

|                                                              |            |
|--------------------------------------------------------------|------------|
| <b>CAPITULO 8. RECOMPOSICIÓN DE LA MUESTRA.....</b>          | <b>275</b> |
| <b>8.1 RECOMPOSICIÓN SOBRE VARIABLES CUANTITATIVAS .....</b> | <b>275</b> |
| 8.1.1. ESTIMADOR DE RAZÓN .....                              | 276        |
| <i>Recomposición de la media.....</i>                        | <i>276</i> |
| <i>Recomposición del total .....</i>                         | <i>277</i> |
| <i>Recomposición por el cociente .....</i>                   | <i>277</i> |
| <i>Varianza aproximada de la estimación de razón .....</i>   | <i>282</i> |
| <i>Sesgo en la estimación de razón .....</i>                 | <i>285</i> |
| <i>Estimador de razón en el muestreo estratificado .....</i> | <i>286</i> |
| 8.1.2. ESTIMADOR POR MÍNIMOS CUADRADOS .....                 | 287        |
| 8.1.3 VARIAS VARIABLES REALES .....                          | 288        |
| <b>8.2 RECOMPOSICIÓN SOBRE VARIABLES CATEGÓRICAS .....</b>   | <b>289</b> |
| 8.2.1 ESTIMADOR DE POST ESTRATIFICACIÓN .....                | 289        |
| 8.2.2 VARIAS VARIABLES CATEGÓRICAS: MÉTODO RAS.....          | 290        |
| <b>PROBLEMAS Y CASOS DE ESTUDIO .....</b>                    | <b>293</b> |
| CASO 8.1 POBLACIÓN EN LA PROVINCIA DE CÓRDOBA .....          | 293        |
| <b>BIBLIOGRAFÍA.....</b>                                     | <b>293</b> |



# CAPITULO 8. RECOMPOSICIÓN DE LA MUESTRA

En este capítulo se estudia uno de los problemas habituales del muestreo: la recomposición de la muestra. En una investigación econométrica basada en fuentes de información primaria, luego de realizar el muestreo es importante ajustar la estimación inicial de los parámetros. Para ello, generalmente, se dispone de información adicional que se puede utilizar para llevar a cabo dicho ajuste. Esto deviene en una mejor estimación de los parámetros poblacionales, ya sea que estos representen medidas obtenidas sobre variables cualitativas o cuantitativas de las unidades de observación. En el proceso de investigación econométrica, se realizó el muestreo aleatorio y se calculó el estimador usual  $\bar{y}$  de la media  $\mu_y$  de la variable de interés  $Y$ . Pero, se dispone sobre la población de una información auxiliar sobre una media  $\mu_x$  de una variable cuantitativa  $X$  o sobre la forma de la distribución de la población según una variable categórica y no se vuelve a encontrar sobre la muestra la misma media o la misma distribución. Se proponen entonces nuevos estimadores de  $\mu_y$  con mejores propiedades que las del estimador usual. Para la mayoría de ellos se trata de ponderar, otra vez, las unidades de observación de la muestra. Esta es la razón por la cual esos métodos se llaman métodos de recomposición.

## 8.1 Recomposición sobre variables cuantitativas

En esta etapa del proceso de investigación econométrica, si se ha trabajado con fuentes primarias de información, es importante llevar a cabo la recomposición de la muestra. Esto obliga a tener, desde el momento en que se planifica el trabajo, referencias sobre variables relacionadas.

Existen diferentes métodos para llevar a cabo el proceso de recomposición pero en la mayoría de ellos se trata de ponderar otra vez a los individuos de la muestra. En muestreo es frecuente usar información adicional para mejorar la precisión de las estimaciones, a estos se los conoce como métodos indirectos y los más comunes son los de razón (o cociente) y los de mínimos cuadrados. Estos métodos requieren la observación de una variable auxiliar ( $X$ ) que se encuentra correlacionada con la variable de estudio ( $Y$ ) y además conocer el total de  $X$ . La correlación entre  $X$  e  $Y$  es lo que permite el incremento de precisión al estimar los parámetros de  $Y$ .

### 8.1.1. Estimador de Razón

La realización de un muestreo aleatorio simple arroja información sobre  $\bar{y}$ , el estimador de la media  $\mu_y$  de la variable de interés Y. Adicionalmente, se dispone de una *variable* auxiliar que se puede considerar proporcional a Y, la variable X. Las dos variables, X e Y, están relacionadas para cada unidad de la muestra; se conoce con precisión la media  $\mu_x$  en la población y el valor que tienen en la variable X los integrantes de la muestra. A este último concepto se lo identifica con  $x$ ; mientras que, a los valores de la variable de interés en los integrantes de la muestra,  $y$ .

Droesbeke y Fine (1997) establecen que si se conociera, mediante la base del muestreo, el valor de X sobre todos los individuos de la población, se podría realizar un muestreo con *probabilidades desiguales* proporcionales a X, o también, un *muestreo estratificado* sobre la variable X *discretizada*; como sólo se conoce la media de X sobre la población, solo se puede utilizar un método de recomposición denominado de la razón.

La recomposición se puede hacer para tres indicadores: la media, el total y el cociente.

#### Recomposición de la media

El estimador de razón para la media ( $\bar{y}_R$ ), se define por:

$$\bar{y}_R = \bar{y} \frac{\mu_x}{\bar{x}} = \frac{\bar{y}}{\bar{x}} \mu_x \quad (8.1)$$

Este estimador tiene un ligero sesgo pero, si se verifica la condición de proporcionalidad,  $\bar{y}_R$  estima  $\mu_y$  de manera más precisa que  $\bar{y}$ . Para comprobar que la condición de proporcionalidad está vigente, sobre los datos muestrales se calculan los coeficientes de la recta de regresión de  $y$  sobre  $x$

$$y = ax + b \quad (8.2)$$

Donde

$$a = \frac{s_{xy}}{s_x} = \frac{\text{covarianza } xy}{\text{varianza } x} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$b = \bar{y} - a\bar{x}$$

Si se observa que  $b \leq \frac{\bar{y}}{2}$ , entonces  $\bar{y}_R$  es mejor que  $\bar{y}$ .

Si la cantidad que se quiere estimar es el valor de la media de la población  $\bar{Y}$  para una variable Y, la estimación de razón es

$$\bar{Y}_R = \frac{y}{x} \bar{X} \quad (8.3)$$

### Recomposición del total

La estimación por la razón para el total de la población ( $\hat{Y}_R$ ) -es decir, para  $\sum Y_i \quad \forall i = 1, \dots, N$ -, se obtiene haciendo el siguiente cálculo:

$$\hat{Y}_R = \frac{y}{x} X = \frac{\bar{y}}{\bar{x}} X \quad (8.4)$$

Donde

- $x$  e  $y$  son los totales en la muestra;  $y = \sum y_i \quad \forall i = 1, \dots, n$  y  $x = \sum x_i \quad \forall i = 1, \dots, n$
- $X$  es el total en la población para la variable auxiliar;  $X = \sum X_i \quad \forall i = 1, \dots, N$
- $\bar{x}$  y  $\bar{y}$  son las medias de la variable de interés y de la variable auxiliar en la muestra, es decir,  $\bar{x} = \sum x_i/n \quad \forall i = 1, \dots, n$   $\bar{y} = \sum y_i/n \quad \forall i = 1, \dots, n$

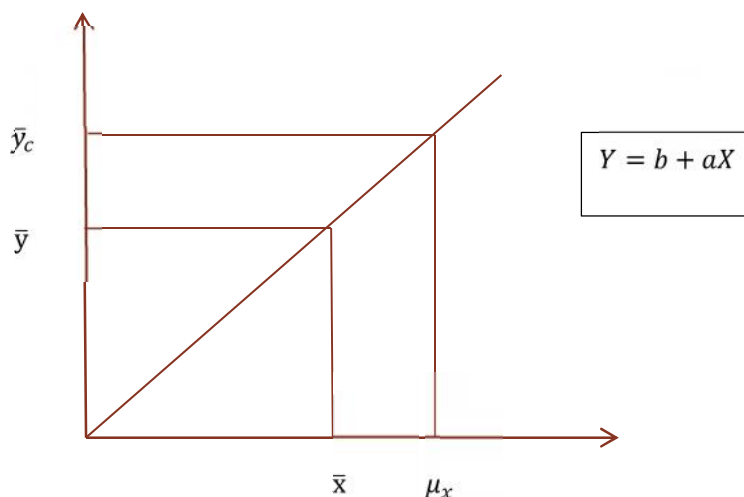
Si  $x_i$  es el valor de  $y_i$  en un momento anterior, el método de la razón utiliza la muestra para estimar el cambio relativo  $Y/X$  ocurrido desde el momento anterior. El cambio relativo estimado  $y/x$  se multiplica por el total de población ( $X$ ) conocido anteriormente, para proporcionar una estimación del total de la población actual ( $Y$ ). Si la razón  $y_i/x_i$  es casi la misma en todas las unidades del muestreo, los valores de  $y/x$  varían poco entre las muestras y la estimación de razón se vuelve altamente precisa.

### Recomposición por el cociente

Si se quiere estimar una razón y no un total o una media -por ejemplo, la razón de hectáreas de soja respecto a las hectáreas de trigo, la razón del consumo debido al ingreso o la razón de activos corrientes a la de activos totales-; la estimación muestral de la razón ( $\hat{R}$ ) es:

$$\hat{R} = \frac{y}{x} \quad (8.5)$$

En este caso no es necesario conocer el total de la variable auxiliar  $X$ . Para Scheaffer et al (2007) el estimador de la razón es más apropiado cuando la relación entre  $x$  e  $y$  es lineal a través del origen, como en la figura 8.1.



**Figura 8.1: Recta de regresión de Y sobre X**

Ejemplo. Considérese una población de explotaciones agropecuarias,  $N$ ; sea  $Y$  la variable de interés que representa la producción de soja en quintales y  $X$  el tamaño de las explotaciones en has. Sobre una muestra de explotaciones, se observa una producción media de soja igual a 8000 quintales. Cuando se mide el tamaño de las explotaciones, se observa en la muestra una media  $\bar{x}$  de 600 hectáreas; pero se sabe en realidad que el tamaño medio  $\mu_x$  de las explotaciones agrícolas es de 500 hectáreas.

De esta forma se piensa que  $8000 \frac{500}{600} = 6666,67$  es una estimación de la producción media de soja en la población mejor que 8000.

Ejemplo. El objetivo es estimar el total de Población del Departamento Río Cuarto a partir de una muestra de localidades.

Se trabaja con una situación hipotética donde no se conoce el resultado del Censo de Población de 2001 sino el resultado de una muestra de 10 localidades. La información disponible da cuenta de la cantidad de población en el año 2001 y en el año 1991, para cada uno de los asentamientos urbanos integrantes de la muestra. Además, se conoce el total de población existente en el Departamento Río Cuarto en el año 1991 y en el resto de asentamientos urbanos que no integran la muestra.

En la Figura 8.2 se tiene la tabla de datos con la totalidad de localidades del Departamento Río Cuarto. La variable  $X$  representa la cantidad de población en cada localidad según el Censo de Población del año 1991.

Las localidades 1, 11,12, 13, 18, 23, 25, 27, 28 y 32 integran la muestra que fue seleccionada del conjunto de asentamientos urbanos; a partir de esta muestra se debe calcular el total de población del Departamento para el año 2001.

| Observación | Localidad                      | Departamento | Población X | Muestra |         |
|-------------|--------------------------------|--------------|-------------|---------|---------|
|             |                                |              |             | y       | x       |
| 1           | Río Cuarto                     | Río Cuarto   | 134.355     | 144.021 | 134.355 |
| 2           | Las Higueras                   | Río Cuarto   | 4.498       |         |         |
| 3           | Vicuña Mackenna                | Río Cuarto   | 7.181       |         |         |
| 4           | Coronel Moldes                 | Río Cuarto   | 7.695       |         |         |
| 5           | Sampacho                       | Río Cuarto   | 7.160       |         |         |
| 6           | Berrotarán                     | Río Cuarto   | 6.187       |         |         |
| 7           | Adelia María                   | Río Cuarto   | 5.294       |         |         |
| 8           | Alcira (Est. Gigena)           | Río Cuarto   | 4.948       |         |         |
| 9           | Santa Catalina (Est. Holmberg) | Río Cuarto   | 3.035       |         |         |
| 10          | San Basilio                    | Río Cuarto   | 2.510       |         |         |
| 11          | Elena                          | Río Cuarto   | 2.670       | 2.815   | 2.670   |
| 12          | Achiras                        | Río Cuarto   | 2.123       | 2.173   | 2.123   |
| 13          | Las Acequias                   | Río Cuarto   | 1.984       | 2.116   | 1.984   |
| 14          | Coronel Baigorria              | Río Cuarto   | 1.147       |         |         |
| 15          | Bulnes                         | Río Cuarto   | 802         |         |         |
| 16          | Las Vertientes                 | Río Cuarto   | 671         |         |         |
| 17          | Alpa Corral                    | Río Cuarto   | 344         |         |         |
| 18          | La Cautiva                     | Río Cuarto   | 663         | 685     | 663     |
| 19          | Chaján                         | Río Cuarto   | 484         |         |         |
| 20          | Washington                     | Río Cuarto   | 511         |         |         |
| 21          | Monte de los Gauchos           | Río Cuarto   | 376         |         |         |
| 22          | Tosquitas                      | Río Cuarto   | 378         |         |         |
| 23          | Las Albahacas                  | Río Cuarto   | 183         | 292     | 183     |
| 24          | Suco                           | Río Cuarto   | 290         |         |         |
| 25          | Chucul                         | Río Cuarto   | 176         | 236     | 176     |
| 26          | Malena                         | Río Cuarto   | 129         |         |         |
| 27          | La Carolina                    | Río Cuarto   | 158         | 165     | 158     |
| 28          | Las Peñas                      | Río Cuarto   | 148         | 139     | 148     |
| 29          | Villa El Chacay                | Río Cuarto   | 37          |         |         |
| 30          | Paso del Durazno               | Río Cuarto   |             |         |         |
| 31          | La Gilda                       | Río Cuarto   | 80          |         |         |
| 32          | Villa Santa Eugenia            | Río Cuarto   |             | 58      |         |
| Sumas       |                                |              | 196.217     | 152.700 | 142.460 |

Figura 8.2. Población en el Departamento Río Cuarto

Para las localidades que integran la muestra se tiene información de la cantidad de habitantes:

1. que surge del relevamiento muestral en el año 2001, la variable  $y$
2. la existente en el Censo de población del año 1991, la variable  $x$

La variable  $X$  es la variable auxiliar o asociada a la variable de interés  $Y$ .

La recomposición por el total se realiza haciendo

$$\hat{Y}_R = \frac{y}{x} X = \frac{\bar{y}}{\bar{x}} X$$

$$\hat{Y}_R = \frac{152700}{142460} 196217 = 210321$$

La media de la muestra es de

$$\bar{y} = \frac{\sum y_i}{n} = \frac{152700}{10} = 15270$$

El Departamento tiene 32 asentamientos. Aplicando el cálculo tradicional:

$$Y = \bar{y}N = 15270 * 32 = 488640$$

El Censo de Población del año 2001 arrojó una población total para el Departamento de 215541.

Ejemplo. La Figura 8.3 informa el número de habitantes (en miles) para 49 ciudades, extraídas en una muestra aleatoria simple de la población de 196 ciudades de un país. El problema es estimar el número total de habitantes en las 196 ciudades en el momento  $t$ . Se conoce el total para el período  $t-1$ ,  $X_{t-1} = 22919$ .

La mayoría de las ciudades en la muestra presentan un aumento en tamaño de  $t-1$  a  $t$  del orden de 20%. De los datos de la muestra se tiene que

$$y = \sum y_i = 6262 \quad x = \sum x_i = 5054$$

La estimación de razón del total del período  $t$  para las 196 ciudades es:

$$\hat{Y}_R = \frac{y}{x} X = \frac{6262}{5054} 22919 = 28397$$

La estimación basada en la media de la muestra por ciudad es

$$\hat{Y} = N\bar{y} = 196 \frac{6262}{49} = 25048$$

| CIUDAD | $y = 2001$<br><small>2001</small> | $x = 1991$<br><small>1991</small> |
|--------|-----------------------------------|-----------------------------------|
| 1      | 76                                | 80                                |
| 2      | 138                               | 143                               |
| 3      | 67                                | 67                                |
| 4      | 29                                | 50                                |
| 5      | 381                               | 464                               |
| 6      | 23                                | 48                                |
| 7      | 37                                | 63                                |
| 8      | 120                               | 115                               |
| 9      | 61                                | 69                                |
| 10     | 387                               | 459                               |

*Continúa...*

Continuación

| CIUDAD | 2011 | 2012 |
|--------|------|------|
| 11     | 93   | 104  |
| 12     | 172  | 183  |
| 13     | 78   | 106  |
| 14     | 66   | 86   |
| 15     | 60   | 57   |
| 16     | 46   | 65   |
| 17     | 2    | 50   |
| 18     | 507  | 634  |
| 19     | 179  | 260  |
| 20     | 121  | 113  |
| 21     | 50   | 64   |
| 22     | 44   | 58   |
| 23     | 77   | 89   |
| 24     | 64   | 63   |
| 25     | 64   | 77   |
| 26     | 56   | 142  |
| 27     | 40   | 60   |
| 28     | 40   | 64   |
| 29     | 38   | 52   |
| 30     | 136  | 139  |
| 31     | 116  | 130  |
| 32     | 46   | 53   |
| 33     | 243  | 291  |
| 34     | 87   | 105  |
| 35     | 30   | 111  |
| 36     | 71   | 79   |
| 37     | 256  | 288  |
| 38     | 43   | 61   |
| 39     | 25   | 57   |
| 40     | 94   | 85   |
| 41     | 43   | 50   |
| 42     | 298  | 317  |
| 43     | 36   | 46   |
| 44     | 161  | 232  |
| 45     | 74   | 93   |
| 46     | 45   | 53   |
| 47     | 36   | 54   |
| 48     | 50   | 58   |
| 49     | 48   | 75   |

Figura 8.3. Tamaño de ciudades (en miles) en el período t y en el período t-1

Donde  $\hat{Y}$  es el total estimado de la población para el momento  $t$ . El verdadero valor para el momento  $t$  era de 29351, lo que permite confirmar la precisión con la que estima el método de la razón.

---

### Varianza aproximada de la estimación de razón

La distribución de la estimación de razón ha sido siempre un problema, porque tanto  $y$  como  $x$  varían de muestra en muestra. Los resultados teóricos conocidos no corresponden a lo que se necesita saber para aplicaciones prácticas. Los resultados principales se exponen primero sin prueba.

La estimación de razón es consistente. Tiene sesgo, excepto para algunos tipos especiales de poblaciones, aunque el sesgo es despreciable en muestras grandes. La distribución límite de la estimación de razón, conforme  $n$  se hace grande, es normal, sujeta a algunas restricciones menores sobre el tipo de población de la cual se está muestreando. En muestras de tamaño moderado la distribución presenta una tendencia hacia asimetría positiva en las clases de poblaciones para las cuales se usa el método más a menudo.

No existen fórmulas exactas para el sesgo y la varianza de muestreo de la estimación sino sólo aproximaciones que son válidas en muestras grandes.

Estos resultados equivalen a decir que no hay dificultad si la muestra es lo suficientemente grande para que:

- a) la razón sea casi normalmente distribuida y
- b) sea válida la fórmula de la muestra grande para su varianza.

Las estimaciones de razón del total de la población  $Y$ , la media de la población  $\bar{Y}$  y la razón de la población  $Y/X$ , son, respectivamente

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}} X, \quad \bar{Y}_R = \frac{\bar{y}}{\bar{x}} \bar{X}, \quad \hat{R} = \frac{\bar{y}}{\bar{x}}$$

En una muestra aleatoria simple de tamaño  $n$ , donde  $n$  es grande, las varianzas se definen como:

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n} \left[ \frac{\sum_{i=1}^N (y_i - R x_i)^2}{N-1} \right] \quad (8.6)$$

$$V(\bar{Y}_R) = \frac{(1-f)}{n} \left[ \frac{\sum_{i=1}^N (y_i - R x_i)^2}{N-1} \right] \quad (8.7)$$

$$V(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^N [(y_i - R x_i)]^2}{N-1} \quad (8.8)$$

Donde:  $x_i$  e  $y_i$  representan a cada una de las observaciones en la muestra -de la variable auxiliar  $X$  y la variable de interés  $Y$ -,  $R = \bar{Y}/\bar{X}$  es la razón de las medias de la población,  $f = n/N$  es la fracción de muestreo.

Observación: Si las variables  $x_i$  e  $y_i$  son medidas para cada unidad de observación de una muestra aleatoria simple de tamaño  $n$  grande, el error cuadrático medio (ECM) y la varianza de  $\hat{R}$  son, aproximadamente, iguales a

$$ECM(\hat{R}) \cong V(\hat{R}) \cong \frac{(1-f)}{n\bar{X}^2} \left[ \frac{\sum_{i=1}^N (y_i - Rx_i)^2}{N-1} \right] \quad (8.9)$$

En donde,  $R = \bar{Y}/\bar{X}$  es la razón de las medias de la población y  $f = n/N$  es la fracción de muestreo

Este resultado se puede probar partiendo de

$$\hat{R} - R \cong \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}}$$

Si  $n$  es grande, la media de la muestra ( $\bar{x}$ ) no debe diferir mucho de la media de la población ( $\bar{X}$ ), por lo que puede remplazarse  $\bar{x}$  por  $\bar{X}$  en el denominador:

$$\hat{R} - R \cong \frac{\bar{y} - R\bar{x}}{\bar{x}}$$

Al tomar esperanza matemática, el promedio entre todas las muestras aleatorias simples de tamaño  $n$  es

$$E(\hat{R} - R) = \frac{E(\bar{y} - R\bar{x})}{\bar{X}} = \frac{\bar{Y} - R\bar{X}}{\bar{X}}$$

$$\text{Pero } R = \frac{\bar{Y}}{\bar{X}} \rightarrow E(\hat{R} - R) = \frac{\bar{Y} - R\bar{X}}{\bar{X}} = \frac{\bar{Y} - \frac{\bar{Y}}{\bar{X}}\bar{X}}{\bar{X}} = 0$$

Este resultado demuestra que  $\hat{R}$  es un estimador insesgado de  $R = \bar{Y}/\bar{X}$ , de lo que se deduce que

$$ECM(\hat{R}) = E(\hat{R} - R)^2 \cong \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2$$

La cantidad  $(\bar{y} - R\bar{x})$  es la media de la muestra de la variable  $(y_i - Rx_i)$ , cuya media en la población es  $(\bar{Y} - R\bar{X}) = 0$ . Por lo que se puede encontrar la varianza,

$$V(\hat{R}) = \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2 = \frac{1}{\bar{X}^2} \frac{S_{y_i - Rx_i}^2}{n} (1-f)$$

$$= \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^N [(y_i - Rx_i) - (\bar{Y} - R\bar{X})]^2}{N-1} = \frac{1-f}{n\bar{X}^2} \frac{\sum_{i=1}^N (y_i - Rx_i)^2}{N-1}$$

Como consecuencia de los motivos esgrimidos en esta observación, la varianza de los otros dos estimadores por el cociente o razón son las siguientes

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n} \left[ \frac{\sum_{i=1}^N (y_i - Rx_i)^2}{N-1} \right] \quad (8.10)$$

$$V(\bar{Y}_R) = \frac{(1-f)}{n} \left[ \frac{\sum_{i=1}^N (y_i - Rx_i)^2}{N-1} \right] \quad (8.11)$$

Hay varias formas alternativas del resultado; dado  $\bar{Y} = R\bar{X}$ , se puede escribir

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n(N-1)} \sum_{i=1}^N [(y_i - \bar{Y}) - R(x_i - \bar{X})]^2$$

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n(N-1)} \left[ \sum_{i=1}^N (y_i - \bar{Y})^2 + \sum_{i=1}^N R^2(x_i - \bar{X})^2 - 2R \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) \right]$$

El coeficiente de correlación  $\rho$  entre  $y_i$  y  $x_i$  en la población finita está definido por la ecuación

$$\rho = \frac{\sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X})}{(N-1)S_y S_x}$$

Esto conduce al resultado

$$V(\hat{Y}_R) = \frac{N^2(1-f)}{n} [S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x] \quad (8.12)$$

Como  $R = \bar{Y}/\bar{X}$ , entonces se puede escribir  $R^2 = \bar{Y}^2/\bar{X}^2$ , con lo que una forma equivalente de escribir (8.12) es

$$V(\hat{Y}_R) = (1-f) \frac{Y^2}{n} \left[ \frac{S_y^2}{\bar{Y}^2} + \frac{S_x^2}{\bar{X}^2} - \frac{2S_{yx}}{\bar{Y}\bar{X}} \right] \quad (8.13)$$

Donde  $S_{yx} = \rho S_y S_x$  es la covarianza entre  $y_i$  y  $x_i$ . Esta relación puede escribirse como

$$V(\hat{Y}_R) = (1-f) \frac{Y^2}{n} [C_{yy} + C_{xx} - 2C_{yx}] \quad (8.14)$$

Donde  $C_{yy}$ ,  $C_{xx}$  son los cuadrados de los coeficientes de variación (CV) de  $y_i$  y  $x_i$ , respectivamente, y  $C_{yx}$  es la covarianza relativa.

Dado que  $\hat{Y}_R$ ,  $\bar{y}_R$  y  $\hat{R}$  difieren solamente por multiplicadores conocidos, el coeficiente de variación (por ejemplo, el error estándar dividido entre la cantidad que se está estimando) es el mismo para las tres estimaciones. De (8.14) el cuadrado de este CV es

$$(CV)^2 = \frac{V(\hat{Y}_c)}{Y^2} = \frac{(1-f)}{n} [C_{yy} + C_{xx} - 2C_{yx}] \quad (8.15)$$

Esta última cantidad recibe el nombre de *varianza relativa*, su uso evita la repetición de las fórmulas de varianza para cantidades relacionadas como la media o el total de población estimados.

W. Cochran (1987) estudia el caso de los estimadores por el producto. Si una variable auxiliar  $X$  tiene correlación negativa con  $Y$ , donde ambas variables sólo toman valores positivos, un análogo del estimador por el cociente es el estimador por el producto, para el cual

$$\bar{y}_P = \bar{y} \frac{\bar{x}}{\mu_x}$$

Por expansión usual en serie de Taylor, el análogo de (8.16) para el estimador por el producto en una muestra aleatoria simple, de tamaño grande, es

$$cv^2 = \frac{(1-f)}{n} (C_{YY} + C_{XX} + 2C_{YX}) \quad (8.17)$$

### Sesgo en la estimación de razón

De acuerdo a lo anterior los estimadores de razón tienen un sesgo aproximado del orden de  $1/n$ . En este sentido

$$E(\hat{R}) \neq R.$$

Sin embargo, cuando la muestra es grande, el sesgo resulta insignificante. En otras palabras, en la medida que  $n \rightarrow N$ ;  $P[\hat{R} \rightarrow R] \rightarrow 1$ . Por lo que es un estimador consistente aunque sesgado.

La pregunta que se puede formular es cuándo se considera que  $n$  es lo suficientemente grande para que la distribución del estimador de razón se aproxime a la distribución normal y resulte aceptable la fórmula de la varianza que se ha estudiado. Como regla práctica se aconseja un tamaño de la muestra superior a 60 o cuando los coeficientes de variación de  $X$  e  $Y$  son inferiores al 10%.

Cuando el tamaño de la muestra es superior a 60, el sesgo es despreciable, pero sesgo no es lo mismo que precisión, por lo tanto una muestra de ese tamaño no garantiza suficiente precisión.

### Estimador de razón en el muestreo estratificado

Para estimar el total de una población utilizando el estimador de razón, a partir de una muestra con selección aleatoria estratificada, hay dos maneras de hacerlo según W. Cochran (1987) pp. 211-217. Estas maneras se denominan estimador por razón separada y estimador por razón combinada.

En el *estimador por razón* -de un total de una población- *de manera separada* se debe conocer el valor poblacional de la variable auxiliar para cada estrato,

$$\hat{Y}_{Rs} = \sum_{h=1}^L \frac{y_h}{x_h} X_h = \sum_{h=1}^L \frac{\sum_{i=1}^{n_h} y_{hi}}{\sum_{i=1}^{n_h} x_{hi}} X_h = \sum_{h=1}^L \frac{\bar{y}_h}{\bar{x}_h} X_h \quad (8.18)$$

Donde,  $h$  representa el  $h$  –ésimo estrato de la población dividida en  $L$  estratos;

$X_h = \sum_{i=1}^{N_h} X_{hi}$ , es el total poblacional de la variable auxiliar para el  $h$  –ésimo estrato.

En el *estimador de razón del total de manera combinada*, se obtiene una sola razón combinada con datos muestrales y se multiplica por el total poblacional de la variable auxiliar,

$$\hat{Y}_{RC} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} X = \frac{\bar{y}_{st}}{\bar{x}_{st}} X \quad (8.19)$$

Donde,  $\hat{Y}_{st} = \sum_h N_h \bar{y}_h$  y  $\hat{X}_{st} = \sum_h N_h \bar{x}_h$  se calculan a partir de los datos obtenidos en una muestra estratificada y representan las estimaciones de los totales de población  $Y$  y  $X$ .

$\bar{y}_{st} \approx \hat{Y}_{st}/N$  y  $\bar{x}_{st} \approx \hat{X}_{st}/N$  son las medias de la población estimadas de una muestra estratificada.

$\hat{Y}_{RC}$  es la estimación del total poblacional para la variable  $Y$  a través de un estimador de razón combinado aplicado a información surgida de un muestreo estratificado.

La ventaja del método combinado sobre el de separación es que no se necesita conocer el total de la variable  $X$  para cada estrato.

En el caso de que la muestra sea grande es conveniente utilizar el primer método. Para muestras pequeñas conviene utilizar el estimador por razón combinada.

### 8.1.2. Estimador por mínimos cuadrados

Si  $X$  e  $Y$  son correlacionadas pero no necesariamente proporcionales, para la estimación de  $\bar{Y}$  se puede utilizar el estimador por la regresión lineal  $\bar{y}_{rl}$ . Para Scheaffer et al (2007), si existe relación lineal entre los valores observados de las variables  $X$  e  $Y$ , pero no se comportan como una recta que pase por el origen, la información extra proporcionada por la variable auxiliar  $X$  puede tomarse en cuenta mediante un estimador de regresión de la media  $\bar{Y}$ .

Este estimador se define con:

$$\bar{y}_{rl} = a\mu_x + b = a(\bar{X} - \bar{x}) + \bar{y} \quad (8.20)$$

Donde  $a = \frac{s_{xy}}{s_x^2}$  y  $b = \bar{y} - a\bar{x}$ , ambos se calculan sobre los datos muestrales  $(y, x)$  y son los coeficientes de una recta de regresión simple.

$\mu_x$  es el valor medio que asume la variable asociada en la población.

Si  $a$  y  $b$  son los coeficientes de la recta de regresión, de  $Y$  en  $X$  calculados sobre la muestra, se define

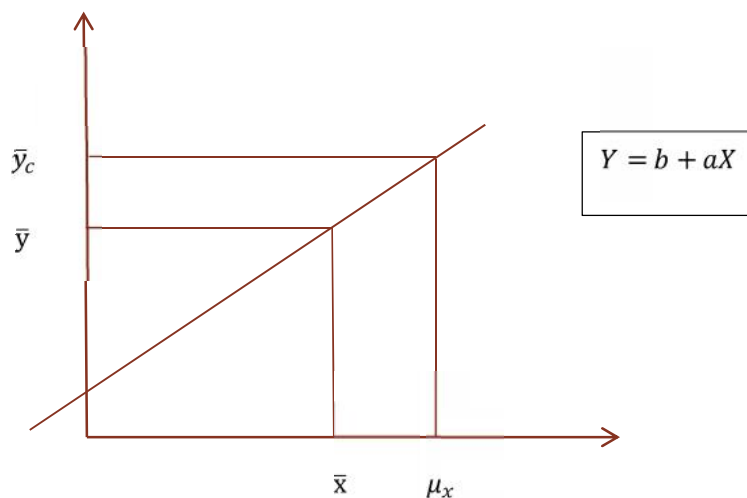


Figura 8.4 Recta de regresión de  $Y$  sobre  $X$

Se muestra que  $\bar{y}_{rl}$  es un estimador de  $\bar{Y}$  casi sin sesgo y que  $V(\bar{y}_{rl})/V(\bar{y}) = 1 - \rho^2$ , donde  $\rho$  es el coeficiente de correlación lineal de  $X$  e  $Y$ .

Si el valor de  $\rho$  es próximo a 1 (o a -1) –lo que está indicando alta correlación entre las variables- entonces  $\bar{y}_{rl}$  es un estimador de la media de la población ( $\bar{Y}$ ) mucho más preciso que  $\bar{y}$ .

La estimación de regresión es consistente en el sentido de que cuando la muestra comprende a toda la población  $\bar{x} = \bar{X}$  y la estimación de regresión se reduce a  $\bar{Y}$ . Con valores de  $b$  adecuados, la estimación por regresión incluye, como casos particulares a la media por unidad y la estimación de razón:

Si  $b = 0 \rightarrow \bar{y}_{rl} = b(\bar{X} - \bar{x}) + \bar{y} = \bar{y}$ , la media por unidad

Si  $b = \frac{\bar{y}}{\bar{x}} \rightarrow \bar{y}_{rl} = \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) + \bar{y} = \frac{\bar{y}}{\bar{x}}\bar{X} - \frac{\bar{y}}{\bar{x}}\bar{x} + \bar{y} = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{\bar{Y}}_R$ , la estimación de razón.

Si  $b = 1 \rightarrow \bar{y}_{rl} = b(\bar{X} - \bar{x}) + \bar{y} = \bar{X} - \bar{x} + \bar{y}$ , la estimación por diferencia

### 8.1.3 Varias variables reales

Este método se lo conoce como de regresión múltiple. Si se dispone de varias variables auxiliares reales ( $X_1, X_2, \dots, X_p$ ) asociadas a la variable de interés ( $Y$ ), se formula la ecuación:

$$y = b_1x_1 + b_2x_2 + \dots + b_kx_k + a \quad (8.21)$$

para calcular por mínimos cuadrados los valores de  $b_1, b_2, \dots, b_p$  y  $a$ . Tanto los valores de  $y$  como los de  $x_1, x_2, \dots, x_k$  provienen del relevamiento muestral.

El valor total de  $Y$  se conoce a partir de hacer

$$Y_{rm} = b_1X_1 + b_2X_2 + \dots + b_kX_k + a = \sum_{k=1}^p b_k X_k + a \quad (8.22)$$

Donde  $Y_{rm}$  es el total de la variable de interés ( $Y$ ) por el método de regresión múltiple y  $X_1, X_2, \dots, X_k$  son los totales en la población para las variables asociadas.

Si se conoce con precisión las medias de las variables asociadas en la población ( $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p$ ), el estimador de  $\bar{Y}$  por la regresión múltiple es

$$\bar{Y}_{rm} = \sum_{k=1}^p b_k \bar{X}_k + a = \sum_{k=1}^p b_k (\bar{X}_k - \bar{x}_k) + \bar{y} \quad (8.23)$$

Donde  $\bar{x}_k$  es la media de  $X_k$  sobre la muestra.

## 8.2 Recomposición sobre variables categóricas

### 8.2.1 Estimador de post estratificación

Ocasionalmente, aparecen problemas de muestreo en los que se quiere estratificar en función de una variable clave, pero no es posible ubicar las unidades de muestreo en sus estratos correctos sólo hasta después de haber seleccionado la muestra. Es decir, no hay manera de identificar a cada individuo de la población con su estrato a la hora de hacer el muestreo.

La estratificación después de la selección de la muestra (o postestratificación), es frecuentemente apropiada cuando la muestra aleatoria simple no está correctamente equilibrada de acuerdo con los segmentos principales de la población. En estos casos se propone el estimador de post estratificación:

$$\bar{y}_{post} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h \quad (8.24)$$

Donde  $\bar{y}_h = \sum_{i \in h} y_i / n_h$  es la media de  $Y$  sobre los individuos de la muestra que pertenecen a la categoría  $h$ .

Cuando se realiza un muestreo estratificado, se fija de antemano el tamaño que tendrá la muestra en cada estrato  $(n_1, n_2, \dots, n_h)$  de modo que la suma de esos tamaños da lugar al tamaño de la muestra

$$\sum_{h=1}^H n_h = n \quad (8.25)$$

En la post estratificación el número de individuos  $(n_h)$  que pertenecen a cada estrato  $(h)$  es una variable aleatoria. Se sabe que si  $n$  es bastante grande, la proporción de individuos de la muestra que pertenecen a la categoría  $h$   $(n_h/n)$  es un estimador sin sesgo de la proporción correspondiente en la población

$$E\left(\frac{n_h}{n}\right) = \frac{N_h}{N} \quad (8.26)$$

siendo la varianza de  $\bar{y}_{post}$  similar a la del estimador del muestreo aleatorio estratificado proporcional:

$$E(\bar{y}_{post}) \approx \mu \quad (8.27)$$

$$V(\bar{y}_{post}) = \frac{\sigma_{dentro}^2}{n} \left(1 - \frac{n}{N}\right) \quad (8.28)$$

**Ejemplo.** Se encuestaron 800 docentes para conocer el consumo de un producto determinado arrojando el resultado de \$194,25. La población de docentes se particiona entre quienes imparten sus conocimientos en el ámbito primario y quienes lo hacen en el secundario; la constitución de la muestra no ha tenido en cuenta (o no ha podido tener en cuenta) esta situación. En la Figura 8.5 se observa la participación de cada grupo en la población y en la muestra y el resultado que asume la variable de interés en cada estrato muestral.

| Ámbito escolar | Población  | Muestra    |               |
|----------------|------------|------------|---------------|
|                | Proporción | Proporción | Consumo medio |
| Primaria       | 0.30       | 0.35       | \$165.00      |
| Secundaria     | 0.70       | 0.65       | \$210.00      |

**Figura 8.5 Participación en la población y en la muestra**

La muestra no respeta exactamente las proporciones de docentes en el primario y en el secundario que existen en la población bajo estudio, para solucionar esto se propone una estimación del consumo medio por post estratificación haciendo

$$\bar{y}_{post} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h = 0.30 * 165 + 0.70 * 210 = 196,50$$

Si se hubiera considerado el resultado arrojado por la muestra, sin corregir por el tamaño del estrato en la población, se habría incurrido en una subestimación.

## 8.2.2 Varias variables categóricas: método RAS

Cuando en una encuesta se tiene varias variables categóricas y se conoce cuál es la frecuencia en la población, de cada modalidad integrante de cada variable, es posible construir una única variable categórica donde las modalidades se originan a partir de la combinación de modalidades de las variables originales. A partir de aquí, el proceso de recomposición de la muestra se aplica utilizando la post estratificación sobre esta variable.

**Ejemplo.** Una muestra aleatoria de 1200 casos presenta la distribución conjunta de las variables Sexo y Grupo de edad con la frecuencia observada en la Figura 8.6. Se anexa también, la frecuencia esperada en la población de acuerdo a la composición demográfica conocida.

El problema que se presenta es que la composición en edades y sexo de la muestra no refleja la existente en la población. Por esto es necesario recomponer la muestra.

|      |           | Edad |         |     | Frecuencia |           |
|------|-----------|------|---------|-----|------------|-----------|
|      |           | <25  | 25 a 65 | >65 | Muestra    | Población |
| Sexo | Hombre    | 150  | 150     | 200 | 500        | 600       |
|      | Mujer     | 180  | 220     | 300 | 700        | 600       |
| Frec | Muestra   | 330  | 370     | 500 | 1200       |           |
|      | Población | 300  | 500     | 400 | 1200       |           |

**Figura 8.6 Información obtenida de la muestra**

Para recomponer la muestra se utiliza el método iterativo RAS (Raking Adjusted Statistics) consistente en:

### 1. Ajuste sobre las líneas

Con los datos de la Figura 8.6, se multiplica la primera línea por  $600/500$  y la segunda por  $600/700$ ; es decir, se multiplica por la frecuencia de la población y se divide por la hallada en la muestra. El resultado es la Figura 8.7.

|      |           | Edad  |         |       | Frecuencia |           |
|------|-----------|-------|---------|-------|------------|-----------|
|      |           | <25   | 25 a 65 | >65   | Muestra    | Población |
| Sexo | Hombre    | 180   | 180     | 240   | 600        | 600       |
|      | Mujer     | 154,3 | 188,6   | 257,1 | 600        | 600       |
| Frec | Muestra   | 334,3 | 368,6   | 497,1 | 1200       |           |
|      | Población | 300   | 500     | 400   | 1200       |           |

**Figura 8.7 Ajuste sobre líneas – primer ajuste**

### 2. Ajuste sobre las columnas

Con los datos de la Figura 8.7, se multiplica la primera columna por  $300/334.3$ , la segunda por  $500/368.6$ , la tercera por  $400/497.1$ . El resultado es la Figura 8.8.

|      |           | Edad  |         |       | Frecuencia |           |
|------|-----------|-------|---------|-------|------------|-----------|
|      |           | <25   | 25 a 65 | >65   | Muestra    | Población |
| Sexo | Hombre    | 161.5 | 244.2   | 193.1 | 598.8      | 600       |
|      | Mujer     | 138.5 | 255.8   | 206.9 | 601.2      | 600       |
| Frec | Muestra   | 300   | 500     | 400   | 1200       |           |
|      | Población | 300   | 500     | 400   | 1200       |           |

**Figura 8.8 Ajuste sobre columnas**

### 3. Ajuste sobre las líneas

Con los datos de la Figura 8.8, se multiplica la primera línea por  $600/598.8$  y la segunda por  $600/601.2$ . El resultado es la Figura 8.9 en la cual la composición de la muestra responde a la esperada según la composición de la población.

|      |           | Edad  |         |       | Frecuencia |           |
|------|-----------|-------|---------|-------|------------|-----------|
|      |           | <25   | 25 a 65 | >65   | Muestra    | Población |
| Sexo | Hombre    | 161.8 | 244.7   | 193.5 | 600        | 600       |
|      | Mujer     | 138.2 | 255.3   | 206.5 | 600        | 600       |
| Frec | Muestra   | 300   | 500     | 400   | 1200       |           |
|      | Población | 300   | 500     | 400   | 1200       |           |

**Figura 8.9 Ajuste sobre líneas – segundo ajuste**

La Figura 8.9 indica la cantidad de personas que debieron ser encuestadas en cada grupo y en la Figura 8.6 la cantidad que efectivamente ha sido entrevistada. Esta diferencia hace que las opiniones de algunos grupos estén sobredimensionadas y las de otros subdimensionadas. Para darle a las opiniones de cada grupo el peso que le corresponden, se asignará un peso en función de la clase a la cual pertenece cada uno de los 1200 individuos de la muestra. Este peso, que relaciona la participación del estrato en la muestra y en la población, combina los resultados de la Figura 8.9 y la Figura 8.6:

| Clase                       | Peso              |
|-----------------------------|-------------------|
| Hombres de menos de 25 años | 161.8/150=1.07867 |
| Hombres de 25 a 65 años     | 244.7/150=1.63133 |
| Hombres de más 65 años      | 193.5/200=0.9675  |
| Mujeres de menos de 25 años | 138.2/180=0.76778 |
| Mujeres de 25 a 65 años     | 255.3/220=1.16045 |
| Mujeres de más 65 años      | 206.5/300=0.68833 |

La suma de los pesos sobre los 1200 individuos es igual a 1200 y el estimador de la media poblacional por recomposición RAS es:

$$\bar{y}_{RAS} = \frac{1}{1200} \sum_{i=1}^{1200} p_i y_i \quad (8.29)$$

En definitiva, se trata de volver a ponderar los individuos de la muestra asignándole un peso distinto de 1. Por esto se los denomina métodos de recomposición de muestras.

Estos métodos son muy utilizados por los politólogos para estimar las intenciones de votos. Los resultados en bruto de los muestreos tendrían un alto error al no utilizar las informaciones auxiliares conocidas. Droesbeke y Fine (1997) advierten que las personas encuestadas que tienen opiniones extremas no confiesan sus intenciones reales de voto a un encuestador.

Entre los numerosos especialistas en muestreo se afirma que “recomponer no es engañar” pero “no recomponer cuando es posible constituye una falta profesional para el estadístico”.

## PROBLEMAS Y CASOS DE ESTUDIO

### Caso 8.1 Población en la Provincia de Córdoba

La tabla de datos localidades.xls tiene información de cantidad de personas para 526 localidades de la Provincia de Córdoba, según los censos de población de 1991 y 2001. Seleccione una muestra aleatoria de 50 localidades y aplique el estimador de la razón para corregir el resultado muestral.

---

---

### Bibliografía

- **Cochran, William G.** *Técnicas De Muestreo*. Editorial CECSA, 1987.
- **Droesbeke, Jean-Jacques y Fine, Jeanne.** "Metodología De La Encuesta," L. d. M. d. T. d. D. Université Libre de Bruxelles, Belgique, *Programme de Recherche et d'enseignement en Statistique Appliquée*. Concepción, Chile: Universidad de Concepción, 1997.
- **Rodriguez, Norberto.** "Curso Taller Sobre Diseño De Muestras Probabilísticas," *XXIII Coloquio Argentino de Estadística*. Cordoba: Sociedad Argentina de Estadística, 1995.
- **Scheaffer, Richard; William Mendenhall III y R.Lyman Ott.** *Elementos De Muestreo*. Madrid: Editorial Thomson, 2007.