Muestreo Simple Aleatorio, Binomial, Estimación de Razón y Estratificado: Descripción y Análisis Comparativo

(Simple Random Samplings, Binomial, Ratio Estimation, and Stratified Sampling: Description and Comparative Analysis)

Badii, M.H., A. Guillen, J. [1]Valenzuela, E. [1]Cerna & J.L. Abreu *

Resumen. Se describen los muestreos de tipo simple aleatorio, binomial, estimación de razón e estratificado. Se hace un análisis comparativo entre el muestreo simple aleatorio y el muestreo estratificado, detallando e ejemplificando la relevancia del factor de ponderación en el caso de la estratificación. Se presentan por medio de ejemplos reales, las ecuaciones pertinentes y también las de tamaño óptimo de la muestra para cada uno de éstos tipos de muestreo.

Palabras claves. Análisis comparativo, clases de muestreo, tamaño óptimo de la muestra.

Abstract. Simple random sampling, binomial sampling, ratio estimation sampling and stratified sampling are described. A detailed comparison between simple random sampling and stratified sampling highlighting the relevance of weight factor for stratification is carried out. General equations and those for estimation of optimal sample size for each case is provided using real case examples.

Keywords. Comparative analysis, optimal sample size, sample types.

Muestreo simple aleatorio (MSA)

Los fundamentos y las matemáticas del muestreo han sido reportados (Cochran, 1977; Cornfield, 1951; Deming, 1960; Hansen et al., 1953; Kish, 1965; Mendenhall, 1971). Este tipo de diseño se usa cuando el ambiente es homogéneo. Se establece esta homogeneidad en base del procedimiento de ANOVA. Con este tipo de diseño, el investigador selecciona de forma al azar un número óptimo de las unidades muestreales de la totalidad del cuadro y luego cuantifica los elementos de cada U.M. y en base a las ecuaciones pertinentes se estiman los parámetros poblacionales de interés. Los conceptos del muestreo simple aleatorio, estratificado, binomial y estimación de razón han sido analizados detalladamente (Armitrage, 1947; Birnbaum, 1950; Brewere, 1963; Burstein, 1975, Chatterjee, 1967, 1968, 1972; Cochran, 1946; David & Sukhtme, 1974).

Ejemplos del MSA

Ejemplo A. Estimar la media y 95% de intervalo de confianza (IC)

Suponemos que nos interesa investigar en base de un estudio de auditoría, las cuentas bancarias de los pacientes de un hospital con el objetivo de estimar el saldo promedio de cuenta por cada paciente. Se toma una muestra simple aleatoria del tamaño de 200 cuentas de un total de 1000 cuentas y según este muestreo el promedio es de $94.22 ($m$) pesos y la varianza es de $445.21 pesos ($V$).
La pregunta sería el estimar el promedio per cápita por la población (μ) con su intervalo de confianza o el error de estimación ($L = 2EE$), donde, L es el error de estimación y EE es el error estándar.

Solución

$\mu = m \pm L$
$L = 2\ EE$
$EE = (V/n)^{1/2} * (1- \varphi)^{1/2}$

Donde,

$(1- \varphi)$ = la corrección para la población finita y φ = la fracción de la muestra. Para los valores de $\varphi<10\%$, la $(1- \varphi)$ es igual a 1.

$EE = (445.21/200)^{1/2} * (1- 200/1000)^{1/2}$
$EE = 1.334$ $

$\mu = m \pm L = 94.22 \pm 2(1.334)$

Por tanto,

$\mu_i = 94.22 - 2(1.334) = 91.552$ límite inferior en pesos
$\mu_s = 94.22 + 2(1.334) = 96.888$ límite superior en pesos

Nota

En la ecuación $EE = (V/n)^{1/2}(1- \varphi)^{1/2}$, la variable $\varphi = n/N$, se denomina de fracción de la muestra, es decir, la fracción de la población que esta incluida en la muestra. El factor $(1- \varphi)$ es la *corrección para la población finita*, y es igual a cero cuando $n = N$. En una población con una cantidad fija de varianza, la EE depende principalmente en el tamaño de la muestra (n) y tiene poca dependencia sobre la φ. Por ejemplo, para un valor fijo de la varianza, un tamaño de la muestra igual al 100

es tan preciso para una población del tamaño de 2,000 elementos comparada con otra del tamaño de 2, 000,000 elementos. Suponer $V = 4$:

$$EE = (V/n)^{1/2}(1- \varphi)^{1/2}$$

Para $N = 2000$: $EE = (4/100)^{1/2} (1 – 100/2000)^{1/2} = 0.1949$
Para $N = 2000,000$: $EE = (4/100)^{1/2} (1–100/2000,000)^{1/2} = 0.1999$

Es obvia la cercanía de estos dos valores de *EE*, a pesar de que una población es 1000 veces más grande que la otra.

Ejemplo B. Estimar el total de la población y 95% d IC

Vamos a suponer que una empresa esta preocupada por el número de las horas perdidas semanalmente por los empleados en hacer otras actividades fuera de sus obligaciones laborales y como consecuencia, estas horas significan pérdidas por la empresa.
En base de un MSA de 50 empleados resulta que cada empleado semanalmente desperdicia, en término medio (m), 10.31 horas con una varianza (V) igual a 2.25 horas^2. Si el total de los empleados de la empresa es 750 personas, cual sería la totalidad ($T=Nm$) de horas perdidas semanalmente por los trabajadores de esta empresa, con su grado de error de estimación $L = 2\ EE$.

Solución

$T = N\,m \pm L$
$L = 2EE$
$EE = (N^2V/n)^{1/2} * (1- \varphi)^{1/2}$

Donde,

N = Total de los elementos en la población, y los demás anotaciones como arriba mencionadas
$T = N\,m = 750\ (10.31) = 7{,}732.5$ horas
$L = 2EE$
$EE = (N^2V/n)^{1/2} * (1- \varphi)^{1/2} = [750^2\ (2.25/50)]^{1/2} = (1 – 50/750)^{1/2} = 153.704$ pesos
$T = N\,m \pm L$

$T_i = 7{,}732.5 – 2(153.704) = 7{,}425.1$ límite inferior en horas
$T_s = 7{,}732.5 – 2(153.704) = 8{,}039.9$ límite inferior en horas

C. Estimar el tamaño óptimo de la muestra

Suponemos que deseamos estimar el aumento en el peso corporal de 1000 pollos sujetos a una dieta específica ($N = 1000$).
Según un muestreo piloto, ya sabemos que la varianza del aumento del pesos corporal de los pollos es igual a 36 gramos^2 ($V = 36$).
La pregunta sería cual es el tamaño óptimo de la muestra para esta población con un error de estimación igual a 1000 gramos (L =1000).

Solución

$N_{opt} = 4N^2*(V/L^2)$

Donde,

N_{opt} = Tamaño óptimo de la muestra, y todas las notaciones como antes mencionadas.
$N_{opt} = 4N^2*(V/L^2) = 4(1000^2)*(36/1000^2) = 144$ individuos

La ecuación de N_{opt} para estimar el promedio poblacional: $N_{opt} = 4\ (V/L^2)$
La ecuación de N_{opt} para estimar el total poblacional: $N_{opt} = 4N^2(V/L^2)$

Por ejemplo

La N_{opt} para estimar el promedio poblacional del *ejemplo A* es igual a:

$$N_{opt} = 4(V/L^2) = 4(445.21/2.668^2) = 250.174 \approx 250 \text{ cuentas}$$

La N_{opt} para estimar el total poblacional del *ejemplo B* es igual a:

$$N_{opt} = 4N^2*(V/L^2) = 4\ (750^2)(2.25/307.408^2) = 53.571 = 54 \text{ empleados}$$

Muestreo binomial (MB)

Este clase de muestreo se utiliza para los datos en la escala discreta y se trata de las situaciones en donde la respuesta puede tener uno de las dos siguientes respuestas: favorable o exitoso con su probabilidad de éxito o la ocurrencia (p_i) y no favorable o

fracaso con su probabilidad de fracaso o la no ocurrencia igual a 1 - p_i = q_i. Las ecuaciones para n_{opt} son:

Población finita: $n_{opt} = (N_t\, pq) / \{[(N_t - 1)b^2/Z^2] + pq\}$
Población infinita: $n_{opt} = (Z^2 pq) / b^2$

Donde,

n_{opt} = Tamaño óptimo de la muestra
N_t = Número total de las unidades de la muestra
p = Probabilidad de la ocurrencia
q = Probabilidad de no ocurrencia
Z = 1.96 = Valor de la tabla para IC a nivel 95%
b = El margen de error o error de estimación

Ejemplo para la población finita

Suponer los datos siguientes: una inspección de un cultivo que contiene 500 plantas individuales resulta en que 50% de estas plantas son enfermas. Suponiendo que aceptamos un margen de error igual a 10%, cuál sería la n_{opt}?

N_t = 500
p = 0.5
q = 0.5
Z = 1.96
b = 0.10

Calcular n_{opt}

$n_{opt} = (N_t\, pq) / \{[(N_t - 1)b^2 / Z^2] + pq\}$
$n_{opt} = (500 * 0.5 * 0.5) / \{[(500 - 1)0.10^2 / 1.96^2] + (0.5*0.5)\} = 80.7$

Nota: Debido a que 80.7 es mayor que el 10% (o 50) de número total de las unidades de la muestra, se calcula un n_{opt} ajustado (n_{opta}) según la siguiente ecuación:

$n_{opta} = n_{opt} / (1 + \varphi)$

Donde,
$\varphi = n_{opt} / N_t$

$n_{opta} = 80.7 / (1 + 80.7/500) = 69.48 \approx 70$

Ejemplo para la población infinita

Estimar n_{opt} para datos siguientes de una población infinita.
$p = 0.5$
$q = 0.5$
$Z = 1.96$
$b = 0.10$

Calcular n_{opt}

$n_{opt} = (Z^2pq) / b^2$
$n_{opt} = (1.96^2\ (0.5*0.5) / 0.10^2 = 0.9604 / .01 = 96.04$

Nota

Debido a que en el muestreo binomial (MB) se necesita tener una idea previa sobre el valor de "*p*", sería necesario abundar más sobre este variable. Si el valor de la *p* cae entre 35% y 65%, el producto (*pq*) varía muy poco. Por tanto, el cálculo de n_{opt} requiere una intuición aproximada (no precisa) sobre el valor de *p*. Sin embargo, si el valor de *p* está cerca de 0% o 100%, la estimación de n_{opt} requiere una adivinanza más precisa.

Estimación de razón (ER)

Para clarificar más el concepto de MSA, se puede usar este tipo de diseño muestreal de una manera distinta denominada muestreo con el diseño de estimación de razón (ER). Se utiliza este diseño cuando existen las siguientes condiciones:

1. La respuesta esta correlacionada con una variable auxiliar.

2. La varianza de la variable de respuesta esta proporcional a la de variable auxiliar.

3. Cuando se puede medir las 2 variables sobre cada elemento de la muestra.

La estimación de razón genera estimaciones de la media y el total poblacional de mayor precisión comparado con el muestreo simple aleatorio, y esto

se debe a la que ER utiliza información adicional (relación entre la variable de respuesta y la de auxiliar). Además, la estimación de razón es más efectiva cuando hay una correlación de más de 50% entre las dos variables.

Ecuaciones para estimación de razón

$$r = \sum Y_i / \sum X_i$$
$$T_y = r\, T_x$$
$$\mu_x = \sum X_i / n$$
$$\mu_y = r\, \mu_x$$

$$V_r = [(N - n) / nN]\, (1/(\mu_x)^2\, [\sum (Y_i - rX_i)^2 / (n - 1)]$$
$$EE_r = [V_r]^{1/2}$$
$$L = 2EE_r$$

$$V_{(\mu y)} = V_r * (\mu_x)^2$$
$$EE_\mu = [V_{(\mu y)}]^{1/2}$$
$$L = 2\, EE_\mu$$

$$V_{(Ty)} = V_r * (T_x)^2$$
$$EE_{(Ty)} = [V_{(Ty)}]^{1/2}$$
$$L = 2EE_{(Ty)}$$

Tamaño de la muestra para estimar la media de la población:

$$n_\mu = N\, [\sum (Y_i - rX_i)^2/(n - 1)]/ND + [\sum (Y_i - rX_i)^2/(n - 1)]$$
$$D = L^2/4$$

Tamaño de la muestra para estimar el total de la población:

$$n_{(Ty)} = N\, [\sum (Y_i - rX_i)^2/(n - 1)]/ND + [\sum (Y_i - rX_i)^2/(n - 1)]$$
$$D = L^2/4N^2$$

Donde,

r = Razón
T_y = Tamaño total de la población
T_x = Media de la población
V_r = Varianza de la razón
$V_{(\mu y)}$ = Varianza para estimar la media poblacional

$V_{(Ty)}$ = Varianza para estimar la total poblacional
Los demás anotaciones son claramente lógicos y como antes descritas.

Ejemplo 1 de ER

Los datos de la Tabla 1 se usan para ilustrar este tipo de muestreo.

Tabla 1. Peso y contenido de azúcar en kg.

i **(naranja)**	Y_i **(contenido de azúcar)**	X_i **(peso)**
1	0.021	0.40
2	0.30	0.48
3	0.025	0.43
4	0.022	0.42
5	0.033	0.50
6	0.027	0.46
7	0.019	0.39
8	0.021	0.41
9	0.023	0.42
10	0.025	0.44
$n = 10$	$\sum Y_i = 0.246$	$\sum X_i = 4.35$

Solución

$V_r = [(N - n) / nN] (1/(\mu_x)^2 [\sum(Y_i - rX_i)^2 / (n - 1)]$

$r = \sum Y_i / \sum X_i = 0.246 / 4.35 = 0.05788235$
$T_y = r\, T_x = 0.05788235\ (1800) = 101.78$
$\mu_x = \sum X_i / n = 4.35 / 10 = 0.435$
$\mu_y = r\, \mu_x = 0.05788235\ (0.435) = 0.025178822$
$V_r = [(N-n)/nN](1/(\mu_x)^2[\sum(Y_i - rX_i)^2/(n1)] = [(4,13810)/10(4,138)](1/(0.435)^2[(0.021 - 0.05788235*0.40) +.(0.030-0.05788235*0.408)]^2/(10-1)] = 0.0000030627$

$V_{(Ty)} = V_r * (T_x)^2 = 0.0000030627\ (1,800)^2 = 9.9231$
$L = 2EE_{(Ty)} = 2*(9.9231)^{1/2} = 3.42$

Por tanto, el tamaño total de la población y su error de estimación es como sigue:

$T_y \pm L = 101.78 \pm 3.42$

Límite inferior = 98.36

Límite superior = 105.20

$\mu_y = 0.025178822$

$V_{(\mu y)} = V_r * (\mu_x)^2$

$V_{(\mu y)} = 0.05788235 * (0.435)^2 = 5.80944*10^{-7}$

$L = 2EE_{(\mu y)} = 2[5.80944*10^{-7}]^{1/2} = 1.52439*10^{-3}$

Por consiguiente, la media de la población y su error de estimación son como sigue:

$\mu_y \pm L = 4.35 \pm 0.025$

Límite inferior = 4.325

Límite superior = 4.375

Calcular el tamaño óptimo de la muestra para estimar la media poblacional.

Ecuación del tamaño óptimo de la muestra: $n = N\delta^2 / ND + \delta^2$

Ejemplo 2 de ER

Suponemos que deseamos estimar la media poblacional (μ_y) de los árboles por hectárea de una huerta con $N = 1000$ hectáreas. Basado en fotos aéreas, se conoce el número total de los árboles y por tanto, la media de la variable auxiliar, es decir, la μ_x. Cuál sería el tamaño óptimo de la muestra para estimar la media poblacional (μ_y) con un límite de error igual a 1 árbol ($L = 1$).

Solución

En la ecuación del tamaño óptimo de la muestra ($n = N\delta^2/ND+\delta^2$), no se conoce el valor de la δ^2 y por tanto, hay que hacer un muestreo piloto, con la información que en un día fácilmente, se puede contar 10 parcelas ($n_a = 10$) según la tabal siguiente (Tabla 2).

Tabla 2. Estimación aérea (X) y número de árboles (Y) en el muestreo piloto.

***i* (parcela)**	**X_i (estimación áerea)**	**Y_i (# actual de árboles)**
1	23	25
2	14	15
3	20	22
4	25	24
5	12	13
6	18	18
7	30	35
8	27	30
9	08	10
10	31	29
$n = 10$	**$\sum X_i = 208$**	**$\sum Y_i = 221$**

$r = \sum Y_i / \sum X_i = 221/208 = 1.06$

$\delta^2 = [\sum(Y_i - rX_i)^2/(n - 1)] = [\sum(Y_i)^2 - 2\, r\sum X_i Y_i + r^2\sum(X_i)^2]/(n - 1)$
$= [5{,}469 - 2(1.06)\ 208*221 + (1.06)^2\ (208)^2] / 9 = 4.21$

Ahora se pude estimar el tamaño óptimo de la muestra de manera siguiente:

$n = N\delta^2/ND + \delta^2$
$D = L^2 / 4 = 1^2/4 = 0.25$
$n = (1000*4.21) / (1000*0.25)+4.21 = 16.56 \approx 17$ parcelas a muestrar.

Muestreo estratificado (ME)

Si existe un gradiente de variabilidad en el hábitat, entonces se divide el medio en estratos que reflejen este gradiente, de esta manera se divide la población en varias subpoblaciones o estratos y en cada estrato se procede con el muestreo simple aleatorio. Se determina la existencia de la heterogeneidad mediante el Análisis de Varianza (ANOVA).

Ejemplos del ME

Suponemos que una compañía desea tener información sobre qué tanto debe enfatizar una comercial en la tele en una ciudad, sobre sus productos que puede vender en esta ciudad. En otras palabras, la empresa va a formular un plan de muestreo para estimar el número de horas por semana que cada familia en la ciudad observa la tele.

La ciudad tiene 3 áreas diferentes. La sección **A** que está cerca de una empresa grande y la mayoría de las gentes que habitan las casas en esta zona son trabajadores de la empresa con niños de edad escolar. La zona **B** es un área exclusiva que contiene gentes de mayor edad y con pocos niños de edad escolar. La zona **C** es un área rural. Por tanto, en forma general, sospechamos que la ciudad esta compuesto de 3 estratos o secciones o sub-poblaciones.

Hay 155 (N_1) familias en la zona A, 62 (N_2) en B y 93 (N_3) en C. Por tanto, en total hay 310 (N_t) familias en toda la ciudad. Suponemos que hay recursos financieros para realizar un muestreo de tamaño de 40 familias.

Una forma racional de muestreo sería un esquema de muestreo estratificado. Para poder asignar los recursos de forma proporcional, se divide 40 (n) entre 310 (N_t) y se arroja un resultado igual a 0.129. Se utiliza esta proporción para estimar el número inicial de las familias a muestrar (tamaño de la muestra) en cada estrato dentro de un esquema de muestreo estratificado de forma siguiente:

Número inicial (n_1) de familias a muestrear en estrato **A** será:

A: (0.129)*(155 = N_1) = 20 casas

Número inicial (n_2) de familias a muestrear en estrato **B** será:

B: (0.129)*(62 = N_2) = 8 casas

Número inicial (n_3) de familias a muestrear en estrato **C** será:

C: (0.129)*(93 = N_3) = 12 casas

Hasta este momento la división de la ciudad en 3 estratos ha sido resultado de la intuición, sin embargo, se debe comprobar estadísticamente la existencia de estos estratos, es decir, se debe conducir una ANOVA para este fin, de forma siguiente.

Procedimiento

Se realiza un muestreo simple aleatorio por cada estrato y se arroja los siguientes resultados (Tabla 3 y Tabla 4).

Tabla 3. Número de horas que cada familia (casa) se gasta en observar tele.

Estrato A				**Estrato B**				**Estrato C**			
35	28	26	41	27	4	49	10	8	15	21	7
43	29	32	37	15	41	25	30	14	30	20	11
36	25	29	31					12	32	34	24
39	38	40	45								
28	27	35	34								
n_1 20	m_1 33.9	V_1 35.358	N_1 155	n_2 8	m_2 25.125	V_2 232.411	N_2 62	n_3 12	m_3 19	V_3 87.636	N_3 93

Tabla 4. Tabla de ANOVA para los datos de la Tabla 3.

FV	Gl	SC	CM	F_{cal} ($F_{tab,\ 2,\ 37} = 3.23$)
Estratos	2	1,730.100	865.0500	9.81*
Error	37	3,262.678	88.1804	
Total	39	4,992.775		9.8 (F_{cal}) > 3.23 (F_{tab})

Por tanto, debido al valor significativo de F para los tratamientos (estratos), concluimos que existen distintos estratos.

Una vez establecida estadísticamente los estratos, ya podemos seguir con los cálculos siguientes.

A. Estimar la media poblacional con 95% de intervalo de confianza

$\mu = m_{Est} = (1/N)\sum N_i m_i$
$EE(m_{Est}) = [(1/N_t^2)*\sum(V_i/n_i)\{(N_i)^2(1-\varphi_i)\}]^{1/2}$

Donde,

μ = Media de la población
m_{Est} = Media estratificad
N_i = Tamaño del estrato "*i*"
m_i = Media del estrato "*i*"
$EE(m_{Est})$ = Error estándar de la media estratificada
N_t = Total de la población
V_i = Varianza del estrato "*i*"
n_i = Tamaño de la muestra del estrato "*i*"
φ_i = proporción de la muestra del estrato "*i*"

$\mu = m_{Est}=(1/N)\sum N_i m_i = (1/310)[(155*33.9)+(62*25.125)+(93*87.636)] = 27.675$

Es decir, en término medio cada familia gasta 27.675 horas semanales en observar televisión.

La estimación de error estándar de la media del estrato "*i*"

$EE(m_{Est}) = [(1/N_t^2)*\sum(V_i/n_i)\{(N_i)^2(1-\varphi_i)\}]^{1/2}$
$EE(m_{Est})=[(1/310^2)\{[35.358/20(155^2(1\text{-}20/155)]*[232.411/8(62^2(1\text{-}8/62)]*$
$[87.636/12\ (93^2(1\text{-}12/93)]\}^{1/2} = 1.4035$

Por tanto, la media estratificada y su límite de error (error de estimación) son como sigue.

$\mu = m_{Est} \pm L,$

Donde,
$L = 2\ EE(m_{Est})$

$\mu = 27.675 \pm 2(1.4035)$
Límite inferior = 24.868
Límite superior = 30.482

De la misma manera se puede estimar las medias de cada estrato y su nivel de error de estimación (L), usando la ecuación siguiente para estimar el EE de cada estrato:

$EE\text{est}_i = [V_i/n_i(1 - \varphi_i)]^{1/2}$
$EE\text{est}_1 = 1.2510$
$EE\text{est}_2 = 5.0270$
$EE\text{est}_3 = 2.5206$

B. Estimar el total de la población (T) con 95% de IC

Procedimiento

$T = N_t\ m_{Est} = \sum N_i\ m_{Est}$
$T = N_t\ m_{Est} = 310\ (27.675) = 8{,}579.25$ horas totales por la población.
$EE_{(T)} = (N_t)\ EE(m_{Est}) = 310\ (1.4035) = 435.085$
$T \pm L = 8{,}579.25 \pm [2(435.085)]$

C. Estimar el tamaño óptimo de la muestra (n_{opt})

C_1. Modalidad de asignación proporcional

En esta modalidad se asigna el mismo factor de ponderación (W_i) a cada uno de los estratos.

Ecuaciones

$n = \sum(N_i V_i/W_i)\ /(N_t)^2 D + \sum N_i\ V_i$

Donde,

n = Tamaño óptimo de la muestra
N_i = Tamaño del estrato "i"
V_i = Varianza del estrato "i"
W_i = Fracción de n asignado al estrato "i" = factor de ponderación para "i"
N_t = Total de la población
$D = L^2/ N^2$: En caso de estimar la media poblacional
$D = L^2/4N^2$: En caso de estimar la total de la población

Ejemplo

A. Calcular n_{opt} para estimar la *media poblacional*:

$V_1 = 25$
$V_2 = 225$
$V_3 = 100$
$W_1 = W_2 = W_3 = 0.333$
***L* = 2**

Calcular el tamaño óptimo (n_{opt}) de la muestra para estimar la media poblacional, para todos y cada uno de los estratos:

$$n_{opt} = \sum(N_i V_i/W_i) /(N_t)^2 D + \sum N_i V_i$$
$$D = L^2/4$$

$n_{ópt}$=[(155*25)/0.333+(62*225)/0.333+(93*100)/0.333]/{$310^2(2^2/4)$+[(155*25)+(62*225)+(93*100)]}=56.7 ≈ 57

Por tanto, el tamaño óptimo (n_{opt}) para estimar la media poblacional para cada estrato es:

$n_1 = n\ (W_1) = 57\ (0.333) = 19$
$n_2 = n\ (W_2) = 57\ (0.333) = 19$
$n_3 = n\ (W_3) = 57\ (0.333) = 19$

B. Calcular $n_{ópt}$ para estimar el *total de la población*:

$V_1 = 25$
$V_2 = 225$
$V_3 = 100$
$W_1 = W_2 = W_3 = 0.333$

***L* = 400**

Calcular el tamaño óptimo (n_{opt}) de la muestra para estimar el tamaño total de la población, para todos y cada uno de los estratos:

$$n_{opt} = \sum(N_i V_i/W_i) /(N_t)^2 D + \sum N_i V_i$$
$$D = L^2/4N^2$$

n_{opt}
$=[(155*25)/0.333+(62*225)/0.333+(93*100)/0.333]/\{310^2$
$(400^2/4[310^2])+[(155*25)$
$+(62*225) + (93*100)]\} = 105$

Por tanto, el tamaño óptimo (n_{opt}) para cada estrato para estimar el total de la población es:

$n_1 = n\ (W_1) = 105\ (0.333) = 35$
$n_2 = n\ (W_2) = 105\ (0.333) = 35$
$n_3 = n\ (W_3) = 105\ (0.333) = 35$

C_2. Modalidad de asignación óptima

Hay que recordar que el objetivo del muestreo es obtener información lo más representativa de la población con el mínimo costo y mínima varianza. Dentro de este contexto, el mejor plan del muestreo estratificado es aquel que este rejido por los siguientes factores:

1. El número total de los elemento de cada estrato (N_i)

2. El nivel de la varianza de cada estrato (V_i)

3. El costo de obtener un elemento de cada estrato (C_i)

Por tanto, si el costo de obtener el elemento varía entre los estratos, debemos tomar una muestra del tamaño pequeño del estrato con alto costo. Una asignación óptima es aquel que minimiza el costo del muestreo por cada unidad de varianza o minimiza la varianza por cada unidad del costo.

Calcular el tamaño óptimo de la muestra (n_{opt}) para estimar la media poblacional, para todos y cada uno de los estratos:

$$N_{opt} = \sum([N_i]^2 V_i/W_i) / (N_t)^2 D + \sum N_i V_i$$
$$D = L^2 / 4N^2$$
$$W_i = [N_i V_i^{1/2} / C_i^{1/2}] / \sum[N_i V_i^{1/2} / C_i^{1/2}]$$

Donde, todas las notaciones como arriba escrita.

A. Datos para calcular n_{opt} para estimar la *media poblacional*

$V_1 = 25$
$V_2 = 225$
$V_3 = 100$
$C_1 = C_2 = 9$ pesos
$C_3 = 16$ pesos
$\boldsymbol{L = 2}$

Es decir, es más costoso obtener información del área rural que los otros 2 sectores.

Ecuaciones y procedimiento

$$W_i = [N_i V_i^{1/2} / C_i^{1/2}] / \sum[N_i V_i^{1/2} / C_i^{1/2}]$$

$W_1 = [155 * 25^{1/2} / 9^{1/2}] / \sum[155 * 25^{1/2} / 9^{1/2}] = 0.32$
$W_2 = [62 * 225^{1/2} / 9^{1/2}] / \sum[62 * 225^{1/2} / 9^{1/2}] = 0.39$
$W_3 = [93 * 100^{1/2} / 16^{1/2}] / \sum[93 * 100^{1/2} / 16^{1/2}] = 0.29$

$$N_{opt} = \sum([N_i]^2 V_i/W_i) / (N_t)^2 D + \sum N_i V_i$$
$$D = L^2/4$$

$n_{opt} = \{(155^2 * 25 / 0.32) + (62^2 * 225 / 0.39) + (93^2 * 100 / 0.29)\}/(310)^2(2^2/4) + \{(155*25) + (62*225) + (93*100)\} = 57.43 \approx 58$ = tamaño óptimo de la muestra (n_{opt}) para estimarla media poblacional.

Por tanto: el tamaño óptimo (n_{opt}) para estimar la media poblacional para cada estrato es:

$n_1 = n_{opt}\,(W_1) = 58\,(0.32) = 18.5 \approx 18$
$n_2 = n_{opt}\,(W_2) = 58\,(0.39) = 22.6 \approx 23$

$n_3 = n_{opt}\,(W_3) = 58\ (0.29) = 16.8 \approx 17$

B. Datos para calcular n_{opt} para estimar el *total de la población*

$V_1 = 25$
$V_2 = 225$
$V_3 = 100$
$C_1 = C_2 = 9$ pesos
$C_3 = 16$ pesos
$\boldsymbol{L = 2}$

Es decir, es más costoso obtener información del área rural que los otros 2 sectores.

Procedimiento e ecuaciones

$$W_i = [N_i\ V_i^{1/2} / C_i^{1/2}] \,/\, \sum[N_i\ V_i^{1/2} / C_i^{1/2}]$$

$$W_1 = [155*25^{1/2}/9^{1/2}]/\sum[155 * 25^{1/2}/9^{1/2}]*[62*225^{1/2}/9^{1/2}]*[93*100^{1/2}/16^{1/2}] = 0.32$$

$$W_2 = [62*225^{1/2}/9^{1/2}]/\sum[155* 25^{1/2}/9^{1/2}]*[62*225^{1/2}/9^{1/2}]*[93*100^{1/2}/16^{1/2}] = 0.39$$

$$W_3 = [93*100^{1/2}/16^{1/2}]/\sum[155 * 25^{1/2}/9^{1/2}]*[62*225^{1/2}/9^{1/2}]*[93*100^{1/2}/16^{1/2}] = 0.29$$

$$n_{opt} = \sum([N_i]^2 V_i/W_i)\,/(N_t)^2 D + \sum N_i\ V_i$$
$$D = L^2/\,4N^2$$

$n_{opt} = \{(155^2 * 25/0.32) + (62^2*225/0.39) + (93^2 * 100/0.29)\}/(310)^2(2^2/4[310^2]) + \{(155*25) + (62*225) + (93*100)\} \approx 261$ Tamaño de la muestra (n_{opt}) para estimarla total de la población.

Por tanto: el tamaño óptimo de la muestra (n_{opt}) para estimar la total de la población para cada estrato es:

$n_1 = n_t\,(W_1) = 261\ (0.32) = 83.52 \approx 83$
$n_2 = n_t\,(W_2) = 261\ (0.39) = 101.79 \approx 102$
$n_3 = n_t\,(W_3) = 261\ (0.29) = 75.69 \approx 76$

Nota

Si los costos son iguales ($C_1 = C_2 = C_3$), entonces en la ecuación del factor de ponderación: $W_i = [N_i V_i^{1/2}/C_i^{1/2}] / \sum[N_i V_i^{1/2}/C_i^{1/2}]$, se eliminan los costos y consecuentemente, la ecuación anterior de n_{opt} se convierte a la siguiente ecuación más simple: $W_i = [N_i V_i^{1/2}] / \sum[N_i V_i^{1/2}]$. Substituyendo los valores de las varianzas y los tamaños de los estratos, se generan los valores de los factores de ponderación:

$W_1 = [155*25^{1/2}] / \sum[155 * 25^{1/2}]* [62*225^{1/2}]*[93*100^{1/2}] = 0.29$
$W_2 = [62 *225^{1/2}] / \sum[155* 25^{1/2}]* [62*225^{1/2}]*[93*100^{1/2}] = 0.35$
$W_3 = [93 * 100^{1/2}] / \sum[155 * 25^{1/2}]*[62*225^{1/2}]*[93*100^{1/2}] = 0.35$

$$n_{opt} = \sum([N_i]^2 V_i/W_i) /(N_t)^2 D + \sum N_i\ V_i,\ D = L^2/\ 4$$

n_{opt}={$(155^2*25/0.29)+(62^2*225/0.35)+(93^2*100/0.35)$}/$(310)^2(2^2/4)$+{$(155*25)+(62*22)+(93*100)$}=56.7≈57 tamaño óptimo de la muestra (n_{opt}) para estimar la la media poblacional.
Por tanto, el tamaño óptimo (n_{opt}) para estimar la media poblacional para cada estrato es:

$n_1 = n_{opt}\ (W_1) = 57\ (0.29) = 16.53 \approx 17$
$n_2 = n_{opt}\ (W_2) = 57\ (0.35) = 19.95 \approx 20$
$n_3 = n_{opt}\ (W_3) = 57\ (0.35) = 19.95 \approx 20$

Comparación del sesgo entre MSA y ME

A. Caso de MSA

Supongamos que tenemos 6 objetos ($N = 6$) con los valores siguientes: **a** = 1, **b** = 2, **c** = 4, **d** = 6, **e** = 7, y **f** = 16. La suma de estos 6 valores es igual a $T = 36$. La idea es que deseamos estimar la T mediante un MSA del tamaño $n = 3$.
Debido a que $N = 6$ y la $n = 3$, entonces, la manera más simple sería multiplicar el resultado de la muestra por 2.
Ahora, podemos escribir todas las posibles combinaciones de las muestras del tamaño de 3, y hacer estimación de cada muestra y luego ver qué tan cercana están estas estimaciones en comparación con el valor verdadero de $T = 36$.
Usando la formula de la combinación, existe un total de 20 diferentes muestras del tamaño 3 cada una:

$$_6C_3 = 6!\ /\ \{3!(6-3)!\} = (6*5*4*3*2*1)\ /\ (3*2*1)\ (3*2*1*) = 20$$

Los resultados de estas 20 muestras se observan en la Tabla A. Algunas muestras como la de **abf**, y **cde** son muy buenas en el sentido de que estiman muy bien la realidad, mientras que otras muestras como **abc** dan resultado muy pobre y lejos de la realidad.

De antemano no sabemos cómo cada muestra individual funciona; algunos bien y algunos mal.

La media del error de estimación (tomando en cuenta los signos) se denomina el sesgo de la estimación o del plan del muestreo. Un sesgo positivo significa sobreestimación y viceversa.

En base a los resultados de la Tabla A, podemos concluir que nuestro plan del muestreo es *sin sesgo*, debido a que el promedio de las 20 estimación es exactamente 36, y por consecuencia, el error de estimación es igual a cero.

En un esquema del MSA, este resultado se consigue para cualquier población de cualquier tamaño. Por tanto, *una característica favorable del MSA es estimaciones sin sesgo*, aunque poco sesgo es también aceptable.

Para estimar la *exactitud* del plan de muestreo de MSA, utilizamos el *cuadrado medio del error de estimación* (*CME*); es decir, $CME = \sum(\text{error de estimaciones})^2/20 = \sum(X_i - m_i)^2/n = \sum(X_i)^2 - (\sum X_i)^2/n = [14^2+18^2+\ldots+58^2] - [(14+18+\ldots+58)^2/20] = 175.2$, y el error estándar de la estimación o *EE* de estimación $= (\text{CME})^{1/2} = (175.2)^2 = 13.2$.

Tabla A. Resultados de las 20 muestras del tamaño de $n = 3$ del MSA.

Composición de la muestra	**Total de la muestra [T_{MSA}]**	**Estimación del total T_E = [T_{MSA}] * 2**	**Error de estimación T_E – 36**
Abc	7	14	-22
Abd	9	18	-18
Abe	10	20	-16
Abf	19	38	+02
Acd	11	22	-14
Ace	12	24	-12
Acf	21	42	+06
Ade	14	28	-08
Adf	23	46	+10
Aef	24	48	+12
Bcd	12	24	-12
Bce	13	26	-10
Bcf	22	44	+08
Bde	15	30	-06
Bdf	24	48	+12
Bef	25	50	+14

Cde	17	34	-02
Cdf	26	52	+16
Cef	27	54	+18
Def	29	58	+22
Media del funcionamiento del plan de muestreo		**18 (2) = 36**	**0**

Como resultado, podemos comentar que el plan de MSA proporciona una estimación del total de la población que esta *sin sesgo* y que tiene un *EE* = 13.2. Este valor constituye el 37% del tamaño total (36) de la población y por consecuencia, este plan de muestreo *no es muy exacto* para la población ya que nos ofrece los siguientes límites inferiores y superiores:

$T \pm EE = 36 \pm 13.2$
Límite inferior: 36 – 13-2 = 22.8
Límite superior: 36 + 13-2 = 49.2

B. Caso de ME

Vamos a suponer que antes de planear la muestra, esperemos que el elemento **f**, da un valor más alto que cualquier otro miembro de la población. ¿Cómo podemos utilizar esta información de buena manera? Es obvio que la estimación de la muestra depende principalmente, si el elemento **f** cae o no dentro de la muestra. Según la Tabla A se puede verificar que cualquier muestra que contiene el elemento **f** produce una sobreestimación y viceversa.
Dentro de este contexto, el mejor plan del muestreo sería aquel que asegure que este elemento caiga dentro de cada muestra sin excepción. Se puede conseguir este objetivo mediante dividir la población en dos estratos o dos subpoblaciones. El estrato **I** que consiste solamente del elemento **f**, y el estrato **II** que contiene los demás elementos; es decir, **a**, **b**, **c**, **d**, y **e**.
Tomamos una MSA del tamaño $n = 2$ del estrato **II** y recordar que el estrato **I** solamente contiene el elemento **f**, para de esta manera mantengamos el tamaño de muestra como en el caso de la MSA de tamaño $n = 3$; es decir, dos elemento del estrato **II** y el único elemento del estrato **I**.
Se necesita alguna intuición para ver de qué manera podamos estimar la población total bajo este esquema de ME. Por ejemplo, si multiplicamos el total de la estimación por 2 (como el caso anterior de MSA), esto pondrá mucho peso sobre el elemento **f**, y por ende, siempre terminamos con una sobreestimación del total de la población. Para manejar este problema, tenemos que utilizar el muestreo estratificado con su lógica de ponderación adecuada.

Para el estrato **I**, conocemos el total correcto que es igual a 16, ya que siempre involucramos al elemento **f**. En el caso del estrato **II**, donde 2 de los 5 elementos van a estar seleccionadas, la lógica de la ponderación sería el multiplicar el total de la muestra de este estrato por 5/2 = 2.5. Por tanto, la estimación apropiada del total de la población sería como sigue: 16 + 2.5 (total de la muestra del estrato **II**). Estas estimaciones están dadas para 10 posibles siguientes combinaciones en la Tabla B.

$$_5C_2 = 5! / 2!\,(5-2)! = (5*4*3*2*1) / (2*1)\,(3*2*1*) = 10$$

Tabla B. Resultados de las 10 muestras del tamaño de $n = 3$ del ME.

Composición de la muestra	**∑ de muestra en estrato II [T_2]**	**Estimación del total** $T_E = 16 + (2.5[T_2])$	**Error de estimación T_E – 36**
Abf	3	23.5	-12.5
Acf	5	28.5	-07.5
Adf	7	33.5	-02.5
Aef	8	36.0	0.00
Bcf	6	31.0	5.00
Bdf	8	36.0	0.00
Bef	9	38.5	+2.5
Cdf	10	41.0	+5.0
Cef	11	43.5	+7.5
Def	13	48.5	+12.5
Media del funcionamiento del plan de muestreo		**36**	**0.0**

Se puede observar (Tabal B) que la estimación es *sin sesgo* y su CME = $[(23.5^2 + 28.5^2 + \ldots + 48.5^2) - (23.5+28.5+\ldots+48.5)^2/10] = 48.75$. Por tanto, el $EE = (48.74)^{1/2} \approx 7$. Esta valor (7) constituye el 19% del total (36) de la población y claramente, es un mejoramiento considerable en comparación con el EE =13.2 del caso de MSA.

Por tanto, el muestreo en el plan de ME se realiza con las fracciones desiguales de la muestra, es decir, en el caso del ejemplo arriba, el estrato **I** esta muestreada al 100%, mientras que el estrato **II**, se muestra "2" elementos del total del "5", es decir con una probabilidad igual al 40%.

La estratificación nos permite dividir la población en estratos que son más homogéneos comparado con la población original. En el muestreo estratificado la selección de los elementos se realiza con factor de ponderación distinta para diferentes estratos.

Conclusiones

Cuando los datos de muestreo proceden de una población con una distribución normal y dependiendo en las condiciones específicas se utilizaría los esquemas siguientes del muestreo. En caso que existe una uniformidad en la población, entonces el diseño óptimo del muestreo sería muestreo simple aleatorio (MSA). Sin embargo, en caso de detectar (por medio de un procedimiento estadístico correcto, por ejemplo ANOVA) heterogeneidad o variación entre diferentes secciones de la población, entonces el esquema óptimo del muestreo sería el muestreo estratificado (ME). Para los mismo datos normales, tanto el MSA como ME proveen estimaciones no-sesgadas, sin embargo empleando el ME mejoraría considerablemente la estimación del error estándar y por ende incrementa el nivel de exactitud en comparación con el MSA.

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.
Armitrage, P. 1947. A comparison of stratified with unrestricted random sampling. Biometrika, 34: 273-280.
Birnbaum, Z.W. 1950. Bias due to nonavailability in sampling surveys. J. Am. Stat. Assoc. 45: 98-111.
Brewer, K.W.R. 1963. Ratio estimation in finite populations: Some results deducible from the assumption of an underlying stochastic process. Australian J. Stat. 5: 93-105.
Burstein, H. 1975. Finite population correction for binomial confidence limits. J. Am. Stat. Assoc. 70: 67-69.
Chatterjee, S. 1967. A note on optimum stratification. Skand. Akt., 50: 40-44.
Chatterjee, S. 1968. Multivariate stratified surveys. J. Am. Stat. Assoc., 63: 530-534.
Chatterjee, S. 1972. A study of optimum allocation in multivariate stratified surveys. Skand. Akt., 55: 73-80.
Cochran, W.G. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. Ann. Math. Stat., 17: 167-177.
David, I.P. & B.V. Sukhtme. 1974. On the bias and mean square error of the ratio estimator. J. Am. Stat. Assoc., 69: 464-466.

*** Acerca de los Autores**

Badii M.H. Profesor investigador del área de posgrado, UANL, México, mhbadiiz@gmail.com
Guillen A. Profesor investigador del área de posgrado, UANL, México
Valenzuela J. Profesor investigador del área de posgrado, UAAAN, Saltillo, Coah., México
Cerna E. Profesor investigador del área de posgrado, UAAAN, Saltillo, Coah., México
Abreu José Luis. Profesor investigador del área de posgrado, UANL, México.
UANL, San Nicolás, N.L., México, mhbadiiz@gmail.com, UAAAN, Saltillo, Coah., México