right)\left(\widehat{Y}_{R_jS} - Y\right)\right] \qquad (3.27)$$

y sustituyendo en la expresión 3.27 los valores:

$$ECM\left(\widehat{Y}_{R_iS}\right) = \sum_{h=1}^{L}\frac{1-f_h}{n_h}Y_h^2\left(C_{yh}^2 + C_{x_ih}^2 - 2C_{yx_ih}\right)$$

$$E\left[\left(\widehat{Y}_{R_iS} - Y\right)\left(\widehat{Y}_{R_jS} - Y\right)\right] = \sum_{h=1}^{L}\frac{1-f_h}{n_h}Y_h^2\left(C_{yh}^2 - \right.$$

$$\left. - \rho_{jh}C_{yh}C_{x_jh} - \rho_{ih}C_{yh}C_{x_ih} + \rho_{x_ix_jh}C_{x_ih}C_{x_jh}\right)$$

obtenemos:

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right) = \sum_{i=1}^{k} w_i^2 \sum_{h=1}^{L}\frac{1-f_h}{n_h}Y_h^2\left(C_{yh}^2 + C_{x_ih}^2 - 2C_{yx_ih}\right) +$$

$$+\sum_{i\neq j} w_i w_j \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left(C_{yh}^2 -$$

$$-\rho_{jh} C_{yh} C_{x_j h} - \rho_{ih} C_{yh} C_{x_i h} + \rho_{x_i x_j h} C_{x_i h} C_{x_j h}\right) =$$

$$= \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \left\{ \sum_{i=1}^{k} w_i^2 \left(C_{yh}^2 + C_{x_i h}^2 - 2C_{y x_i h}\right) + \right.$$

$$\left. + \sum_{i\neq j} w_i w_j \left(C_{yh}^2 - \rho_{jh} C_{yh} C_{x_j h} - \rho_{ih} C_{yh} C_{x_i h} + \rho_{x_i x_j h} C_{x_i h} C_{x_j h}\right) \right\}$$

La expresión para $ECM\left(\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}\right)$ es inmediata sin más que considerar que $ECM\left(\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}\right) = N^{-2} ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right)$.

Considerando ahora las matrices:

$$W = (w_1, \ldots, w_k)'$$

$$A_h = \left(a_{ij}^h\right)_{(k\times k)}$$

donde

$a_{ij}^h = C_{yh}^2 - \rho_{jh} C_{yh} C_{x_j h} - \rho_{ih} C_{yh} C_{x_i h} + \rho_{x_i x_j h} C_{x_i h} C_{x_j h}$, $i = 1, \ldots, k$; $j = 1, \ldots, k$

podemos escribir la fórmula 3.26 en la forma:

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \sum_{i,j=1}^{k} w_i w_j a_{ij}^h = \sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 W' A_h W$$

y llamando $B_h = \dfrac{1-f_h}{n_h} Y_h^2 A_h$, y $A = \sum_{h=1}^{L} B_h$, B_h y A son matrices $(k \times k)$ y la expresión 3.26 adquiere la forma:

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right) = W' A W$$

Teorema 3.2.25 *El error cuadrático medio del estimador* $\widehat{Y}_{R_S}^{1,2,\ldots,k}$ *es mínimo si* $W = \dfrac{A^{-1}e}{e'A^{-1}e}$ *y vale:*

$$mín\ ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right) = \left(e'A^{-1}e\right)^{-1}$$

donde $e_{(k\times 1)} = (1,1,\ldots,1)'$.

Demostración.-

Veamos en primer lugar que la matriz A es definida positiva. Las matrices A_h, $h = 1,\ldots,L$ son definidas positivas pues se pueden expresar de la forma $A_h = LC_hL'$ donde

$$L_{(k\times(k+1))} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

y C_h es la matriz de coeficientes de variación en el estrato h de orden $(k+1)\times(k+1)$, que asumimos definida positiva:

$$C_h = \begin{bmatrix} C_{yh}^2 & \rho_{1h}C_{yh}C_{x_1h} & \rho_{2h}C_{yh}C_{x_2h} & \cdots & \rho_{kh}C_{yh}C_{x_kh} \\ \rho_{1h}C_{yh}C_{x_1h} & C_{x_1h}^2 & \rho_{12h}C_{x_1h}C_{x_2h} & \cdots & \rho_{1kh}C_{x_1h}C_{x_kh} \\ \rho_{2h}C_{yh}C_{x_2h} & \rho_{12h}C_{x_1h}C_{x_2h} & C_{x_2h}^2 & \cdots & \rho_{2kh}C_{x_2h}C_{x_kh} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{ih}C_{yh}C_{x_ih} & \rho_{1ih}C_{x_1h}C_{x_ih} & \rho_{2ih}C_{x_2h}C_{x_ih} & \cdots & \rho_{ikh}C_{x_ih}C_{x_kh} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{kh}C_{yh}C_{x_kh} & \rho_{1kh}C_{x_1h}C_{x_kh} & \rho_{2kh}C_{x_2h}C_{x_kh} & \cdots & C_{x_kh}^2 \end{bmatrix}$$

Puesto que C_h es definida positiva y L de rango k, LC_hL' es también definida positiva, según la proposición 3.1.11. Por tanto $B_h = \dfrac{1-f_h}{n_h}Y_h^2A_h$ también lo es $\forall h$ por ser el producto de una constante por una matriz definida positiva. Por último, la matriz A es suma de matrices definidas positivas y en consecuencia ella será también definida positiva.

Estamos pues en las condiciones de aplicar la desigualdad extendida de Cauchy–Schwartz, 3.16, según la cual:

$$\frac{(W'e)^2}{(e'A^{-1}e)} \leq W'AW = ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right)$$

y el mínimo valor de $ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,k}\right)$ es $(e'A^{-1}e)^{-1}$ ya que $W'e = \sum_{i=1}^{k} w_i = 1$.

Además, como ya comprobamos en el caso del muestreo aleatorio simple, el mínimo se alcanza para

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

como queríamos demostrar.

Comparación con el caso univariante.

El estimador separado multivariante es como mínimo igual de preciso que el estimador separado univariante, como muestra el siguiente teorema:

Teorema 3.2.26 *En muestreo estratificado aleatorio, el estimador de razón separado que utiliza q variables auxiliares es como mínimo igual de preciso que el estimador de razón separado que utiliza $p < q$ variables auxiliares, si los pesos son los óptimos.*

Demostración.-
Según el teorema 3.2.25:

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,p}\right) = \frac{1}{e_p' A_p^{-1} e_p}$$

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,q}\right) = \frac{1}{e_q' A_q^{-1} e_q}$$

donde

$$A_p = \sum_{h=1}^{L} B_h^{(p)} \; ; \quad A_q = \sum_{h=1}^{L} B_h^{(q)}$$

Siguiendo el mismo procedimiento que en caso del muestreo aleatorio simple, se tiene:

$$A_q = \begin{pmatrix} A_p & B \\ B' & D \end{pmatrix}$$

y

$$A_q^{-1} = \begin{pmatrix} A_p^{-1} + FH^{-1}G & FH^{-1} \\ H^{-1}G & H^{-1} \end{pmatrix}$$

siendo

$A_pF = -B$

$GA_p = -B'$ y

$H = D - B'A_p^{-1}B.$

Entonces

$$e_q'A_q^{-1}e_q \geq e_p'A_p^{-1}e_p$$

si y sólo si

$$e_q'A_q^{-1}e_q - e_p'A_p^{-1}e_p \geq 0$$

Pero

$$e_q'A_q^{-1}e_q - e_p'A_p^{-1}e_p =$$

$$= e_q'\begin{pmatrix} A_p^{-1}BH^{-1}B'A_p^{-1} & -A_p^{-1}BH^{-1} \\ -H^{-1}B'A_p^{-1} & H^{-1} \end{pmatrix}e_q =$$

$$= e_q'\begin{pmatrix} A_p^{-1}B \\ -I \end{pmatrix}H^{-1}\begin{pmatrix} B'A_p^{-1} & -I \end{pmatrix}e_q \geq 0$$

por ser H^{-1} definida positiva, puesto que la matriz A_q^{-1} lo es.

Por tanto

$$ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,q}\right) \leq ECM\left(\widehat{Y}_{R_S}^{1,2,\ldots,p}\right)$$

Corolario 3.2.27 *El estimador de razón separado multivariante es al menos tan preciso como el estimador separado univariante.*

Demostración.-

Es inmediata sin más que considerar el caso $p = 1$ en el teorema anterior.

3.2.3 El estimador de razón bivariante combinado.

Otra forma de estimar el total de la variable Y utilizando dos variables auxiliares x_1 y x_2 en un muestreo estratificado consiste en calcular los estimadores de los totales Y, X_1 y X_2, a partir de los datos de la muestra:

$$\widehat{Y}_{st} = \sum_{h=1}^{L} N_h \overline{y}_h \; ; \quad \widehat{X}_{1st} = \sum_{h=1}^{L} N_h \overline{x}_{1h} \; ; \quad \widehat{X}_{2st} = \sum_{h=1}^{L} N_h \overline{x}_{2h}$$

y definir el estimador de razón combinado bivariante, como sigue:

$$\widehat{Y}_{R_{12}C} = w_1 \frac{\widehat{Y}_{st}}{\widehat{X}_{1st}} X_1 + w_2 \frac{\widehat{Y}_{st}}{\widehat{X}_{2st}} X_2$$

$$\widehat{\overline{Y}}_{R_{12}C} = \frac{1}{N} \widehat{Y}_{R_{12}C}$$

Este estimador tiene la gran ventaja sobre el estimador separado de que no necesita el conocimiento de los totales X_{1h} y X_{2h} en cada estrato, sino sólo de los totales X_1 y X_2. Sin embargo su uso será conveniente en el caso de que las razones $\frac{y}{x_1}$ e $\frac{y}{x_2}$ permanezcan constantes no solo dentro de cada estrato, sino también de estrato a estrato.

Sesgo.

Proposición 3.2.28 *El estimador de razón bivariante combinado es sesgado y una aproximación del sesgo viene dada por la expresión:*

$$sesgo\left(\widehat{Y}_{R_{12}C}\right) \simeq N \sum_{h=1}^{L} \frac{1-f_h}{n_h} \left(\frac{w_1}{\overline{X}_1} \left(R_1 S_{x_1 h}^2 - \rho_{1h} S_{yh} S_{x_1 h} \right) + \right.$$

$$\left. + \frac{w_2}{\overline{X}_2} \left(R_2 S_{x_2 h}^2 - \rho_{2h} S_{yh} S_{x_2 h} \right) \right)$$

Demostración.-

Sustituyendo en la expresión:

$$sesgo\left(\widehat{Y}_{R_{12}C}\right) = w_1 sesgo\left(\widehat{Y}_{R_1 C}\right) + w_2 sesgo\left(\widehat{Y}_{R_2 C}\right)$$

$$sesgo\left(\widehat{Y}_{R_iC}\right) \simeq N \sum_{h=1}^{L} \frac{1-f_h}{n_h \overline{X}_i} \left(\frac{\overline{Y}}{\overline{X}_i} S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \tag{3.28}$$

se tiene:

$$sesgo\left(\widehat{Y}_{R_{12}C}\right) = w_1 N \sum_{h=1}^{L} \frac{1-f_h}{n_h \overline{X}_1} \left(\frac{\overline{Y}}{\overline{X}_1} S_{x_1 h}^2 - \rho_{1h} S_{yh} S_{x_1 h} \right) +$$

$$+ w_2 N \sum_{h=1}^{L} \frac{1-f_h}{n_h \overline{X}_2} \left(\frac{\overline{Y}}{\overline{X}_2} S_{x_2 h}^2 - \rho_{2h} S_{yh} S_{x_2 h} \right) =$$

$$= N \sum_{h=1}^{L} \frac{1-f_h}{n_h} \left(\sum_{i=1}^{2} \frac{w_i}{\overline{X}_i} \left(R_i S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right)$$

En consecuencia

$$sesgo\left(\widehat{\overline{Y}}_{R_{12}C}\right) = \sum_{h=1}^{L} \frac{1-f_h}{n_h} \left(\sum_{i=1}^{2} \frac{w_i}{\overline{X}_i} \left(R_i S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right)$$

Error cuadrático medio.

Proposición 3.2.29 *Una expresión aproximada del error cuadrático medio del estimador del total viene dada por la expresión:*

$$ECM\left(\widehat{Y}_{R_{12}C}\right) \simeq \sum_{h=1}^{L} \frac{1-f_h}{n_h} N_h^2$$

$$\left[w_1^2 \left(S_{yh}^2 + R_1^2 S_{x_1 h}^2 - 2 R_1 \rho_{1h} S_{x_1 h} S_{yh} \right) + \right.$$

$$+ w_2^2 \left(S_{yh}^2 + R_2^2 S_{x_2 h}^2 - 2 R_2 \rho_{2h} S_{x_2 h} S_{yh} \right) +$$

$$\left. + 2 w_1 w_2 \left(S_{yh}^2 - R_1 \rho_{1h} S_{x_1 h} S_{yh} - R_2 \rho_{2h} S_{x_2 h} S_{yh} + R_1 R_2 \rho_{12h} S_{x_1 h} S_{x_2 h} \right) \right]$$

y para el estimador de la media por

$$ECM\left(\widehat{\overline{Y}}_{R_{12}C}\right) = \frac{1}{N^2} ECM\left(\widehat{Y}_{R_{12}C}\right)$$

Demostración.-
Partiendo de:

$$ECM\left(\widehat{Y}_{R_{12}C}\right) = w_1^2 ECM\left(\widehat{Y}_{R_1C}\right) + w_2^2 ECM\left(\widehat{Y}_{R_2C}\right) +$$

$$+2w_1w_2E\left[\left(\widehat{Y}_{R_1C} - Y\right)\left(\widehat{Y}_{R_2C} - Y\right)\right]$$

puesto que conocemos :

$$ECM\left(\widehat{Y}_{R_iC}\right) =$$

$$= \sum_{h=1}^{L} N^2 \frac{1-f_h}{n_h}\left(S_{yh}^2 + R_i^2 S_{x_ih}^2 - 2R_i\rho_{ih}S_{x_ih}S_{yh}\right) \quad (i=1,2) \tag{3.29}$$

sólo necesitamos saber el valor de $E\left[\left(\widehat{Y}_{R_1C} - Y\right)\left(\widehat{Y}_{R_2C} - Y\right)\right]$

Para ello partimos de la siguiente aproximación:

$$\frac{\widehat{Y}_{st}}{\widehat{X}_{ist}}X_i - Y = \frac{\overline{y}_{st}}{\overline{x}_{ist}}X_i - Y =$$

$$= \frac{\overline{y}_{st}X_i - Y\overline{x}_{ist}}{\overline{x}_{ist}} = X_i\frac{\overline{y}_{st} - R_i\overline{x}_{ist}}{\overline{x}_{ist}} =$$

$$= N\overline{X}_i\frac{\overline{y}_{st} - R_i\overline{x}_{ist}}{\overline{x}_{ist}} \simeq N\left(\overline{y}_{st} - R_i\overline{x}_{ist}\right) \quad (i=1,2)$$

si el tamaño es grande.

Con ella:

$$E\left[\left(\widehat{Y}_{R_1C} - Y\right)\left(\widehat{Y}_{R_2C} - Y\right)\right] =$$

$$= N^2E\left[\left(\overline{y}_{st} - R_1\overline{x}_{1st}\right)\left(\overline{y}_{st} - R_2\overline{x}_{2st}\right)\right]$$

Ahora bien, si consideramos la variable $u_i = y - R_ix_i$ $(i=1,2)$, entonces $\overline{y}_{st} - R_i\overline{x}_{ist} = \overline{u}_{ist}$ es la media de la variable u_i en la muestra estratificada, y además $\overline{u}_{ist} = 0$.

Entonces:

$$E\left[\left(\widehat{Y}_{R_1C} - Y\right)\left(\widehat{Y}_{R_2C} - Y\right)\right] = N^2\,\mathrm{Cov}\left(\overline{u}_{1st}, \overline{u}_{2st}\right)$$

$$\mathrm{Cov}\left(\overline{u}_{1st}, \overline{u}_{2st}\right) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2\frac{1-f_h}{n_h}S_{u_1u_2h}$$

donde:

$$S_{u_1u_2h} = \sum_{i=1}^{N_h} \frac{(u_{1hi} - \overline{u}_{1h})(u_{2hi} - \overline{u}_{2h})}{N_h - 1} =$$

$$= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left[(y_{hi} - R_1 x_{1hi}) - \left(\overline{Y}_h - R_1 \overline{X}_{1h} \right) \right]$$

$$\left[(y_{hi} - R_2 x_{2hi}) - \left(\overline{Y}_h - R_2 \overline{X}_{2h} \right) \right] =$$

$$= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left[\left(y_{hi} - \overline{Y}_h \right) - R_1 \left(x_{1hi} - \overline{X}_{1h} \right) \right]$$

$$\left[\left(y_{hi} - \overline{Y}_h \right) - R_2 \left(x_{2hi} - \overline{X}_{2h} \right) \right] =$$

$$= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left[\left(y_{hi} - \overline{Y}_h \right)^2 - R_1 \left(x_{1hi} - \overline{X}_{1h} \right) \left(y_{hi} - \overline{Y}_h \right) - \right.$$

$$\left. - R_2 \left(x_{2hi} - \overline{X}_{2h} \right) \left(y_{hi} - \overline{Y}_h \right) + R_1 R_2 \left(x_{1hi} - \overline{X}_{1h} \right) \left(x_{2hi} - \overline{X}_{2h} \right) \right] =$$

$$= S_{yh}^2 - R_1 \rho_{1h} S_{x_1h} S_{yh} - R_2 \rho_{2h} S_{x_2h} S_{yh} + R_1 R_2 \rho_{12h} S_{x_1h} S_{x_2h}$$

Por tanto:

$$E \left[\left(\widehat{Y}_{R_1C} - Y \right) \left(\widehat{Y}_{R_2C} - Y \right) \right] =$$

$$= \sum_{h=1}^{L} N_h^2 \frac{1 - f_h}{n_h} \left[S_{yh}^2 - R_1 \rho_{1h} S_{x_1h} S_{yh} - R_2 \rho_{2h} S_{x_2h} S_{yh} + R_1 R_2 \rho_{12h} S_{x_1h} S_{x_2h} \right] \tag{3.30}$$

Resumiendo:

$$ECM \left(\widehat{Y}_{R_{12}C} \right) = \sum_{h=1}^{L} \frac{1 - f_h}{n_h} N_h^2$$

$$\left[w_1^2 \left(S_{yh}^2 + R_1^2 S_{x_1h}^2 - 2 R_1 \rho_{1h} S_{x_1h} S_{yh} \right) + \right.$$

$$+ w_2^2 \left(S_{yh}^2 + R_2^2 S_{x_2h}^2 - 2 R_2 \rho_{2h} S_{x_2h} S_{yh} \right) +$$

$$\left. + 2 w_1 w_2 \left(S_{yh}^2 - R_1 \rho_{1h} S_{x_1h} S_{yh} - R_2 \rho_{2h} S_{x_2h} S_{yh} + R_1 R_2 \rho_{12h} S_{x_1h} S_{x_2h} \right) \right]$$

Así, el ECM del estimador $\widehat{\overline{Y}}_{R_{12}C}$ viene dado por la expresión:

$$ECM\left(\widehat{\overline{Y}}_{R_{12}C}\right) = \frac{1}{N^2} ECM\left(\widehat{Y}_{R_{12}C}\right)$$

3.2.4 El estimador combinado con $k > 2$ variables auxiliares.

Definición del estimador.

Se definen los estimadores de razón multivariante del total y de la media de la forma:

$$\widehat{Y}_{R_C}^{1,2,\ldots,k} = \sum_{i=1}^{k} w_i \frac{\widehat{Y}_{st}}{\widehat{X}_{ist}} X_i \tag{3.31}$$

$$\widehat{\overline{Y}}_{R_C}^{1,2,\ldots,k} = \sum_{i=1}^{k} w_i \frac{\widehat{Y}_{st}}{\widehat{X}_{ist}} \overline{X}_i \tag{3.32}$$

Estos estimadores tienen la gran ventaja sobre $\widehat{Y}_{R_S}^{1,2,\ldots,k}$ y $\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}$ de no necesitar el conocimiento de los totales X_{ih}, $i = 1, \ldots, k$ en cada estrato, sino sólo de los totales X_i.

Vamos a continuación a estudiar sus propiedades en cuanto a sesgo y error cuadrático medio.

Sesgo.

Proposición 3.2.30 *El estimador de razón combinado multivariante es sesgado y una aproximación del sesgo viene dada por la expresión:*

$$sesgo\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) \simeq N \sum_{h=1}^{L} \frac{1-f_h}{n_h} \left(\sum_{i=1}^{k} \frac{w_i}{\overline{X}_i} \left(R_i S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right)$$

$$sesgo\left(\widehat{\overline{Y}}_{R_C}^{1,2,\ldots,k}\right) \simeq \sum_{h=1}^{L} \frac{1-f_h}{n_h} \left(\sum_{i=1}^{k} \frac{w_i}{\overline{X}_i} \left(R_i S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right)$$

Demostración.-
Puesto que:

$$sesgo\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) = \sum_{i=1}^{k} w_i sesgo\left(\widehat{Y}_{R_i C}\right)$$

sin más que sustituir los valores de $sesgo\left(\widehat{Y}_{R_iC}\right)$ que ya usamos en el capítulo 2, se obtiene:

$$sesgo\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) \simeq \sum_{i=1}^{k} w_i \left(N \sum_{h=1}^{L} \frac{1-f_h}{n_h \overline{X}_i} \left(\frac{\overline{Y}}{\overline{X}_i} S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right) =$$

$$= N \sum_{h=1}^{L} \frac{1-f_h}{n_h} \left(\sum_{i=1}^{k} \frac{w_i}{\overline{X}_i} \left(R_i S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right)$$

Es inmediato pues que

$$sesgo\left(\widehat{\overline{Y}}_{R_C}^{1,2,\ldots,k}\right) \simeq \sum_{h=1}^{L} \frac{1-f_h}{n_h} \left(\sum_{i=1}^{k} \frac{w_i}{\overline{X}_i} \left(R_i S_{x_i h}^2 - \rho_{ih} S_{yh} S_{x_i h} \right) \right)$$

Error cuadrático medio.

Proposición 3.2.31 *Una expresión aproximada del error cuadrático medio del estimador* $\widehat{\overline{Y}}_{R_C}^{1,2,\ldots,k}$ *viene dada por la expresión:*

$$ECM\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) \simeq \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} \left\{ \sum_{i=1}^{k} w_i^2 \left(S_{yh}^2 + R_i^2 S_{x_i h}^2 - 2R_i \rho_{ih} S_{x_i h} S_{yh} \right) + \right.$$

$$\left. + \sum_{i \neq j} w_i w_j \left(S_{yh}^2 - R_j \rho_{jh} S_{yh} S_{x_j h} - R_i \rho_{ih} S_{yh} S_{x_i h} + R_i R_j \rho_{ijh} S_{x_i h} S_{x_j h} \right) \right\}$$

Demostración.-

$$ECM\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) = ECM\left(\sum_{i=1}^{k} w_i \frac{\widehat{Y}_{st}}{\widehat{X}_{ist}} X_i \right) =$$

$$= E\left[\sum_{i=1}^{k} w_i \frac{\widehat{Y}_{st}}{\widehat{X}_{ist}} X_i - Y \right]^2 = E\left[\sum_{i=1}^{k} w_i \left(\frac{\widehat{Y}_{st}}{\widehat{X}_{ist}} X_i - Y \right) \right]^2 =$$

$$= E\left[\sum_{i=1}^{k} w_i^2 \left(\widehat{Y}_{R_iC} - Y \right)^2 + \sum_{i \neq j} w_i w_j \left(\widehat{Y}_{R_iC} - Y \right) \left(\widehat{Y}_{R_jC} - Y \right) \right]$$

donde $\widehat{Y}_{R_iC} = \dfrac{\widehat{Y}_{st}}{\widehat{X}_{ist}} X_i$.

Entonces:

$$ECM\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) = \sum_{i=1}^{k} w_i^2 ECM\left(\widehat{Y}_{R_iC}\right) + \sum_{i\neq j} w_i w_j E\left[\left(\widehat{Y}_{R_iC} - Y\right)\left(\widehat{Y}_{R_jC} - Y\right)\right] \tag{3.33}$$

Ahora bien, según vimos en las proposiciones 3.29 y 3.30, se tiene:

$$ECM\left(\widehat{Y}_{R_iC}\right) = \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h}\left(S_{yh}^2 + R_i^2 S_{x_ih}^2 - 2R_i\rho_{ih}S_{x_ih}S_{yh}\right) \ (i=1,2)$$

$$E\left[\left(\widehat{Y}_{R_iC} - Y\right)\left(\widehat{Y}_{R_jC} - Y\right)\right] =$$

$$= \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h}\left[S_{yh}^2 - R_i S_{yx_ih} - R_j S_{yx_jh} + R_i R_j S_{x_ix_jh}\right]$$

y sustituyendo se obtiene el resultado deseado.

La fórmula para $ECM\left(\widehat{\widehat{Y}}_{R_C}^{1,2,\ldots,k}\right)$ se obtiene sin más que considerar la igualdad:

$$ECM\left(\widehat{\widehat{Y}}_{R_C}^{1,2,\ldots,k}\right) = \frac{1}{N^2} ECM\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right)$$

Determinación de los pesos óptimos.

Expresando el error cuadrático medio en forma matricial, tenemos:

$$ECM\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) = w' A_C w$$

donde $w = (w_1, \ldots, w_k)'$ y $A_C = \left(a_{ij}^C\right)_{k\times k}$ con

$$a_{ij}^C = \sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h}\left(S_{yh}^2 - R_i\rho_{ih}S_{yh}S_{x_ih} - R_j\rho_{jh}S_{yh}S_{x_jh} + R_iR_jS_{x_ix_jh}\right)$$

Entonces podemos calcular el vector w de pesos óptimos en el sentido de minimizar el error cuadrático medio del estimador $\widehat{Y}_{R_C}^{1,2,\ldots,k}$.

Teorema 3.2.32 *El error cuadrático medio del estimador* $\widehat{Y}_{R_C}^{1,2,\ldots,k}$ *es mínimo si*

$$w = \frac{A_C^{-1} e}{e' A_C^{-1} e}$$

y vale

$$\text{mín } ECM\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) = \frac{1}{e' A_C^{-1} e}$$

donde $e_{k\times 1} = (1, 1, \ldots, 1)'$.

Demostración.-

La matriz A_C es como mínimo semidefinida positiva, pues para todo x el valor $x' A_C x$ es el error cuadrático medio del estimador $\sum_{i=1}^{k} x_i \widehat{Y}_{R_i C}$ y por tanto es siempre mayor o igual a cero.

Si asumimos que es definida positiva, aplicando la desigualdad extendida de Cauchy–Schwartz obtenemos:

$$\text{mín } ECM\left(\widehat{Y}_{R_C}^{1,2,\ldots,k}\right) = \frac{1}{e' A_C^{-1} e}$$

y este mínimo se alcanza para

$$w = \frac{A_C^{-1} e}{e' A_C^{-1} e}$$

Comparación con el caso univariante.

Todos los resultados obtenidos respecto al estimador separado son ciertos también para el estimador combinado, como muestra el siguiente teorema:

Teorema 3.2.33 *El estimador de razón multivariante combinado es como mínimo igual de preciso que el estimador de razón combinado univariante, en un muestreo estratificado aleatorio.*

Demostración.-

Es la misma que la realizada para el teorema 3.2.26, sin más que cambiar la matriz A por la nueva matriz A_C.

§3.3 El estimador de razón multivariante en el muestreo con probabilidades desiguales.

Definición del estimador.

Sean $x_1, x_2, \ldots, x_k$ las variables auxiliares que suponemos correlacionadas con la variable objeto de estudio y.

Supongamos que queremos estimar el total o media poblacional mediante una muestra de tamaño n con reemplazo y probabilidades p_i, $i = 1, \ldots, N$.

Consideremos las variables:

$$z^i = \frac{y^i}{Np_i} \quad v_j^i = \frac{x_j^i}{Np_i} \quad i = 1, \ldots, N \quad j = 1, \ldots, k$$

Sean:

$$\widehat{\overline{Y}}_{R_j} = \frac{\overline{z}}{\overline{v}_j}\overline{X}_j$$

Definimos el estimador de razón multivariante como:

$$\widehat{\overline{Y}}_{RD}^{1,2,\ldots,k} = w_1\widehat{\overline{Y}}_{R_1D} + w_2\widehat{\overline{Y}}_{R_2D} + \cdots + w_k\widehat{\overline{Y}}_{R_kD}$$

siendo $w_1, w_2, \ldots, w_k$ pesos definidos de forma que $\sum_{j=1}^{k} w_j = 1$.

Si interesa estimar el total utilizaremos:

$$\widehat{Y}_{RD}^{1,2,\ldots,k} = N\widehat{\overline{Y}}_{RD}^{1,2,\ldots,k}$$

Para el estudio del sesgo y error cuadrático medio de estos estimadores, comenzamos con el caso de dos variables, para ver después el caso $k > 2$.

3.3.1 El estimador bivariante.

Comenzamos con el estudio del estimador de la media:

$$\widehat{\overline{Y}}_{R_{12}D} = w_1\widehat{\overline{Y}}_{R_1D} + w_2\widehat{\overline{Y}}_{R_2D}$$

Proposición 3.3.34 *Si la muestra ha sido elegida con reemplazo y probabilidades p_i $i = 1, \ldots, N$, entonces una cota para el sesgo del estimador $\widehat{\overline{Y}}_{R_{12}D}$ viene dada por la expresión:*

$$|\ sesgo\left(\widehat{\overline{Y}}_{R_{12}D}\right)| \leq \frac{1}{\sqrt{n}}\ \underset{i}{máx}\left(C_{v_i} \cdot \sigma_{\widehat{\overline{Y}}_{R_iD}}\right)$$

Demostración.-

Según vimos en el **capítulo 1**, una cota del sesgo del estimador de razón univariante en un esquema de muestreo con probabilidades desiguales es:

$$\mid sesgo\left(\widehat{\overline{R}}_{iD}\right) \mid \leq \sigma_{\widehat{\overline{R}}_{i}D} C_{\overline{v}_i}$$

y por tanto

$$\mid sesgo\left(\widehat{\overline{Y}}_{R_iD}\right) \mid = \mid \overline{X}_i sesgo\left(\widehat{\overline{R}}_{iD}\right) \mid \leq \overline{X}_i \sigma_{\widehat{\overline{R}}_{i}D} C_{\overline{v}_i} = \sigma_{\widehat{\overline{Y}}_{R_iD}} C_{\overline{v}_i} \; i = 1,2$$

Así pues

$$\mid sesgo\left(\widehat{\overline{Y}}_{R_{12}D}\right) \mid = \mid w_1 sesgo\left(\widehat{\overline{Y}}_{R_1D}\right) + w_2 sesgo\left(\widehat{\overline{Y}}_{R_2D}\right) \mid \leq$$

$$\leq w_1 \mid sesgo\left(\widehat{\overline{Y}}_{R_1D}\right) \mid + w_2 \mid sesgo\left(\widehat{\overline{Y}}_{R_2D}\right) \mid \leq \underset{i}{máx}\left(\sigma_{\widehat{\overline{Y}}_{R_iD}} \cdot C_{\overline{v}_i}\right) =$$

$$= \frac{1}{\sqrt{n}} \underset{i}{máx}\left(\sigma_{\widehat{\overline{Y}}_{R_iD}} \cdot C_{v_i}\right)$$

Por otra parte, hay veces que es más util una expresión aproximada que una cota.

Proposición 3.3.35 *Una expresión aproximada de orden* $O(n^{-2})$ *para el sesgo del estimador viene dada por la expresión:*

$$sesgo\left(\widehat{\overline{Y}}_{R_{12}D}\right) \simeq \frac{\overline{Y}}{n}\left(\sum_{i=1}^{2} w_i \left(\frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{\rho_i \sigma_{v_i} \sigma_z}{\overline{X}_i \overline{Y}}\right)\right)$$

Demostración.-

$$sesgo\left(\widehat{\overline{Y}}_{R_{12}D}\right) = w_1 sesgo\left(\widehat{\overline{Y}}_{R_1D}\right) + w_2 sesgo\left(\widehat{\overline{Y}}_{R_2D}\right) \tag{3.34}$$

según la proposición 2.2.20:

$$sesgo\left(\widehat{\overline{Y}}_{R_iD}\right) \simeq \frac{\overline{Y}}{n}\left(\frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{\rho_i \sigma_{v_i} \sigma_z}{\overline{X}_i \overline{Y}}\right)$$

donde $\rho_i = \rho_{v_i,z}$.

Entonces sustituyendo esta expresión en la expresión 3.34 se obtiene el resultado deseado.

De aquí se obtiene:

$$sesgo\left(\widehat{\overline{Y}}_{R_{12}D}\right) \simeq \frac{\overline{Y}}{n}\left(\sum_{i=1}^{2} w_i\left(\frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}}\right)\right) =$$

$$= \frac{\overline{Y}}{n}\sum_{i=1}^{2} w_i\left(C_{v_i}^2 - \rho_i C_{v_i} C_z\right)$$

Error cuadrático medio del estimador.

Proposición 3.3.36 *Los errores cuadráticos medios de los estimadores* $\widehat{\overline{Y}}_{R_{12}D}$ *y* $\widehat{Y}_{R_{12}D}$ *vienen dados por las expresiones:*

$$ECM\left(\widehat{\overline{Y}}_{R_{12}D}\right) = \frac{\overline{Y}^2}{n}\left[w_1^2\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_1}^2}{\overline{X}_1^2} - \frac{2\rho_1\sigma_{v_1}\sigma_z}{\overline{X}_1\overline{Y}}\right) + \right.$$

$$+w_2^2\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_2}^2}{\overline{X}_2^2} - \frac{2\rho_2\sigma_{v_2}\sigma_z}{\overline{X}_2\overline{Y}}\right) +$$

$$\left.+2w_1w_2\left(\frac{\sigma_z^2}{\overline{Y}^2} - \frac{\rho_2\sigma_{v_2}\sigma_z}{\overline{X}_2\overline{Y}} - \frac{\rho_1\sigma_{v_1}\sigma_z}{\overline{X}_1\overline{Y}} + \frac{\rho_{12}\sigma_{v_1}\sigma_{v_2}}{\overline{X}_1\overline{X}_2}\right)\right]$$

$$ECM\left(\widehat{Y}_{R_{12}D}\right) = \frac{Y^2}{n}\left[w_1^2\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_1}^2}{\overline{X}_1^2} - \frac{2\rho_1\sigma_{v_1}\sigma_z}{\overline{X}_1\overline{Y}}\right) + \right.$$

$$+w_2^2\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_2}^2}{\overline{X}_2^2} - \frac{2\rho_2\sigma_{v_2}\sigma_z}{\overline{X}_2\overline{Y}}\right) +$$

$$\left.+2w_1w_2\left(\frac{\sigma_z^2}{\overline{Y}^2} - \frac{\rho_2\sigma_{v_2}\sigma_z}{\overline{X}_2\overline{Y}} - \frac{\rho_1\sigma_{v_1}\sigma_z}{\overline{X}_1\overline{Y}} + \frac{\rho_{12}\sigma_{v_1}\sigma_{v_2}}{\overline{X}_1\overline{X}_2}\right)\right]$$

Demostración.-

Partiendo de la expresión:

$$ECM\left(\widehat{\overline{Y}}_{R_{12}D}\right) = w_1^2 ECM\left(\widehat{\overline{Y}}_{R_1D}\right) + w_2^2 ECM\left(\widehat{\overline{Y}}_{R_2D}\right) +$$

$$+2w_1w_2E\left[\left(\widehat{\overline{Y}}_{R_1D}-\overline{Y}\right)\left(\widehat{\overline{Y}}_{R_2D}-\overline{Y}\right)\right]$$

y dado que una aproximación del error cuadrático medio del estimador $\widehat{\overline{Y}}_{R_iD}$, según la proposición 2.2.22, viene dada por la expresión:

$$ECM\left(\widehat{\overline{Y}}_{R_iD}\right)\simeq\frac{1}{n}\left[\sum_{j=1}^{N}p_j(z^j-Rv_i^j)^2\right]=\overline{Y}^2\frac{1}{n}\left[\frac{\sigma_z^2}{\overline{Y}^2}+\frac{\sigma_{v_i}^2}{\overline{X}_i^2}-\frac{2\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}}\right]$$

sólo queda calcular el término de los productos cruzados.

De la aproximación:

$$\widehat{\overline{Y}}_{R_iD}-\overline{Y}=\overline{Y}\left[\frac{\overline{z}-\overline{Y}}{\overline{Y}}-\frac{\overline{v}_i-\overline{X}_i}{\overline{X}_i}\right]$$

se deduce:

$$E\left[\left(\widehat{\overline{Y}}_{R_1D}-\overline{Y}\right)\left(\widehat{\overline{Y}}_{R_2D}-\overline{Y}\right)\right]=$$

$$=\overline{Y}^2\left\{\frac{E\left[\left(\overline{z}-\overline{Y}\right)^2\right]}{\overline{Y}^2}-\frac{E\left[\left(\overline{z}-\overline{Y}\right)\left(\overline{v}_2-\overline{X}_2\right)\right]}{\overline{Y}\,\overline{X}_2}-\right.$$

$$\left.-\frac{E\left[\left(\overline{z}-\overline{Y}\right)\left(\overline{v}_1-\overline{X}_1\right)\right]}{\overline{Y}\,\overline{X}_1}+\frac{E\left[\left(\overline{v}_2-\overline{X}_2\right)\left(\overline{v}_1-\overline{X}_1\right)\right]}{\overline{X}_1\overline{X}_2}\right\}=$$

$$=\overline{Y}^2\left\{\frac{\sigma_z^2}{n\overline{Y}^2}-\frac{\rho_2\sigma_z\sigma_{v_2}}{n\overline{Y}\,\overline{X}_2}-\frac{\rho_1\sigma_z\sigma_{v_1}}{n\overline{Y}\,\overline{X}_1}+\frac{\rho_{12}\sigma_z\sigma_{v_2}}{n\overline{X}_1\overline{X}_2}\right\}$$

Por tanto, sustituyendo los valores de $ECM\left(\widehat{\overline{Y}}_{R_1D}\right)$, $ECM\left(\widehat{\overline{Y}}_{R_2D}\right)$ y $E\left[\left(\widehat{\overline{Y}}_{R_1D}-\overline{Y}\right)\left(\widehat{\overline{Y}}_{R_2D}-\overline{Y}\right)\right]$, se tiene:

$$ECM\left(\widehat{\overline{Y}}_{R_{12}D}\right)=\frac{\overline{Y}^2}{n}\left[w_1^2\left(\frac{\sigma_z^2}{\overline{Y}^2}+\frac{\sigma_{v_1}^2}{\overline{X}_1^2}-\frac{2\rho_1\sigma_{v_1}\sigma_z}{\overline{X}_1\overline{Y}}\right)+\right.$$

$$\left.+w_2^2\left(\frac{\sigma_z^2}{\overline{Y}^2}+\frac{\sigma_{v_2}^2}{\overline{X}_2^2}-\frac{2\rho_2\sigma_{v_2}\sigma_z}{\overline{X}_2\overline{Y}}\right)+\right.$$

$$+2w_1w_2\left(\frac{\sigma_z^2}{\overline{Y}^2}-\frac{\rho_2\sigma_{v_2}\sigma_z}{\overline{X}_2\overline{Y}}-\frac{\rho_1\sigma_{v_1}\sigma_z}{\overline{X}_1\overline{Y}}+\frac{\rho_{12}\sigma_{v_1}\sigma_{v_2}}{\overline{X}_1\overline{X}_2}\right)\Bigg]$$

La fórmula para $ECM\left(\widehat{Y}_{R_{12}D}\right)$ se obtiene sin más que considerar la igualdad

$$ECM\left(\widehat{Y}_{R_{12}D}\right)=N^2ECM\left(\widehat{\overline{Y}}_{R_{12}D}\right)$$

Por último, vamos a determinar los pesos óptimos en el sentido de que produzcan menor error cuadrático medio del estimador de razón.

Proposición 3.3.37 *Los pesos que minimizan el* $ECM\left(\widehat{\overline{Y}}_{R_{12}D}\right)$ *vienen dados por las expresiones siguientes:*

$$w_1=\frac{\frac{\sigma_{v_2}^2}{\overline{X}_2^2}-\frac{\rho_2\sigma_z\sigma_{v_2}}{\overline{Y}\,\overline{X}_2}-\frac{\rho_1\sigma_z\sigma_{v_1}}{\overline{Y}\,\overline{X}_1}+\frac{\rho_{12}\sigma_{v_1}\sigma_{v_2}}{\overline{X}_1\overline{X}_2}}{\frac{\sigma_{v_1}^2}{\overline{X}_1^2}+\frac{\sigma_{v_2}^2}{\overline{X}_2^2}-2\frac{\rho_{12}\sigma_{v_1}\sigma_{v_2}}{\overline{X}_1\overline{X}_2}}$$

$$w_2=1-w_1$$

Demostración.-

Si derivamos la expresión del $ECM\left(\widehat{\overline{Y}}_{R_{12}D}\right)$ en función de w_1 e igualamos a cero, obtenemos:

$$w_1=\frac{ECM\left(\widehat{\overline{Y}}_{R_2D}\right)-E\left[\left(\widehat{\overline{Y}}_{R_1D}-\overline{Y}\right)\left(\widehat{\overline{Y}}_{R_2D}-\overline{Y}\right)\right]}{ECM\left(\widehat{\overline{Y}}_{R_1D}\right)+ECM\left(\widehat{\overline{Y}}_{R_2D}\right)-2B}=$$

$$=\frac{\left(\frac{\sigma_z^2}{\overline{Y}^2}-\frac{\sigma_{v_2}^2}{\overline{X}_2}-\frac{2\rho_2\sigma_z\sigma_{v_2}}{\overline{Y}\,\overline{X}_2}\right)-\left(\frac{\sigma_z^2}{\overline{Y}^2}-\frac{\rho_2\sigma_z\sigma_{v_2}}{\overline{Y}\,\overline{X}_2}-\frac{\rho_1\sigma_z\sigma_{v_1}}{\overline{Y}\,\overline{X}_1}+\frac{\rho_{12}\sigma_{v_1}\sigma_{v_2}}{\overline{X}_1\overline{X}_2}\right)}{\left(\frac{\sigma_z^2}{\overline{Y}^2}+\frac{\sigma_{v_1}^2}{\overline{X}_1^2}-2\frac{\rho_1\sigma_z\sigma_{v_1}}{\overline{Y}\,\overline{X}_1}\right)+\left(\frac{\sigma_z^2}{\overline{Y}^2}+\frac{\sigma_{v_2}^2}{\overline{X}_2^2}-2\frac{\rho_2\sigma_z\sigma_{v_2}}{\overline{Y}\,\overline{X}_2}\right)-2B}=$$

$$= \frac{\frac{\sigma_{v_2}^2}{\overline{X}_2^2} - \frac{\rho_2 \sigma_z \sigma_{v_2}}{\overline{Y}\,\overline{X}_2} + \frac{\rho_1 \sigma_z \sigma_{v_1}}{\overline{Y}\,\overline{X}_1} - \frac{\rho_{12} \sigma_{v_1} \sigma_{v_2}}{\overline{X}_1 \overline{X}_2}}{\frac{\sigma_{v_1}^2}{\overline{X}_1^2} + \frac{\sigma_{v_2}^2}{\overline{X}_2^2} - 2\frac{\rho_{12} \sigma_{v_1} \sigma_{v_2}}{\overline{X}_1 \overline{X}_2}}$$

donde

$$B = E\left[\left(\hat{\overline{Y}}_{R_1 D} - \overline{Y}\right)\left(\hat{\overline{Y}}_{R_2 D} - \overline{Y}\right)\right] =$$

$$= \frac{\sigma_z^2}{\overline{Y}^2} + \frac{\rho_2 \sigma_z \sigma_{v_2}}{\overline{Y}\,\overline{X}_2} - \frac{\rho_1 \sigma_z \sigma_{v_1}}{\overline{Y}\,\overline{X}_1} + \frac{\rho_{12} \sigma_{v_1} \sigma_{v_2}}{\overline{X}_1 \overline{X}_2}$$

En el caso particular

$$\rho_1 = \rho_2 = \rho \quad \text{y} \quad C_{v_1} = C_{v_2} = C$$

se obtiene:

$$w_1 = \frac{C^2 - \rho C + \rho C^2 - \rho_{12} C^2}{C^2 + C^2 - 2\rho_{12} C^2} = \frac{C^2 (1 - \rho_{12})}{2C^2 (1 - \rho_{12})} = \frac{1}{2}$$

y el error cuadrático medio adopta la forma:

$$ECM\left(\hat{\overline{Y}}_{R_{12} D}\right) = \frac{\overline{Y}^2}{n}\left[\frac{1}{4}\left(\frac{\sigma_z^2}{\overline{Y}^2} + C^2 - 2\rho C_z C\right) +\right.$$

$$\left. + \left(\frac{\sigma_z^2}{\overline{Y}^2} + C^2 - 2\rho C_z C\right) + \frac{1}{2}\left(\frac{\sigma_z^2}{\overline{Y}^2} - 2\rho \frac{\sigma_z}{\overline{Y}} C + \rho_{12} C^2\right)\right] =$$

$$= \frac{\overline{Y}^2}{n}\left[\frac{\sigma_z^2}{\overline{Y}^2} + \frac{C^2}{2} - 2\rho C_z C + \frac{\rho_{12}}{2} C^2\right] =$$

$$= \frac{\overline{Y}^2}{n}\left[C_z^2 + \frac{C^2}{2}(1 + \rho_{12}) - 2\rho C_z C\right]$$

Comparación con el estimador de Hansen y Hurwitz.

Si no se utiliza ninguna variable auxiliar, el estimador que se suele utilizar para estimar la media, es el estimador de *Hansen* y *Hurwitz* (1943), cuya varianza vale:

$$V(\overline{y}_{HH}) = \frac{\sigma_z^2}{n} = \sum_{j=1}^{N} p_j (z_j - \overline{Y})^2 = \frac{C_z^2 \overline{Y}^2}{n}$$

por lo que el estimador $\widehat{\overline{Y}}_{R_{12}D}$ será más eficaz si:

$$V(\overline{y}_{HH}) - ECM\left(\widehat{\overline{Y}}_{R_{12}D}\right) = -\frac{C^2}{2}(1+\rho_{12}) + 2\rho C_z C > 0 \Leftrightarrow$$

$$\Leftrightarrow \frac{C^2}{2}(1+\rho_{12}) < 2\rho C_z C \Leftrightarrow \frac{C}{C_z}\,\frac{1+\rho_{12}}{\rho} < 4$$

Si la correlación entre x_1 y x_2 es perfecta ($\rho_{12} = 1$) el resultado anterior se traduce en:

$$\rho \frac{C_z}{C} > \frac{1}{2}$$

que coincide con el resultado obtenido en el caso de utilizar sólo una variable auxiliar.

3.3.2 Caso de $k > 2$ variables auxiliares.

En el muestreo con probabilidades desiguales hemos definido el estimador de razón multivariante como

$$\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k} = \sum_{i=1}^{k} w_i \widehat{\overline{Y}}_{R_i D}$$

Este estimador en general no es consistente, pues los estimadores $\widehat{\overline{Y}}_{R_i D}$ no lo son. Estudiamos a continuación sus propiedades de sesgo y error cuadrático medio.

Sesgo.

Proposición 3.3.38 *Si la muestra ha sido elegida con reemplazo y probabilidades p_i, $i = 1, \ldots, N$, una cota para el sesgo del estimador $\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}$ viene dada por*

$$\left|sesgo\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right)\right| \leq \frac{1}{\sqrt{n}} \max_i \left(C_{vi} \cdot \sigma_{\widehat{\overline{Y}}_{R_i D}}\right)$$

Proposición 3.3.39 *Una expresión aproximada de orden* $O(n^{-2})$ *para el sesgo del estimador* $\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}$ *viene dada por la expresión:*

$$sesgo\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right) \simeq \frac{\overline{Y}}{n}\left[\sum_{i=1}^{k} w_i\left(\frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{\rho_i \sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}}\right)\right]$$

La demostración de las dos proposiciones es inmediata a partir de las proposiciones 3.3.34 y 3.3.35.

Error cuadrático medio.

Vamos a obtener una expresión del error cuadrático medio del estimador $\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}$, para posteriormente calcular los pesos $w_1, \ldots, w_k$ que lo hacen mínimo.

Proposición 3.3.40 *Expresiones aproximadas del error cuadrático medio de los estimadores* $\widehat{Y}_{R_D}^{1,2,\ldots,k}$ *e* $\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}$ *son:*

$$ECM\left(\widehat{Y}_{R_D}^{1,2,\ldots,k}\right) = \sum_{i=1}^{k} w_i^2 \frac{Y^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{2\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}}\right) +$$

$$+ \sum_{i\neq j} w_i w_j \frac{Y^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} - \frac{\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}} - \frac{\rho_j\sigma_{v_j}\sigma_z}{\overline{X}_j\overline{Y}} + \frac{\rho_{ij}\sigma_{v_i}\sigma_{v_j}}{\overline{X}_i\overline{X}_j}\right)$$

$$ECM\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right) = \sum_{i=1}^{k} w_i^2 \frac{\overline{Y}^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{2\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}}\right) +$$

$$+ \sum_{i\neq j} w_i w_j \frac{\overline{Y}^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} - \frac{\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}} - \frac{\rho_j\sigma_{v_j}\sigma_z}{\overline{X}_j\overline{Y}} + \frac{\rho_{ij}\sigma_{v_i}\sigma_{v_j}}{\overline{X}_i\overline{X}_j}\right)$$

Demostración.-

$$ECM\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right) = E\left[\sum_{i=1}^{k} w_i \widehat{\overline{Y}}_{R_i D} - \sum_{i=1}^{k} w_i \overline{Y}\right]^2 =$$

$$= E\left[\sum_{i=1}^{k} w_i \left(\hat{\overline{Y}}_{R_iD} - \overline{Y}\right)\right]^2 =$$

$$= E\left[\sum_{i=1}^{k} w_i^2 \left(\hat{\overline{Y}}_{R_iD} - \overline{Y}\right)^2 + \sum_{i\neq j} w_i w_j \left(\hat{\overline{Y}}_{R_iD} - \overline{Y}\right)\left(\hat{\overline{Y}}_{R_jD} - \overline{Y}\right)\right] =$$

$$= \sum_{i=1}^{k} w_i^2 E\left[\hat{\overline{Y}}_{R_iD} - \overline{Y}\right]^2 + \sum_{i\neq j} w_i w_j E\left[\left(\hat{\overline{Y}}_{R_iD} - \overline{Y}\right)\left(\hat{\overline{Y}}_{R_jD} - \overline{Y}\right)\right] =$$

$$= \sum_{i=1}^{k} w_i^2 \frac{\overline{Y}^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{2\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}}\right) +$$

$$+ \sum_{i\neq j} w_i w_j \frac{\overline{Y}^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} - \frac{\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}} - \frac{\rho_j\sigma_{v_j}\sigma_z}{\overline{X}_j\overline{Y}} + \frac{\rho_{ij}\sigma_{v_i}\sigma_{v_j}}{\overline{X}_i\overline{X}_j}\right)$$

sustituyendo en la expresión los valores:

$$ECM\left(\hat{\overline{Y}}_{R_iD}\right) = \frac{\overline{Y}^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} + \frac{\sigma_{v_i}^2}{\overline{X}_i^2} - \frac{2\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}}\right)$$

$$E\left[\left(\hat{\overline{Y}}_{R_iD} - \overline{Y}\right)\left(\hat{\overline{Y}}_{R_jD} - \overline{Y}\right)\right] = \frac{\overline{Y}^2}{n}\left(\frac{\sigma_z^2}{\overline{Y}^2} - \frac{\rho_i\sigma_{v_i}\sigma_z}{\overline{X}_i\overline{Y}} - \frac{\rho_j\sigma_{v_j}\sigma_z}{\overline{X}_j\overline{Y}} + \frac{\rho_{ij}\sigma_{v_i}\sigma_{v_j}}{\overline{X}_i\overline{X}_j}\right)$$

se obtiene el resultado deseado.

La expresión para $ECM\left(\hat{Y}_{R_D}^{1,2,\ldots,k}\right)$ es inmediata sin más que considerar que $ECM\left(\hat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right) = N^{-2}ECM\left(\hat{Y}_{R_D}^{1,2,\ldots,k}\right)$.

Considerando ahora las matrices:

$$W = (w_1, \ldots, w_k)'$$

$$A_{(k\times k)} = \left(a_{ij}\right)$$

donde

$$a_{ij} = \frac{\sigma_z^2}{\overline{Y}^2} - \frac{\rho_i \sigma_{v_i} \sigma_z}{\overline{X}_i \overline{Y}} - \frac{\rho_j \sigma_{v_j} \sigma_z}{\overline{X}_j \overline{Y}} + \frac{\rho_{ij} \sigma_{v_i} \sigma_{v_j}}{\overline{X}_i \overline{X}_j} \ \forall i, j$$

podemos escribir $ECM\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right)$ de la forma:

$$ECM\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right) = \frac{\overline{Y}^2}{n} \sum_{i,j=1} w_i w_j a_{ij}$$

$$ECM\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right) = \frac{\overline{Y}^2}{n} W' A W$$

Teorema 3.3.41 *El error cuadrático medio del estimador* $\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}$ *es mínimo si* $W = \dfrac{A^{-1}e}{e'A^{-1}e}$ *y vale:*

$$mín\ ECM\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right) = \frac{\overline{Y}^2}{n}\left(e'A^{-1}e\right)^{-1}$$

donde $e_{(k\times 1)} = (1, 1, \ldots, 1)'$.

Demostración.-

Veamos en primer lugar que la matriz A es definida positiva. Llamando C a la matriz de orden $(k+1) \times (k+1)$:

$$C = \begin{bmatrix} C_z^2 & \rho_1 C_z C_{v_1} & \rho_2 C_z C_{v_2} & \cdots & \rho_k C_z C_{v_k} \\ \rho_1 C_z C_{v_1} & C_{v_1}^2 & \rho_{12} C_{v_1} C_{v_2} & \cdots & \rho_{1k} C_{v_1} C_{v_k} \\ \rho_2 C_z C_{v_2} & \rho_{12} C_{v_1} C_{v_2} & C_{v_2}^2 & \cdots & \rho_{2k} C_{v_2} C_{v_k} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_i C_z C_{v_i} & \rho_{1i} C_{v_1} C_{v_i} & \rho_{2i} C_{v_2} C_{v_i} & \cdots & \rho_{ik} C_{v_i} C_{v_k} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_k C_z C_{v_k} & \rho_{1k} C_{v_1} C_{v_k} & \rho_{2k} C_{v_2} C_{v_k} & \cdots & C_{v_k}^2 \end{bmatrix}$$

que asumimos definida positiva y L:

$$L_{(k\times(k+1))} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

de rango k, LCL' es también definida positiva, según la proposición 3.1.11.

Estamos pues en las condiciones de aplicar la desigualdad extendida de Cauchy–Schwartz, 3.16, según la cual:

$$\frac{(W'e)^2}{(e'A^{-1}e)} \leq W'AW = ECM\left(\widehat{\overline{Y}}_{R_S}^{1,2,\ldots,k}\right)$$

y el mínimo valor de $ECM\left(\widehat{\overline{Y}}_{R_D}^{1,2,\ldots,k}\right)$ es $(e'A^{-1}e)^{-1}$ ya que $W'e = \sum_{i=1}^{k} w_i = 1$.

Además, como ya comprobamos en el caso del muestreo aleatorio simple, el mínimo se alcanza para

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

como queríamos demostrar.

Comparación con el caso univariante.

Al igual que en caso del muestreo aleatorio simple, el estimador de razón multivariante es más preciso que el univariante.

Teorema 3.3.42 *En muestreo con reemplazo y probabilidades desiguales el estimador de razón multivariante que utiliza q variables auxiliares tiene un error cuadrático medio inferior o igual al estimador de razón multivariante que utiliza $p < q$ variables auxiliares.*

Demostración.-

La demostración es la misma que para el caso del muestreo aleatorio simple sin más que cambiar la matriz A por la correspondiente en el muestreo con probabilidades desiguales.

§3.4 Estimadores condensados de razón.

3.4.1 Introducción.

En muchas investigaciones se dispone de información de varias variables relacionadas con la variable de estudio y.

En las secciones anteriores se utilizaban estas variables para construir estimadores de razón (bien para los valores totales o para los estratos) y posteriormente se combinaban de forma que hicieran mínimo el error cuadrático medio del estimador multivariante final. Sin embargo, este procedimiento tiene una serie de inconvenientes, como son:

1. Los estimadores así construidos son sesgados.

2. Los estimadores óptimos son función de matrices construidas a partir de parámetros poblacionales desconocidos.

3. El cálculo de los estimadores óptimos y de sus errores cuadráticos medios es más complejo que el de los correspondientes estimadores de expansión simple y sin embargo, no tienen por qué ser más precisos que estos.

En esta sección se propone una nueva idea acerca de como utilizar la información que suministran las variables auxiliares mediante estimadores de tipo razón que van a resolver algunos de los inconvenientes antes citados.

La idea básica es la siguiente: si se dispone de varias variables auxiliares se pueden condensar en otra variable que sea una función lineal de las primeras y que haga máxima la correlación entre la variable principal y esta función lineal. Una vez obtenida esta variable "condensada" podemos utilizarla para construir un estimador de tipo razón "univariante" utilizando esta nueva variable como variable auxiliar. De esta forma se utiliza óptimamente la información que proporcionan todas las variables auxiliares disponibles pero con un procedimiento mucho más simple que en el caso el estimador multivariante propuesto anteriormente, que conlleva la determinación de k estimadores de razón (siendo k el número de variables auxiliares).

Se pueden utilizar distintos criterios para construir la variable condensada. Nosotros vamos a utilizar dos criterios distintos que originarán estimadores en principio distintos.

3.4.2 El estimador de razón condensado (Estimador RC).

La primera condición que utilizamos para construir la variable condensada viene determinada por el hecho, tal y como se estudió en el capítulo 1, que cuanto mayor sea la correlación entre la variable principal y la auxiliar que se estudia para construir el estimador de razón, mayor será la precisión del

estimador obtenido y más ventajoso será el método de razón comparado con el de expansión simple. Vamos a proceder a desarrollar esta idea.

Determinación de la variable condensada.

Supongamos que tenemos información para cada unidad de la población de $x_1, x_2, \ldots, x_k$ variables auxiliares.

Sea x_{ij} el valor de la j-ésima variable auxiliar en la i-ésima unidad de la población.

Definimos las matrices:

$$S_{(k\times k)} = \left(S_{ij}\right) \; ; \quad S_{ij} = \mathrm{Cov}\left(x_i, x_j\right)$$

$$S^0_{(k\times 1)} = \left(S_i\right) \; ; \quad S_i = \mathrm{Cov}\left(y, x_i\right)$$

$$x_{(k\times 1)} = (x_1, x_2, \ldots, x_k)'$$

Para obtener una variable que condense las k variables auxiliares que utilizaremos después para construir el estimador de razón, consideramos la siguiente función de $x_1, x_2, \ldots, x_k$:

$$z = \sum_{i=1}^{k} a_i x_i$$

o en forma matricial:

$$z = a'x$$

donde $a = (a_1, \ldots, a_k)'$ es el vector de coeficientes que elegimos de forma que la correlación entre z e y sea máxima. Es decir, maximizamos:

$$\rho\left(z, y\right) = \frac{\mathrm{Cov}\left(y, z\right)}{S_y S_z} = \frac{a'S^0}{S_y\left(a'Sa\right)^{\frac{1}{2}}}$$

La matriz S asumimos que es definida positiva por lo que aplicando la desigualdad extendida de Cauchy–Schwartz a la expresión:

$$E = \frac{\left(a'S^0\right)^2}{a'Sa}$$

obtenemos que esta función es máxima en

$$a = CS^{-1}S^0$$

para alguna constante C y por tanto, al ser $\rho(z,y)$ una función monótona creciente de E, tomará su valor máximo en el mismo punto.

La constante C la determinamos imponiendo la condición:

$$V(z) = a'Sa = 1$$

obteniendo entonces

$$C = \left(S^{0\prime}S^{-1}S^0\right)^{-\frac{1}{2}}$$

es decir:

$$a = \frac{S^{-1}S^0}{\left(S^{0\prime}S^{-1}S^0\right)^{\frac{1}{2}}}$$

Definimos la variable condensada como:

$$z = \frac{S^{0\prime}S^{-1}x}{\left(S^{0\prime}S^{-1}S^0\right)^{\frac{1}{2}}}$$

Definición del estimador.

Dadas y y z, podemos construir un estimador de razón que utilice la variable z como auxiliar, de la forma:

$$\widehat{\overline{Y}}_R^{cond} = \frac{\overline{y}}{\overline{z}}\overline{Z}$$

siendo

$\overline{z}$ la media muestral de la variable z y

$\overline{Z}$ la media poblacional de la variable z.

Este estimador en función de x adopta la forma:

$$\widehat{\overline{Y}}_R^{cond} = \frac{\overline{y}}{S^{0\prime}S^{-1}\overline{x}}S^{0\prime}S^{-1}\overline{X}$$

donde

$\overline{x}_{(k\times 1)} = (\overline{x}_1, \ldots, \overline{x}_k)'$ y

$\overline{X}_{(k\times 1)} = \left(\overline{X}_1, \ldots, \overline{X}_k\right)'$.

Del forma análoga se define el estimador del total:

$$\hat{Y}_R^{cond} = \frac{\overline{y}}{S^{0\prime} S^{-1} \overline{x}} S^{0\prime} S^{-1} X$$

Pasamos a continuación a estudiar las propiedades de los estimadores definidos.

Sesgo.

Como en anteriores casos daremos una expresión aproximada del sesgo del estimador.

Proposición 3.4.43 *Una expresión aproximada de orden* $O(n^{-2})$ *del sesgo del estimador* $\widehat{\overline{Y}}_R^{cond}$ *viene dada por la expresión:*

$$sesgo\left(\widehat{\overline{Y}}_R^{cond}\right) \simeq \overline{Y} \frac{1-f}{n} \frac{S^{0\prime} S^{-1} S^0}{S^{0\prime} S^{-1} \overline{X}} \left[\frac{1}{S^{0\prime} S^{-1} \overline{X}} - \frac{1}{\overline{Y}}\right]$$

Demostración.-

Considerando las variables:

$$o_1 = \frac{\overline{y} - \overline{Y}}{\overline{Y}} \; ; \quad o_2 = \frac{\overline{z} - \overline{Z}}{\overline{Z}}$$

podemos expresar el estimador $\widehat{\overline{Y}}_R^{cond}$ de la forma:

$$\widehat{\overline{Y}}_R^{cond} = \overline{Y}\,(1 + o_1)\,(1 + o_2)^{-1}$$

Desarrollando $(1 + o_2)^{-1}$ en serie de Taylor, se obtiene:

$$\widehat{\overline{Y}}_R^{cond} - \overline{Y} = \overline{Y}\left[(o_1 - o_2) - o_2\,(o_1 - o_2) + o_2^2\,(o_1 - o_2) - \ldots\right]$$

y quedándonos sólo con los términos de orden igual o inferior a dos en o_1 y o_2, obtenemos la aproximación:

$$\widehat{\overline{Y}}_R^{cond} - \overline{Y} \simeq \overline{Y}\left[o_1 - o_2 - o_2 o_1 + o_2^2\right]$$

Al tomar esperanzas nos queda:

$$sesgo\left(\widehat{\overline{Y}}_R^{cond}\right) \simeq \overline{Y}\left[E\left(o_2^2\right) - E\left(o_2 o_1\right)\right]$$

Ahora bien:

$$E\left(o_2^2\right) = \frac{V\left(\overline{z}\right)}{\overline{Z}^2} \; ; \quad E\left(o_2 o_1\right) = \frac{\mathrm{Cov}\left(\overline{z}, \overline{y}\right)}{\overline{Y}\,\overline{Z}}$$

y en un muestreo aleatorio simple:

$$V\left(\overline{z}\right) = \frac{1-f}{n} V\left(z\right) \; ; \quad \mathrm{Cov}\left(\overline{z}, \overline{y}\right) = \frac{1-f}{n} \mathrm{Cov}\left(z, y\right)$$

con

$$V\left(z\right) = 1 \; ; \quad \mathrm{Cov}\left(z, y\right) = \left(S^{0\prime} S^{-1} S^0\right)^{\frac{1}{2}}$$

Por tanto

$$sesgo\left(\widehat{\overline{Y}}_R^{cond}\right) \simeq \overline{Y}\frac{1-f}{n}\left[\frac{S^{0\prime} S^{-1} S^0}{\left(S^{0\prime} S^{-1}\overline{X}\right)^2} - \frac{S^{0\prime} S^{-1} S^0}{\overline{Y}\left(S^{0\prime} S^{-1}\overline{X}\right)}\right] =$$

$$= \overline{Y}\frac{1-f}{n}\frac{S^{0\prime} S^{-1} S^0}{S^{0\prime} S^{-1}\overline{X}}\left[\frac{1}{S^{0\prime} S^{-1}\overline{X}} - \frac{1}{\overline{\overline{Y}}}\right]$$

Condiciones bajo las cuales el estimador de razón condensado es insesgado.

Aplicando el teorema 1.3.15 para $x = z$ obtenemos las condición bajo la cual el estimador de razón condensado es insesgado.

Proposición 3.4.44 *Si la regresión de y sobre z es lineal y pasa por el origen, en una población finita, los estimadores* $\widehat{\overline{Y}}_R^{cond}$ *e* $\widehat{Y}_R^{cond}$ *no tienen sesgo en una muestra aleatoria simple de tamaño n.*

Este resultado nos da idea que cuanto más próximo a cero esté la ordenada en el origen de la recta de regresión de y sobre z, más pequeño será el sesgo de los estimadores de razón. Esta idea nos va permitir construir otro estimador condensado que va a ser insesgado.

3.4.3 El estimador de razón condensado e insesgado (Estimador RCI).

Consideramos una variable z' de la forma:

$$z' = a_0 + a'x$$

que haga máxima la correlación con y. Puesto que esta correlación es independiente de a_0, obtenemos que z' es de la forma:

$$z' = a_0 + \frac{S^{0'}S^{-1}x}{\left(S^{0'}S^{-1}S^0\right)^{\frac{1}{2}}}$$

y podemos determinar a_0 de forma que la ordenada en el origen de la recta de regresión de y sobre z' sea cero:

$$\overline{Y} - \frac{a'S^0}{a'Sa}\overline{Z'} = \overline{Y} - \left(S^{0'}S^{-1}S^0\right)^{\frac{1}{2}}\left(a_0 + \frac{S^{0'}S^{-1}\overline{X}}{\left(S^{0'}S^{-1}S^0\right)^{\frac{1}{2}}}\right) = 0$$

Despejando a_0, obtenemos:

$$a_0 = \frac{\overline{Y} - S^{0'}S^{-1}\overline{X}}{\left(S^{0'}S^{-1}S^0\right)^{\frac{1}{2}}}$$

y por tanto podemos considerar la variable condensada:

$$z' = \frac{\overline{Y} + S^{0'}S^{-1}\left(x - \overline{X}\right)}{\left(S^{0'}S^{-1}S^0\right)^{\frac{1}{2}}}$$

Esta nueva variable z' verifica:

$$\overline{Z'} = \frac{\overline{Y}}{\left(S^{0\prime} S^{-1} S^{0}\right)^{\frac{1}{2}}}$$

y

$$\overline{z'} = \frac{\overline{Y} + S^{0\prime} S^{-1} \left(\overline{x} - \overline{X}\right)}{\left(S^{0\prime} S^{-1} S^{0}\right)^{\frac{1}{2}}}$$

Podemos así definir un nuevo estimador de razón:

$$\widehat{\overline{Y}}_R^{cond\,\prime} = \frac{\overline{y}}{\overline{z'}} \overline{Z'}$$

Sesgo del estimador $\widehat{\overline{Y}}_R^{cond\,\prime}$.

El nuevo estimador considerado es insesgado, como prueba la siguiente proposición.

Proposición 3.4.45 *En un muestreo aleatorio simple el estimador $\widehat{\overline{Y}}_R^{cond\,\prime}$ tiene sesgo cero para muestras de tamaño n.*

Demostración.-
Partimos de la expresión:

$$\widehat{\overline{Y}}_R^{cond\,\prime} = \overline{Y}\left(1 + o_1'\right)\left(1 + o_2'\right)^{-1}$$

donde

$$o_1' = \frac{\overline{y} - \overline{Y}}{\overline{Y}} \; ; \quad o_2' = \frac{\overline{z'} - \overline{Z'}}{\overline{Z'}}$$

Desarrollando $\left(1 + o_2'\right)^{-1}$ en serie de Taylor, se obtiene:

$$\widehat{\overline{Y}}_R^{cond\,\prime} - \overline{Y} = \overline{Y}\left[(o_1' - o_2') - o_2'(o_1' - o_2') + o_2'^2(o_1' - o_2') - \ldots\right]$$

y quedándonos sólo con los términos de orden igual o inferior a dos en o_1' y o_2', obtenemos la aproximación:

$$\widehat{\overline{Y}}_R^{cond\,\prime} - \overline{Y} \simeq \overline{Y}\left[o_1' - o_2' - o_2'o_1' + o_2'^2\right]$$

Al tomar esperanzas nos queda:

$$sesgo\left(\widehat{\overline{Y}}_R^{cond\,\prime}\right) \simeq \overline{Y}\left[E\left(o_2'^2\right) - E\left(o_2'o_1'\right)\right]$$

Ahora bien:

$$E\left(o_2'^2\right) = \frac{V\left(\overline{z'}\right)}{\overline{Z'}^2} \;;\quad E\left(o_2'o_1'\right) = \frac{\mathrm{Cov}\left(\overline{z'}, \overline{y}\right)}{\overline{Y}\;\overline{Z'}}$$

y en un muestreo aleatorio simple:

$$V\left(\overline{z'}\right) = \frac{1-f}{n}V\left(z'\right) \;;\quad \mathrm{Cov}\left(\overline{z'}, \overline{y}\right) = \frac{1-f}{n}\mathrm{Cov}\left(z', y\right)$$

con

$$V\left(z'\right) = V\left(z\right) = 1 \;;\quad \mathrm{Cov}\left(z', y\right) = \mathrm{Cov}\left(z, y\right) = \left(S^{0\prime}S^{-1}S^{0}\right)^{\frac{1}{2}}$$

Por tanto

$$sesgo\left(\widehat{\overline{Y}}_R^{cond\,\prime}\right) \simeq \overline{Y}\frac{1-f}{n}\left[\frac{1}{\overline{Z'}^2} - \frac{\left(S^{0\prime}S^{-1}S^{0}\right)^{\frac{1}{2}}}{\overline{Y}\;\overline{Z'}}\right] =$$

$$= \overline{Y}\frac{1-f}{n}\left[\frac{S^{0\prime}S^{-1}S^{0}}{\overline{Y}^2} - \frac{S^{0\prime}S^{-1}S^{0}}{\overline{Y}^2}\right] = 0$$

Error cuadrático medio.

Calculemos el error cuadrático medio de los estimadores $\widehat{\overline{Y}}_R^{cond\,'}$ y $\widehat{\overline{Y}}_R^{cond}$.

En primer lugar consideremos el caso del estimador $\widehat{\overline{Y}}_R^{cond\,'}$.

Proposición 3.4.46 *En un muestreo aleatorio simple, una primera aproximación de la varianza del estimador* $\widehat{\overline{Y}}_R^{cond\,'}$ *viene dada por la expresión:*

$$V\left(\widehat{\overline{Y}}_R^{cond\,'}\right) \simeq \frac{1-f}{n}\left[S_y^2 - S^{0\prime} S^{-1} S^0\right]$$

Demostración.-

$$\widehat{\overline{Y}}_R^{cond\,'} - \overline{Y} = \overline{Y}\left[(o_1' - o_2') - o_2'(o_1' - o_2') + o_2'^{\,2}(o_1' - o_2') - \ldots\right]$$

donde

$$o_1' = \frac{\overline{y} - \overline{Y}}{\overline{Y}} \;;\quad o_2' = \frac{\overline{z'} - \overline{Z'}}{\overline{Z'}}$$

Entonces, elevando al cuadrado y quedándonos sólo con los términos de orden inferior o igual a dos en o_1' y o_2', obtenemos:

$$V\left(\widehat{\overline{Y}}_R^{cond\,'}\right) = E\left[\widehat{\overline{Y}}_R^{cond\,'} - \overline{Y}\right]^2 \simeq \overline{Y}^2 E\left[o_1'^2 - o_2'^2 - 2o_2'o_1'\right] =$$

$$= \overline{Y}^2\left[\frac{V(\overline{y})}{\overline{Y}^2} + \frac{V\left(\overline{z'}\right)}{\overline{Z'}^2} - \frac{2\,\mathrm{Cov}\left(\overline{y}, \overline{z'}\right)}{\overline{Y}\,\overline{Z'}}\right]$$

y como

$$V\left(\overline{z'}\right) = V(\overline{z}) = \frac{1-f}{n}$$

$$\text{Cov}\left(\overline{y},\overline{z'}\right) = \frac{1-f}{n}\text{Cov}\left(y,z\right) = \frac{1-f}{n}\left(S^{0'}S^{-1}S^{0}\right)^{\frac{1}{2}}$$

Sustituyendo:

$$V\left(\widehat{\overline{Y}}_R^{cond\,'}\right) = \overline{Y}^2\frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{1}{\overline{Z'}^2} - \frac{2}{\overline{Y}\,\overline{Z'}}\left(S^{0'}S^{-1}S^{0}\right)^{\frac{1}{2}}\right] =$$

$$= \overline{Y}^2\frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{S^{0'}S^{-1}S^{0}}{\overline{Y}^2} - \frac{2}{\overline{Y}^2}S^{0'}S^{-1}S^{0}\right] =$$

$$= \frac{1-f}{n}\overline{Y}^2\left[\frac{S_y^2}{\overline{Y}^2} - \frac{S^{0'}S^{-1}S^{0}}{\overline{Y}^2}\right]$$

como queríamos demostrar.

Siguiendo el mismo procedimiento para z se obtiene fácilmente:

$$ECM\left(\widehat{\overline{Y}}_R^{cond}\right) \simeq \frac{1-f}{n}\overline{Y}^2\left[\frac{S_y^2}{\overline{Y}^2} + \frac{S^{0'}S^{-1}S^{0}}{\left(S^{0'}S^{-1}\overline{X}\right)^2} - \frac{2S^{0'}S^{-1}S^{0}}{S^{0'}S^{-1}\overline{X}\,\overline{Y}}\right]$$

Comparación con el muestreo aleatorio simple.

Proposición 3.4.47 *El estimador de razón condensado* $\widehat{\overline{Y}}_R^{cond\,'}$ *es como mínimo igual de eficiente que el estimador de expansión simple para muestras aleatorias simples de tamaño* n.

Demostración.-
Como es sabido

$$V\left(\overline{y}\right) = \frac{1-f}{n}S_y^2$$

y por la proposición anterior

$$V\left(\widehat{\overline{Y}}_R^{cond\,'}\right) \simeq \frac{1-f}{n}\left[S_y^2 - S^{0'}S^{-1}S^{0}\right]$$

Entonces, el estimador $\widehat{\overline{Y}}_R^{cond\,'}$ será al menos igual de preciso que el estimador $\overline{y}$ si

$$V\left(\widehat{\overline{Y}}_R^{cond\,'}\right) \leq V(\overline{y})$$

si y sólo si

$$S^{0\prime} S^{-1} S^0 \geq 0$$

Pero $S^{0\prime} S^{-1} S^0 \geq 0$ siempre por ser S^{-1} la inversa de la matriz de covarianzas.

Por tanto el estimador $\widehat{\overline{Y}}_R^{cond\,'}$ es más eficiente que $\overline{y}$ a no ser que

$$S^{0\prime} S^{-1} S^0 = 0$$

en cuyo caso será igual.

Esto último ocurre si y sólo si

$$S^0 = (0, \ldots, 0)' \Leftrightarrow \mathrm{Cov}(y, x_i) = 0 \quad \forall i$$

es decir, si todas las variables auxiliares están incorreladas con la variable principal.

Comparación con el caso de $p < q$ variables auxiliares.

Sean x_1, x_2, ..., x_q variables auxiliares positivamente correlacionadas con la variable de estudio.

Teorema 3.4.48 *El estimador de razón condensado que se puede construir a partir de q variables auxiliares,* $\widehat{\overline{Y}}_R^{cond\,'}(q)$, *es como mínimo igual de preciso que el estimador de razón condensado que se puede construir a partir de* $p < q$ *variables auxiliares,* $\widehat{\overline{Y}}_R^{cond\,'}(p)$.

Demostración.-

Según la proposición 3.4.46, hemos visto que

$$V\left(\hat{\bar{Y}}_R^{cond\,'}(q)\right) = \frac{1-f}{n}\left[S_y^2 - S_q^{0\prime} S_q^{-1} S_q^0\right]$$

$$V\left(\hat{\bar{Y}}_R^{cond\,'}(p)\right) = \frac{1-f}{n}\left[S_y^2 - S_p^{0\prime} S_p^{-1} S_p^0\right]$$

siendo

$S_q^0 = \left(S_{yx_1}, S_{yx_2}, \ldots, S_{yx_q}\right)'$

$S_q = \left(S_{x_i x_j}\right) \quad i,j = 1,\ldots,q.$

Entonces el estimador $\hat{\bar{Y}}_R^{cond\,'}(q)$ será al menos igual de preciso que el estimador $\hat{\bar{Y}}_R^{cond\,'}(p)$ si

$$V\left(\hat{\bar{Y}}_R^{cond\,'}(q)\right) \leq V\left(\hat{\bar{Y}}_R^{cond\,'}(p)\right)$$

si y sólo si

$$S_p^{0\prime} S_p^{-1} S_p^0 \leq S_q^{0\prime} S_q^{-1} S_q^0$$

Ahora bien, podemos escribir la matriz S_q en función de la matriz S_p de la forma:

$$S_q = \begin{pmatrix} S_p & B \\ B' & D \end{pmatrix}$$

donde S_p y D son matrices no singulares por lo que

$$S_q^{-1} = \begin{pmatrix} S_p^{-1} + FH^{-1}G & FH^{-1} \\ H^{-1}G & H^{-1} \end{pmatrix}$$

con

$S_p F = -B$

$GS_p = -B'$

$H = D - B' S_p^{-1} B$

Por tanto

$$S_q^{0\prime} S_q^{-1} S_q^0 - S_p^{0\prime} S_p^{-1} S_p^0 =$$

$$= S_q^{0\prime} \begin{pmatrix} S_p^{-1} + S_p^{-1} B H^{-1} B' S_p^{-1} & -S_p^{-1} B H^{-1} \\ -H^{-1} B' S_p^{-1} & H^{-1} \end{pmatrix} S_q^0 - S_q^{0\prime} \begin{pmatrix} S_p^{-1} & 0 \\ 0 & 0 \end{pmatrix} S_q^0 =$$

$$= S_q^{0\prime} \begin{pmatrix} S_p^{-1} B H^{-1} B' S_p^{-1} & -S_p^{-1} B H^{-1} \\ -H^{-1} B' S_p^{-1} & H^{-1} \end{pmatrix} S_q^0 =$$

$$= S_q^{0\prime} \begin{pmatrix} S_p^{-1} B \\ -I \end{pmatrix} H^{-1} \begin{pmatrix} B' S_p^{-1}, & -I \end{pmatrix} S_q^0$$

Pero esta última expresión es siempre mayor o igual que cero puesto que H^{-1} es siempre definida positiva al serlo la matriz S_p^{-1}.

Por tanto

$$S_p^{0\prime} S_p^{-1} S_p^0 \leq S_q^{0\prime} S_q^{-1} S_q^0$$

y así

$$V\left(\widehat{\overline{Y}}_R^{cond\,\prime}(q)\right) \leq V\left(\widehat{\overline{Y}}_R^{cond\,\prime}(p)\right)$$

como queríamos demostrar.

3.4.4 El estimador de razón condensado de mínima varianza (Estimador RCMV).

En el apartado anterior hemos construido un estimador de razón a partir de una variable auxiliar que llamábamos condensada pues "condensa" la información que proporcionan todas las variables auxiliares disponibles $x_1, x_2, \ldots, x_k$. Esta variable condensada, z, se determina previamente como la función lineal de $x_1, x_2, \ldots, x_k$ que hace máxima la correlación entre la variable principal, y, y la variable condensada.

Sin embargo se pueden utilizar otros criterios para determinar la variable condensada z. En este apartado vamos a determinar la variable condensada como aquella función lineal de las variables auxiliares que haga mínimo el error cuadrático medio del estimador de razón univariante que puede construirse

utilizando esta variable como auxiliar. Es decir, el procedimiento es distinto al antes expuesto pues primero se construye el estimador de razón y posteriormente se determina la variable condensada. Así vamos a obtener un nuevo estimador de razón univariante que utiliza toda la información disponible de la variables auxiliares.

Definición del estimador.

Consideremos la variable aleatoria condensada h definida de la forma:

$$h = \sum_{i=1}^{k} a_i x_i = a'x$$

donde

$a_{(k\times 1)} = (a_1, \ldots, a_k)'$

$x_{(k\times 1)} = (x_1, \ldots, x_k)'$

Por tanto, la media muestral y poblacional de la variable h vienen dadas por:

$$\overline{h} = a'\overline{x} \; ; \quad \overline{H} = a'\overline{X}$$

donde

$\overline{x}_{(k\times 1)} = (\overline{x}_1, \ldots, \overline{x}_k)'$

$\overline{X}_{(k\times 1)} = \left(\overline{X}_1, \ldots, \overline{X}_k\right)'$

A partir de estos valores construimos el estimador de razón:

$$\overline{y}_R^h = \frac{\overline{y}}{\overline{h}}\overline{H}$$

Vamos a determinar el vector a como aquél que haga mínimo el error cuadrático medio del estimador $\overline{y}_R^h$ construido a partir de la variable condensada.

Determinación del vector a.

Consideremos las variables

$$c_1 = \frac{\overline{y} - \overline{Y}}{\overline{Y}} \; ; \quad c_2 = \frac{\overline{h} - \overline{H}}{\overline{H}}$$

En función de estas variables podemos expresar el estimador $\overline{y}_R^h$ en la forma:

$$\overline{y}_R^h = \overline{Y}\,(1 + c_1)\,(1 + c_2)^{-1} = \overline{Y}\,(1 + c_1)\left(1 - c_2 + c_2^2 - \cdots\right)$$

Entonces

$$ECM\left(\overline{y}_R^h\right) = E\left(\overline{y}_R^h - \overline{Y}\right)^2 = \overline{Y}^2 E\left[(1+c_1)\left(1+c_2^{-1}\right) - 1\right]^2$$

Desarrollando $(1+c_2)^{-1}$ en serie de Taylor, elevando al cuadrado y quedándonos sólo con los términos en c_1 y c_2 de orden menor o igual a dos, tenemos:

$$ECM\left(\overline{y}_R^h\right) \simeq \overline{Y}^2 E\left(c_1^2 + c_2^2 - 2c_1c_2\right) =$$

$$= \overline{Y}^2\left[E\left(c_1^2\right) + E\left(c_2^2\right) - 2E\left(c_1c_2\right)\right] = \overline{Y}^2\left[\frac{V\left(\overline{Y}\right)}{\overline{Y}^2} + \frac{V\left(\overline{h}\right)}{\overline{H}^2} - \frac{2\,\mathrm{Cov}\left(\overline{y},\overline{h}\right)}{\overline{H}\,\overline{Y}}\right]$$

Ahora bien, en un muestreo aleatorio simple

$$V\left(\overline{h}\right) = \frac{1-f}{n}V_h = \frac{1-f}{n}a'Sa$$

y

$$\mathrm{Cov}\left(\overline{y},\overline{h}\right) = \frac{1-f}{n}\mathrm{Cov}\left(y,h\right) = \frac{1-f}{n}a'S^0$$

donde

$S_{k\times k} = \left(S_{ij}\right)$ con $S_{ij} = \mathrm{Cov}\left(x_i,x_j\right)$

$S^0_{k\times 1} = \left(S_{0i}\right)$ con $S_{0i} = \mathrm{Cov}\left(y,x_i\right)$.

Entonces, al sustituir estos valores se obtiene:

$$ECM\left(\overline{y}_R^h\right) \simeq \overline{Y}^2\frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} - \frac{2a'S^0}{\overline{Y}a'\overline{X}}\right]$$

Ahora bien:

$$\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} - \frac{2a'S^0}{\overline{Y}a'\overline{X}}\right] =$$

$$= a'\left[\frac{S_y^2\overline{X}\,\overline{X}'}{\overline{Y}^2\left(a'\overline{X}\right)^2} + \frac{S}{\left(a'\overline{X}\right)^2} - \frac{2S^0\overline{X}'}{\overline{Y}a'\overline{X}\,\overline{X}'a}\right]a =$$

$$= \frac{a'\left[\dfrac{S_y^2 \overline{X}\ \overline{X}'}{\overline{Y}^2} + S - \dfrac{2S^0\overline{X}'}{\overline{Y}}\right] a}{\left(a'\overline{X}\right)^2}$$

La matriz

$$B = \frac{S_y^2 \overline{X}\ \overline{X}'}{\overline{Y}^2} + S - \frac{2S^0\overline{X}'}{\overline{Y}}$$

es definida positiva, salvo en el caso que $\overline{y}_R^h$ tomase un único valor, y aplicando la desigualdad extendida de Cauchy–Schwartz, tenemos:

$$\frac{\left(a'\overline{X}\right)^2}{a'Ba} \leq \overline{X}'B^{-1}\overline{X}$$

y por tanto

$$ECM\left(\overline{y}_R^h\right) = \overline{Y}^2 \frac{1-f}{n} \frac{a'Ba}{\left(a'\overline{X}\right)^2} \geq \overline{Y}^2 \frac{1-f}{n} \frac{1}{\overline{X}'B^{-1}\overline{X}}$$

y este mínimo se alcanza para $a = CB^{-1}\overline{X}$ donde C es una constante distinta de cero.

Así, definimos:

$$h = C\overline{X}'B^{-1}x$$

con lo que obtenemos:

$$\overline{h} = C\overline{X}'B^{-1}\overline{x}$$

$$\overline{H} = C\overline{X}'B^{-1}\overline{X}$$

y así elegimos el estimador de razón

$$\overline{y}_R^h = \frac{\overline{y}}{\overline{X}'B^{-1}\overline{x}} \overline{X}'B^{-1}\overline{X}$$

cuyo error cuadrático medio viene dado por la expresión:

$$ECM\left(\overline{y}_R^h\right) = \overline{Y}^2 \frac{1-f}{n} \left[\frac{1}{\overline{X}'B^{-1}\overline{X}}\right]$$

Sesgo.

Proposición 3.4.49 *Una expresión aproximada del sesgo del estimador de razón* $\overline{y}_R^h$ *viene dada por la expresión:*

$$sesgo\left(\overline{y}_R^h\right) \simeq \overline{Y}\frac{1-f}{n}\left[\frac{\overline{X}'B^{-1}SB^{-1}\overline{X}}{\left(\overline{X}'B^{-1}\overline{X}\right)^2} - \frac{\overline{X}'B^{-1}S^0}{\overline{Y}\,\overline{X}'B^{-1}\overline{X}}\right]$$

Demostración.-

$$sesgo\left(\overline{y}_R^h\right) = E\left(\overline{y}_R^h\right) - \overline{Y} = \overline{Y}E\left[(1+c_1)(1+c_2)^{-1} - 1\right] \simeq$$

$$\simeq \overline{Y}E\left[c_1 - c_2 - c_1c_2 + c_2^2\right] = \overline{Y}\left[E\left(c_2^2\right) - E\left(c_1c_2\right)\right] =$$

$$= \overline{Y}\left[\frac{V\left(\overline{h}\right)}{\overline{H}^2} - \frac{\mathrm{Cov}\left(\overline{h},\overline{y}\right)}{\overline{H}\,\overline{Y}}\right] = \overline{Y}\frac{1-f}{n}\left[\frac{V\left(h\right)}{\overline{H}^2} - \frac{\mathrm{Cov}\left(h,y\right)}{\overline{H}\,\overline{Y}}\right] =$$

$$= \overline{Y}\frac{1-f}{n}\left[\frac{\overline{X}'B^{-1}SB^{-1}\overline{X}}{\left(\overline{X}'B^{-1}\overline{X}\right)^2} - \frac{\overline{X}'B^{-1}S^0}{\overline{Y}\,\overline{X}'B^{-1}\overline{X}}\right]$$

Así, el estimador $\overline{y}_R^h$ tiene un sesgo de expresión compleja. En la parte final del capítulo modificaremos el estimador de forma que sea insesgado.

Comparación con el estimador de razón multivariante.

Proposición 3.4.50 *En muestreo aleatorio simple, el estimador de razón condensado,* $\overline{y}_R^h$*, tiene la misma expresión aproximada del error cuadrático medio que el estimador de razón multivariante de Olkin.*

Según vimos en el teorema 3.1.12, el error cuadrático medio del estimador de razón multivariante viene dado por la expresión:

$$ECM\left(\widehat{\overline{Y}}_R^{1,2,\ldots,k}\right) = \frac{1-f}{n}\overline{Y}^2\left(\frac{1}{e'A^{-1}e}\right)$$

donde

$A_{(k\times k)} = \left(a_{ij}\right)$ y $e = (1, 1, \ldots, 1)'$ con

$$a_{ij} = C_y^2 - \rho_j C_y C_{x_j} - \rho_i C_y C_{x_i} + \rho_{ij} C_{x_i} C_{x_j}$$

Por otra parte, podemos escribir el error cuadrático medio del estimador $\overline{y}_R^h$ de la forma:

$$ECM\left(\overline{y}_R^h\right) = \overline{Y}^2 \frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} - \frac{2a'S^0}{\overline{Y}a'\overline{X}}\right] =$$

$$= \overline{Y}^2 \frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(a'\overline{X}\right)^2} - \frac{a'S^0}{\overline{Y}a'\overline{X}} - \frac{S^{0\prime}a}{\overline{Y}a'\overline{X}}\right] =$$

$$= \overline{Y}^2 \frac{1-f}{n} \frac{a'\left[\frac{S_y^2}{\overline{Y}^2}\overline{X}\,\overline{X}' + S - \frac{S^{0\prime}\overline{X}'}{\overline{Y}} - \frac{\overline{X}S^{0\prime}}{\overline{Y}}\right]a}{\left(a'\overline{X}\right)^2}$$

La matriz

$$T = \frac{S_y^2}{\overline{Y}^2}\overline{X}\,\overline{X}' + S - \frac{S^{0\prime}\overline{X}'}{\overline{Y}} - \frac{\overline{X}S^{0\prime}}{\overline{Y}}$$

es definida positiva y podemos aplicar la desigualdad extendida de Cauchy–Schwartz, y obtenemos que el valor mínimo de $ECM\left(\overline{y}_R^h\right)$ adopta la expresión:

$$ECM\left(\overline{y}_R^h\right) = \overline{Y}^2 \frac{1-f}{n} \frac{1}{\overline{X}'T^{-1}\overline{X}}$$

Ahora bien, $T = CAC'$ siendo

$$C_{k\times k} = \begin{pmatrix} \overline{X}_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \overline{X}_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \cdots & \overline{X}_i & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \overline{X}_k \end{pmatrix}$$

Por tanto

$$\overline{X}'T^{-1}\overline{X} = \overline{X}'C^{-1}A^{-1}C^{-1}\overline{X}$$

y al ser $\overline{X}'C^{-1} = (1, \ldots, 1) = e'$, se obtiene

$$ECM\left(\overline{y}_R^h\right) = \overline{Y}^2 \frac{1-f}{n} \frac{1}{e'A^{-1}e}$$

Es decir, el error cuadrático medio del estimador $\overline{y}_R^h$ coincide con el error cuadrático medio del estimador multivariante y así ambos tendrán la misma precisión.

3.4.5 El estimador de razón condensado de mínima varianza e insesgado (Estimador RCMVI).

En el apartado anterior hemos definido un estimador de razón a partir de una variable condensada h que se determina como aquella que minimiza el error cuadrático medio del estimador de razón así construido. Como se observa en 3.4.49 la expresión del sesgo de este estimador es bastante compleja y no permite su comparación con el sesgo del estimador de razón que se obtiene a partir de cualquiera de las variables auxiliares. Esto lleva a plantearse como construir otro estimador de razón que minimice el error cuadrático medio pero que tenga un sesgo más pequeño. El procedimiento que se va a seguir es el mismo que el utilizado con anterioridad para construir el estimador $\widehat{\overline{Y}}_R^{cond\,'}$ que como vimos era insesgado.

Definición del estimador.

Consideremos la variable condensada

$$h' = a_0 + \sum_{i=1}^{k} a_i x_i = a_0 + a'x$$

Vamos a determinar los valores de a_0 y $a_1, \ldots, a_k$ de forma que minimicen el error cuadrático medio del estimador de razón que se puede construir a partir de h' y haga cero la ordenada en el origen de la recta de regresión de y sobre h'.

Comencemos imponiendo esta última restricción. Tenemos:

$$\overline{Y} - \frac{\mathrm{Cov}\,(y, h')}{V\,(h')}\overline{H}' = \overline{Y} - \frac{a'S^0}{a'Sa}\left(a_0 + a'\overline{X}\right)$$

Igualando a cero esta expresión y despejando a_0, obtenemos:

$$a_0 = \frac{\overline{Y}a'Sa}{a'S^0} - a'\overline{X}$$

con lo que la media poblacional y la media muestral de la variable h' valen:

$$\overline{H}' = \overline{Y}\frac{a'Sa}{a'S^0}\ ;\ \ \overline{h}' = \overline{Y}\frac{a'Sa}{a'S^0} + a'\left(\overline{x} - \overline{X}\right)$$

Consideramos ahora el estimador

$$\overline{y}_R^{h\,\prime} = \frac{\overline{y}}{\overline{h}'}\overline{H}'$$

y determinemos el vector a que minimice el error cuadrático medio de este entimador:

$$ECM\left(\overline{y}_R^{h\,\prime}\right) = E\left(\overline{y}_R^{h\,\prime} - \overline{Y}\right)^2 = \overline{Y}^2 E\left[(1+c_1')(1+c_2')^{-1} - 1\right]^2$$

donde

$$c_1' = \frac{\overline{y} - \overline{Y}}{\overline{Y}}\ ;\ \ c_2' = \frac{\overline{h}' - \overline{H}'}{\overline{H}'}$$

Procediendo como en el caso anterior, tenemos:

$$ECM\left(\overline{y}_R^{h\,\prime}\right) \simeq \overline{Y}^2 E\left[c_1'^{\,2} + c_2'^{\,2} - 2c_1'c_2'\right] =$$

$$= \overline{Y}^2\left[E\left(c_1'^{\,2}\right) + E\left(c_2'^{\,2}\right) - 2E\left(c_1'c_2'\right)\right] =$$

$$= \overline{Y}^2\left[\frac{V\left(\overline{y}\right)}{\overline{Y}^2} + \frac{V\left(\overline{h}'\right)}{\overline{H}'^2} - \frac{2\,\mathrm{Cov}\left(\overline{y},\overline{h}'\right)}{\overline{Y}\ \overline{H}'}\right] =$$

$$= \overline{Y}^2\frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\overline{H}'^2} - \frac{2a'S^0}{\overline{Y}\ \overline{H}'}\right] =$$

$$= \overline{Y}^2 \frac{1-f}{n} \left[\frac{S_y^2}{\overline{Y}^2} + \frac{a'Sa}{\left(\frac{\overline{Y} a'Sa}{a'S^0} \right)^2} - \frac{2a'S^0}{\overline{Y}\,\overline{Y} \frac{a'Sa}{a'S^0}} \right] =$$

$$= \frac{1-f}{n} \left[S_y^2 + \frac{\left(a'S^0\right)^2}{a'Sa} - \frac{2\left(a'S^0\right)^2}{a'Sa} \right] = \frac{1-f}{n} \left[S_y^2 - \frac{\left(a'S^0\right)^2}{a'Sa} \right]$$

Por tanto hemos de minimizar en a la expresión:

$$\frac{1-f}{n} \left[S_y^2 - \frac{\left(a'S^0\right)^2}{a'Sa} \right] \tag{3.35}$$

o equivalentemente, maximizar

$$\frac{\left(a'S^0\right)^2}{a'Sa}$$

Si asumimos que S es definida positiva, aplicando la desigualdad extendida de Cauchy–Schwartz, obtenemos:

$$\max_a \frac{\left(a'S^0\right)^2}{a'Sa} = S^{0\prime} S^{-1} S^0$$

y el máximo se alcanza en el punto $a = CS^{-1}S^0$.

Por tanto definimos

$$h' = C\left(\overline{Y} + S^{0\prime} S^{-1} \left(x - \overline{X} \right) \right)$$

y así:

$$\overline{h}' = C\left(\overline{Y} + S^{0\prime} S^{-1} \left(\overline{x} - \overline{X} \right) \right)$$

$$\overline{H}' = C\overline{Y}$$

Por tanto, el estimador de razón se define como:

$$\overline{y}_R^{h\,\prime} = \frac{\overline{y}}{\left(\overline{Y} + S^{0\prime} S^{-1}\left(\overline{x} - \overline{X}\right)\right)} \overline{Y}$$

que coincide con el estimador $\widehat{\overline{Y}}_R^{cond\,\prime}$ y por tanto tiene las mismas propiedades que éste y que ya han sido estudiadas anteriormente.

Obtenemos pues un resultado importante: los estimadores de razón insesgados obtenidos minimizando el error cuadrático medio del estimador y maximizando la correlación entre la variable principal y la condensada, son el mismo.

3.4.6 Distribución asintótica de los estimadores condensados.

Consideremos la población finita U como una sucesión de poblaciones $\{U_\nu\}$ donde n_ν y N_ν tienden a infinito de forma que $N_\nu - n_\nu$ también tienda a infinito cuando ν lo haga.

Teorema 3.4.51 *Si las variables $\{y_\nu\}$ y $\{x_{i\nu}\}$ $(i = 1, \ldots, k)$ satisfacen la condición de Lindeberg–Hájek dada en 1.14, entonces para un muestreo aleatorio simple los estimadores $\widehat{\overline{Y}}_R^{cond}$, $\widehat{\overline{Y}}_R^{cond\,\prime}$, $\overline{y}_R^h$ y $\overline{y}_R^{h\prime}$ son asintóticamente normales.*

Demostración.-

Según el Teorema Central del Límite para poblaciones finitas se tiene que $\overline{y}_\nu$, $\overline{x}_{1\nu}$, $\overline{x}_{2\nu}$, ..., $\overline{x}_{k\nu}$ se distribuyen normalmente cuando $\nu \to \infty$ pues todas cumplen la condición de Lindeberg–Hájek.

Consideremos ahora la función de $\overline{y}_\nu$, $\overline{x}_{1\nu}$, $\overline{x}_{2\nu}$, ..., $\overline{x}_{k\nu}$ dada por

$$H\left(\overline{y}_\nu, \overline{x}_{1\nu}, \overline{x}_{2\nu}, \ldots, \overline{x}_{k\nu}\right) = \frac{\overline{y}_\nu}{\overline{b}_\nu} \overline{B}_\nu$$

siendo

$$\overline{b}_\nu = \textstyle\sum_{i=1}^k a_{i\nu} \overline{x}_{i\nu} \quad \text{ó} \quad \overline{b}_\nu = a_0 + \sum_{i=1}^k a_{i\nu} \overline{x}_{i\nu}$$

y $\overline{B}_\nu$ una constante distinta de cero.

Entonces H es una función continua con derivadas parciales de primer y segundo orden, continuas en un entorno de centro $\left(\overline{Y}_\nu, \overline{X}_{1\nu}, \overline{X}_{2\nu}, \ldots, \overline{X}_{k\nu}\right)$ y

aplicando el resultado obtenido por *Cramer* (1946), $H(\overline{y}_\nu, \overline{x}_{1\nu}, \overline{x}_{2\nu}, \ldots, \overline{x}_{k\nu})$ es asintóticamente normal.

Si consideramos

$$\overline{b} = \sum_{i=1}^{k} a_i x_i$$

obtenemos la normalidad asintótica de los estimadores $\widehat{\overline{Y}}_R^{cond}$ y $\overline{y}_R^h$ y considerando

$$\overline{b} = a_0 + \sum_{i=1}^{k} a_i x_i$$

obtenemos la normalidad asintótica de $\widehat{\overline{Y}}_R^{cond\,\prime}$ y $\overline{y}_R^{h\prime}$.

Resaltar la importancia de que estos dos nuevos estimadores $\widehat{\overline{y}}_R^{h\prime}$ y $\widehat{\overline{Y}}_R^{cond\,\prime}$ son insesgados.

Como vimos en el capítulo 1, la obtención de estimadores de razón insesgados es bastante compleja en el caso de una sola variable auxiliar. En la bibliografía que hemos manejado no se conocían, para más de una variable auxiliar, estimadores insesgados simples. *Olkin* (1958) indica en su trabajo que la combinación lineal de estimadores de *Hartley* y *Ross* proporcionaría un estimador insesgado, pero no estudia el problema por su complejidad.

También resaltar que estos estimadores son siempre como mínimo igual de precisos que el de expansión simple. Por tanto, el método condensado nos ha llevado a estimadores que resuelven dos de los inconvenientes del estimador de *Olkin*.

§3.5 Estimadores iterados de razón.

En este capítulo se ha planteado el problema de qúe hacer en caso de estar disponibles más de una variable auxiliar positivamente correlada con la variable principal. El problema se ha intentado resolver por dos caminos:

1. Construyendo los estimadores de razón a partir de cada variable y haciendo una combinación lineal de ellos.

2. Combinando, según ciertos criterios, las variables auxiliares en una nueva variable y construyendo el estimador de razón univariante considerando esta nueva variable "condensada" como auxiliar.

Sin embargo, cualquiera de estos procedimientos, que como hemos demostrado mejoran los estimadores univariantes, tiene el inconveniente de que los coeficientes de las combinaciones lineales (bien de los estimadores de razón, bien de las variables) para ser óptimos deben tomar ciertos valores determinados que dependen de valores desconocidos. En la práctica, el inconveniente se resuelve tomando los valores muestrales, en vez de los poblacionales, y los estimadores así construidos suelen dar buenos resultados.

Este inconveniente nos lleva a plantearnos como utilizar más de una variable auxiliar para construir un estimador de tipo razón que se pueda calcular siempre y no dependa de ningún parámetro desconocido. Una solución a este problema es un nuevo estimador de razón que proponemos con el nombre de estimador de razón iterado y que procedemos a estudiar.

3.5.1 Definición del estimador.

Sean $x_1, x_2, \ldots, x_k$ las variables auxiliares disponibles.

Definición 3.5.52 *Llamamos estimador de razón iterado de la media a partir de las variables $x_1, x_2, \ldots, x_k$ al estimador*

$$\overline{y}^{it}_{R_{1,2,\ldots,k}} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\cdots\frac{\overline{X}_k}{\overline{x}_k}$$

Análogamente, llamamos estimador de razón iterado del total a partir de las variables $x_1, x_2, \ldots, x_k$ al estimador

$$\widehat{Y}^{it}_{R_{1,2,\ldots,k}} = N\overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\cdots\frac{\overline{X}_k}{\overline{x}_k}$$

donde N es el tamaño de la población.

La motivación de este estimador es la siguiente:

Dada la variable auxiliar x_1, el estimador estimador de razón que se construye a partir de ella, $\overline{y}_{R_1} = \overline{y}\dfrac{\overline{X}_1}{\overline{x}_1}$ es mejor (bajo ciertas condiciones) que el

estimador de expansión simple, $\overline{y}$. Entonces para construir el estimador de razón utilizando la variable auxiliar x_2, podríamos considerar $\overline{y}_{R_1}$ en vez de $\overline{y}$, y por tanto definiríamos

$$\overline{y}^{it}_{R_{1,2}} = \overline{y}_{R_1}\frac{\overline{X}_2}{\overline{x}_2} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}$$

Ahora bien, si este estimador es mejor que $\overline{y}$, podemos repetir el procedimiento sustituyéndolo en la expresión del estimador que se construye con la variable auxiliar x_3, obteniéndose así:

$$\overline{y}^{it}_{R_{1,2,3}} = \overline{y}^{it}_{R_{1,2}}\frac{\overline{X}_3}{\overline{x}_3} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\frac{\overline{X}_3}{\overline{x}_3}$$

y continuando el proceso hasta las k variables auxiliares, obtendríamos el estimador:

$$\overline{y}^{it}_{R_{1,2,\ldots,k}} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\cdots\frac{\overline{X}_k}{\overline{x}_k}$$

3.5.2 Sesgo.

El estimador de razón iterado es sesgado y su sesgo viene dado por la siguiente proposición:

Proposición 3.5.53 *Una aproximación del sesgo del estimador* $\overline{y}^{it}_{R_{1,2,\ldots,k}}$ *viene dada por la expresión:*

$$sesgo\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) = \frac{1-f}{n}\overline{Y}\left[-\sum_{i=1}^{k}C_{yx_i} + \sum_{i\neq j}C_{x_ix_j} + \sum_{i=1}^{k}C^2_{x_i}\right]$$

Demostración.-
Consideramos las variables:

$$a = \frac{\overline{y}-\overline{Y}}{\overline{Y}}\ ;\quad e_i = \frac{\overline{x}_i-\overline{X}_i}{\overline{X}_i}\ i = 1,\ldots,k$$

con las que el estimador de razón iterado adopta la forma:

$$\overline{y}^{it}_{R_{1,2,\ldots,k}} = \overline{Y}(1+a)(1+e_1)^{-1}(1+e_2)^{-1}\cdots(1+e_k)^{-1}$$

Si $|e_i| < 1$, $\forall i$, podemos desarrollar en serie de Taylor cada término $(1+e_i)^{-1}$

$$\overline{y}^{it}_{R_{1,2,\ldots,k}} = \overline{Y}(1+a)\prod_{i=1}^{k}\left(1-e_i+e_i^2+\cdots\right)$$

y conservando sólo los términos de grado inferior o igual a dos en a y e_i $(i=1,\ldots,k)$, obtenemos la aproximación:

$$\overline{y}^{it}_{R_{1,2,\ldots,k}} \simeq \overline{Y}\left(1+a-\sum_{i=1}^{k}e_i-\sum_{i=1}^{k}ae_i+\sum_{i\neq j}^{k}e_ie_j+\sum_{i=1}^{k}e_i^2\right)$$

Entonces

$$sesgo\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) = E\left[\overline{y}^{it}_{R_{1,2,\ldots,k}}-\overline{Y}\right] \simeq \overline{Y}\left[-\sum_{i=1}^{k}E(ae_i)+\right.$$

$$\left.+\sum_{i\neq j}E\left(e_ie_j\right)+\sum_{i=1}^{k}E\left(e_i^2\right)\right] =$$

$$= \overline{Y}\left[-\sum_{i=1}^{k}\frac{\mathrm{Cov}\left(\overline{y},\overline{x}_i\right)}{\overline{Y}\,\overline{X}_i}+\sum_{i\neq j}\frac{\mathrm{Cov}\left(\overline{x}_i,\overline{x}_j\right)}{\overline{X}_i\overline{X}_j}+\sum_{i=1}^{k}\frac{V\left(\overline{x}_i\right)}{\overline{X}_i^2}\right] =$$

$$= \frac{1-f}{n}\overline{Y}\left[-\sum_{i=1}^{k}\frac{S_{yx_i}}{\overline{Y}\,\overline{X}_i}+\sum_{i\neq j}\frac{S_{x_ix_j}}{\overline{X}_i\overline{X}_j}+\sum_{i=1}^{k}\frac{S_{x_i}^2}{\overline{X}_i^2}\right] =$$

$$= \frac{1-f}{n}\overline{Y}\left[-\sum_{i=1}^{k}C_{yx_i}+\sum_{i\neq j}C_{x_ix_j}+\sum_{i=1}^{k}C_{x_i}^2\right]$$

3.5.3 Error cuadrático medio.

A continuación damos una expresión del error cuadrático medio del estimador.

Proposición 3.5.54 *Una expresión aproximada del error cuadrático medio del estimador de razón iterado viene dada por la expresión:*

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) \simeq \overline{Y}^2\frac{1-f}{n}\left[C_y^2+\sum_{i=1}^{k}C_{x_i}^2+\sum_{i\neq j}C_{x_ix_j}-2\sum_{i=1}^{k}C_{yx_i}\right]$$

Demostración.-

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) = E\left[\overline{y}^{it}_{R_{1,2,\ldots,k}} - \overline{Y}\right]^2 \simeq$$

$$\simeq \overline{Y}^2 E\left[a - \sum_{i=1}^{k} e_i - \sum_{i=1}^{k} a e_i + \sum_{i\neq j}^{k} e_i e_j + \sum_{i=1}^{k} e_i^2\right]^2 \simeq$$

$$\simeq \overline{Y}^2 E\left[a^2 + \sum_{i=1}^{k} e_i^2 - 2\sum_{i=1}^{k} a e_i + \sum_{i\neq j}^{k} e_i e_j\right]$$

y reteniendo sólo los términos de orden inferior o igual a dos en a y e_i, obtenemos:

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) \simeq \overline{Y}^2\left[\frac{V(\overline{y})}{\overline{Y}^2} + \sum_{i=1}^{k}\frac{V(\overline{x}_i)}{\overline{X}_i^2} - 2\sum_{i=1}^{k}\frac{\mathrm{Cov}(\overline{y},\overline{x}_i)}{\overline{Y}\,\overline{X}_i} + \right.$$

$$\left. + \sum_{i\neq j}\frac{\mathrm{Cov}\left(\overline{x}_i,\overline{x}_j\right)}{\overline{X}_i\overline{X}_j}\right] =$$

$$= \overline{Y}^2\frac{1-f}{n}\left[\frac{S_y^2}{\overline{Y}^2} + \sum_{i=1}^{k}\frac{S_{x_i}^2}{\overline{X}_i^2} - 2\sum_{i=1}^{k}\frac{S_{yx_i}}{\overline{X}_i\overline{X}_j} + \sum_{i\neq j}\frac{S_{x_ix_j}}{\overline{X}_i\overline{X}_j}\right] =$$

$$= \overline{Y}^2\frac{1-f}{n}\left[C_y^2 + \sum_{i=1}^{k} C_{x_i}^2 + \sum_{i\neq j} C_{x_ix_j} - 2\sum_{i=1}^{k} C_{yx_i}\right]$$

Si llamamos $O_{k+1} = (1,-1,-1,\ldots,-1)'_{k+1}$ y

$$A_{(k+1)\times(k+1)} = \begin{pmatrix} C_y^2 & C_{yx_1} & C_{yx_2} & \cdots & C_{yx_k} \\ C_{yx_1} & C_{x_1}^2 & C_{x_1x_2} & \cdots & C_{x_1x_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{yx_k} & C_{x_kx_1} & C_{x_kx_2} & \cdots & C_{x_k}^2 \end{pmatrix}$$

entonces podemos escribir el error cuadrático medio de la forma

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) \simeq \overline{Y}^2\frac{1-f}{n}O'_{k+1}AO_{k+1}$$

3.5.4 Distribución asintótica del estimador iterado.

Considerando la población finita U como una sucesión de poblaciones $\{U_\nu\}$ donde n_ν y N_ν tienden a infinito de forma que $N_\nu - n_\nu$ también tienda a infinito cuando ν lo haga, obtenemos el siguiente teorema:

Teorema 3.5.55 *Si las variables* $\{y_\nu\}$ *y* $\{x_{i\nu}\}$ $(i = 1, \ldots, k)$ *satisfacen la condición de Lindeberg-Hájek dada en 1.14, entonces para un muestreo aleatorio simple el estimador* $\overline{y}^{it}_{R_{1,2,\ldots,k}}$ *es asintóticamente normal.*

Demostración.-

Es análoga a la dada en la demostración del teorema 3.4.51 considerando la función

$$H\left(\overline{y}_\nu, \overline{x}_{1\nu}, \overline{x}_{2\nu}, \ldots, \overline{x}_{k\nu}\right) = \overline{y}_\nu \frac{\overline{X}_{1\nu}}{\overline{x}_{1\nu}} \frac{\overline{X}_{2\nu}}{\overline{x}_{2\nu}} \cdots \frac{\overline{X}_{k\nu}}{\overline{x}_{k\nu}}$$

3.5.5 Comparación con el caso de una sóla variable auxiliar.

Consideramos $x_1, x_2, \ldots x_k$ variables positivamente relacionadas con la variable principal.

Hemos construido los estimadores:

$$\overline{y}^{it}_{R_1} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}$$

$$\overline{y}^{it}_{R_{1,2}} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}$$

$$\overline{y}^{it}_{R_{1,2,\ldots,k}} = \overline{y}\frac{\overline{X}_1}{\overline{x}_1}\frac{\overline{X}_2}{\overline{x}_2}\cdots\frac{\overline{X}_k}{\overline{x}_k}$$

En este apartado se estudia el problema de determinar si es mejor utilizar todas las variables auxiliares o habrá algún paso en el que no convenga hacer una nueva iteración. El problema es importante pues es claro que el estimador de razón no siempre mejora el estimador de expansión simple, y por tanto el pasar de $\overline{y}^{it}_{R_{1,2,\ldots,k-1}}$ a $\overline{y}^{it}_{R_{1,2,\ldots,k}}$ no tiene por que producir estimadores con más precisión. El teorema siguiente da las condiciones para las cuales el estimador iterado con k variables es mejor que el que utiliza sólo $k-1$ variables.

Teorema 3.5.56 *Sean* $x_1, x_2, \ldots x_k$ *variables auxiliares y sean* $\overline{y}^{it}_{R_{1,2,\ldots,k-1}}$ *e* $\overline{y}^{it}_{R_{1,2,\ldots,k}}$ *los estimadores de razón iterados construidos a partir de las variables*

x_1, x_2, ... x_{k-1} y x_1, x_2, ... x_k, respectivamente. Entonces el estimador $\overline{y}^{it}_{R_{1,2,\ldots,k}}$ es más preciso que el estimador $\overline{y}^{it}_{R_{1,2,\ldots,k-1}}$ si se verifica:

$$2\sum_{i=1}^{k-1} C_{x_i x_k} + C^2_{x_k} < 2C_{yx_k}$$

Demostración.-
Como ya hemos visto

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) \simeq \overline{Y}^2 \frac{1-f}{n} O'_{k+1} A O_{k+1}$$

Si llamamos $O_k = (1, -1, -1, \ldots, -1)'_k$ y

$$B_{(k\times k)} = \begin{pmatrix} C^2_y & C_{yx_1} & C_{yx_2} & \cdots & C_{yx_{k-1}} \\ C_{yx_1} & C^2_{x_1} & C_{x_1x_2} & \cdots & C_{x_1x_{k-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{yx_{k-1}} & C_{x_{k-1}x_1} & C_{x_{k-1}x_2} & \cdots & C^2_{x_{k-1}} \end{pmatrix}$$

obtendremos

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k-1}}\right) \simeq \overline{Y}^2 \frac{1-f}{n} O'_k B O_k$$

Será útil realizar una nueva iteración si

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) < ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k-1}}\right)$$

si y sólo si

$$O'_{k+1} A O_{k+1} - O'_k B O_k \leq 0$$

Pero escribiendo

$$A = \begin{pmatrix} B & C \\ C' & D \end{pmatrix}$$

obtenemos:

$$O'_{k+1} A O_{k+1} - O'_k B O_k = O'_{k+1} A O_{k+1} - O'_{k+1} \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} O_{k+1} =$$

$$= O'_{k+1}\begin{pmatrix} 0 & C \\ C' & D \end{pmatrix} O_{k+1} = -C'O_k - O'_k C + D$$

Ahora bien

$$C'O_k = C_{yx_k} - \sum_{i=1}^{k-1} C_{x_i x_k}$$

y así

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k}}\right) < ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,k-1}}\right)$$

si y sólo si

$$2\left(-C_{yx_k} + \sum_{i=1}^{k} C_{x_i x_k}\right) + C^2_{x_k} < 0$$

si y sólo si

$$2\sum_{i=1}^{k-1} C_{x_i x_k} + C^2_{x_k} < 2C_{yx_k}$$

3.5.6 Elección de la mejor variable auxiliar para la iteración.

Otro problema importante a considerar es qué variable auxiliar elegir entre las disponibles para realizar una nueva iteración. Es decir, si $\overline{y}^{it}_{R_{1,2,\ldots,i}} = \overline{y}\prod_{j=1}^{i}\frac{\overline{X}_j}{\overline{x}_j}$, ¿cuál de las $k-i$ variables que no han sido utilizadas será la mejor para construir el nuevo estimador?

Proposición 3.5.57 *Será más conveniente el uso de la variable x_p que de la variable x_q si*

$$C^2_{x_q} + 2\sum_{j=1}^{i} C_{x_j x_q} - 2C_{yx_q} > C^2_{x_p} + 2\sum_{j=1}^{i} C_{x_j x_p} - 2C_{yx_p}$$

Demostración.-
Sean

$$\overline{y}^{it}_{R_{1,2,\ldots,i,p}} = \overline{y}\left(\prod_{j=1}^{i}\frac{\overline{X}_j}{\overline{x}_j}\right)\frac{\overline{X}_p}{\overline{x}_p}$$

$$\overline{y}^{it}_{R_{1,2,\ldots,i,q}} = \overline{y}\left(\prod_{j=1}^{i}\frac{\overline{X}_j}{\overline{x}_j}\right)\frac{\overline{X}_q}{\overline{x}_q}$$

Entonces, según la proposición 3.5.54:

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,i,p}}\right) = \overline{Y}^2\frac{1-f}{n}\left[C_y^2 + \sum_{j=1}^{i}C_{x_j}^2 + C_{x_p}^2 + \right.$$

$$\left. + \sum_{j\neq l}^{i}C_{x_jx_l} - 2\sum_{j=1}^{i}C_{yx_j} + 2\sum_{j=1}^{i}C_{x_jx_p} - 2C_{yx_p}\right]$$

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,i,q}}\right) = \overline{Y}^2\frac{1-f}{n}\left[C_y^2 + \sum_{j=1}^{i}C_{x_j}^2 + C_{x_q}^2 + \right.$$

$$\left. + \sum_{j\neq l}^{i}C_{x_jx_l} - 2\sum_{j=1}^{i}C_{yx_j} + 2\sum_{j=1}^{i}C_{x_jx_q} - 2C_{yx_q}\right]$$

Entonces será mejor utilizar x_p si

$$ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,i,q}}\right) - ECM\left(\overline{y}^{it}_{R_{1,2,\ldots,i,p}}\right) > 0$$

$$C_{x_q}^2 - C_{x_p}^2 + 2\sum_{j=1}^{i}\left(C_{x_jx_q} - C_{x_jx_p}\right) - 2\left(C_{yx_q} - C_{yx_p}\right) > 0$$

$$C_{x_q}^2 + 2\sum_{j=1}^{i}C_{x_jx_q} - 2C_{yx_q} > C_{x_p}^2 + 2\sum_{j=1}^{i}C_{x_jx_p} - 2C_{yx_p}$$

Por tanto serían tanto más deseables las variables x_p en cuanto que

1. Tengan un menor coeficiente de variación.
2. La covarianza con el resto de las variables auxiliares sea más pequeña.
3. La covarianza con la variable principal sea más grande.

3.5.7 Método Forward.

En vista de los resultados anteriores parece lógico que para plantear el método iterado de razón de forma óptima en el sentido de menor error cuadrático medio, hemos de considerar dos cuestiones:

1. En qué orden seleccionar las variables
2. Hasta cuando continuar el proceso

Esto nos ha llevado al siguiente método forward:

Para conseguir el estimador iterado de forma óptima en el sentido de menor error cuadrático medio, debemos proceder de la forma siguiente:

1. Antes de realizar una nueva iteración, elegir la variable x_p tal que la función

$$F\left(x_p\right) = C_{x_p}^2 + 2\sum_{j=1}^{i} C_{x_j x_p} - 2C_{y x_p}$$

sea menor (siendo x_j, $j = 1, \ldots, i$, las variables utilizadas en las iteraciones anteriores)

2. Una nueva iteración será conveniente si

$$F\left(x_p\right) < 0$$

Esta función $F\left(x_p\right)$ depende de los coeficientes de variación que en general son desconocidos. No obstante se pueden estimar por sus valores muestrales.

Este método tiene una serie de ventajas, como son:

1. El estimador iterado no depende de ningún parámetro desconocido y por tanto se puede calcular siempre.
2. El procedimiento iterativo para construir el estimador y evaluar su error cuadrático medio es computacionalmente sencillo.
3. Es siempre al menos tan preciso como el de expansión simple.
4. No utiliza todas las variables disponibles sino sólo las que producen una mejora en la precisión de la estimación.

Apéndice A

Ejemplos numéricos.

Se incluyen aquí los resultados de la aplicación de los métodos de razón multivariantes definidos en el capítulo 3, para cuatro ejemplos clásicos de la literatura de muestreo en poblaciones finitas.

En cada ejemplo se especifican las variables principal y auxiliares consideradas y se da una tabla con el error cuadrático medio de cada estimador, su eficiencia respecto al de expansión simple y si tiene o no sesgo.

1. *Olkin* (1958)

   Se quiere estimar el número de habitantes de las 200 ciudades más grandes de Estados Unidos a partir de una muestra de tamaño 50.

   y = Número de habitantes en 1950
   x_1 = Número de habitantes en 1940
   x_2 = Número de habitantes en 1930

Estimador		ECM	Eficiencia	Sesgo
Expansión simple	$N\overline{y}$	$N^2\frac{1-f}{n}\cdot 40260.4$	1	NO
Multivariante	$\widehat{Y}_R^{1,2}$	$N^2\frac{1-f}{n}\cdot 918.8$	43.8	SI
RCMVI	$\widehat{Y}_R^{h\ \prime}$	$N^2\frac{1-f}{n}\cdot 842.9$	47.7	NO
Iterado	$\widehat{Y}_{R_{1,2}}^{it}$	$N^2\frac{1-f}{n}\cdot 2303$	17.4	SI

2. *Sukhatme* (1984)

Se quiere estimar la cosecha de guayaba en el distrito de Allahabad (India) a partir de una muestra de 27 pueblos elegida mediante muestreo aleatorio simple de entre 153 pueblos.

y = producción total de guayaba
x_1 = número de árboles de guayaba plantados
x_2 = superficie cultivada de guayaba

Estimador		ECM	Eficiencia	Sesgo
Expansión simple	$N\overline{y}$	$N^2\frac{1-f}{n}\cdot 1268$	1	NO
Multivariante	$\widehat{Y}_R^{1,2}$	$N^2\frac{1-f}{n}\cdot 540$	2.35	SI
RCMVI	$\widehat{Y}_R^{h\,\prime}$	$N^2\frac{1-f}{n}\cdot 302.4$	4.19	NO
Iterado	$\widehat{Y}_{R_{1,2}}^{it}$	$N^2\frac{1-f}{n}\cdot 403.8$	3.1	SI

3. *Srivastava* (1989)

Los datos corresponden a una muestra de 25 niños de Varanasi (India).

y = perímetro del antebrazo
x_1 = peso
x_2 = perímetro craneal

Estimador		ECM	Eficiencia	Sesgo
Expansión simple	$N\overline{y}$	$N^2\frac{1-f}{n}\cdot 0.7363$	1	NO
Multivariante	$\widehat{Y}_R^{1,2}$	$N^2\frac{1-f}{n}\cdot 0.3476$	2.11	SI
RCMVI	$\widehat{Y}_R^{h\,\prime}$	$N^2\frac{1-f}{n}\cdot 0.1916$	3.48	NO
Iterado	$\widehat{Y}_{R_{1,2}}^{it}$	$N^2\frac{1-f}{n}\cdot 0.3528$	2.08	SI

4. *Singh* (1983)

Los datos provienen del Instituto Indio de Investigación Estadística para la Agricultura y corresponden a

y = peso por parcela de azúcar de caña
x_1 = altura de las cañas
x_2 = número de cañas por parcela
x_3 = diámetro de las cañas

Estimador		ECM	Eficiencia	Sesgo
Expansión simple	$N\overline{y}$	$N^2\frac{1-f}{n}\cdot 51.744$	1	NO
Multivariante	$\widehat{Y}_R^{1,2,3}$	$N^2\frac{1-f}{n}\cdot 25.278$	2.046	SI
RCMVI	$\widehat{Y}_R^{h\ \prime}$	$N^2\frac{1-f}{n}\cdot 24.924$	2.076	NO
Iterado	$\widehat{Y}_{R_{1,2,3}}^{it}$	$N^2\frac{1-f}{n}\cdot 34.413$	1.503	SI

Bibliografía

[1] Ayachit, G. R. (1953), *Some aspects of large–scale sample surveys with particular reference to the ratio method of estimation*, M. Sc. Thesis, Bombay University, Bombay.

[2] Azorín, F. y Sánchez–Crespo, J. L. (1986), *Métodos y Aplicaciones del Muestreo*, Alianza Universidad Textos. Madrid.

[3] Barnet, V. (1982), *Elements of Sampling Theory*, Hodder and Stoughton, London.

[4] Beale, E. (1962), *Some uses of computers in operations research*, Industrielle Organization, 31.

[5] Bose, C. (1943) *Note of the sampling error in the method of double sampling*, Sankhya, 6.

[6] Cassel, Sarndall & Wretman (1977), *Foundations of Inference in Survey Sampling*, Wiley.

[7] Chaubey (1991), *A study of ratio and product estimators under a super population model*, Commun. Statist. Theory and Meth., 20(5 & 6).

[8] Chaubey & Singh (1984), *An efficiency comparison of produt and ratio estimator*, Commun. Statist. Theory and Meth., 13(6).

[9] Chaudhuri & Adhikari (1989), *On efficiency of the ratio estimator*, Metrika, 36.

[10] Cochran, W. G. (1946) *Relative accuracy of sistematic and stratified random samples for a certain class of populations*, Ann. Math. Statist., 17.

[11] Cochran, W. G. (1963), *Sampling Techniques*, Wiley & Sons, New York.

[12] Cochran, W. G. (1978), *Laplace' s ratio estimator.* In: H. A. David, ed., *Contributions to Survey Sampling and Applied Statistics*, Academic Press, New York.

[13] Cramer, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.

[14] David & Sukhatme (1973), *A note on Koop' s approach for finding the bias of the ratio estimate*, JASA, Volume 68, Number 342.

[15] David & Sukhatme (1974), *On the bias and mean square error of the ratio estimator*, JASA, Volume 69, Number 346.

[16] Deming (1950), *Some Theory of Sampling*, Wiley, New York.

[17] Feiveson (1984) *Weighted ratio estimation in agricultural surveys*, Commun. Statist. Theory and Meth, 13(23).

[18] Fieller (1932), *The distribution of the index in a normal bivariate population*, Biometrika, 24.

[19] Goodman, L. A. and Hartley, H. O. (1958), *The precision of unbiased ratio–type estimators*, JASA, 53.

[20] Grosbras (1987), *Methodes Statistiques des Sondages*, Económica, Paris.

[21] Hájek, K. J. (1960), *Limiting distributions in simple random sampling from a finite population*, Pub. Math. Inst. Hungarian Acad. Sci., 5.

[22] Hansen & Hurwitz (1943), *On the theory of sampling from finite populations*, Ann. Math. Statist., 14.

[23] Hansen, Hurwitz & Gurney (1946), *Problems and methods of the sample surveys of business*, JASA, 41.

[24] Hansen, Hurwitz & Madow (1953), *Sample Survey Methods and Theory*, Vol 1, Methods and Applications, New York and London, Wiley Publications.

[25] Hansen, Hurwitz & Madow (1953), *Sample Survey Methods and Theory*, Vol 2, Theory, New York and London, Wiley Publications.

[26] Hartley & Ross (1954), *Unbiased ratio estimates*, Nature, 174.

[27] Hwang & Tsay (1988), *On the asymtotic distribution of the ratio and regression estimators*, Sankhya, Series B, Volume 50.

[28] Kendall & Stuart (1958), *The Advanced Theory of Statistics*, Volume 1, Griffin, London.

[29] Kendall & Stuart (1958), *The Advanced Theory of Statistics*, Volume 2, Griffin, London.

[30] Kish, L. (1965), *Survey Sampling*, John Wiley ans Sons, New York.

[31] Kish, Namboodiri & Pillai, (1962), *The ratio bias in surveys*, JASA, 57.

[32] Krishnaiah & Rao (1988), *Handbook of Statistics*, Vol 6, North–Holland.

[33] Lahiri (1951), *A method of sample selection providing unbiased ratio estimates*, Bulletin of the international Statistical Institute, 33.

[34] Menéndez y Ferrales (1989), *El estimador de razón generalizado*, Trabajos de Estadística, Vol 4, Number 1.

[35] Mickey (1959), *Some finite population unbiased ratio and regression estimators*, JASA, 54.

[36] Midzuno (1952), *On the sampling system with probability proportional to the sum of the sizes*, Annals of the Institute of Statistical Mathematics, 3.

[37] Naik & Gupta (1991), *A general class of estimators for estimating population mean using auxiliary information*, Metrika, 38.

[38] Nieto de Pascual (1961), *Unbiased ratio estimators in stratified sampling*, JASA, 56.

[39] Olkin (1958), *Multivariate ratio estimation for finite population*, Biometrika.

[40] Pandey & Dubey (1989), *On almost unbiased ratio estimators*, Metron 47, no. 1–4.

[41] Prasad (1986), *Some unbiased estimators versus mean per unit and ratio estimators in finite population sample surveys*, Commun. Statist. Theory and Meth., 15(12).

[42] Prasad (1989), *Some improved ratio type estimators of population mean and ratio in finite population sample surveys*, Commun. Statist. Theory and Meth., 18(1).

[43] Quenouille, M. H. (1965), *Notes on bias in estimation*, Biometrika, 43.

[44] Rao, J. N. K. (1964), *Unbiased ratio and regression estimators in multi–stage sampling*, Journal of the Indian Society of Agricultural Statistics, 14.

[45] Rao, PSRS (1969), *Comparison of four ratio–type estimates under a model*, JASA, 64.

[46] Rao, PSRS (1975), *Hartley–Ross type estimator in finite populations*, JASA, 70.

[47] Rao, PSRS (1975), *On the two–phase ratio estimator in finite population*, JASA, 70.

[48] Rao, PSRS (1981), *Efficiencies of the nine two–phase ratio estimators for the mean*, JASA, 76.

[49] Rao, T. J. (1967), *Contributions to the theory of sampling strategies*, Ph. D. thesis, ISI., Calcuta.

[50] Rao, T. J. (1991), *On certain methods of improving ratio and regression estimators*, Commun. Statist. Theory and Meth., 20(10).

[51] Rao & Mudholkar (1967), *Generalized multivariate estimator for the mean of finite populations*, JASA, 62.

[52] Rao & Pereira (1967), *On double ratio estimators*, Sankhya, Series A.

[53] Rao & Ramachandran (1974), *Comparison of the separate and combined ratio estimates*, Sankhya, Series C, 36.

[54] Rao & Webster (1966), *On two methods of bias reduction in the estimation of ratios*, Biometrika, 53, 3.

[55] Raj, Des. (1972), *The Design of Sample Surveys*, MacGraw–Hill, New York.

[56] Raj, Des. (1968), *Sampling Theory*, New York, MacGraw–Hill.

[57] Ray, S. K. and Sahai, A. (1980), *Efficient families of ratio and product–type estimators*, Biometrika, 67.

[58] Ray & Singh, (1981) *A product type estimator in double sampling*, Biom. Jour., 23, 7.

[59] Robson, D. S. (1957), *Application of multivariante polykeys to the theory of unbiased ratio–type estimators*, JASA, 52.

[60] Royall & Cumberland (1981), *An Empirical Study of the Ratio Estimator ad estimator of its variance*, JASA, Volume 76, number 373.

[61] Ruiz, M. (1991), *Comparación de estimadores óptimos de razón, producto y regresión*, Trabajos de Estadística, Vol. 6, número 1.

[62] Ruiz & Santos (1989), *Unbiased mean of the ratio estimators*, Statistica, 49, number 4.

[63] Sahoo & Mishra (1989), *Classes of transformed ratio and product estimators in two stage sampling*, Statistica, 49, number 2.

[64] Scheaffer, Mendenhall & Ott (1987), *Elementos de Muestreo*, Grupo Editorial Iberoamérica, México.

[65] Scott & Wu (1981), *On the asymtotic distribution of ratio and regression estimators*, JASA, 76.

[66] Singh (1965), *On the estimation of ratio and product of the population parameters*, Sankhya, Series B.

[67] Singh (1966), *Efficient use of systematic sampling in ratio and product estimation*, Metrika, 10.

[68] Singh, Kumar & Chandak (1983), *Use of multi-auxiliary variables as a condensed auxiliary variable in selecting a sample*, Commun. Statist. Theory and Meth., 12(14).

[69] Som (1989), *A Manual of Sampling Techniques*, Heinemann, London, 1973.

[70] Srivastava (1967), *An estimator using auxiliary information in sample surveys*, Cal. Stat. Assoc. Bull., 16.

[71] Srivastava, Srivastava & Khare (1989), *Chain ratio type estimator for ratio of two population means using auxiliary characters*, Commun. Statist. Theory and Meth. 18(10).

[72] Srivenkataramana (1980), *A dual to ratio estimator in sample surveys*, Biometrika, 67, 1.

[73] Sukhatme & Sukhatme (1984), *Sampling Theory of Surveys Applications*, Iowa State.

[74] Swain (1964), *The use of systematic sampling in ratio estimate*, Journal of the Indian Statistical Association, 2.

[75] Tankou & Dharmadhikari (1989), *Improvement of ratio-type estimators*, Biometrika, 5.

[76] Williams (1961), *Generating unbiased ratio and regression estimators*, Biometrics, 17.

[77] Williams (1962), *On two methods of unbiased estimation with auxiliary variates*, JASA, 57.

[78] Yates (1971), *Sampling Methods for Censuses & Surveys*, Griffin, London.

C/M

UNIVERSIDAD DE GRANADA
FACULTAD DE CIENCIAS

Núm. 5021

Tengo el gusto de remitirle 1 ejemplar de la Tesis Doctoral de D.ª MARIA DEL MAR RUEDA GARCIA, para su archivo en la Biblioteca de esta Facultad.

Granada, 14 de Junio de 1993

EL SECRETARIO,

Fdo.: Gabriel Cardenete Hernández

Sr. Director de la Biblioteca de esta Facultad.