Métodos indirectos de estimación: razón, regresión y diferencia

Contenido

1. Introducción, definición de estimadores indirectos
2. Estimador de razón, propiedades y varianzas. Límites de confianza.
3. Tamaño de la muestra en los estimadores de razón
4. Eficiencia de los estimadores de razón
5. Estimadores de razón en el muestreo estratificado: razón separado y razón combinado.
6. Estimadores de regresión con “b” pre-asignada y con “b” estimada.
7. Comparación entre los estimadores de razón, de regresión y simple.
8. Estimadores de regresión en el muestreo estratificado: separado y combinado.
9. Estimador de diferencia.

Métodos indirectos de estimación

Es frecuente en el muestreo usar información adicional para mejorar la precisión de las estimaciones, esto se conoce como métodos indirectos y los más comunes son los de razón y los de regresión.

Los métodos indirectos requieren el poder observar una variable auxiliar X apareada con la variable en estudio Y (y_i,x_i) y además conocer el total X.

La correlación existente entre X e Y es lo que permite el incremento de la precisión al estimar los parámetro de Y.

Los estimadores indirectos se usan para estimar el total, la media y la razón propiamente.

Métodos indirectos de estimación

Para el total $\hat{Y}$ la forma general de los estimadores indirectos es:

$$\hat{Y}_I = \hat{Y} + \alpha(X - \overline{X})$$

Donde α es el coeficiente de corrección para mejorar el estimador directo $\hat{Y}$

Si $\alpha = 0$	$\hat{Y}_I \equiv \hat{Y}$	estimador directo
Si $\alpha = \hat{R} = \frac{\hat{Y}}{\hat{X}}$	$\hat{Y}_I = \hat{R}\,X$	estimador de razón
Si $\alpha = 1$	$\hat{Y}_I = \hat{Y} + X - \hat{X}$	estimador de diferencia
Si $\alpha = b$	$\hat{Y}_I = Y + b(X - \hat{X})$	estimador de regresión

b coeficiente de regresión lineal Y/X

Estimador de razón en el Muestreo Aleatorio Simple.

En el mas el estimador de razón del total es: $\hat{Y}_R = \hat{R}X = \frac{\bar{y}}{\bar{x}}X$

X_i debe ser una variable apareada con Y_i y correlacionada entre si, X_i puede ser el valor de Y_i en una ocasión anterior ($X_t=Y_{t-s}$)
Si la razón y_i/x_i es la misma en todas las unidades de muestreo, $\bar{y}/\bar{x}$ es estable de muestra en muestra y $\hat{Y}_R$ será mas precisa. El método se justifica al suponer que Y_i varía proporcional a X_i $(y_i = kx_i \rightarrow \bar{y} = k\bar{x})$

El estimador de razón $\hat{Y}_R (o \quad \bar{y}_R)$ es:

1. Consistente en el sentido de cochran
2. Sesgado, pero el sesgo es despreciable en muestras grandes
3. En muestras grandes se puede utilizar la distribución normal (grandes si n>30 y $CV(\bar{x}) < 10\%$ y $CV(\bar{y}) < 10\%$), en muestras medianas la distribución es asimétrica positiva

Estimadores de razón del total, la media, y la razón.

Estimadores	Sus varianzas (aproximadas)
$\hat{Y}_R = \hat{R}\,X$	$V\left(\hat{Y}_R\right) = \frac{N^2(1-f)}{n}\left[\frac{\sum(y_i - Rx_i)^2}{N-1}\right]$
$\bar{y}_R = \hat{R}\,\bar{x}$	$V\left(\bar{y}_R\right) = \frac{1-f}{n}\left[\frac{\sum(y_i - Rx_i)^2}{N-1}\right]$
$\hat{R} = \frac{\bar{y}}{\bar{x}}$	$V\left(\hat{R}\right) = \frac{1-f}{n\bar{X}^2}\left[\frac{\sum(y_i - Rx_i)^2}{N-1}\right]$

Los estimadores son “aproximadamente” insesgados y las pruebas de las varianzas se vieron en el tema 2 para $\hat{R}$ la de los otros dos sigue inmediato.

Formas de expresión de la varianza

Como

$$\sum^{N}(y_i - Rx_i)^2 = \sum^{N}\left(y_i - Rx_i - \overline{Y} - R\overline{X}\right)^2 = \sum^{N}\left(\left(y_i - \overline{Y}\right) - R\left(x_i - \overline{X}\right)\right)^2$$

$$= \sum^{N}\left(y_i - \overline{Y}\right)^2 + R^2\sum^{N}\left(x_i - \overline{X}\right)^2 - 2R\sum^{N}\left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right)$$

Y $S_Y^2 = \dfrac{\sum^{N}\left(y_i - \overline{Y}\right)^2}{N-1}$; $S_X^2 = \dfrac{\sum^{N}\left(x_i - \overline{X}\right)^2}{N-1}$ y $\rho = \dfrac{\sum^{N}\left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right)}{(N-1)S_X S_Y} = \dfrac{S_{XY}}{S_X S_Y}$

Así $V\left(\hat{Y}_R\right) = \dfrac{N^2(1-f)}{n}\left(S_Y^2 + R^2 S_X^2 - 2R\rho S_X S_Y\right)$

$$= (1-f)\frac{\overline{Y}^2}{n}\left(\frac{S_Y^2}{\overline{Y}^2} + \frac{S_X^2}{\overline{X}^2} - \frac{2S_{XY}}{\overline{Y}\,\overline{X}^2}\right) = (1-f)\frac{\overline{Y}^2}{n}\left(C_{YY} + C_{YY} - 2C_{XY}\right)$$

donde C_{YY} y C_{XX} son los cuadrados de los coeficientes de variación y C_{XY} es la covarianza relativa.

El cuadrado del coeficiente de variación de $\hat{Y}_R$ se denomina varianza relativa $\left(CV(\hat{Y}_R)\right)^2 = \dfrac{V(\hat{Y}_R)}{Y^2} = \dfrac{(1-f)}{n}\left(C_{YY} + C_{YY} - 2C_{XY}\right)$

Estimaciones de la varianza

Las estimaciones $\hat{V}(\hat{R})$, $\hat{V}(\hat{Y}_R)$ y $\hat{V}(\bar{y}_R)$ se obtienen al sustituir $\frac{\sum^N (y_i - Rx_i)^2}{N-1}$ por $\frac{\sum^n (y_i - Rx_i)^2}{n-1}$ en las formulas de $V(\bullet)$.

Cuando $\bar{X}$ o R son desconocidos se sustituyen por sus estimaciones $\bar{x}$ o $\hat{R}$.

Límites de confianza

Si n es suficientemente grande y, $\mathrm{CV}(\bar{x})$ y $\mathrm{CV}(\bar{y})$ son ambos <0,1 se puede aplicar razonablemente la aproximación normal. Obteniendo los límites de confianza para Y, $\bar{Y}$ y R.

R $\hat{R} \mp t_\alpha \sqrt{\hat{V}(\hat{R})}$ Y $\hat{Y}_R \mp t_\alpha \sqrt{\hat{V}(\hat{Y}_R)}$ $\bar{Y}$ $\hat{\bar{Y}} \mp t_\alpha \sqrt{\hat{V}(\hat{\bar{Y}}_R)}$

Cuando las condiciones exigidas no se cumplen la distribución puede ser de notoria asimetría. En estos casos existen métodos alternos para calcular los límites de confianza.

Tamaño de la muestra en los estimadores de razón

El procedimiento para obtener el tamaño de la muestra para estimar $R, \quad (Y \quad o \quad \overline{Y})$ en el muestreo aleatorio simple es igual al ya estudiado; fijado el riesgo (α) que estamos dispuestos a afrontar de que el error real supere a **e** en $\Pr\left(\left|\hat{R} - R\right| \geq e\right) = \alpha$ y suponiendo que $\hat{R} \sim n(R, \sigma_R)$,

de $e = t_\alpha \sqrt{\hat{V}(\hat{R})} \rightarrow \frac{e^2}{t_\alpha^2} = \frac{N-n}{nN} \frac{s_d^2}{\overline{X}^2}$ donde $s_d^2 = \frac{\sum (y_i - \hat{R}\, x_i)^2}{n-1}$

Así $n = \frac{n_o}{1 + \frac{n_o}{N}}$ donde $n_o = \frac{t_\alpha^2 s_d^2}{e^2 \overline{X}^2}$

s_d^2 es calculada a priori de acuerdo a información disponible.

Tamaño de la muestra en los estimadores de razón

De igual manera se pueden obtener los tamaños de muestras requeridos para estimar Y o $\overline{Y}$ usando estimadores de razón

Obteniendo $n_o = \frac{t_\alpha^2 s_d^2}{e^2}$ para $\overline{Y}_R$

y $n_o = \frac{N t_\alpha^2 s_d^2}{e^2}$ para Y_R, en este caso $n = \frac{n_o}{1 + \frac{n_o}{N}}$

Eficiencia de los estimadores.

La eficiencia de los estimadores de razón (en el mas), se mide en relación con la de los estimadores directos. En muestras grandes (mas) el estimador de razón del total $\hat{Y}_R$ tiene una varianza menor que el estimador directo $\hat{Y} = N\bar{y}$ si: $\rho > \frac{CV(X)}{2CV(Y)}$ así:

$$V\left(\hat{Y}\right) > V\left(\hat{Y}_R\right) \rightarrow \frac{N^2(1-f)}{n}S_Y^2 > \frac{N^2(1-f)}{n}\left(S_Y^2 + R^2S_X^2 - 2R\rho S_Y S_X\right)$$

$$\rightarrow 2R\rho S_Y S_X > R^2 S_X^2 \quad \text{si} \quad R = \frac{\bar{y}}{\bar{x}} > 0 \quad \rightarrow \rho > \frac{RS_X}{2S_Y} \rightarrow \rho > \frac{CV(X)}{2CV(Y)}$$

Se demuestra que los estimadores de razón son estimadores óptimos dentro de la clase de estimadores insesgados si:

1. Existe una relación lineal entre X,Y que pasa por el origen y se mantiene para “todo” par (X_i, Y_i)
2. La varianza alrededor de la línea recta es proporcional a X_i, es decir, $V(Y/X) = X_i V(\ .\ )$

Se pueden comprobar estas 2 condiciones haciendo un grupo de dispersión (X_i, Y_i) (caso particular $X_t = Y_{t-1}$)

Estimadores de razón en el muestreo estratificado.

En el muestreo estratificado, los estimadores de razón pueden ser más eficientes que los estimadores directos si los tamaños de muestra en cada estrato son lo suficientemente grandes como para que las varianzas sean las apropiadas y se cumple las condiciones de optimización.

Hay dos tipos de estimadores de razón (para Y o $\overline{Y}$) en el muestreo estratificado: de razón separada y de razón combinada.

Estimadores de razón separada

En las estimaciones de razón separada, la razón R_h en cada estrato h de la población se estima por separado, se pondera por los respectivos totales X_h y luego se suma sobre todos los estratos $\hat{Y}_{RS} = \sum \frac{\bar{y}_h}{\bar{x}_h} X_h$

Si se supone que la verdadera razón permanece constante al pasar de un estrato a otro y además se requiere conocer X_h h=1,2..,L

Como la selección en cada estrato es aleatoria simple, independiente y tamaño suficientemente grande se cumple que:

$$V(\hat{Y}_{RS}) > \sum \frac{N_h^2(1-f_h)}{n_h}\left(S_{Y_h}^2 + R_h^2 S_{X_h}^2 - 2R_h\rho_h S_{Y_h} S_{X_h}\right)$$

La prueba es inmediata

Cuando los n_h son pequeños $\hat{Y}_{RS}$ puede poseer un sesgo no despreciable

Estimación de razón combinada

Este estimador consiste en estimar previamente $\hat{Y}_{st}$ y $\hat{X}_{st}$ y luego hacer una sola estimación estimada.

Con los datos muestrales se calcula: $\hat{Y}_{st} = \sum N_h \bar{y}_h$ y $\hat{X}_{st} = \sum N_h \bar{x}_h$

La estimación de razón combinada del total estimado Y es:

$$\hat{Y}_{RC} = \frac{\hat{Y}_{st}}{\hat{Y}_{st}} X = \frac{\bar{y}_{st}}{\bar{x}_{st}} X \qquad \text{donde} \qquad \bar{y}_{st} = \frac{\hat{Y}_{st}}{N} \quad \text{y} \quad \bar{x}_{st} = \frac{\hat{X}_{st}}{N}$$

La estimación de $\hat{Y}_{RC}$ no requiere del conocimiento de los totales por estrato X_h sino del total general X, esta estimación está menos sujeta a riesgo de sesgo que de razón separada.

Se comprueba: $V(\hat{Y}_{RC}) > \sum \frac{N_h^2 (1 - f_h)}{n_h} \left(S_{Y_h}^2 + R^2 S_{X_h}^2 - 2R\rho_h S_{Y_h} S_{X_h} \right)$

para las estimaciones muéstrales de estas varianzas se sustituye los R_h y , y los S_{Y_h} S_{X_h} y S_{XY} por sus respectivas estimaciones muestrales.

Estimadores de regresión

Los estimadores de regresión (al igual que los de razón) se justifican en el incremento de la precisión de las estimaciones, que brindan ante los estimadores directos.

Estos estimadores se utilizan propiamente cuando:

a) la relación entre Y y X es lineal y no pasa por el origen
b) es menos costoso observar x_i que y_i
c) se quieren predecir valores particulares de y_i.

El estimador de $\overline{Y}$ de regresión es: $\overline{y}_{rl} = \overline{y} + b\left(\overline{X} - \overline{x}\right)$

y para el total $\hat{Y}_{rl} = N\overline{y}_{rl}$

donde **b** es una estimación del coeficiente de regresión lineal de y en x.

Estimadores de regresión

El fundamento de esta estimación es que si $\bar{x}$ está por debajo del promedio, también deberíamos esperar que $\bar{y}$ lo estuviera por la cantidad $b\left(\overline{X} - \bar{x}\right)$ debido a la regresión de y_i en x_i. Los estimadores de regresión tienen las mismas propiedades que los de razón, son consistentes en el sentido de cochran pero suelen ser sesgados, pero la razón del sesgo al error estándar se reduce cuando n es grande y en este caso se puede usar la aproximación normal.

b se obtiene por estimaciones muéstrales, o se obtiene de los valores muéstrales o con información previa a la muestra, así, se distinguen dos casos:

1. Estimadores de regresión con "**b**" preasignada
2. Estimadores de regresión con "**b**" calculada en la muestra

Estimador de regresión con b presasignada

Si b_0 es una constante calculada con información pre-muestral, el estimador de regresión lineal es: $\bar{y}_{rl} = \bar{y} + b_o\left(\bar{X} - \bar{x}\right)$
el cual es insesgado con varianza:

$$V\left(\bar{y}_{rl}\right) = \frac{1-f}{n}\left(\sum\left[\left(y_i - \bar{Y}\right) - b_o\left(x_i - \bar{X}\right)\right]^2\right)$$

$$V\left(\bar{y}_{rl}\right) = \frac{1-f}{n}\left(S_Y^2 - 2b_o S_{YX} + b_o^2 S_X^2\right)$$

cuyo estimador insesgado es $\hat{V}\left(\bar{y}_{rl}\right) = \frac{1-f}{n}\left(s_Y^2 - 2b_o s_{YX} + b_o^2 s_X^2\right)$

El valor de b_o que minimiza $\hat{V}\left(\bar{y}_{rl}\right)$ es $b_o = B = \frac{S_{YX}}{S_X S_Y}$ calculado sobre la población.

Estimador de regresión con b estimada en la muestra

Si **b** no es conocida previo a la muestra, debe calcularse con los datos muéstrales por: $$b = \frac{\sum \left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right)}{\sum \left(x_i - \overline{X}\right)^2}$$

b así calculada, es un estimador, una variable aleatoria y para poderlo utilizar adecuadamente debemos exigirle que la relación entre Y y X sea aproximadamente lineal, la varianza alrededor de la línea de regresión sea constante (homoscedasticidad), que el tamaño de la muestra sea grande y el muestreo sea aleatorio simple. En este caso: $\overline{y}_{rl} = \overline{y} + b\left(\overline{X} - \overline{x}\right)$

cuya varianza es $V\left(\overline{y}_{rl}\right) = \frac{1-f}{n} S_Y^2 \left(1 - \rho^2\right)$ donde $\rho = \frac{S_{YX}}{S_X S_Y}$

cuya estimación muestral válida para muestras grandes es:

$$V\left(\overline{y}_{rl}\right) = \frac{1-f}{n(n-2)} \frac{\sum \left[\left(y_i - \overline{y}\right) - b\left(x_i - \overline{x}\right)\right]^2}{\sum \left(x_i - \overline{x}\right)^2}$$

Comparación entre los estimadores de razón, regresión y simples (n grande)

Si n es suficientemente grande de modo que $V(\bar{y}_{rl})$ y $V(\bar{y}_R)$ aproximadas sean válidas, se cumple que bajo ciertas condiciones $V(\bar{y}_{rl}) < V(\bar{y})$ $V(\bar{y}_{rl}) < V(\bar{y}_R)$ y como vimos $V(\bar{y}_R) < V(\bar{y})$

Recordemos $V(\bar{y}_{rl}) = \frac{N-n}{Nn} S_Y^2 (1-\rho^2)$ $V(\bar{y}) = \frac{N-n}{Nn} S_Y^2$

$$V(\bar{y}_R) = \frac{N-n}{Nn}\left(S_Y^2 + R^2 S_X^2 - 2R\rho S_Y S_X\right)$$

Entonces

$V(\bar{y}_{rl}) < V(\bar{y})$ Se cumple para todo ρ y es igual para $\rho = 0$

$V(\bar{y}_{rl}) < V(\bar{y}_R)$ si $-\rho^2 S_Y^2 < R^2 S_X^2 - 2R\rho S_Y S_X$ o equivalente

$$(\rho S_Y - R S_X)^2 > 0 \quad \text{o} \quad (B-R)^2 > 0$$

así la $\bar{y}_{rl}$ es mas precisa que $\bar{y}_R$ a menos que B=R y esto ocurre si la regresión pasa por el origen.