7. ESTIMADORES CON PROBABILIDADES PROPORCIONALES AL TAMAÑO (PPT) CON REEMPLAZO.

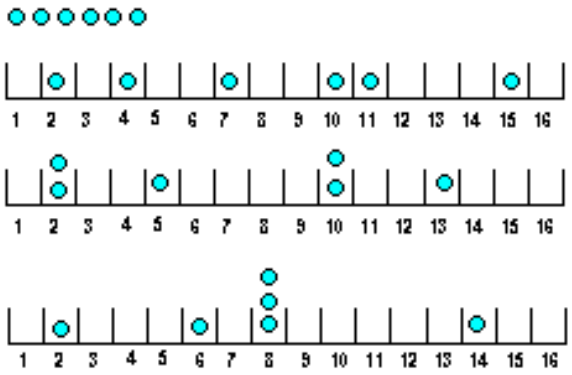
7.1 Introducción

Supongamos que se desea seleccionar una muestra de empresas dedicadas al ramo textil para conocer el valor de la producción en el ramo. Si se dispone de una lista de empresas se puede proceder a seleccionar una muestra aleatoria a partir de la lista. Las empresas de este ramo, en forma semejante a lo que sucede con otros ramos, suelen tener distribuciones muy asimétricas en lo que se refiere a su tamaño. Las empresas de gran tamaño y medianas, que de hecho son las que dominan el mercado, son poco numerosas y muy numerosas las pequeñas y microempresas. En consecuencia una muestra aleatoria de la lista estaría dominada por empresas pequeñas y es factible que ninguna de las 10 empresas más grandes apareciera en muestra. Desde luego, si se dispone de datos sobre su capital social o del número de trabajadores que trabajan en ellas, es posible adoptar esa información para estratificarlas o efectuar una estimación indirecta vía un estimador de razón, pues lógicamente estas variables se espera que guarden una fuerte correlación positiva con la producción. Otra alternativa frecuentemente utilizada es adoptar la variable como medida de tamaño y utilizarla para definir probabilidades de selección proporcionales a esa medida de tamaño (PPT). Esto es las probabilidades de selección serán desiguales.

7.2 Estimador de Hansen y Hurwitz

El diseño de muestras con reemplazo y probabilidades desiguales fue propuesto inicialmente por Hansen y Hurwitz y se plantea en los siguientes términos:

Una vez definido un tamaño de muestra n, si la selección se hace con reemplazo, cada una de las N unidades en la población puede ser seleccionada 0,1,2,……n veces. Esta situación es análoga a tener una serie de N cajas y arrojarles n bolas. En cada caja puede suceder que no caiga una sola bola, que caiga una o más de 1. El caso extremo sería que las n bolas arrojadas cayeran en la misma caja. En otra perspectiva, cada una de las N cajas puede acumular 0,1,2,…..n bolas.



Sea Pi la probabilidad de que una bola caiga en la caja i-ésima en cada evento de arrojar una bola y sea Xi el número de bolas acumuladas en cada caja. La distribución conjunta de las Xi corresponde a una multinomial:

$$f\left(x_1, x_2, \ldots\ldots x_N\right) = \frac{n}{x_1! x_2! \ldots\ldots x_N!} P_1^{x_1} P_2^{x_2} \ldots\ldots P_N^{x_N}$$

La función de probabilidad multinomial tiene los siguientes valores esperados:

$$E(X_i) = nP_i \qquad V(X_i) = nP_i(1-P_i) \qquad Cov(X_i, X_j) = -nP_iP_j$$

El estimador del total para la variable Y propuesto por Hansen y Hurwitz adopta la siguiente forma:

$$\hat{Y}_{HH} = \sum_{i=1}^{N} \frac{Y_i x_i}{E(X_i)}$$

$$= \sum_{i=1}^{N} \frac{Y_i x_i}{nP_i}$$

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

Una vez eliminadas las unidades no seleccionadas y con inclusión de posibles repeticiones de las seleccionadas.

Considérese el caso particular en el que $P_i = \frac{1}{N}$, esto es que las probabilidades para todas las cajas son homogéneas. El estimador adoptará entonces la forma del conocido estimador del total:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{P_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{1/N} = \frac{N}{n} \sum_{i=1}^{n} Y_i = Ny$$

Si Yi tiene probabilidades que guardan una relación de proporcionalidad con el total, esto es PiY=Yi, entonces el estimador coincide con el parámetro para cualquier muestra.

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{P_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{Y_i/Y} = \frac{1}{n} \sum_{i=1}^{n} Y = Y$$

Esta propiedad sugiere que si se tiene una variable de tamaño correlacionada con la variable objetivo que guarde cierta relación de proporcionalidad con la variable objetivo y como consecuencia con una fuerte correlación positiva, entonces esa variable de tamaño se puede utilizar para definir probabilidades proporcionales al tamaño y que redundaría en una mejor estimación del total que con una muestra aleatoria simple.

7.3 Varianza del Estimador de Hansen y Hurvitz

La varianza del estimador de Hansen y Hurvitz se obtiene a continuación:

$$V(\hat{Y}_{HH}) = E(\hat{Y}_{HH} - Y)^2 = E(\hat{Y}^2) - Y^2$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{N}\frac{Y_i X_i}{P_i}\right)^2 - Y^2 = E\left(\frac{1}{n^2}\sum_{i=1}^{N}\frac{Y_i^2 X_i^2}{P_i^2} + \frac{2}{n^2}\sum_{i}^{N-1}\sum_{<j}^{N}\frac{Y_i X_i}{P_i}\frac{Y_j X_j}{P_j}\right) - Y^2 =$$

$$= \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 E(X_i^2) + \frac{2}{n^2}\sum_{i}^{N-1}\sum_{<j}^{N}\frac{Y_i Y_j}{P_i P_j}E(X_i X_j) - Y^2 =$$

Ahora se vuelve la atención a las propiedades de la multinomal:

$$V(X_i) = E(X_i^2) - E(X_i)^2 = E(X_i^2) - n^2 P_i^2 = nP_i(1-P_i)$$

De donde

$$E(X_i^2) = nP_i(1-P_i) + n^2 P_i^2 = nP_i - nP_i^2 + n^2 P_i^2$$

$$Cov(X_i, X_j = E((X_i - E(X_i)(X_i - E(X_i)) = E(X_i X_j) - E(X_i)E(X_j) = E(X_i X_j) - n^2 P_i P_j = -nP_i P_j$$

De donde

$$E(X_i X_j) = n^2 P_i P_j - nP_i P_j$$

Por lo tanto al sustituir en la expresión de la varianza

$$V(\hat{Y}_{HH}) = \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 E(X_i^2) + \frac{2}{n^2}\sum_{i}^{N-1}\sum_{<j}^{N}\frac{Y_i Y_j}{P_i P_j}E(X_i X_j) - Y^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 \left(nP_i - nP_i^2 + n^2 P_i^2\right) + \frac{2}{n^2}\sum_{i}^{N-1}\sum_{<j}^{N}\frac{Y_i Y_j}{P_i P_j}\left(n^2 P_i P_j - nP_i P_j\right) - Y^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 nP_i - \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 nP_i^2 + \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 n^2 P_i^2 + \frac{2}{n^2}\sum_{i}^{N-1}\sum_{<j}^{N}\frac{Y_i Y_j}{P_i P_j}n^2 P_i P_j - \frac{2}{n^2}\sum_{i}^{N-1}\sum_{<j}^{N}\frac{Y_i Y_j}{P_i P_j}nP_i P_j - Y^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 nP_i - \frac{1}{n}\sum_{i=1}^{N}Y_i^2 + \sum_{i=1}^{N}Y_i^2 + 2\sum_{i}^{N-1}\sum_{<j}^{N}Y_i Y_{j\,j} - \frac{2}{n}\sum_{i}^{N-1}\sum_{<j}^{N}Y_i Y_j - Y^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 nP_i - \frac{1}{n}\sum_{i=1}^{N}Y_i^2 + Y^2 - \frac{2}{n}\sum_{i}^{N-1}\sum_{<j}^{N}Y_i Y_j - Y^2$$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 P_i - \frac{1}{n}\sum_{i=1}^{N} Y_i^2 - \frac{2}{n}\sum_{i}^{N-1}\sum_{<j}^{N} Y_i Y_j$$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 P_i - \frac{1}{n}\left(\sum_{i=1}^{N} Y_i^2 + 2\sum_{i}^{N-1}\sum_{<j}^{N} Y_i Y_j\right)$$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i}\right)^2 P_i - \frac{Y^2}{n}$$

De donde finalmente se tiene la varianza:

$$V(\hat{Y}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y\right)^2 P_i$$

Si se torna al estimador con probabilidades iguales Pi= 1/N y con reemplazo, su varianza estaría dada por:

$$V\left(\hat{Y}_{HH}\right) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{1/N} - Y\right)^2 \frac{1}{N}$$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(NY_i - Y\right)^2 \frac{1}{N}$$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(NY_i - Y\right)^2 \frac{1}{N}$$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(NY_i - \frac{NY}{N}\right)^2 \frac{1}{N}$$

$$= \frac{1}{n}\sum_{i=1}^{N}\left(Y_i - \frac{Y}{N}\right)^2 \frac{N^2}{N}$$

$$= \frac{N^2}{n}\frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}{N} = N^2\frac{\sigma^2}{n}$$

Varianza que resulta familiar, ya que corresponde al estimador del total para una muestra aleatoria con reemplazo.

Un estimador insesgado de la varianza del estimador de Hansen y Hurwitz se calcula mediante la siguiente fórmula:

$$\hat{V}(\hat{Y}_{HH}) = \frac{\sum_{i=1}^{n}\left(\frac{Y_i}{P_i} - \hat{Y}\right)^2}{n(n-1)}$$

Este es un estimador muy fácil de calcular y como se mencionó, es insesgado, lo cual se verifica a continuación:

$$\hat{V}(\hat{Y}_{HH}) = \frac{\sum_{i=1}^{n}\left(\frac{Y_i}{P_i} - \hat{Y}_{HH}\right)^2}{n(n-1)}$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{Y_i}{P_i}\right)^2 - n\hat{Y}_{HH}^2\right]$$ Por una relación de frecuente uso en estadística.

Como la varianza de una variable más la suma algebráica de una constante no se altera, V(X)=V(X±K). Se suma y resta el valor parametral del total en la expresión anterior.

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{Y_i}{P_i} - Y\right)^2 - n(\hat{Y}_{HH} - Y)^2\right]$$

A continuación se agrega la variable indicadora de selección de la unidad correspondiente y el recorrido de la suma se extiende a N

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y\right)^2 X_i - n(\hat{Y}_{HH} - Y)^2\right]$$

Enseguida se procede a tomar esperanza matemática de toda la expresión y se concluye el insesgamiento.

$$E(\hat{V}(\hat{Y}_{HH})) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y\right)^2 E(X_i) - nE\left[(\hat{Y}_{HH} - Y)^2\right]\right]$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y\right)^2 nP_i - nV\left(\hat{Y}_{HH}\right)\right]$$

$$= \frac{1}{n(n-1)}\left[n^2 \frac{\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y\right)^2 P_i}{n} - nV\left(\hat{Y}_{HH}\right) \right]$$

$$= \frac{1}{n(n-1)}\left[n^2 V(\hat{Y}_{HH}) - nV\left(\hat{Y}_{HH}\right) \right]$$

$$= V(\hat{Y}_{HH})$$