Análisis y Aplicación de Muestreo Multietápico, Estimación de Submuestreo y Muestreo de Respuesta Aleatoria

(Analyses and Application of Multi stage Sampling, Sub-sample Estimation and Random Response Sampling)

Badii, M.H., A. Guillen, J. Landeros & E. Cerna *

Resumen. Se analizan de forma detallada las bases del muestreo multietápico, la estimación de submuestreo y el muestreo de respuesta aleatoria. Se presentan las ecuaciones pertinentes ejemplificando por medio de casos reales. También se argumentan las situaciones en donde se debe utilizar estos casos específicos.

Palabras claves. Caso multietápico, muestreo de respuesta aleatoria, submuestreo.

Abstract. The basics of multi-stage sampling, sub-sampling and random response sampling are analyzed. Detailed equations regarding these types of sampling are provided. The applications of these types of samplings are denoted by means of real case studies.

Keywords. Multi-stage sampling, sub-sampling, random response sampling.

Muestreo multietápico (MM)

Múltiples autores han estudiado y reportado los fundamentos del muestreo (Cochran, 1977; Cornfield, 1951; Deming, 1960; Hansen et al., 1953; Kish, 1965; Mendenhall, 1971). Se emplea este tipo de muestreo cuando la estructura del hábitat es compleja, por ejemplo, cuando se desea estimar la densidad poblacional de un organismo en hojas de las ramas de los árboles de las huertas; en este ejemplo, el árbol sería la unidad muestreal primaria, la rama la unidad muestreal secundaria y la hoja la unidad muestreal terciaria y en este caso se trata de un muestreo en tres etapas. Las fórmulas y mayor información para estos cinco tipos de diseños se encuentran en Cochran (1977). Para ilustrar este tipo de muestreo, vamos a demostrar un muestreo de 2 etapas. Suponemos que una población de insectos que habita los frutos de una planta ocupa una superficie planteada y esta superficie se divide en cuadrantes. La primera etapa es seleccionar un número óptimo de estos cuadrantes denominados unidades primarias de muestreo (etapa 1). Luego, de cada cuadrante seleccionado, escoger una sub-muestra, los cuales puedan considerarse varias plantas. A éstas plantas denominan las unidades segundarias de muestreo (etapa 2).

Tabla 1. Número de elemento por planta (X_i) por cuadrante.

Planta	**cuadrantes**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	2	1	0	1	0	0	0	0	2	1
2	1	1	0	2	0	1	0	0	0	0
3	2	2	2	0	0	0	0	0	0	1
4	1	0	0	0	0	0	0	0	2	0
5	0	0	1	1	0	0	0	0	1	0
n	5	5	5	5	5	5	5	5	5	5
$\boldsymbol{T_{cuad}}$	6	4	3	4	0	1	0	0	5	2
$\boldsymbol{m_{cuad}}$	1.2	0.8	0.6	0.8	0	0.2	0	0	1	0.4

Cabe destacar que los diseños de muestreo multietápico, submuestreo y muestreo de respuesta aleatoria han sido estudio intensivamente (Brewer & Hanif, 1970, Brooks, 1955; Chapman, 1952; Durbin, 1959; Jones, 1956; Warner, 1965).

Datos**:** Se contabiliza el número de los insectos por cada una de las 5 plantas seleccionadas dentro de cada uno de los 10 cuadrantes (Tabla 1).

$N_{general} = 50$

$T_{general} = 25$

Conducir una ANOVA de forma siguiente (Tabla 2):

Tabla 2. Tabla de ANOVA para los datos de la Tabla 9*.

FV	**Gl**	**SC**	**CM**	**F_{obs}** F_{tabla}=2.45
Tratamiento (cuadrantes)	K -1=10-1 = 9	$\sum[(T_{cuad})^2/n_{cuad}]-FC$ = 21.4 – 12.5 = 8.9	8.9/9 = 0.98	0.98/0.44 = 2.24 NS
Error	(Kn–1)-(K-1) = 49-9 = 40	$SC_{total} - Sc_{Tratamiento}$ = 26.5 – 8.9 = 17.6	17.6/40 = 0.44	2.24 < 2.45
Total	Kn-1(10*5) - 1= 49	$\sum(X_i)^2 - FC$ =39 -12.5= 26.5		
*: FC = Factor de corrección = $(T_{general})^2 / N_{general} = (25)^2 / 50 = 12.5$				

Donde,

n_1 = Número de cuadrantes (unidades primarias de muestreo) muestreadas = 10
n_2 = Número de plantas muestreadas por cada cuadrante = 5

$V_1 = (CM_{trat} - CM_{error}) / n_2 = (0.98 - 0.44) / 5 = 0.108$

CM_{trat} = Cuadrado medio de tratamientos, CM_{error} = Cuadrado medio de error

$V_2 = CM_{error} = 0.44$

$$EE_m = [(V_1/n_1)+(V_2/n_1*n_2)]^{1/2} = [(0.98/10)+(0.44/10*5)] = 0.14$$

Según Southwood (1978) El valor adecuado de EE_m para el caso de investigación es igual a 0.10 y para el caso de aplicación es igual a 0.25. En este trabajo el valor de EE_m resultó igual a 0.14.

Para el caso de aplicación debemos seleccionar aquella combinación de n_1 y n_1 para que la EE_m sea igual a 0.25.

Después de una serie de error y ensayo, llegamos con los valores de $n_1 = 3$ y $n_2 = 5$ para tener un EE_m igual a 0.2556.

Por tanto, los tamaños óptimos de la muestra para las unidades primarias (n_1) y secundarias (n_2) serían 3 y 5, respectivamente.

Estimación de submuestreo (ESM)

Hay que acordar que al usar diseños de tipo muestreo simple aleatorio, muestreo estratificado, muestreo sistemático, o muestreo conglomerado, el investigador asume que los datos se registran de forma correcta, y hay un cuadro (una lista completa de las unidades muestreales) disponible. Bajo estos supuestos uno puede hacer estimaciones correctas con sus límites de errores o intervalo de confianza.
Sin embargo, hay muchas situaciones en las cuales:

1. Los datos registrados no son correctas debido a la presencia del sesgo o equipo de muestreo.

2. No se cuenta con el cuadro para el muestreo.

3. No se puede conseguir datos correctos debido a la naturaleza delicada (secretos involucrados) de las preguntas.

En este caso, se utiliza el procedimiento de la estimación de submuestreo (ESM) cuando hay sesgo en los datos o cuando no existe un cuadro.

Aplicación

Suponemos que nos interesa tener información a base de un muestreo simple aleatorio (MSA) de *n* personas seleccionadas de una población del tamaño de *N*.

Tenemos *k* entrevistadores para hacer el trabajo, pero estos entrevistadores difieren en su manera de hacer las entrevistas y por tanto, su trabajo producirá sesgo en el muestreo. Por ejemplo, hacer entrevistas sobre el estado de salud (en una escala de Likert de 1 a 5, lo cual es subjetivo) de la gente requiere mucha capacidad, de tal forma que una entrevista detallada sobrestima las respuestas y viceversa.

Para obtener estimaciones correctas en estas situaciones se puede usar la estimación del submuestreo (ESM) de forma siguiente. Dividir de forma aleatorio la muestra del tamaño *n* elementos, en *k* submuestras, cada uno del tamaño *m* y asignar una entrevistador a cada una de *k* submuestras. Por tanto, $m = n/k$, y se puede seleccionar la *n* de tal forma para que la *m* sea un valor integro. La primera submuestra sería una muestra simple aleatoria del tamaño *m* seleccionada de *n* elementos de la muestra. La segunda submuestra sería una del tamaño de *m* tomadas de "*n-m*" elementos restantes. Seguir asi hasta que todos los elementos han sido divididos de forma al alzar en *k* submuestras. A las *k* submuestras se denominan también *submuestras interpenetrantes*.

Esperemos que algunos entrevistado subestiman y otras sobreestiman los datos, sin embargo, la media de todas mediciones de la muestra debe aproximarse el resultado verdadero, es decir, el sesgo de los entrevistadores tiene una media muy cercana al cero. Por tanto, la media de la muestra "i" o $M_i = 1/m(\sum X_{ij})$, es una estimación sin sesgo de la media poblacional, donde "*j*" denota el elemento en la muestra ($J = 1,\ldots, m$) y la $i = 1,\ldots, K$ submuestras).

$$\text{Media de la población: } M_p = 1/K\sum M_i$$

$$\text{Varianza de la muestra: } V_{(Mp)} = (N - n/N)\,[\sum(M_i - M_p)^2/K(K\text{-}1)]$$

$$\text{Error de estimación: } L = 2\sqrt{V_{(Mp)}}$$

Ejemplo

Queremos estimar la media de la longitud de flores (en centímetros) de la rosa de una comunidad (parque central de una ciudad) que tiene 800 flores, es decir, la totalidad de la población consiste en 800 elementos ($N = 800$). Hay 10 asistentes para apoyar al proyecto, cada uno con su equipo de medición. Debido a que sospechamos que las personas que conducen el muestreo vayan a generar sesgo en el trabajo, dividimos una muestra del tamaño de 80 en 10 submuestras del tamaño de 8 cada uno.

Supongamos que llegamos con las 10 medias siguientes para las 10 submuestras (Tabla 3).

Tabla 3. Las medias de 10 submuestras.

Submuestra	1	2	3	4	5	6	7	8	9	10
$\mathbf{M_i}$	5.9	5.8	6.1	6	6.1	5.7	5.8	5.6	5.9	6

Estimar la media de la población con su error de estimación.

$$\text{Media de la población: } M_p = 1/K\sum M_i = 1/10\ (5.9+5.8+\ldots+6) = 5.89$$

$$\text{Varianza de la muestra: } V_{(Mp)} = (N–n/N)[\sum(M_i-M_p)^2/K(K\text{-}1)] = (800\text{-}(80/800)[0.25/90] = 0.0025$$

$$\text{Error de estimación: } L = 2\sqrt{V}_{(Mp)} = 2\ \sqrt{0.0025} = 0.10$$

Por tanto, la media de la población es igual a 5.89 con un error de estimación igual a 0.10.

Muestreo de respuesta aleatoria (MRA)

Este tipo de muestreo está diseñado para obtener la respuesta correcta cuando los elementos (personas) no quieren dar una contestación correcta a la pregunta recibida. Por ejemplo, en caso de las tendencias políticas algunas gentes no desean contestar con franqueza la pregunta ¿Usted es una fascista? O en caso de las preferencias sexuales lo mismo puede ocurrir, es decir, la persona puede que no proporcione la respuesta correcta a la pregunta ¿Usted es un homosexual?
Muestreo de respuesta aleatoria (MRA) es una técnica de muestreo para estimar el porciento o la proporción de las personas que comparten cierto rasgo en común, sin conseguir de ellas respuestas directas a las preguntadas emitidas.
Primero hay que designar las personas de la población quienes *poseen* o *no* el rasgo de interés, en grupos *A* y *B*, respectivamente. Por tanto, cada persona de la población está ubicado en el grupo *A* o *B*. Déjale que *P* sea la proporción de la población que pertenecen al grupo *A*.
El objetivo de este muestreo es el estimar la *P* sin preguntar directamente a cada persona si pertenece o no al grupo *A*. Empecemos con un conjunto de cartas que son idénticos, con la excepción de que una fracción de ellas *F* es marcadas con la letra *A* y el resto (1 - *F*) esta marcadas con la letra *B*.
Se solicita de cada persona en la muestra que seleccione de forma aleatoria una carta de la conjunta de las cartas y luego diga “sí” si la letra en la carta concuerda con el grupo al cual esta persona pertenece, o diga “no” si la letra en la carta no concuerda con el grupo al cual esta persona pertenece. Se remplaza la carta antes de seguir con el siguiente sujeto. El entrevistador (quien realiza el muestreo) no observa la carta y solamente registra si la persona dice “sí” o “no”. Déjale que n_1

sea el número de personas quienes contestaron "sí". Una estimación sin sesgo de la *P* se obtiene por la siguiente ecuación:

Estimación de la proporción:

$$P = (F\text{-}1)/(2F\text{-}1) + [n_1/(2F\text{-}1)n]$$

El valor de *F* no debe ser ni menor que 0.5 ni igual a 1. Usualmente, un valor entre 0.5 y 1, por ejemplo, 0.75 es un valor adecuado.

Varianza estimada:

$$V_{(P)} = 1/n[1/16\ (F\text{-}1/2)^2 - (P\text{-}1/2)^2]$$

Límite de error de estimación:

$$L = 2\sqrt{V_{(P)}}$$

Ejemplo

Deseamos estimar la proporción de la gente quienes dan reportes falsos sobre sus declaraciones a la hacienda. Debido a que el entrevistado no va a admitir que sus declaraciones son falsas, debemos usra la técnica de muestreo de respuesta aleatoria (MRA). Suponer que 75% de las cartas están marcadas con la *F* (respuesta falsa) y el 25% con la letra *C* (respuesta correcta). Se realiza un muestreo simple aleatorio con el tamaño de la muestra igual a n = 400 de la población que debe pagar su impuesto a la hacienda.
En una entrevista se solicita de cada persona en la muestra que selecciona una carta de forma al azar y responde "si" si la letra concuerda al grupo al cual esta persona pertenece. El entrevistador arroja un valor para n_1 = 120. En base a estos datos, estimar la proporción de las personas quienes falsifican sus declaraciones ante la hacienda con un límite de error igual a 5%.

Solución

$$P = (F\text{ -1}) / (2F\text{ -1}) + [n_1 / (2F\text{ -1})n]$$

$$P = (0.75\text{ -1}) / (2*0.75\text{ -1}) + [120 / (2*0.75\text{ -1})\ 400 = 0.10$$

$$V_{(P)} = 1/n\ [1/16\ (F\text{ -1/2})^2 - (P\text{ -1/2})^2]$$

$V_{(P)} = 1/400\ [1/16\ (0.75 - ½)^2 - (0.10 - 1/2)^2] = 0.002$

$L = 2\sqrt{V_{(P)}}$

$L = 2\sqrt{0.002} = 0.092$

Concluimos que 10% de la población falsifica sus declaraciones con un límite de error igual a 0.092. En otras palabras, $P \pm L = 0.10 \pm 0.092$.

Conclusiones

Existen varios métodos de análisis de muestreo. En la selección de un método el investigador debe determinar si la distribución de los datos obedece al modelo normal; en caso de que la respuesta sea afirmativa, se puede utilizar alguno de los métodos apropiados según las condiciones de la población y la naturaleza de los propósitos del muestreo. Por ejemplo si existe una heterogeneidad en la población muy compleja, es decir, compuesta de varios niveles o etapas, entonces, el diseño óptimo de la muestra será el muestreo multietápico. En caso de las siguientes condiciones: a) Los datos registrados no son correctas debido a la presencia del sesgo o equipo de muestreo. b) No se cuenta con el cuadro para el muestreo. O c) No se puede conseguir datos correctos debido a la naturaleza delicada (secretos involucrados) de las preguntas, entonces, el diseño óptimo de la muestra sería por medio de la estimación de sub-muestreo. Y cuando los elementos (personas) no quieren dar una contestación correcta a la pregunta recibida, entonces el esquema óptimo del muestreo será muestreo de respuesta aleatoria.

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. 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.
Brewer, K.W.R. & M. Hanif. 1970. Dubin´s new multistage variance estimator. J. Roy. Stat. Soc. B32: 302-311.
Brooks, S. 1955. The estimation of an optimum subsampling numer. J. Am. Stat. Assoc. 50: 398-415.
Chapman, D.G. 1952. Inverse, multiple and sequential sample censuses. Biometrics, 8: 286-306.
Durbin, J. 1959. Design of multistage surveys for the estimation of smpling errors. Appl. Stat. 16: 152-164.

Jones, H.L. 1956. Investigations of the properties of a sample mean by employing random subsample means. J. Am. Stat. Assoc. 51: 54-83.
Warner, S.L. 1965. Randomized response: A survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. 60: 63-69.

*** Acerca de los Autores**

Badii, M.H. Es Profesor investigador del área de posgrado, UANL, México, mhbadiiz@gmail.com

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

Landeros, L. Es profesor investigador del área de posgrado, UAAAN, Saltillo, Coah., México

Cerna, E. Es profesor investigador del área de posgrado, UAAAN, Saltillo, Coah., México
UANL, San Nicolás, N.L., México, mhbadiiz@gmail.com, UAAAN, Saltillo, Coah., México