MUESTREO POR CONGLOMERADOS

Ing. MSc . Ángel Gómez

Es un método en el cual la unidad de muestreo consiste de un grupo de unidades elementales. Es decir, que cada grupo o conglomerado es un agregado de unidades elementales. Cada conglomerado es considerado como una unidad de muestreo de diferente rango a las unidades elementales que son las de interés.

Entonces:

$$\text{Sea } C = \text{Conglomerado} \Longleftrightarrow C = \left\{ Y_i \right\}_{i=1}^{M}$$

En muestreo por conglomerados se tienen 2 tipos de unidades:

1) Unidades elementales (de interés)
2) Conglomerados (Unidades de Muestreo)

Puesto que en Muestreo lo más caro es llegar a la unidad de muestreo, entonces ¿No sería posible en lugar de enumerar una sola de ellas al llegar al lugar de localización, enumerar uno solo de ellos?

Esto es lo que se hace en el Muestreo por Conglomerados; es decir, se incrementa con bajo costo el tamaño de muestra.

La principal razón para usar el muestreo por conglomerados es que no hay una lista confiable de elementos en la población y seria costoso elaborar dicha lista sobre todo cuando la población es grande y por otro lado, aún cuando se tenga la lista la búsqueda de la información es muy costosa.

VENTAJAS DEL MUESTREO POR CONGLOMERADOS

1. A una precisión y confiabilidad predefinida, resulta más barato.

2. Permite estudiar Universos donde no se conozca el marco de las unidades elementales, solamente se requiere el marco de conglomerados.

3. El uso de conglomerados facilita la supervisión de las entrevistas y la administración del trabajo de campo.

4. Es conveniente aclarar que para facilidades de enseñanza, es necesario plantear el caso de conglomerados de igual tamaño.

Notación:

Considérese el caso de conglomerados de igual tamaño y supóngase que la población está compuesta de N conglomerados de M elementos cada uno, y una muestra de n conglomerados es seleccionada por el método de Muestreo Aleatorio Irrestricto.

Denote por

Y_{ij} el valor de la característica bajo estudio del j-ésimo elemento del i-ésimo conglomerado, $(j=1,2,\ldots,M)$ y $(i=1,2,\ldots,N)$;

$\overline{Y}_{i.} = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}$ promedio por elemento del i-ésimo conglomerado,

$\bar{\overline{Y}}_{N.} = \frac{1}{N}\sum_{i=1}^{N} \overline{Y}_{i.}$ promedio de medias de los conglomerados en la población,

Si una muestra de nM elementos fuese seleccionada de la población de NM elementos por Muestreo Aleatorio Irrestricto, la varianza del promedio por elemento está dada por:

$$V\left(\bar{Y}_{nM}\right) = \frac{NM - nM}{NM} \cdot \frac{S^2}{nM}$$

Con esto podemos ver que la eficiencia del muestreo por conglomerados comparada con la del Muestreo Aleatorio está dada por:

$$\text{Eficiencia Relativa: } \frac{V\left(\bar{Y}_{nM}\right)}{V\left(\bar{\bar{Y}}_{n.}\right)} = \frac{S^2}{M\, S_b^2} \quad \ldots\ldots\ldots(1)$$

Se quiere que la Eficiencia Relativa sea > 1 y esto puede ocurrir si M es pequeño y S5 es pequeña, donde M es el tamaño de los conglomerados y Sb es la variar^ za entre promedios de conglomerados. Es posible plantear el Muestreo por Conglomerados en términos de Análisis de la Varianza.

ANÁLISIS DE LA VARIANZA

Entre conglomerados	N-1	$\frac{M}{N-1}\Sigma(\bar{Y}i. - \bar{\bar{Y}}_{N.})^2 = M\, S_b^2$
Dentro de conglomerados	N(M-1)	$\frac{1}{N(M-1)}\Sigma\Sigma(Yij - \bar{Y}i.)^2 = \bar{S}_W^2$
T o t a l	NM-1	$\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(Yij - \bar{\bar{Y}}_{N.})^2 = S^2$

En este caso, Eficiencia Relativa será:

$$\left[\frac{\text{Conglom.}}{\text{M.A.I.}}\right] = \frac{\text{CM Total}}{\text{CM entre Congl.}} = \frac{S^2}{M\, S_b^2}$$

Se observa entonces que la varianza entre conglomerados debe ser pequeña y el tamaño de los conglomerados debe ser pequeño. En general, la recomendación es: Conglomerados pequeños, muchos y dispersos.

ESTIMACIÓN DE LA EFICIENCIA RELATIVA DE UNA MUESTRA DE CONGLOMERADOS

Datos de una población completa rara vez se obtienen en la práctica. Lo que se obtiene es una muestra de conglomerados y el análisis de varianza de los elementos en la muestra.

La muestra consiste de n conglomerados y el análisis está dado por:

ANÁLISIS DE LA VARIANZA

FUENTE DE VARIACION	GL	CM
Entre Conglomerados	n-1	$\frac{1}{n-1}\sum_{i=1}^{n} M\,(\bar{Y}_{i.} - \bar{\bar{Y}}_{n.})^2 = M\,s_b^2$
Dentro de Conglomerados	n(M-1)	$\frac{1}{n(M-1)}\sum_{i=1}^{n}\sum_{j=1}^{M}(Y_{ij} - \bar{Y}_{i.})^2 = \bar{s}_w^2$
T o t a l	nM-1	$\frac{1}{nM-1}\sum_{i=1}^{n}\sum_{j=1}^{M}(Y_{ij} - \bar{Y}_{n.})^2 = s^2$

En una muestra aleatoria de conglomerados, s_b^2 y s_w^2, son estimadores insesgados de los Correspondientes parámetros y s^2 no es un estimador insesgado de S^2, ya que los elementos no se consideran como una muestra aleatoria de elementos de la población de NM unidades. Sin embargo, un estimador insesgado de S^2 es posible obtenerlo ya que se sabe que:

$$(NM-1)S^2 = (N-1)\,M\,S^2 + N(M-1)\,\bar{S}_w^2;$$

donde un estimador de S^2 está dado por:

$$\text{Est. } S^2 = \frac{(N-1)\,M\,s_b^2 + N(M-1)\,s_w^2}{NM-1}$$

De esta forma La Eficiência Relativa está dada por:

$$\text{Est. (Eficiencia Relativa)} = \frac{(N-1)\ M\ s^2_b + N(M-1)\ \bar{s}^2_w}{(NM-1)\ M\ s^2_b}$$

$$\leq \frac{1}{M} + \frac{M-1}{M}\ \frac{s^2_w}{M\ s^2_b} \quad \ldots\ldots\ldots (2)$$

Para **N** grande.

De la ecuación (2) se puede ver que cuando crece la varianza dentro de conglomerados crece la Eficiencia Relativa y también aumentará la Eficiencia Relativa cuando el tamaño del conglomerado es pequeño.

MUESTREO POR CONGLOMERADOS EN TÉRMINOS POR CORRELACIÓN INTRACLASE

En la práctica los conglomerados no están constituidos por elementos aleatorios, lo que es aleatorio son los grupos de elementos; entonces, es interesante medir la asociación de unidades elementales dentro de conglomerados. Esto se hace a través del coeficiente de correlación intraclase (p).

Así, la varianza del estimador de la media poblacional será una función de p.

$$V(\bar{\bar{Y}}_{n.}) = f(\rho, M)$$

$$\rho = \frac{E\left\{(Y_{ij} - \bar{\bar{Y}}_{n.})\ (Y_{ik} - \bar{\bar{Y}}_{n.})\right\}}{E\ (Y_{ij} - \bar{\bar{Y}}_{n.})^2} \quad \ldots\ldots\ldots\ldots (3)$$

donde $E\left\{(Y_{ij} - \bar{\bar{Y}}_{n.})\ (Y_{ik} - \bar{\bar{Y}}_{n.})\right\}$ es la covarianza entre Y_{ij} y Y_{ik} $\forall i$.

Hansen, Horwitz y Madow (1) consideran a p como una medida de homogeneidad del conglomerado.

Generalmente la varianza de un estimador en el muestreo por conglomerado tiende a ser mayor que una muestra equivalente de elementos seleccionados individualmente al azar.

La varianza de la media de n conglomerados en términos de correlación intraclase (p) es dada por:

$$V(\bar{Y}_{n.}) = \frac{N-M}{N} \cdot \frac{NM-1}{M(N-1)} \cdot \frac{S^2}{nM} \left\{ 1+(M-1)\rho \right\} \quad \ldots\ldots\ldots \quad (4)$$

$$\cong \frac{S^2}{nM} \left\{ 1+(M-1)\rho \right\} \ldots\ldots\ldots \quad (5)$$

Si N es Suficientemente Grande.

También, la Eficiencia Relativa en términos de correlación intraclase (p) esta dada por:

$$E.R. = \frac{M(N-1)}{NM-1} \cdot \frac{1}{\left\{ 1+(M-1)\rho \right\}}$$

$$E.R. \cong \frac{1}{1 + (M-1)\rho}$$

Si N es suficientemente grande.

En la ecuación (4) de la varianza del estimador Yn., se tiene que el factor:

$$\left\{ 1 + (M-1)\rho \right\}$$

Es el cambio de la varianza con el uso de un conglomerado como unidad de muestreo en lugar de un elemento como unidad de muestreo.

Generalmente p es positivo y decrece si el tamaño del conglomerado incrementa, pero a una rata de decrecimiento relativamente pequeña con grandes incrementos de M; tal que ordinariamente, incrementos en el tamaño del conglomerado trae sustancial incremento en la varianza muestral del estimador muestral.

Este punto puede ser ilustrado a través de la siguiente tabla.

TABLA 1

CAMBIO RELATIVO EN VARIANCIA CON EL INCREMENTO DE TAMAÑO DE CONGLOMERADO

TAMAÑO DEL CONGLOMERADO (M)	2	4	8	16
ρ	0.28	0.22	0.18	0.14
$(M-1)\rho$	0.28	0.66	1.26	2.10

Si p = o, esto significa que las unidades elementales dentro de conglomerados son heterogéneas.

Por lo general, p es positivo, ya que los conglomerados normalmente se forman uniendo granjas, establecimientos, familias, etc., geográficamente contiguas.

Es posible dar una alternativa para el cálculo de esto es;

$$\rho = \frac{\frac{N-1}{N} S_b^2 - \frac{S_w^{-2}}{N}}{\frac{NM-1}{NM} S^2}$$

TAMAÑO DE MUESTRA EN MUESTREO POR CONGLOMERADO

Ordinariamente se puede decir que muestreo por conglomerados disminuye precisión pero reduce el costo.

Problema: Determinar el tamaño óptimo de conglomerado de acuerdo a un presupuesto disponible.

El costo de un estudio basado en n conglomerados dependerá, además del costo de

planeación y análisis, del costo de viaje para enumerar todos los elementos dentro de un conglomerado y del costo de viaje de conglomerado a conglomerado.

Asimismo, el costo del estudio puede ser expresado por una función de costo; considérese la siguiente función de costo:

$$C = C_1 nM + C_2 d; \quad \ldots\ldots\ldots \quad (6)$$

Donde C representa el costo de enumerar un elemento incluyendo el costo del viaje de un elemento a otro dentro de un conglomerado; C2 es el costo de viajes una unidad de distancia de un conglomerado a otro, d es la distancia total entre conglomerados.

Relacionado con esto está el problema de determinar como se afecta la varianza del estimador de la media al variar el tamaño del conglomerado. Esto es posible determinarlo si se expresa Sb2 como una función de M, donde:

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}i. - \bar{\bar{Y}}n.)^2$$

El primero en hacer este intento fue Fairfield Smith. El argumento que si el conglomerado fuera una muestra aleatoria de M elementos; entonces:

$$S_b^2 = \frac{S^2}{M};$$

Pero unidades dentro de conglomerados ordinariamente poseen correlación intraclase; entonces, medias entre conglomerados difieren una de otras más que sí los conglomerados estuviesen formados de elementos seleccionados al azar. Así:

$$S_b^2 = \frac{S^2}{M}$$

Smith propuso la relación:

$$S_b^2 = \frac{S^2}{M^g} \quad \ldots\ldots\ldots \quad (7)$$

Donde g < 1 y es calculado de la muestra; pero (7) es fácilmente linealizable

$$\log S_b^2 = \log S^2 - g \log M \quad \ldots\ldots\ldots (8)$$

Y se puede observar en (8) que para varias M se pueden tener varios valores de s_b^2 así, por análisis de regresión puede fácilmente obtenerse un estimador de g.

También, Mahalanobis y Jessen demostraron que algunas características económicas siguen una ecuación ligeramente distinta a la de Fairfield Smith.

Establecen una suposición:

$$S_W^2 = aM^b; \quad (b > 0);$$

Del análisis de la varianza se tiene:

$$S_b^2 = \frac{(NM-1)S^2 - N(M-1)aM^b}{M(N-1)}$$

Los valores S2, a, b pueden evaluarse de los datos de tenemos S_w^2 tamaño

Si se considera que toda la población es un solo conglomerado,

Entonces:

$$S^2 = a(NM)^b;$$

Esto nos dice que es posible estimar a y b de un estudio en el cual únicamente un tamaño de conglomerado (M) es utilizado.

Entonces:

$$S_b^2 = \frac{(NM-1)\, a(NM)^b - N(M-1)aM^b}{M(N-1)} \quad \ldots\ldots\ldots (9)$$

Por otro lado, Mahalanobis encontró que si se tienen n puntos al azar, la distancia esperada entre todos ellos se puede aproximar por:

$$E(d) \quad \alpha \; n^{1/2} - \bar{n}^{1/2}$$

Jessen por medio del trabajo experimental encontró que $n^{/2}$ trabaja bien en la práctica.

Así, el costo del muestreo referido a la ecuación (6) esta dado por:

$$C \quad = \quad C_1 nM + \quad C_2 n^{1/2}$$

Además vimos que $S_w = aM$ y la varianza del estimador del promedio por elemento era:

$$V(\bar{\bar{Y}}_{n.}) = \frac{N-M}{N} \cdot \frac{S_b^2}{n} \quad \ldots\ldots\ldots \quad (10)$$

Donde:

$$S_b^2 = \frac{(NM-1)S^2 - N(M-1)aM^b}{M(N-1)} \quad \ldots\ldots\ldots \quad (11)$$

Sustituyendo en (10) la ecuación (11) se tiene:

$$V(\bar{\bar{Y}}_{n.}) \cong \frac{1}{n}\left\{S^2 - (M-1)aM^{b-1}\right\} \quad \ldots\ldots\ldots \quad (12)$$

Donde el multiplicador finito es ignorado El problema seria seleccionar n y m tal que la varianza dada por (12) sea minimizada para un costo especificado. La solución de esto es dada por Cochran, para seleccionar M y n tal que la varianza dada por (12) sea minimizada para un valor del Costo Total

$C = CO$ entonces se debe encontrar $\{n, M\}$ tal que $V(\bar{\bar{Y}}_{n.})$ sea mínima en un espacio restringido por:

$$C = C_1 nM + C_2 n^{1/2}$$

$$Co = C_1 nM + C_2 n^{1/2}$$

Considere función tipo lagrangiana:

$$\phi(n,M) = (V(\bar{\bar{Y}}_{n.}) + \lambda(C_1 nM + C_2 n^{1/2} - C_0);$$

Donde:

λ = mutliplicador de Lagrange.

$$V(\bar{\bar{Y}}_{n.}) = \frac{N-n}{nM} \cdot S_b^2$$

Así, siguiendo la técnica de Lagrange:

$$\frac{\delta\phi}{\delta n} = o \qquad \frac{\delta\phi}{\delta M} = o$$

$$\Rightarrow \quad si \quad \frac{\delta\, V(\bar{\bar{Y}}_{n.})}{\delta n} = \frac{-V(\bar{\bar{Y}}_{n.})}{n} = \frac{-V(\bar{\bar{Y}}_{n.})}{n}$$

Donde:

$$\frac{\delta\, V(\bar{\bar{Y}}_{n.})}{\delta n} = -\frac{1}{n^2}\left\{S^2-(M-1)aM^{b-1}\right\} = -\frac{1}{n}\left[\frac{1}{n}\left\{S^2-(M-1)aM^{b-1}\right\}\right] = \frac{V}{n}$$

$$-\frac{V}{n}(\bar{\bar{Y}}_{n.}) + \lambda(C_1 M - {}^1/_2\, C_2 n^{-1/2}) = o \quad \ldots\ldots\ldots\ldots \quad (13)$$

$$\frac{\delta\, V(\bar{\bar{Y}}_{n.})}{\delta\, M} + \lambda\, C_1 n = o \quad \ldots\ldots\ldots \quad (14)$$

$$C_1 nM + C_2 n^{1/2} = C \quad \ldots\ldots\ldots \quad (15)$$

Eliminando λ de (13) y (14) y sustituyendo en (13) se obtiene que

$$-\frac{V(\bar{\bar{Y}}_{n.})}{n} + \left(-\frac{\delta\, V(\bar{\bar{Y}}_{n.})}{\delta\, M} \cdot \frac{1}{C_1 n}\right)\left(C_1 M + {}^1/_2\, C_2 n^{-1/2}\right) = o$$

O bien

$$-\frac{\delta V(\bar{Y}_{n.})}{\delta M}\cdot\frac{M}{V(\bar{Y}_{n.})} = \frac{1}{1+\dfrac{C_2}{2C_1 M n^{1/2}}} \qquad \ldots\ldots\ldots (16)$$

Resolviendo (15) como una cuadrática un n /2 se tiene;

$$n^{1/2} = \frac{-C_2^2 + (C_2 + C_1 C_0 M)^{1/2}}{2C_1 M} \qquad \ldots\ldots\ldots (17)$$

Finalmente, sustituyendo (17) en (16) y simplificando se tiene:

$$\frac{M}{C(\bar{Y}_{n.})}\frac{\delta V(\bar{Y}_{n.})}{\delta M} = -1 + \left(1 + \frac{4\,C_1 C_0 M}{C_2^2}\right)^{1/2} \qquad \ldots\ldots\ldots (18)$$

La ecuación (18) puede ser resuelta directamente para M; sin embargo, la solución no es fácil. Por lo que se sugiere un método de aproximaciones sucesivas. Se inyecta M tal que (18) se satisfaga y ese M se sustituye en (17) y se obtiene X que n es el número de conglomerados en la muestra.

EJEMPLOS DONDE SE APLICA EL MUESTREO POR CONGLOMERADOS

1.- Una firma desea conocer la aceptación de un nuevo producto en el mercado, en un país. Para ello decide vender el producto en una muestra de negocios. Si se piensa que en la aceptación o rechazo del producto no influirán las características particulares de cada región del país, el muestreo por conglomerados seria casi tan preciso como el muestreo aleatorio irrestricto y mucho más barato.

Se pueden seleccionar aleatoriamente una o más ciudades del país y se ofrece a la venta el nuevo producto en todos los negocios de cada una de las ciudades (conglomerados) seleccionada.

2.- Suponga que una empresa desea conocer el consumo promedio anual por familia en una ciudad. Si se dispone de una lista de las familias en la ciudad es posible seleccionar al azar las muestras de familias. Sin embargo, aún cuando exista la lista de familias, es más barato hacer la selección de cuadras en la ciudad y en esa muestra de cuadras (conglomerados) se entrevistarán todas las familias pertenecientes a cada conglomerado.

3.- En una ciudad se quiere saber sobre características de las viviendas; en este caso las unidades elementales serían las viviendas y los conglomerados serían las cuadras o lotes de vivienda.

4.- En una zona se desea saber el promedio de gastos en ropa que hacen las personas que allí viven. En este caso las unidades elementales serían las personas y los conglomerados o unidades de muestreo serían las viviendas.

5.- En un aeropuerto una linea aérea desea saber ciertas características de sus viajes. Las unidades elementales son los pasajeros que llegan y los conglomerados serían los vuelos.

6.- Se quiere determinar en una zona el origen y destino de vehículos en el tránsito anual en un puente. Las unidades de muestreo son los vehículos y los conglomerados son intervalos de 40 minutos.

BIBLIOGRAFIA

HANSEN H. MORRIS, HURWITZ Y MADON G. WILLIAM

Sample Survey Methods and Theory. 1953