5. MUESTREO SISTEMATICO

5.1 Introducción

Suponga que una empresa con 10,000 trabajadores desea extraer una muestra de 500 de ellos para conocer su opinión sobre aspectos contractuales.

Una muestra aleatoria simple puede ser la respuesta inmediata, pero ello implica un tiempo excesivo para la selección. Si el archivo de personal está grabado en medios magnéticos, otra opción es solicitarle al departamento de informática que emita un listado utilizando un salto sistemático cada 20 registros.

El muestreo ofrece ventajas notables por su facilidad de selección, sin embargo, hay que guardar ciertas precauciones con el marco para seleccionar la muestra, pues se puede incurrir en sesgos notables debidos a un ordenamiento relacionado con las variables objetivo de la investigación.

En principio se supone que la muestra es un submúltiplo de la población, esto es que existe una K entera tal que el tamaño de la población se puede expresar por el producto de K y el tamaño de muestra:

$N = K*n$

Cuando la proporcionalidad se cumple en forma estricta, esto es K es un entero, el procedimiento de selección consiste en los siguientes pasos:

- Seleccionar un número aleatorio A entero en el intervalo $1 \leq A \leq K$.
- Tomar el elemento A de la población como primera unidad en muestra
- Sumar K al aleatorio A y el número obtenido será la siguiente unidad en muestra.
- Repetir el procedimiento de suma para extraer la unidad A+2K, A+3K,...,A+(n-1)K

Suponga que se tiene una población con 12 elementos cuyos valores para la variable de interés son:

Y1,Y2,.....Y12

Se toma una muestra de tamaño n = 4 y por tanto K = 3

En función del número aleatorio de arranque A en el intervalo [1, 3]. Existen 3 posibles muestras sistemáticas.

Y1	Y2	Y3
Y4	Y5	Y6
Y7	Y8	Y9
Y10	Y11	Y12

Cada muestra tiene probabilidad 1/K = 1/3 de ser seleccionada. Si A=1 se incluyen en la muestra Y1,Y4,Y7 y Y10.

Si se toma una muestra n = 3 entonces K = 4 y las diferentes muestras se configuran como sigue:

Y1	Y2	Y3	Y4
Y5	Y6	Y7	Y8
Y9	Y10	Y11	Y12

Suponga que en lugar de seleccionar cada 3 o cada 4, elementos se selecciona uno de cada 5, esto es N ahora es diferente del producto nk. Las 5 muestras resultantes serían:

Y1	Y2	Y3	Y4	Y5
Y6	Y7	Y8	Y9	Y10
Y11	Y12			

Cada una de ellas con probabilidad 1/5 y se tendrían entonces dos muestra de tamaño 3 y tres muestras de tamaño 2.

Ahora se verificará que el estimador de la media mediante el muestreo sistemático es insesgado si se cumple la relación de proporcionalidad.

$$E(\bar{y}_{sis}) = \frac{1}{K}(\bar{y}_1 + \bar{y}_2 + \ldots\ldots + \bar{y}_k)$$

$$= \frac{1}{K}\frac{1}{n}(y_1 + y_2 + \ldots\ldots + y_{12})$$

$$= \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y}$$

Claramente si **K** por **n** no es igual a **N** el estimador resulta sesgado. En la práctica el sesgo es muy pequeño y no se suele tomar en cuenta.

Una de las alternativas que se utilizan en la práctica es tomar K no entera.

El procedimiento con la ayuda de una calculadora de bolsillo sería como sigue:

Suponga que se tiene una población de tamaño N = 1000 y se desea tomar una muestra de tamaño n= 145 esto implica K=N/n = 1000/145=6.8955

- Guarde en la memoria de su calculadora el valor de K =6.8955
- Se toma un número aleatorio A en el intervalo de uno a la parte entera de K, esto es en el intervalo [1,6]

- Suponga que el número aleatorio que se obtiene es A = 4.
- Seleccione el elemento que ocupa la posición A en el archivo como primera unidad en muestra..
- Sume al aleatorio A = 4 el valor K=6.8955 lo cual le da el valor 10.8995
- Tome la parte entera del valor obtenido para incluir la unidad 10 en muestra.
- A 10.8995 súmele el valor K=6.8995 que tiene en la memoria y el resultado ahora es 17.7990
- Tome la parte entera (17) y seleccione el elemento correspondiente.
- Continúe sumando hasta que la suma supere el tamaño N. En ese punto habrá concluido la selección y tendrá en muestra n = 145 elementos.
- El procedimiento no tiene saltos sistemáticos estrictamente del mismo tamaño, pero el efecto de sesgo se puede considerar despreciable.

5.2 Varianza del Estimador de la Media

A partir de la definición de varianza se obtiene la expresión para el estimador que parte de una muestra sistemática.

$$V(\bar{y}_{sis}) = \frac{1}{k}\sum_{i=1}^{K}(\bar{y}_i - \bar{\bar{Y}})^2$$

Puesto que hay K muestras distintas, cada una con probabilidad 1/K

Esta sencilla fórmula encierra, sin embargo, la dificultad de no contar con un estimador de la varianza del estimador, pues solamente disponemos de una de las K muestras.

A continuación se procede a analizar la varianza de la media estimada por muestreo sistemático.

Se parte de la suma de cuadrados total, la cual se podrá expresar como la suma de cuadrados dentro de cada muestra sistemática y la suma de cuadrados entre las K muestras sistemáticas.

$$\sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij} - \bar{\bar{Y}}\right)^2 = \sum_{i=1}^{K}\sum_{j=1}^{n}\left[(y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{\bar{Y}})\right]$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n}\left[(y_{ij} - \bar{Y}_i)^2 + (\bar{Y}_i - \bar{\bar{Y}})^2 + 2(y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{\bar{Y}})\right]$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 + \sum_{i=1}^{K}\sum_{j=1}^{n}(\bar{Y}_i-\bar{Y})^2 + 2\sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)(\bar{Y}_i-\bar{Y})$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 + n\sum_{i=1}^{K}(\bar{Y}_i-\bar{Y})^2 + 2\sum_{i=1}^{K}(\bar{Y}_i-\bar{Y})\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 + n\sum_{i=1}^{K}(\bar{Y}_i-\bar{Y})^2$$

Por tanto, si se despeja el segundo sumando se tendrá

$$n\sum_{i=1}^{K}(\bar{Y}_i-\bar{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 - \sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{Y}_i\right)^2$$

El miembro derecho se multiplica y divide por K y se tiene la varianza de la media por muestreo sistemático.

$$nK\frac{1}{K}\sum_{i=1}^{K}(\bar{Y}_i-\bar{Y})^2 = \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 - \sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{Y}_i\right)^2$$

$$nKV\left(\bar{y}_{sis}\right) = \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 - \sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{Y}_i\right)^2$$

$$V\left(\bar{y}_{sis}\right) = \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 \frac{1}{nK}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{Y}_i\right)^2$$

Se multiplica y divide el segundo sumando por N-1 para obtener una expresión en función de S2.

$$= \frac{N-1}{N}\frac{1}{N-1}\sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-\bar{Y}_i)^2 - \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{Y}_i\right)^2$$

$$V\left(\bar{y}_{sis}\right) = \frac{N-1}{N}S^2 - \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(y_{ij}-\bar{Y}_i\right)^2$$

El primer sumando se puede considerar constante en cualquier población. El término que se resta depende de la varianza dentro de cada muestra, entonces en la medida en que cada muestra sea más diversa, esto es, que tenga mayor varianza, tendrá como efecto que la varianza de la media del muestreo sistemático será menor.

Esta característica se suele aprovechar al ordenar las unidades a seleccionar en el marco de muestreo en función de una variable correlacionada con la variable objetivo o la misma variable con datos correspondientes a una medición previa. Entonces se puede proceder con la selección

sistemática. A este procedimiento se le conoce como inducción de una Estratificación Implícita. Es equivalente a tomar una sola observación de los K estratos homogéneos dentro de sí y por tanto no es estimable la varianza dentro de cada estrato.

5.3 Coeficiente de Correlación Intramuestras.

Otra forma de medir la heterogeneidad de las muestras sistemáticas es a través del coeficiente de correlación intramuestras. Este coeficiente se calcula de manera similar al coeficiente de correlación intraclase del muestreo por conglomerados y la varianza del muestreo sistemático se relaciona con el muestreo aleatorio simple y el coeficiente de correlación intramuestras de forma análoga a la relación de muestreo por conglomerados con el aleatorio simple.

$$V(\bar{y}_{sis}) = V(\bar{y}_{MAS})[1 + (n-1)r]$$

De donde se puede despejar el valor de r.

$$r = \frac{\frac{V(\bar{y}_{sis})}{V(\bar{y}_{MAS})} - 1}{\bar{n} - 1}$$ o bien en términos del efecto de diseño $$r = \frac{Deff - 1}{\bar{n} - 1}$$

Se nota fácilmente que si r = 0 la varianza del estimador de la media sistemática es equivalente a la del MAS, pero si r es grande, entonces la varianza del muestreo sistemático será también grande. Pero si r<0 entonces se logra mayor eficiencia.

Para la estimación de varianzas en la práctica se tienen dos alternativas:

Suponer que la muestra se ha extraído por muestreo aleatorio simple y aplicar la fórmula de estimación de varianza que ya conocemos a partir de S^2 estimada con una muestra. Si se adoptó una estimación implícita, esta opción será conservadora, se espera que la varianza sea sensiblemente menor que la del muestreo aleatorio simple.

Tomar varios (m) arranques aleatorios de modo que m = n/L para estimar en cada uno de ellos la media del grupo y a partir de las desviaciones de las medias de grupo respecto de la media global sistemática y así estimar la varianza de la media global como varianza de la media de las m submuestras.

$$\hat{V}(\bar{y}_{sis}) = \frac{\sum_{i=1}^{m} (\bar{y}_i - \bar{y}_{sis})^2}{m(m-1)}$$