Concepto y Aplicación de Muestreo Conglomerado y Sistemático

(Concept and Application of Cluster and Systematic Sampling)

Guillen, A., M.H. Badii, J.L. Prado, J.L. Abreu & J. Valenzuela *

Reumen. Se describen las bases del muestreo conglomerado y muestreo sistemático. Se presentan las ecuaciones pertinentes y aquellas para la estimación del tamaño óptimo de muestreo para cada tipo de muestreo. Se demuestra la aplicación práctica de éstos clases de muestreo por medio de ejemplos reales.

Palabras claves. Muestreo conglomerado, muestreo sistemático, tamaño óptimo de la muestra.

Abstract. Cluster sampling and systematic sampling are described and their equations are provided. Real examples are given in order to show the applications of these types of samplings. Equations to estimate optimal sample sizes are also noted.

Keywords. Cluster sampling, optimal sample size, systematic sampling

Introducción

El muestreo conglomerado se usa cuando hay una conglomeración de las unidades muestreales y cuando se trata de ahorrar el costo del muestreo. En este tipo de muestreo, se selecciona de forma aleatoria un cierto número de los conglomerados del cuadro y luego se hace prácticamente un censo completo de cada uno de los conglomerados (Cochran, 1977; Cornfield, 1951; Deming, 1960; Hansen et al., 1953; Kish, 1965; Mendenhall, 1971)**.**

Ejemplo del Muestreo Conglomerado (MC)

Concepto

Vamos a suponer que deseamos conocer las ganancias por cabeza de las familias de un municipio. Las familias viven en las casas habitacionales y éstas casa están situadas en 415 manzanas dentro del municipio.

Propósito

El objetivo es obtener, en base a un muestreo científico información al respecto. Es claro que el tipo del diseño muestreal adecuados para esta situación por el rasgo del arreglo espacial de tipo agregada sería el muestreo conglomerado (Bellhouse & Rao, 1975; Cochran, 1946; Buckland, 1951).

Procedimiento

De forma al azar se seleccionan 10 manzanas del total de las manzanas del municipio y se arrojan los siguientes datos (Tabla 1).
Además, los tamaños óptimos de la muestra (n_{opt}) para el muestreo conglomerado en función de diferentes valores de "L" se indican en la Tabla 2.

Tabla 1. Resultado de la entrevista con adultos por cada casa en las 10 manzanas.

Manzana (n_i)	**# de adultos (m_i)**	**Ingreso total ($) por manzana (X 1000) ($y_i$)**	**$(m_i)^2$**	**$(y_i)^2$**	**$y_i * m_i$**
1	8	96	8^2	96^2	8 * 96
2	12	121	12^2	121^2	12 *121
3	4	42	4^2	42^2	4 *42
4	5	65	5^2	65^2	5 *65
5	6	52	6^2	52^2	6 *52
6	6	40	6^2	40^2	6 *40
7	7	75	7^2	75^2	7 *75
8	5	65	5^2	65^2	5 *65
9	8	45	8^2	45^2	8 *45
10	3	50	3^2	50^2	3 *50
n =10	**$\sum(m_i)$=64**	**$\sum(y_i)$ = 651**	**$\sum$= 408**	**$\sum$=48,525**	**$\sum$ = 4,625**

$m_{yi} = \sum(y_i) / \sum(m_i) = 651 / 64 = 10.61$ ingreso promedio por adulto

$M = \sum(m_i) / n = 64 / 10 = 6.4$ promedio de adultos por manzana

Ecuaciones:

$$n_{opt} = [N*[(y_i - m_{yi})^2/(n-1)] / [N*D + [(y_i - m_{yi})^2/(n-1)]$$

$$EE(m_{Cong}) = [\{(N-n)/(Nn*M^2)\} * \{\sum(y_i - m_{yi})^2\}/(n-1)]^{1/2}$$

Donde,

n_{opt} = Tamaño óptimo de la muestra

N = Total de las unidades de la muestra

y_i = Ingreso total por manzana "*i*"

m_{yi} = Ingreso promedio por adulto en pesos

$D = L^2 * M / 4$
L = Error de estimación a nivel de 95% de probabilidad = 2

$EE(m_{Cong})$
$EE(m_{Cong})$ = Error estándar del muestreo conglomerado

M = Promedio de adultos por manzana

$$EE(m_{Cong}) = [\{(N - n)/(Nn*M^2)\} * \{\sum(y_i - m_{yi})^2\}/(n - 1)]^{1/2}$$

$$EE(m_{Cong}) = [\{(415–10)/(415*10-(6.4)^2*[(96–10.61)^2+(121–10.61)^2+.+(50–10.61)^2]/(10-1)]^{1/2}=5.59$$

$$VR_{myi} = EE(m_{Cong}) / m_{yi}$$

$$VR_{myi} = 5.59 / 10.16 = 0.55$$

$$L = 2\ VR_{myi} = 2\ (0.55) = 1.10$$

Para diferentes valores de L, se calculan los siguientes tamaños óptimos de la muestra (Tabla 2).

Tabla 2. Tamaños óptimos de muestra (n_{opt}) en base a los valores de L.

Valor hipotético de *L*	**Tamaño óptimo de la muestra (n_{opt})**
0.5	95.41 ≈ 96
1.0	28.85 ≈ 29
2.0	7.6 ≈ 8
3.0	3.41 ≈ 4
4.0	1.92 ≈ 2

5.0	$1.23 \approx 2$
10.0	$0.30 \approx 1$

Por tanto, la estimación de la media poblacional con su intervalo de confianza a nivel de 95% es:

$$m_{yi} \pm L = 10.16 \pm 1.1$$

Límite inferior: 11.26

Límite superior: 10.06

Muestreo sistemático (Msis)

Cuando se trata de un muestreo sencillo y rápido, se usa este tipo de muestreo. Una característica importante del muestreo sistemático es que presenta menos varianza que el muestreo simple aleatorio, esto debido a la presencia de una estratificación innata en el diseño del muestreo sistemático. El muestreo sistemático normalmente se usa en la inspección y el control de calidad debido a la alta rapidez y la baja varianza de este tipo de muestreo. Este tipo de muestreo es adecuado para las situaciones en donde la población es grande y con alto nivel de varianza. El procedimiento es de manera siguiente. En base a un arreglo aleatorio se selecciona un número entre el 1 y el 10, suponemos que este número es el 4. Luego dependiendo del tamaño de la población se seleccione un intervalo que es directamente proporcional al tamaño poblacional. Vamos a suponer que este intervalo es igual a 100. Se selecciona la unidad muestreal (UM) número 4 y luego las unidades muestreales siguientes basado en este intervalo, es decir, las UM´s 104, 204, 304, etc. hasta agotar el cuadro de la muestra. Como ya se mencionó arriba existe una estratificación innata en este tipo de muestreo que permite la reducción de la variabilidad. Sin embargo, a veces, sucede que este intervalo coincide con la presencia de una gradiente de variabilidad natural dentro de la población. Por ejemplo, sucede que cada 10 (largo de intervalo) casas, la casa que se va a muestrear se encuentra en una esquina. Es claro que una casa en la esquina posee información de cuatro direcciones cardinales en comparación con otras casas que no están en la esquina, y obviamente, este gradiente de variabilidad genera sesgo en el muestreo y por tanto, mal representación de la población.

Ejemplo del Msis

A. Estimación de la media poblacional (m_{sis}) con 95% de IC

Vamos a suponer que deseamos estimar la calidad de maple (% de azúcar) en la savia del árbol del maple en una zona específica. El número total de los árboles esta desconocida, por tanto, no se puede hacer un MSA. La alternativa es conducir un muestreo sistemático (Msis), en base a seleccionar 1 de cada 7 árboles. El objetivo es el estimar la media poblacional con su límite de estimación (L) a 95% de confiabilidad. Usando, éste intervalo (muestrear cada 7 árbol), nos arroja los siguientes datos (Tabla 3).

Tabla 3. Datos de maple según un muestreo sistemático.

Árbol muestreado (i)	**Cantidad de azúcar en savia (X_i)**	**$(X_i)^2$**
1	82	$(82)^2$
2	76	$(76)^2$
3	83	$(83)^2$
.	.	.
.	.	.
.	.	.
210	84	$(84)^2$
211	80	$(80)^2$
212	79	$(79)^2$
n = 212	**$\sum (X_i)$ = 17,066**	**$\sum(X_i)^2$ = 1,486,800**

$N_t = n*7$

$N_t = 212*7 = 1{,}848$

$m_{sis} = \sum(X_i) / n$

$m_{sis} = 17{,}066 / 212 = 80.6$

$V=[\sum(X_i)^2-(\sum X_i)^2/n]/(n-1)$

$V = [1{,}486{,}800-(17{,}066/212)]/(212-1) = 535.483$

$EEm_{sis} = [(V/n)(1-\varphi)]^{1/2}$

$EEm_{sis} = [(535.483/212)(1–(212/1.484)]^{1/2} = 1.46$

Donde,

n = Tamaño de la muestra

N_t = Tamaño total de la población

m_{sis} = Media de la muestra sistemática

V = Varianza

EEm_{sis} = Error estándar

Por tanto, la estimación de la media poblacional con su intervalo de confianza a nivel de 95% para el muestreo sistemático es:

$$m_{sis} \pm L = 80.6 \pm 2\ (1.46)$$

Límite inferior: $80.6 – 2(1.46) = 77.68$

Límite superior: $80.6 + 2(1.46) = 83.52$

B. Estimación del total de de la población (N_t) con 95% de IC

Se desea estimar el rendimiento de una huerta de manzano con 1300 árboles (N_t). Suponemos que la media y la varianza del rendimiento en base a un muestreo sistemático con un intervalo de 10 son: m_{sis} = 3.52 cajas por árbol, y V = 0.48. El objetivo es el estimar el tamaño total de la población con su límite de error de estimación (L) a nivel de 95% de confiabilidad.

$n = N_t/10 = 1300/10 = 130$

$T = N_t * m_{sis} = 1300*3.52 = 4{,}576$

$EE_T = [(N_t)^2(V/n)(1-\varphi)]^{1/2}$

$EE_T = [(1300)^2(0.48/130)(1-(130/1300)]^{1/2} = 74.94$

Donde,

n = Tamaño de la muestra

T = Total de la población

EE_T = Error estándar para el total de la población

EEm_{sis} = Error estándar para la media de la población

Por tanto, la estimación de la total poblacional con su intervalo de confianza a nivel de 95% para el muestreo sistemático es:

$$T \pm L = 4{,}576 \pm 2(74.94)$$

Límite inferior: $4{,}576 - 2(74.94) = 4{,}426.12$

Límite superior: $4{,}576 + 2(74.94) = 4{,}725.88$

C. Estimación del tamaño óptimo de la muestra (n_{opt})

Un banco desea estimar el promedio del tiempo que los recibos de los servicios llegan al banco después de la fecha de vencimiento. Este banco hace un muestreo sistemático de $N_t = 2500$ cuentas de clientes con fechas vencidas para los recibos de servicio. Suponemos que basado en un muestreo similar del año anterior fue determinado que $V = 100$. La pregunta es cuál sería el tamaño óptimo de la muestra para estimar la media poblacional con un límite de estimación igual a 2 días ($L = 2$).

$n_{opt} = N_t V/[(N_t - 1)D + V]$

$n_{opt} = 2500(100)/[(2500-1)1+100] = 96.19 \approx 97$

Donde, $D = L^2 / 4 = 2^2 / 4 = 1$

Para diferentes valores de *L*, se puede estimar los siguientes tamaños óptimos de la muestra (Tabla 4):

Tabla 4. Tamaños óptimos de muestra (Msis) en función de los valores de *L**.

Valor hipotético de *L*	**Tamaño óptimo de la muestra (n_{opt})**
1	≈ 345
2	≈ 97
5	≈ 16
10	≈ 4
*: Los valores de n_{opt} están redondeados.	

Conclusiones

Suponiendo que los datos de la muestra proceden de una población con una distribución normal, entonces, si la población tiene un tamaño muy grande y además existe una varianza muy alta también en esta población, el esquema óptimo del muestreo para este caso sería el muestreo sistemático. Hay que notar que para los mismos datos normales, tomar un muestreo de tipo sistemático genera una varianza menor en comparación con conducir una muestra de tipo simple aleatorio, ya que el acto de la selección aleatorio del primer número y por tanto la división de la población en diferentes intervalos, en práctica funciona como generar estratos. Ahora si los datos tienen un arreglo de tipo agregado (amontonado, conglomerado o hacinado) entonces el diseño óptimo del muestreo en este caso sería el muestreo conglomerado.

Referencias

Cochran, W.G. 1977. Sampling Tecniques. 3d. ed., Wiley & Sons, New York.
Cornfield, J. 1951. The determination of simple size. Am. J. Pub. Health. 41: 654-661.
Deming, W.F. 1960. Sample design in Business research. Wiley & Sons, New York.
Hansen, M.H., W.N. Hurwitz & W.G. Madow. 1953. Sample Survey Methods and Theory. Vol. 1. Wiley & Sons, New York.
Kish, L. 1965. Survey Sampling. Wiley & Sons. New York.
Mendenhall, W. 1971. Introduction to Probability and Statistics. 3d. ed., Wadsworth, Belmont.
Bellhouse, D.R. & J.N.K. Rao. 1975. Systematic Sampling in the Presence of a Trend. Biometrika, 62: 694-697.
Cochran, W.G. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. Ann. Math. Stat., 17: 167-177.
Buckland, W.R. 1951. A review of the literature of systematic sampling. J. Roy. Stat. Soc. B13: 208-215.

*** Acerca de los autores**

Guillen, A. Es profesora investigadora del área de posgrado, UANL, México

Badii, M.H. Es profesor investigador del área de posgrado, UANL, México

Prado, JL. Es profesor investigador del área de posgrado, UANL, México

Abreu. José Luis. Es profesor investigador del área de posgrado, UANL, México

Valenzuela, J. Es profesor investigador del área de posgrado, UAAAN, Coah., México
UANL, San Nicolás, N.L., aguillen77@yahoo.com, [1]UAAAN, Saltillo, Coah., México.