6. ESTIMADORES DE RAZON

6.1 Introduccion

Es muy frecuente que a partir de los datos de una encuesta se requiera la estimación de razones de variables que corresponden a la estructura de una clasificación o unión de categorías. Por ejemplo a un banco le interesa conocer del monto total de créditos que otorga, cuánto corresponde a la compra de automóviles. A un economista le interesa saber la proporción del gasto destinado a alimentos respecto del ingreso total de los hogares. A un demógrafo le interesa conocer la razón de trabajadores que ganan dos salarios mínimos o menos entre todos los que se dedican a la construcción. A un organismo de capacitación agropecuaria le interesa conocer la proporción de la superficie de tierras que permanecen ociosas respecto del la superficie total de tierras cultivables. En todos los casos el parámetro que se desea estimar es una razón:

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$$

La estimación del parámetro se realiza a través de la razón muestral de las sumas para las dos variables o de la razón de sus medias:

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

El cociente de estas dos medias se debe analizar en forma especial, pues los valores de las sumas o promedios muestrales, varían de muestra a muestra. Se tiene el cociente de dos variables aleatorias y el estimador resulta ligeramente sesgado. Por otra parte en el cálculo de la varianza de este estimador hay que considerar la presencia de covarianzas entre las variables que se involucran en el cociente.

La estimación de razones, también es un recurso muy utilizado para lograr estimadores más eficientes cuando se dispone de una variable auxiliar fuertemente correlacionada con la variable objetivo. Por ejemplo, se dispone del número de viviendas censadas de un grupo de localidades y se desea estimar el total de viviendas con niños en edad escolar que requieren de becas.

Para comprender de manera más simple a este tipo de estimador, considérense los datos a nivel de AGEB de población total y población económicamente activa (PEA) de la Delegación de Coyoacán según el censo del 2000. La población total de las 153 AGEBs es 643,623 personas y de ellas 287,911 pertenecen a la PEA. La razón de la PEA a la población total es R=0.4495638. Ambas variables están fuertemente correlacionadas. Su correlación es r = 0.993112 .

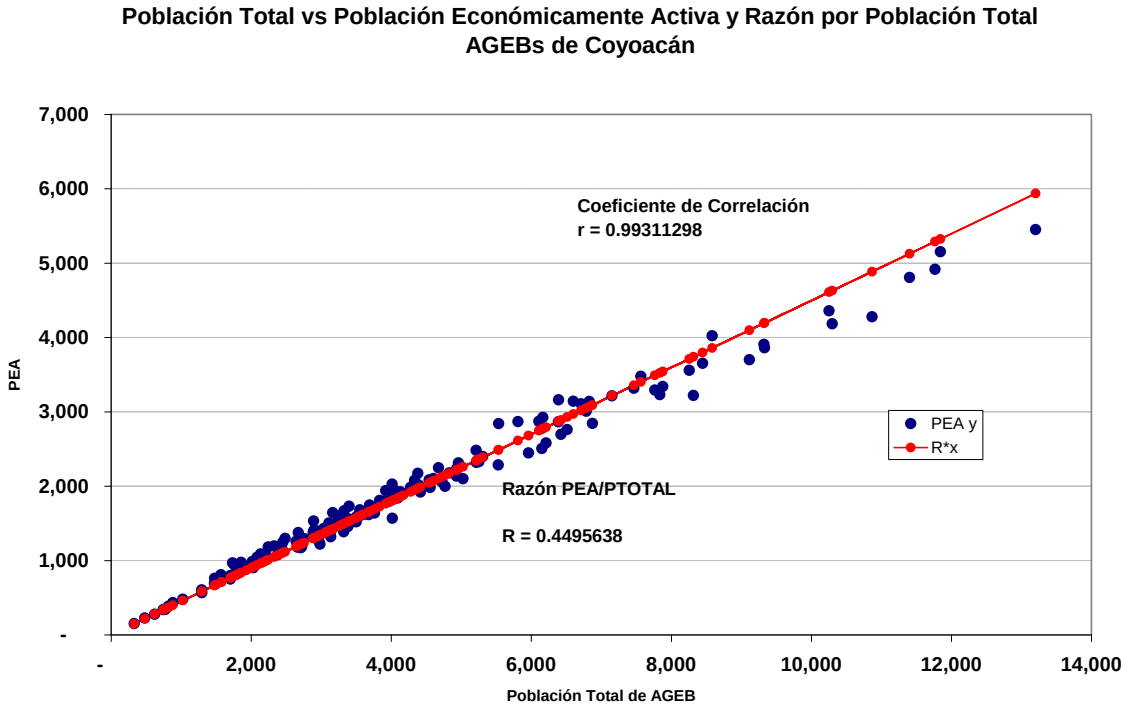
6.2 Interpretación Geométrica del Estimador de Razón.

La razón R se interpreta geométricamente como la pendiente de una recta que pasa por el origen y que describe, en nuestro ejemplo, una relación lineal entre la población total y la PEA. Puesto que la R no es constante para todas las AGEBs, hay que considerar un término de error y el modelo forzado a pasar por el origen se expresa como sigue:

$$y = Rx + \varepsilon$$

Se procede a continuación a estimar el valor de Y correspondiente a cada valor de X observados como el producto de la razón por la población total en cada AGEB.

$\hat{y}_i = Rx_i$ para i = 1,2,……,n. Los valores observados y estimados se presentan en el siguiente gráfico de dispersión:



De acuerdo con el modelo expuesto, es factible obtener un estimador de la razón por mínimos cuadrados ordinarios para el modelo forzado al origen, el cual adopta la siguiente forma:

$$\hat{R} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Pero la varianza de Y_i dado un valor de X_i adopta la forma $\sigma^2 x_i$, esto es, es proporcional a X_i. Para evitar este inconveniente se utilizan mínimos cuadrados ponderados y se adopta la variable

transformada $Y_i / \sqrt{X_i}$, cuya varianza σ^2 no depende de X_i. Así entonces, se minimiza la suma de cuadrados ponderada y estimador resulta ser de mínimos cuadrados ponderados :

$$\sum_{i=1}^{n}\left(y_i - Rx_i\right)^2 \frac{1}{x_i}$$

$$\frac{d}{dR}\sum_{i=1}^{n}\left(y_i - Rx_i\right)^2 \frac{1}{x_i} = -2\sum_{i=1}^{n}\left(x_i y_i - Rx_i^2\right)\frac{1}{x_i} = 0$$

De donde se obtiene el estimador de razón.

$$\sum_{i=1}^{n}\left(y_i - Rx_i\right) = 0 \quad \text{y por lo tanto} \quad \hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

6.3 Sesgo y Varianza del Estimador de Razón.

El estimador de la razón expuesto no es un estimador insesgado, pero su sesgo disminuye sensiblemente para valores moderadamente altos de la muestra n. Se procederá, sin embargo, a un análisis más detallado del sesgo.

Primero considérense las siguientes igualdades:

$$\bar{y} = \bar{Y} + \left(\bar{y} - \bar{Y}\right) = \bar{Y}\left(1 + \frac{\left(\bar{y} - \bar{Y}\right)}{\bar{Y}}\right)$$

$$\bar{x} = \bar{X} + \left(\bar{x} - \bar{X}\right) = \bar{X}\left(1 + \frac{\left(\bar{x} - \bar{X}\right)}{\bar{X}}\right)$$

Entonces el estimador de razón adopta la siguiente forma:

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\bar{Y}}{\bar{X}}\left(1 + \frac{\left(\bar{y} - \bar{Y}\right)}{\bar{Y}}\right)\left(1 + \frac{\left(\bar{x} - \bar{X}\right)}{\bar{X}}\right)^{-1} = R\left(1 + \frac{\left(\bar{y} - \bar{Y}\right)}{\bar{Y}}\right)\left(1 + \frac{\left(\bar{x} - \bar{X}\right)}{\bar{X}}\right)^{-1}$$

El segundo paréntesis se puede expresar en términos de un desarrollo en serie de Taylor

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = R\left(1 + \frac{\left(\bar{y} - \bar{Y}\right)}{\bar{Y}}\right)\left(1 - \frac{\left(\bar{x} - \bar{X}\right)}{\bar{X}} + \frac{\left(\bar{x} - \bar{X}\right)^2}{\bar{X}^2} - \text{.....}\right)$$

De donde al efectuar el producto se obtiene:

$$\hat{R}=\frac{y}{x}=R\left(1+\frac{(y-\bar{Y})}{\bar{Y}}-\frac{(x-\bar{X})}{\bar{X}}+\frac{(x-\bar{X})^2}{\bar{X}^2}-\frac{(y-\bar{Y})(x-\bar{X})}{\bar{Y}\bar{X}}.....\right)$$

A continuación se toma valor esperado de $\hat{R}$ para los primeros 5 términos de la serie

$$E(\hat{R})=E\left(R+R\frac{(y-\bar{Y})}{\bar{Y}}-R\frac{(x-\bar{X})}{\bar{X}}+R\frac{(x-\bar{X})^2}{\bar{X}^2}-R\frac{(y-\bar{Y})(x-\bar{X})}{\bar{Y}\bar{X}}.....\right)$$

$$=\left(R+R\frac{E(y-\bar{Y})}{\bar{Y}}-R\frac{E(x-\bar{X})}{\bar{X}}+R\frac{E(x-\bar{X})^2}{\bar{X}^2}-R\frac{E(y-\bar{Y})(x-\bar{X})}{\bar{Y}\bar{X}}.....\right)$$

$$=\left(R+R(0)-R(0)+\frac{R}{\bar{X}^2}V(\bar{x})-\frac{R}{\bar{Y}\bar{X}}Cov(y,\bar{x}).....\right)$$

$$\approx R+\frac{R}{\bar{X}^2}\left(\frac{N-n}{N}\right)\frac{S_x^2}{n}-\frac{R}{\bar{Y}\bar{X}}\left(\frac{N-n}{N}\right)\rho\frac{S_yS_x}{n}$$

$$\approx R+\frac{R}{\bar{X}^2}\left(\frac{N-n}{N}\right)\frac{S_x^2}{n}-\frac{1}{\bar{Y}\bar{X}}\frac{\bar{Y}}{\bar{X}}\left(\frac{N-n}{N}\right)\rho\frac{S_yS_x}{n}$$

$$\approx R+\frac{R}{\bar{X}^2}\left(\frac{N-n}{N}\right)\frac{S_x^2}{n}-\frac{1}{\bar{X}^2}\left(\frac{N-n}{N}\right)\rho\frac{S_yS_x}{n}$$

$$\approx R+\left(\frac{N-n}{N}\right)\frac{1}{n\bar{X}^2}\left[RS_x^2-\rho S_yS_x\right]$$

Entonces el sesgo aproximado del estimador es:

$$E(\hat{R})-R\approx\left(\frac{N-n}{N}\right)\frac{1}{n\bar{X}^2}\left[RS_x^2-\rho S_yS_x\right]$$

El sesgo se anula si se cumple la siguiente condición o se reduce en la medida que sea mínima la diferencia.

$$RS_x-\rho S_y=0$$

Varianza del estimador de la razón:

La varianza del estimador de razón se puede interpretar como la dispersión de rectas muestrales en torno a la recta poblacional.

$$V(\hat{R})=E(\hat{R}-R)^2$$

Si se supone que la diferencia entre la media muestral de x y la media poblacional de la misma variable es pequeña, entonces $x = \bar{X}$ y la diferencia entre estimador de la razón y el parámetro se puede expresar:

$$\hat{R} - R = \frac{y - Rx}{\bar{X}}$$

$$= \frac{1}{\bar{X}}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - Rx_i)\right]$$

Si ahora se definen nuevas variables como diferencias:

$$d_i = y_i - Rx_i$$

La expresión anterior se expresa como promedio de diferencias.

$$= \frac{1}{\bar{X}}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - Rx_i)\right] = \frac{1}{\bar{X}}\frac{\sum_{i=1}^{n} d_i}{n}$$

La media parametral de las variables de diferencias es nula

$$\bar{D} = \frac{\sum_{i=1}^{N} d_i}{N} = \frac{1}{N}\sum_{i=1}^{N}(y_i - Rx_i) = \bar{Y} - R\bar{X} = 0$$

Entonces la varianza de la razón se puede expresar como la varianza de la media muestral de las diferencias

$$E(\hat{R} - R)^2 = E\left[\frac{1}{\bar{X}}(\bar{d} - \bar{D})\right]^2$$

$$= \frac{1}{\bar{X}^2}\left(\frac{N-n}{N}\right)\frac{\sum_{i=1}^{N}(d_i - \bar{D})^2}{n(N-1)} \qquad \text{Se define} \qquad S_d^2 = \frac{\sum_{i=1}^{N}(d_i - \bar{D})^2}{N-1}$$

$$= \frac{1}{\bar{X}^2}\left(\frac{N-n}{N}\right)\frac{S_d^2}{n}$$

De donde se concluye

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

Una forma alternativa de la varianza de uso frecuente se obtiene en función del coeficiente de correlación. Se parte de una expresión alternativa de la suma de cuadrados.

$$\sum_{i=1}^{N}\left(y_i-\hat{R}x_i\right)^2=\sum_{i=1}^{N}\left(y_i-Rx_i\right)-\left(\overline{Y}-R\overline{X}\right))^2 \qquad \text{Por construcción del parámetro} \quad \overline{Y}-R\overline{X}=0$$

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

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

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

$$=(N-1)\left[S_y^2+R^2S_x^2-2RCov(x,y)\right]=(N-1)\left[S_y^2+R^2S_x^2-2R\rho S_yS_x\right]$$

Por lo tanto la varianza de la razón se expresa alternativamente:

$$V\left(\hat{R}\right)=\frac{1}{\overline{X}^2}\left(\frac{N-n}{N}\right)\frac{1}{n}\left[S_y^2+R^2S_x^2-2R\rho S_yS_x\right]$$

La varianza del estimador de razón se puede estimar con la siguiente fórmula sustituyendo el parámetro por su estimación a partir de la muestra:

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

En forma alternativa, en este caso, en función del coeficiente de correlación, varianzas y desviaciones estándares muestrales:

$$\hat{V}\left(\hat{R}\right)=\frac{1}{\overline{X}^2}\left(\frac{N-n}{N}\right)\frac{1}{n}\left[\hat{S}_y^2+\hat{R}^2\hat{S}_x^2-2\hat{R}\hat{\rho}\hat{S}_y\hat{S}_x\right]$$

6.4 Tamaño de Muestra

Se parte de la expresión de la varianza en función de la S_d^2 de las desviaciones:

$$V(\hat{R})=\frac{1}{\overline{X}^2}\left(\frac{N-n}{N}\right)\frac{S_d^2}{n}$$

Apoyados en el supuesto de normalidad, la varianza se iguala al cociente del cuadrado de la precisión deseada δ entre el cuadrado del coeficiente de confianza Z correspondiente al valor percentilar (1-α/2) de la normal estándar. Este cociente se identifica como varianza deseada D^2.

$$D^2=\frac{\delta^2}{Z^2_{1-\alpha/2}}=\frac{1}{\overline{X}^2}\left(\frac{N-n}{N}\right)\frac{S_d^2}{n}$$

A continuación se despeja el tamaño global de muestra n

$$n=\frac{Z^2_{1-\alpha/2}NS_d^2}{\delta^2\overline{X}^2N+Z^2_{1-\alpha/2}S_d^2}$$

Se divide numerador y denominador entre $\delta^2X^2N+Z^2_{1-\alpha/2}$

$$n=\frac{\frac{Z^2_{1-\alpha/2}S_d^2}{\delta^2\overline{X}^2}}{1+\frac{Z^2_{1-\alpha/2}S_d^2}{\delta^2\overline{X}^2N}}$$ Se define $n_o=\frac{Z^2_{1-\alpha/2}S_d^2}{\delta^2\overline{X}^2}$ el tamaño de muestra para poblaciones no acotadas.

El tamaño de muestra final queda en función de la n_o y del tamaño de la población.

$$n=\frac{n_o}{1+\frac{n_o}{N}}$$

6.5 Efecto de Diseño del Estimador de Razón.

Se ha mencionado que al disponer de una variable auxiliar fuertemente correlacionada con la variable objetivo, se logra una mejor estimación, pero ¿qué tanto se reduce la varianza ante diferentes valores de la correlación? La forma más sencilla de valorarlo es comparar la varianza del estimador de la media utilizando un estimador de razón con la varianza de la media utilizando el estimador simple.

$$V(\hat{R})=\frac{1}{\overline{X}^2}\left(\frac{N-n}{N}\right)\frac{1}{n}\left[S_y^2+R^2S_x^2-2R\rho S_yS_x\right]$$

Si estimamos la media a través del estimador de razón se expresa como producto del estimador por la media de la variable auxiliar.

$$\hat{Y}_R=\overline{X}\hat{R} \qquad V(\hat{Y}_R)=\overline{X}^2V(\hat{R})$$

En consecuencia la varianza de la media a través del estimador de razón será:

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

Se distribuyen los términos en N y n

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

Dentro de los corchetes se obtiene la varianza de la media por muestreo aleatorio simple.

$$V(\hat{Y}_R) = \left[V(\hat{Y}_{MAS}) + R^2 \left(\frac{N-n}{N} \right) \frac{1}{n} S_x^2 - 2 \left(\frac{N-n}{N} \right) \frac{1}{n} R \rho S_y S_x \right]$$

Se expresa la diferencia entre las varianzas de la media obtenida a través del estimador de razón y el obtenido por muestreo aleatorio simple.

$$V(\hat{Y}_R) - V(\hat{Y}_{MAS}) = R^2 \left(\frac{N-n}{N} \right) \frac{1}{n} S_x^2 - 2 \left(\frac{N-n}{N} \right) \frac{1}{n} R \rho S_y S_x$$

Entonces para que la diferencia entre la varianza de la media a través del estimador de razón y el estimador simple sea negativa se debe cumplir:

$$R^2 \left(\frac{N-n}{N} \right) \frac{1}{n} S_x^2 - 2 \left(\frac{N-n}{N} \right) \frac{1}{n} R \rho S_y S_x < 0$$

De donde se obtiene la condición

$$R^2 \left(\frac{N-n}{N} \right) \frac{1}{n} S_x^2 < 2 \left(\frac{N-n}{N} \right) \frac{1}{n} R \rho S_y S_x$$

Al simplificar y expresar la condición para considerar mejor el estimador de razón en función del coeficiente de correlación y de los coeficientes de variación de X,Y se tiene:

$$RS_x < 2 \rho S_y$$

$$\rho > \frac{1}{2} \frac{RS_x}{S_y} = \frac{1}{2} \frac{\bar{Y}}{\bar{X}} \frac{S_x}{S_y}$$

$$\rho > \frac{1}{2} \frac{CV(X)}{CV(Y)}$$

Si se procede en forma análoga con el cociente de la varianza del estimador de razón entre la varianza del estimador simple se obtiene el efecto de diseño:

$$Deff = \frac{V(\hat{Y}_R)}{V(\hat{Y}_{MAS})} = 1 + \frac{CV(X)}{CV(Y)} \left[\frac{CV(X)}{CV(Y)} - 2\rho \right]$$

Entonces si el coeficiente de variación de X es menor o igual que el coeficiente de variación de Y y el coeficiente de correlación es mayor o igual a 0.5, el estimador de razón resulta más eficiente.

6.6 Estimador de Razón en Muestreo Estratificado.

Si la población está estratificada, existen dos alternativas para estimar la razón:

a) Estimador de Razón Combinado

En este caso se estiman las medias poblacionales para ambas variables en forma independiente a partir de las medias por estrato y la razón se calcula como el cociente de estas dos estimaciones:

$$\bar{y} = \sum_{h=1}^{L} W_h \bar{y}_h \qquad \bar{x} = \sum_{h=1}^{L} W_h \bar{x}_h$$

$$\hat{R}_C = \frac{\sum_{h=1}^{L} W_h \bar{y}_h}{\sum_{h=1}^{L} W_h \bar{x}_h} = \frac{\bar{y}_{st}}{\bar{x}_{st}}$$

Cuya varianza se puede aproximar con el supuesto usual de igualdad $x_{st} = \bar{X}_{st}$ de la siguiente forma:

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2}\left[V(y_{st}) + R^2 V(x_{st}) - 2RCov(y_{st}, x_{st})\right]$$

Tamaño de Muestra.

Se procede a desagregar la fórmula para expresar la varianza en términos de $\mathbf{n_h}$

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2}\left[\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{S_{h,y}^2}{n_h} + R^2 \sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{S_{h,x}^2}{n_h} - 2R\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{Cov(y_{hi}, x_{hi})}{n_h}\right]$$

Al factorizar y simplificar se llega a la siguiente fórmula:

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2}\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{1}{n_h}\left[S_{h,y}^2 + R^2 S_{h,x}^2 - 2RCov(y_{hi}, x_{hi})\right]$$

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2}\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{\sum_{i=1}^{N_h}\left(y_{hi} - Rx_{hi}\right)^2}{n_h\left(N_h - 1\right)}$$

Afijación de la Muestra para Razones

Nuevamente partamos de la primera expresión de la varianza para involucrar los criterios de afijación de la muestra.

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2}\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{1}{n_h}\left[S_{h,y}^2 + R^2 S_{h,x}^2 - 2RCov(y_{hi}, x_{hi})\right]$$

Por simplificación se designará: $S_{dh}^2 = S_{h,y}^2 + R^2 S_{h,x}^2 - 2RCov(y_{hi}, x_{hi}) = \sum_{i=1}^{N_h}\left(y_{hi} - Rx_{hi}\right)^2/(N_h - 1)$

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2}\sum_{h=1}^{L} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\frac{S_{dh}^2}{n_h}$$

$$= \frac{1}{\bar{X}^2} \sum_{h=1}^{L} \frac{N_h^2}{N^2} \left(\frac{N_h - n_h}{N_h} \right) \frac{S_{dh}^2}{n_h}$$

$$= \frac{1}{\bar{X}^2} \sum_{h=1}^{L} \frac{N_h}{N^2} (N_h - n_h) \frac{S_{dh}^2}{n_h}$$

Finalmente se tiene una forma general de la varianza del estimador de razón combinado.

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2 N^2} \left[\sum_{h=1}^{L} \frac{N_h^2 S_{dh}^2}{n_h} - \sum_{h=1}^{L} N_h S_{dh}^2 \right]$$

Se adopta esta expresión y a continuación se define una función de $\mathbf{n_h}$ con la adición de un multiplicador de Lagrange aplicado la restricción del tamaño de muestra y se deriva respecto de una $\mathbf{n_h}$ para obtener la expresión para afijación de muestra con el criterio de Neyman.

$$\Phi(n_h) = \frac{1}{\bar{X}^2 N^2} \left[\sum_{h=1}^{L} \frac{N_h^2 S_{dh}^2}{n_h} - \sum_{h=1}^{L} N_h S_{dh}^2 \right] + \lambda \left(\sum_{h=1}^{L} n_h - n \right)$$

$$\frac{d\Phi(n_h)}{dn_h} = -\frac{N_h^2 S_{dh}^2}{\bar{X}^2 N^2 n_h^2} + \lambda = 0$$

Se despeja $\mathbf{n_h}$ de la anterior expresión

$$n_h = \frac{N_h S_{dh}}{\bar{X} N \sqrt{\lambda}} \quad \text{..... (a)}$$

Considerando que siempre se cumple $\sum_{h=1}^{L} n_h = n$

$$n = \sum_{h=1}^{L} \frac{N_h S_{dh}}{\bar{X} N \sqrt{\lambda}} \quad \text{.....(b)}$$

De esta expresión se despeja $\sqrt{\lambda}$ de (b) y se sustituye en (a) para obtener la expresión de $\mathbf{n_h}$

$$\sqrt{\lambda} = \sum_{h=1}^{L} \frac{N_h S_{dh}}{\bar{X} N n}$$

Por lo tanto se obtiene finalmente la forma de $\mathbf{n_h}$ para afijación de Neyman.

$$n_h = \frac{N_h S_{dh}}{\sum_{h=1}^{L} N_h S_{dh}} n$$

Cálculo de la Varianza del Estimador de Razón Combinado con Afijación de Neyman.

S comienza por sustituir la $\mathbf{n_h}$ encontrada en la forma general de la varianza

$$V(\hat{R}_C) = \frac{1}{\bar{X}^2 N^2}\left[\sum_{h=1}^{L} \frac{N_h^2 S_{dh}^2}{n_h} - \sum_{h=1}^{L} N_h S_{dh}^2\right]$$

Finalmente la fórmula para la varianza de la razón bajo el supuesto de afijación de Neyman

$$V(\hat{R}_{Ney}) = \frac{1}{\bar{X}^2 N^2}\left(\sum_{h=1}^{L} N_h S_{dh}\right)^2 \frac{1}{n} - \frac{1}{\bar{X}^2 N^2}\sum_{h=1}^{L} N_h S_{dh}^2$$

Tamaño General de Muestra

Esta expresión de varianza se iguala a una varianza deseada $\mathbf{D^2}$ y se despeja n, se dispondrá entonces de la fórmula para calcular un tamaño de muestra global bajo afijación de Neyman.

$$n = \frac{\left(\sum_{h=1}^{L} N_h S_{dh}\right)^2}{D^2 \bar{X}^2 N^2 + \sum_{h=1}^{L} N_h S_{dh}^2}$$

En caso de utilizar afijación de muestra proporcional al tamaño del estrato, por un procedimiento análogo se obtienen las siguientes fórmulas para afijar la muestra, calcular varianzas y determinar tamaño global de la muestra:

$$n_h = \frac{N_h}{N} n$$

$$V(\hat{R}_{Prop}) = \frac{1}{\bar{X}^2 N}\sum_{h=1}^{L} \frac{N_h S_{dh}^2}{n} - \frac{1}{\bar{X}^2 N^2}\sum_{h=1}^{L} N_h S_{dh}^2$$

$$n = \frac{N\sum_{h=1}^{L} N_h S_{dh}^2}{D^2 \bar{X}^2 N + \sum_{h=1}^{L} N_h S_{dh}^2}$$

b) Estimador de Razón Separado.

El estimador global de razón se obtiene como la suma ponderada de las estimaciones separadas de las razones de los estratos. Es de utilidad cuando los estratos son dominios de estudio y se requieren estimaciones de razones separadas para cada estrato, las cuales pueden diferir sensiblemente.

$$\hat{R}_S = \sum_{h=1}^{L} W_h \hat{R}_h \quad \text{Donde} \quad \hat{R}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{\sum_{i=1}^{n_h} x_{hi}} = \frac{\bar{y}_h}{\bar{x}_h}$$

Puesto que las muestras son independientes

$$V(\hat{R}_S) = \sum_{h=1}^{L} W_h^2 V(\hat{R}_h)$$

Donde

$$V(\hat{R}_h) = \frac{1}{\bar{X}_h^{\,2}} \left(\frac{N_h - n_h}{N_h} \right) \frac{1}{n_h} \left[S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h \rho_h S_{hy} S_{hx} \right]$$

Como se puede observar, la varianza depende de las varianzas, razones y los coeficientes de correlación de cada estrato.

$$V(\hat{R}_S) = \sum_{h=1}^{L} W_h^2 \frac{1}{\bar{X}_h^{\,2}} \left(\frac{N_h - n_h}{N_h} \right) \frac{1}{n_h} \left[S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h \rho_h S_{hy} S_{hx} \right]$$

El sesgo del estimador de razón, que es muy pequeño en la estimación del estimador combinado, puede ser de magnitud peligrosa al estimar cada razón por separado, pues se tiene una acumulación de sesgos en lugar de uno sólo, sobretodo ante la presencia de muchos estratos con pequeños tamaños de muestra. Si se toma la diferencia de varianzas entre ambos estimadores, se pude observar que coinciden, solamente en el caso de que medias, razones y coeficientes de correlación de los estratos, sean iguales a las globales.

$$V(\hat{R}_C) - V(\hat{R}_S) = \sum_{h=1}^{L} \left[\frac{1}{\bar{X}^{\,2}} - \frac{1}{\bar{X}_h^{\,2}} \right] \left(\frac{N_h - n_h}{N_h} \right) \frac{1}{n_h} \left[S_{hy}^2 + \left(R_C - R_h^2 \right) S_{hx}^2 - 2\left(R_C \rho_h - R_h \rho_h \right) S_{hy} S_{hx} \right]$$

En cuanto a cálculos de tamaño de muestra globales y afijación de muestra, se pueden utilizar los mismos resultados vistos para el estimador combinado, excepto que la S^2_{dh} adopta la siguiente forma al estar en función de la razón en cada estrato y no en función de la razón global:

$$S_{dh}^2 = S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h \rho_h S_{hy} S_{hx} = \sum_{i=1}^{N_h} \left(y_{hi} - R_h x_{hi} \right)^2 / (N_h - 1)$$

6.7 Estimador Insesgado de la Razón.

H. Hartley y A. Ross publicaron en 1954 en la revista *Nature* un artículo sobre estimadores insesgados de razón. Su método parte del estimador sesgado calculado como promedio de razones elemento a elemento a partir de una muestra aleatoria simple.

$$\hat{\bar{R}} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}$$

Se verifica a continuación que éste es un estimador sesgado.

$$E\left(\hat{\bar{R}}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}\right)$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{N}\frac{Y_i}{X_i}W_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{N}\frac{Y_i}{X_i}E(W_i)$$

$$= \frac{1}{n}\sum_{i=1}^{N}\frac{Y_i}{X_i}\frac{n}{N}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i}$$ Pero este cociente es en general diferente de $R = \frac{\sum_{i=1}^{N}Y_i}{\sum_{i=1}^{N}X_i}$

El sesgo del estimador promedio de razones queda expresado de la manera siguiente:

$$Sesgo = E\left(\hat{\bar{R}}\right) - \frac{\bar{Y}}{\bar{X}}$$

Por otra parte, un estimador insesgado de

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i} - \frac{\bar{Y}}{\bar{X}}$$

$$= \frac{\bar{X}\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i} - \bar{Y}}{\bar{X}}$$

$$\frac{1}{N-1}\sum_{i=1}^{N}R_i\left(X_i - \bar{X}\right)$$

es (Teorema 2.3 Cochran)

$$\frac{1}{n-1}\sum_{i=1}^{n}R_i\left(x_i - x\right)$$

$$= \frac{\bar{X}\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i}X_i}{\bar{X}}$$

$$= \frac{-\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{X_i}\left(X_i - \bar{X}\right)}{\bar{X}}$$

$$= -\frac{1}{N\bar{X}}\sum_{i=1}^{N}R_i\left(X_i - \bar{X}\right)$$

$$= -\frac{N-1}{N\bar{X}}\frac{1}{N-1}\sum_{i=1}^{N}R_i\left(X_i - \bar{X}\right)$$

$$= \frac{1}{n-1}\frac{n}{n}\sum_{i=1}^{n}R_i\left(x_i - \bar{x}\right)$$

$$= \frac{n}{n-1}\sum_{i=1}^{n}\left(y - \hat{\bar{R}}\bar{x}\right)$$

Entonces un estimador insesgado del sesgo, aunque la frase suene redundante, combinando ambos resultados es:

$$= -\frac{N-1}{N\bar{X}}\frac{n}{n-1}\left(\bar{y} - \hat{\bar{R}}\bar{x}\right)$$

El estimador insesgado de Hartley y Ross se expresa finalmente como la siguiente diferencia:

$$\hat{R}_{HR} = \hat{\bar{R}} - \frac{N-1}{N\bar{X}}\frac{n}{n-1}\left(y - \hat{\bar{R}}\bar{x}\right)$$

Existen otras propuestas para estimadores insesgados y para funciones de razones, tales como cocientes de razones, productos y diferencias.