CAPÍTULO 9: Estimador de Razón y Regresión

por: Enrique Wabo

1 INTRODUCCIÓN

Ya se han visto los diseños de muestreo aleatorio simple (sistemático), y el aleatorio estratificado. En este capítulo introducimos dos nuevos estimadores muestrales: el *Estimador de Razón* y el *Estimador de Regresión*. Estos mecanismos de estimación no están asociados a un diseño de muestreo específico, sino que se pueden usar en diferentes diseños; es decir que son métodos generales. Nosotros veremos su aplicación sólo para el muestreo simple no estratificado.

Ambos estimadores recurren a la observación de una segunda variable en cada unidad de muestreo, como agregado a la variable de interés Y. Esta segunda variable se conoce con el nombre de VARIABLE AUXILIAR, y se la indica con la letra X.

Ambos estimadores operan en base a la relación lineal entre X e Y. En el estimador de razón, la recta pasa por el origen de coordenadas; en el estimador de regresión, la recta no pasa por el origen de coordenadas. En cualquiera de los dos casos, X e Y deben estar alta y positivamente correlacionadas para permitir buenos resultados.

Para estimar la media o el total de Y en la población, estos estimadores requieren que se conozca la media o el total poblacional de X. Hay una excepción a esta regla y es cuando se aplica el estimador de razón para determinar una proporción, situación en la cual se puede usar la media estimada de X.

2 ESTIMADOR DE RAZÓN

El estimador de razón puede ser usado con tres propósitos:

a) estimar la proporción de Y respecto a X;

b) estimar la media poblacional de Y; o

c) estimar el total poblacional de Y.

La razón R es una constante poblacional y se la define como:

$$R = \frac{\mu_y}{\mu_x} = \frac{\sum_1^N y}{\sum_1^N x} = \frac{Y}{X}$$

Si se selecciona una muestra aleatoria sin reemplazo, el estimador $\hat{R}$ de R toma la siguiente forma:

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_1^n y}{\sum_1^n x} \qquad (1)$$

El significado de $\hat{R}$ depende de la definición de X y de Y, la que debe hacerse al comienzo del proceso de estimación. Cualesquiera sean las variables específicas intervinientes, la razón representa el número medio de unidades de Y por cada unidad de X.

Por ejemplo, si Y es definido como volumen por parcela y X es definido como área basal por parcela, R es el volumen medio por unidad de área basal. Si Y es el número de vacas por lote y X la superficie de cada lote, R es el número medio de vacas por unidad de superficie de los lotes.

2.1 ESTIMADOR DE RAZÓN PARA ESTIMAR UNA PROPORCIÓN

Se selecciona una muestra aleatoria de tamaño n de una población de tamaño N y se aplica el estimador (1). La varianza estimada de la razón estimada toma la siguiente forma:

$$s_{\hat{R}}^2 = \frac{1}{\mu_x^2} \times \frac{s_u^2}{n} \times \left(\frac{N-n}{N}\right) \qquad (2)$$

siendo:

$$s_u^2 = \frac{\sum y^2 + \hat{R}^2 \sum x^2 - 2\hat{R} \sum xy}{n-1} \qquad (3a)$$

$$s_u^2 = s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\, s_{xy} \qquad (3b)$$

siendo s_y^2 la varianza de Y, s_x^2 la varianza de X y s_{xy} la covarianza entre X e Y.

El término s_u^2 mide la variabilidad de R dentro de la muestra en forma de varianza. El error estándar de $\hat{R}$ es la raíz cuadrada de (2):

$$s_{\hat{R}} = +\sqrt{s_{\hat{R}}^2} = \frac{s_u}{\mu_x} \times \frac{1}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N}\right)} \qquad (4)$$

Finalmente, el intervalo de confianza del R estimado es:

$$\text{IC (P\%)} = \hat{R} \pm t\, s_{\hat{R}} \qquad (5)$$

siendo t la variable “t” de Student con n-1 grados de libertad y probabilidad P%.

EJEMPLO 1

Un plantador de la zona del Delta de Buenos Aires posee un rodal de álamo de 30 ha de extensión afectado por una enfermedad, y desea saber cuál es la proporción de árboles atacados. Para ello se selecciona una muestra aleatoria de 10 parcelas rectangulares de 10 m de ancho y 25 m de largo. En cada parcela se cuenta el número total de árboles (X) y el número de árboles enfermos (Y). En la siguiente tabla figuran los resultados por parcela

X	30	16	42	28	46	54	25	16	32	30
Y	21	12	34	18	32	39	16	10	26	19

Σx = 319 $\quad \Sigma y$ = 227 $\quad \Sigma xy$ = 8.294

Σx^2 = 11.541 $\quad \Sigma y^2$ = 6.003 $\quad$ n = 10

N = (30 ha)/(0,025 ha/parcela) = 1.200 parcelas

Ahora estimamos R: $\hat{R} = \dfrac{227}{319} = 0{,}711$

Para estimar el error estándar de $\hat{R}$ debemos usar la media poblacional de X, la que desconocemos. En su lugar usamos, como mejor estimación, la media de la muestra: 31,9 . Primero estimamos S_u^2 y luego el error estándar:

$$S_u^2 = \frac{6003 + 0{,}711^2 \times 11541 - 2 \times 0{,}711 \times 8294}{9}$$

4,79

$$S_{\hat{R}}^2 = \frac{1}{(31{,}9)^2} \times \frac{4{,}79}{10} \times \left(\frac{1200 - 10}{1200}\right) =$$

0,000467

y: $\quad S_{\hat{R}} = \sqrt{0{,}000467} = 0{,}0216$

El intervalo de confianza de la razón para un 95 % de probabilidad es:

IC (95%) = 0,711 ± (2) (0,0216

Li = 0,67 $\qquad$ Ls = 0,75

Finalmente, concluimos con una confianza del 95 % que la proporción de árboles enfermos se ubica entre el 67 % y 75 % de todos los árboles.

En el Ejemplo, el objetivo fue la estimación de una proporción, no la estimación de la media o del total de Y; por eso no fue necesario conocer la media poblacional de X.

Cualquiera sea el diseño específico que se utilice, el mecanismo para la estimación de R y su error estándar no varía.

2.2 ESTIMADOR DE RAZÓN PARA ESTIMAR UNA MEDIA POBLACIONAL

Aparece aquí la condición de que la media poblacional de X sea conocida. La media estimada de Y para la población toma la siguiente forma:

$$\hat{\mu}_Y = \bar{y}_R = \hat{R}\,\mu_X \qquad (6)$$

con error estándar:

$$s_{\bar{y}_R} = s_{\hat{R}}\,\mu_X \qquad (7)$$

e intervalo de confianza:

$$\text{IC} = \bar{y}_R \pm t\, s_{\bar{y}_R} \qquad (8)$$

2.3 ESTIMADOR DE RAZÓN PARA LA ESTIMACIÓN DE UN TOTAL POBLACIONAL

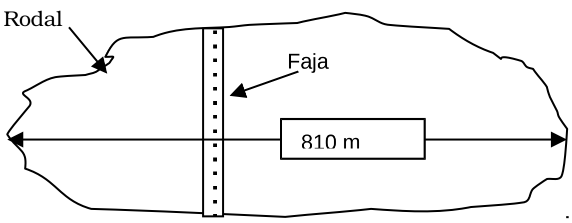
Se siguen los mismos pasos seguidos en el punto 2.2, pero ahora la razón estimada se multiplica por el total de X, que debe conocerse a nivel de población. Si indicamos con X el total de la variable X en la población, tenemos:

Total estimado de Y es: $\quad \hat{Y}_R = \hat{R}\,X \qquad (9)$

con error estándar: $\quad s_{\hat{Y}_R} = s_{\hat{R}}\,X \qquad (10)$

EJEMPLO 2

Se va a realizar un inventario forestal de un bosque para determinar el volumen total de madera existente. Del análisis de las fotos aéreas surge que tiene 49,6 ha de superficie, que su forma es irregular, y que posee 810 m de ancho. El diseño de muestreo recurre a 10 fajas continuas de 10 m de ancho, que van de un borde a otro de la mancha, en sentido perpendicular a su ancho:



Como el rodal tiene forma irregular, algunas fajas van a ser más largas que otras, por lo que no es posible usar el método clásico de expansión de la media de Y. En su lugar se utilizará el estimador de razón.

Se define con X al área ocupada por cada faja, en hectáreas; se define con Y el volumen estimado en cada faja, en m^3. Los datos de las fajas son:

X	1,80	0,62	1,00	0,34	0,68
Y	740	290	450	180	340
X	1,24	0,06	1,66	1,80	1,76
Y	560	36	680	780	820

Σx = 10,96 $\quad \Sigma y$ = 4.876 $\quad \Sigma xy$ = 6.926,76

Σx^2 = 15,8368 $\quad \Sigma y^2$ = 3.040.296 $\quad$ n = 10

N = ancho/faja = 810m/10m = 81 fajas

Estimamos R: $\hat{R} = \dfrac{4876}{10{,}96}$ = 444,89 m^3/ha de faja

Estimamos su error estándar:

$$s_u^2 = 1.281,73$$

$$s_{\hat{R}}^2 = \frac{1}{(0,617284)^2} \times \frac{1281,73}{10} \times \left(\frac{81-10}{81}\right) =$$
2388,27

$$s_{\hat{R}} = \sqrt{2388,27} = 48,87$$

Ahora estimamos el total:

$$\hat{Y} = 50 \text{ ha} \times 444,89 \text{ m}^3/\text{ha} = 22.244,5 \text{ m}^3$$

y su error estándar:

$$s_{\hat{Y}} = 50 \text{ ha} \times 48,87 = 2.443,5 \text{ m}^3$$

El intervalo de confianza para un 95 % es:

IC(95%) = 22.244,5 ± (2) (2.443,5)

Li = 17.358 m^3 Ls = 27.131 m^3

Como comentario final digamos que corresponde el uso del estimador de razón cuando la relación entre X e Y es una línea recta que pasa por el origen de coordenadas. Por otro lado, el estimador de razón es un estimador insesgado de R cuando: a) la relación entre X e Y es lineal y pasa por el origen de coordenadas, y b) cuando la varianza de Y es directamente proporcional a X.

3 ESTIMADOR DE REGRESIÓN

El estimador de Regresión se utiliza para la estimación de medias y totales de una población con la ayuda de una variable auxiliar. La estimación de la media de Y en la población toma la siguiente forma:

$$\hat{\mu}_y = \bar{y}_{reg} = \bar{y} + b\,(\mu_x - \bar{x}) \qquad (11)$$

donde μ_x es la media poblacional de X; mientras que $\bar{y}$, $\bar{x}$ y b son estimados a partir de los datos de la muestra.

El mecanismo de funcionamiento es sencillo: con los datos de la muestra se define una recta de regresión y se determina el valor de Y que corresponde a la media poblacional de X; el valor obtenido es la media estimada de Y.

La constante b se estima con el estimador de mínimos cuadrados:

$$b = \frac{SPXY}{SCDX} \qquad (12)$$

Para estimar los errores estándar necesitamos conocer cómo varían los valores alrededor de la recta de regresión. Esta variación la expresamos mediante la varianza de los desvíos respecto a la regresión, que suele expresarse de distintas formas, a saber: s_E^2, s_{reg}^2, $s_{y.x}^2$. Este último no debe confundirse con el símbolo de la covarianza: S_{xy}. La varianza de la regresión se estima con:

$$s_E^2 = \frac{SCDY - b\,SPXY}{n-2} = \frac{SCDY - \frac{SPXY^2}{SCDX}}{n-2} \qquad (13)$$

3.1 ESTIMADOR DE REGRESIÓN PARA ESTIMAR UNA MEDIA POBLACIONAL POR UNIDAD DE MUESTREO

Para la estimación de la media por unidad de muestreo usamos la fórmula (11). La varianza de la media estimada de Y por unidad toma la siguiente forma:

$$s_{\bar{y}_{reg}}^2 = \frac{s_E^2}{n}\left(1 - \frac{n}{N}\right) \qquad (14)$$

siendo n el tamaño de la muestra y N el tamaño de la población. Su error estándar es la raíz cuadrada de (14):

$$s_{\bar{y}_{reg}} = +\sqrt{s^2_{\bar{y}_{reg}}} \qquad (15)$$

Los Intervalos de confianza para la media estimada es

$$IC = \bar{y}_{reg} \pm t\, s_{\bar{y}_{reg}} \qquad (16)$$

3.2 ESTIMADOR DE REGRESIÓN PARA ESTIMAR UN TOTAL POBLACIONAL

El total de Y estimado para la población es:

$$\hat{Y} = N\,\bar{y}_{reg} \qquad (17)$$

donde $\bar{y}_{reg}$ se estima con la fórmula (11). Su error estándar es:

$$s_{\hat{Y}} = N\,s_{\bar{y}_{reg}} \qquad (18)$$

Los Intervalos de confianza para el total es:

$$IC = \hat{Y}_{reg} \pm t\, s_{\hat{Y}_{reg}} \qquad (19)$$

BIBLIOGRAFÍA

Sampling Techniques for Forest Resource Inventory. Barry D. Shiver and Bruce E. Borders. 1996. John Wiley & Sons INC.

Sampling Theory for Forest Inventory. 1986. Pieter G. de Vries. Springer-Verlag.

Octubre 2002