Tamaño Óptimo de Muestra en Ciencias Sociales y Naturales
Optimal Simple Size (OSS) in Social and Natural Sciences

Badii, M.H., A. Guillen & J.L. Abreu
UANL, Cd. Universitaria, San Nicolás de los Garza, N.L. México
mhbadiiz@gmail.com

Resumen: Tamaño óptimo de la muestra es esencial para poder tener estimaciones confiables y precias de los parámetros poblacionales. En el presente trabajo se describen con ejemplos la forma de estimar tamaños óptimos de la muestra en las investigaciones en campos sociales y de la salud.
Palabras clave: Salud, estudios sociales, parámetros poblacionales, tamaño óptimo de la muestra

Abstract: Optimal simple size is essential for determination of reliable and precise population parameters. In this work, estimation of optimal sample sizes in research studies in healthcare and social areas are described with examples.
Keywords: Healthcare, optimal sample size, population parameters, social studies

Introducción

El muestreo es uno de los conceptos fundamentales en la estadística y de hecho en todas las ramas de ciencias. Las muestras son reflejos imperfectas de la población. Una población es un conjunto total de los individuos (observaciones, unidades, etc.) tangible o conceptual con rasgos similares. La muestra es un subconjunto de una población. El muestreo es el proceso de seleccionar la muestra en base a ciertos criterios (Cochran, 1977, Badii & castillo, 2009, Ryan, 2013).
En el muestreo uno debe definir el tamaño óptimo, la frecuencia y el diseño del muestreo. El diseño del muestreo mayormente depende de los objetivos de muestreo y las características de la población. La frecuencia del muestreo se basa a la tasa de reemplazo o cambio en el rasgo o variable a medir. El tamaño óptimo de la muestra el objetivo de este trabajo se explica a continuación.
TOM depende en 1) la cantidad de recursos para realizar el muestreo tales como capital, energía, tiempo, estructura etc., 2) el grado de precisión deseada y 3) el tipo de la distribución de los datos lo cual nos provee información sobre la magnitud o el promedio y la variabilidad de la población. La precisión es el grado de la diferencia entre diferentes estimaciones tomadas de la misma población y se lo mide vía error estándar de la muestra. El sesgo de la muestra es el grado de diferencia entre el parámetro o la variable poblacional y la estimación o la variable muestreal. Los tipos de distribución de los datos poblacionales en término general y resumida son: a) uniforme (done la razón varianza-media o V/M es estadísticamente menor que la unidad, b) agregada con V/M mayor que la unidad, c) aleatorio, lo cual se presenta

en dos formas: c1) Poisson con V/M igual a la unidad y c2) normal en donde la varianza es independiente de la media. En este trabajo se supone que la distribución de los datos se ajuste a la distribución normal.

TOM en estudios transversales

Para una sola muestra (estudio transversal), TOM está basado en objetivos de la investigación:

1. Determinar la media o la proporción de la población.
2. Determinar más de 2 medias o proporciones poblacionales.

A. Solo un caso

Caso 1.

TOM cuando desea conocer la proporción poblacional

Calcular el TOM para el caso en donde se estudian 15 leyes sobre el tema del medio ambiente y solamente 5 de ellos son aprobadas. Cuál sería el TOM para predecir con mayor precisión la aprobación de las leyes similares (Castillo Serna, 2011).

Ecuación:

TOM = Z^2(pq) / e^2
Donde,
TOM = Tamaño óptimo de la muestra
Z = Valor de la tabal de la distribución normal a una $\alpha = 0.05$ es igual a 1.96.
p = Tasa de aprobación
q = 1 – p
e = Error permisible o error de estimación (*precisión relativa*)

Solución:

Z = 1.96 Valor de Z corresponde a 95% de I.C.
p = 5 /15 = .33
q = 1 - .33 = .67
Con un e = 5%
TOM = Z^2(pq) / e^2 = 1.96^2 (.33 * .67) / $.05^2$ = 340
Si se desea una mayor precisión, por ejemplo una e = 1%, entonces el TOM sería:

$TOM = Z^2(pq) / e^2 = 1.96^2 (.33 * .67) / .01^2 = 8{,}496$
Se puede observar que a mayor precisión deseada, sería mayor el TOM.

Caso 2.

TOM cuando desea conocer la media poblacional:

Ecuación:

$\mathbf{TOM = Z^2(V) / e^2}$
Donde,
TOM = Tamaño óptimo de la muestra
Z = Valor de la tabal de la distribución normal
V = Varianza poblacional
e = Precisión relativa o error permisible o error de estimación

Ejemplo:

En un estudio de bebes recién nacidos se tienen la talla promedia igual a 50.8 cm y una desviación estándar de 2.5 cm (Castillo Serna, 2011). Determinar el TOM para estimar la media de la talla con un error permisible no mayor de 0.5 cm.

Solución:

$Z = 1.96$ Valor de Z corresponde a una $\alpha = 0.05$.
$V = 2.5^2$
$e = .5$ cm
$\mathbf{TOM = Z^2(V) / e^2} = 1.96^2 (2.5^2) / 0.5^2 = 96$

B. Múltiples casos

Caso 1.

Estudio sobre la tendencia de estudiar carrera de política de los hijos provenidos de padres políticos y no políticos. Determinar el tamaño óptimo de la muestra (Castillo y Serna, 2011) para estudiar la situación comparativa, es decir la tendencia de seleccionar o no la carrera de estudios políticos de hijos de cada grupo de políticos y no políticos. Se supone que de un estudio anterior se conoce las medias y las deviaciones estándares para el grupo político (grupo 1) y no político (grupo 2) igual a 3.175 ± 0.362 cm y 3.350 ± 0.345 cm, respectivamente.

Ecuación:

$\mathbf{R = |m_1 - m_2| / (V_1 + V_2)^{1/2}}$
Donde,
R = Valor equivalente para estimar TOM en la Tabla 1 (Natrella, 1963).
| = Tomar el valor absoluto
m_1 = media del grupo 1
m_2 = media del grupo 2
V_1 = varianza del grupo 1
V_2 = Varianza del grupo 2

Tabla 1. TOM en base a los valores R, $\alpha = 0.05$ y poder estadístico $(1 - \beta)$.

R	**1 – β**		
	0.90	**0.95**	**0.99**
0.10	1,051	1,300	838
0.15	500	660	950
0.20	263	325	460
0.25	180	220	325
0.30	130	152	222
0.35	91	111	162
0.40	66	82	115
0.45	53	66	96
0.50	43	54	78
0.55	36	44	64
0.60	30	37	52
0.70	22	28	39
0.80	17	21	29
0.90	14	16	23
1.00	11	13	19
1.10	9	12	16
1.20	8	10	13
1.30	7	9	11
1.40	6	7	10

Solución:

$\mathbf{R = |m1 - m2| / (V1 + V2)^{1/2}}$,

$R = |3.175 - 3.350| / (0.362^2 + .345^2)^{1/2}, = 0.35$
Al consultar la Tabla 1 se observa que para el valor de R = 0.35 y una $\alpha = .05$ y una potencia $(1 - \beta)$ igual a 90% o un error de tipo II igual a 10%, se encuentra un TOM igual a 91 sujetos. Por tanto, hay que considerar mínimamente 91 hijos para el estudio.

TOM en estudios longitudinales

Un estudio longitudinal es aquel en donde los datos se obtienen a través de tiempo en 2 o más muestras.

Ecuación:

$$\mathbf{TOM = [Z^2(q_1) / p_1] + [Z^2(q_2)/ p_2)] \;/ [\ln(1- e)]^2}$$

Donde,
TOM = Tamaño óptimo de la muestra
Z = Valor de la tabal de la distribución normal
p_1 = Tasa de prevalencia en la población expuesta a un factor de riesgo
$\mathbf{q_1 = 1 -}\; p_1$
p_2 = Tasa de prevalencia en la población NO expuesta a un factor de riesgo
$\mathbf{q_2 = 1 -}\; p_2$
e = Precisión relativa o error permisible o error de estimación

Ejemplo:

Se desea en una investigación epidemiológico estudiar a lo largo de tiempo, el cáncer en un grupo de fumadores y otro grupo de no fumadores, y por tanto, se desea estimar el TOM en este estudio (Castillo y Serna, 2011).

Datos:

$p_1 = 0.7$
$p_2 = 0.35$
e = 50%, e = 25% e = 15% y e = 10%

Solución:

Para e = 0.5:
$TOM=[Z^2(q_1)/p_1]+[Z^2(q_2/p_2)]/[\ln(1-e)]^2=[1.96^2(.3)/.7]+[1.96^2(.65/.35)]/[\ln(1- .5)]^2 = 18.27 = 19$

Para e = 0.25:
$TOM=[Z^2(q_1)/p_1]+[Z^2(q_2/p_2)]/[\ln(1-e)]^2=[1.96^2(.3)/.7]+[1.96^2(.65/.35)]/[\ln(1.25)]^2 = 106.09 = 107$

Para e = 0.15:
$TOM=[Z^2(q_1)/p_1]+[Z^2(q_2/p_2)]/[\ln(1-e)]^2=[1.96^2(.3)/.7]+[1.96^2(.65/.35)]/[\ln(1.15)]^2 = 332.45 = 333$

Para e = 0.10:
$TOM=[Z^2(q_1)/p_1]+[Z^2(q_2/p_2)]/[\ln(1-e)]^2=[1.96^2(.3)/.7]+[1.96^2(.65/.35)]/[\ln(1-.10)]^2 = 790.99 = 791$

Conclusión: A medida que se incrementa el error de estimación se reduce el TOM y viceversa.

TOM en estudios de casos y controles

Se refiere a situaciones de estudios transversales, longitudinales, observacionales, prospectivos y comparativos.

Ecuación:

$$\mathbf{TOM = Z_\alpha[(1 + (1/K)\ (p_n\, q_n)]^{1/2} + [Z_\beta(p_1\, q_1) + (p_0\, q_0 /K)]^{1/2} / (p_1 - p_0)^2}$$

Donde,
TOM = Tamaño óptimo de la muestra
$\mathbf{Z_\alpha}$ = Valor de la tabal de la distribución normal para 2 colas (1.96 a una $\alpha = .05$)
$\mathbf{Z_\beta}$ = Valor de la tabal de la distribución normal para 1 colas (0.84 a una $\alpha = 0.2$)
K = # de controles pareados para el estudio
$\mathbf{P_0}$ = Tasa de prevalencia de exposición en controles
$\mathbf{q_0} = 1 - P_0$
$\mathbf{P_1} = P_0R / [1 + P_0(R - 1)]$
R = Razón de momias por demostrar, siendo la mínima 2.5
$\mathbf{q_1} = 1 - P_1$
$P_n = (P_1 + KP_0) / (1 + K)$
$\mathbf{q_n} = 1 - P_n$

Ejemplo:
En un estudio de casos y controles de cáncer gástrico para investigar el efecto de bebidas alcohólicas, en base a la literatura (Castillo y Serna, 2011) se sabe que la

prevalencia de exposición en los controles es $P_0 = 0.25$. Un investigador se propone 3 controles para cada caso de cáncer gástrico.

Procedimiento:

Datos:

$Z_\alpha = 1.96$
$Z_\beta = 0.84$
$K = 3$
$P_0 = 0.25$
$q_0 = 0.75$
$P_1 = 0.45$
$R = 2.5$
$q_1 = 0.55$
$P_n = 0.3$
$q_n = 0.7$

$$\textbf{TOM} = Z_\alpha[(1 + (1/K)\ (p_n\, q_n)]^{1/2} + [Z_\beta(p_1\, q_1) + (p_0\, q_0 /K)]^{1/2} / (p_1 - p_0)^2$$

$$TOM = 1.96[(1 + (1/3)\ (.3 * .7)]^{1/2} + [.84(.45 * .55) + (.25 * .75 /3)]^{1/2} / (.45 - .25)^2 = 57$$

Es decir, 57 casos y por tanto, (57) (3) = 171 controles

TOM en estudios de equivalencia

Para estudios de equivalencia en donde el objetivo es el determinar si la cantidad y la tasa de absorción de dos medicamentos (la de prueba o el medicamento candidato para convertirse a un medicamento genérico y la de referencia o el medicamento de patente o marca) son iguales y no existe una diferencia más de una cantidad predeterminada entre la media de absorción del producto de la prueba y la de referencia, el cálculo del tamaño de la muestra es de forma siguiente. Cabe señalar que el tamaño mínimo de la muestra o el mínimo número de sujetos sanos para estos estudios no puede ser menor de 12. Según Hauschke et al., 2007, se utiliza la siguiente ecuación de simulación para calcular el tamaño necesario de la muestra para estudios de bioequivalencia.

$$n_i \geq 2\ [t_{1-\alpha,n-2} + t_{1-\beta,n-2}]^2\ [VIS / (-\ln L_I - \ln T/R)]^2 \qquad :\text{para } T/R \geq 1$$
$$n_i \geq 2\ [t_{1-\alpha,n-2} + t_{1-\beta,n-2}]^2\ [VIS / (-\ln L_I - \{-\ln T/R\})]^2 \qquad :\text{para } T/R < 1$$

Donde,

α: Probabilidad de cometer error tipo I (0.05)

β: Probabilidad de cometer error tipo II (0.20)

VIS: Variabilidad intra-sujeto

T/R: Razón de media del producto de la prueba entre media de la referencia

L_I: Límite inferior de bioequivalencia (0.80, ya que el intervalo de confianza a 90% en estos estudios normalmente fluctúa entre 80 y 125%)

El objetivo final de esta ecuación es seguir un proceso iterativo comenzando con un tamaño de la muestra determinada hasta tener convergencia, es decir obtener dos valores iguales de "*n*" en iteraciones consecutivas. Los valores de "t" para la ecuación arriba se encuentran en http://www.danielsoper.com/statcalc3/calc.aspx?id=10. Vamos a suponer que para las razones regulatorios el mínimo número de los individuos para estos tipos de estudios es de 12 sujetos.

Ejemplos:

Caso A.

Se supone una razón o cociente entre la media del producto de la prueba (T) y la media del producto de la referencia (R) igual a 110%, un coeficiente de la variabilidad intra-sujeto (VIS) de 31.5%. Determinar el tamaño necesario de la muestra para obtener un poder estadístico de 80% para demostrar equivalencia y que el intervalo de confianza al 90% para el cociente de T/R se ubique dentro de los límites de 75 – 133%.

Datos:

VIS = 0.315
T/R = 1.1
L_I = 0.75

Pasos:

Comenzar con n =12

$$n_i \geq 2[t_{1-0.05,12-2} + t_{1-0.2,12-2}]^2\,[0.315 / (-\ln 0.75 - \{\ln .1.1\})]^2$$
$$n_i \geq 2[1.8125 + 0.8791]^2\,[0.315 / (0.2876 - \{0.0953\})]^2$$
$$n_i \geq 2[2.6916]^2\,[0.315 / (0.1923\,)]^2$$
$$n_i \geq 2[2.6916]^2\,[1.6380]^2$$

$n_i \geq 2[7.2447]$ **2.6830**
$n_i \geq 38.875$
$n_i \geq 39$

<u>Usar n = 39</u>
$n_i \geq 2[t_{1-0.05,39-2} + t_{1-0.2,39-2}]^2$ **2.6830**
$n_i \geq 2[1.6871 + .8514]^2$ **2.6830**
$n_i \geq 2[2.5385]^2$ **2.6830**
$n_i \geq 2[6.4439]$ **2.6830**
$n_i \geq 34.5779$
$n_i \geq 35$

<u>Usar n=35</u>
$n_i \geq 2[t_{1-0.05,35-2} + t_{1-0.2,35-2}]^2$ **2.6830**
$n_i \geq 2[1.6924 + .8526]^2$ **2.6830**
$n_i \geq 2[2.5450]^2$ **2.6830**
$n_i \geq 2[6.4770]$ **2.6830**
$n_i \geq 34.7555$
$n_i \geq 35$

Se obtiene la convergencia y por tanto, n = 35.

Caso B.

Se supone una razón o cociente entre la media del producto de la prueba (T) y la media del producto de la referencia (R) igual a 110%, un coeficiente de la variabilidad intra-sujeto (VIS) de 32%. Determinar el tamaño necesario de la muestra para obtener un poder estadístico de 80% para demostrar bioequivalencia y que el intervalo de confianza al 90% para el cociente de T/R se ubique dentro del rango de 75 – 133%.

Datos:

VIS = 0.32
T/R = 1.1
$L_I = 0.75$

Pasos:

<u>Iniciar con n = 12</u>

$n_i \geq 2[t_{1-0.05,12-2} + t_{1-0.2,12-2}]^2\ [0.32 / (-\ln 0.75 - \{\ln.1.1\})]^2$
$n_i \geq 2[1.8125 + 0.8791]^2\ [0.32 / (0.2876 - \{0.0953\})]^2$
$n_i \geq 2[2.6916]^2\ [0.32 / (0.1923\ )]^2$
$\mathbf{n_i \geq} 2[2.6916]^2\ [1.6640]^2$
$n_i \geq 2[7.2447]\ \mathbf{2.7688}$
$n_i \geq 40.1182$
$\mathbf{n_i \geq 40}$

Usar n = 40

$n_i \geq 2[t_{1-0.05,40-2} + t_{1-0.2,40-2}]^2\ \mathbf{2.7688}$
$n_i \geq 2[1.6860 + .8512]^2\ \mathbf{2.7688}$
$n_i \geq 2[2.5372]^2\ \mathbf{2.7688}$
$n_i \geq 2[6.4373]\ \mathbf{2.7688}$
$n_i \geq 35.6471$
$\mathbf{n_i \geq 36}$

Usar n = 36

$n_i \geq 2[t_{1-0.05,36-2} + t_{1-0.2,36-2}]^2\ \mathbf{2.7688}$
$n_i \geq 2[1.6909 + .8523]^2\ \mathbf{2.7688}$
$n_i \geq 2[2.5432]^2\ \mathbf{2.7688}$
$n_i \geq 2[6.4678]\ \mathbf{2.7688}$
$n_i \geq 35.8160$
$\mathbf{n_i \geq 36}$

Se obtiene la convergencia y por tanto, n = 36.

Conclusión

En la vida actual por las restricciones de los recursos disponibles y accesibles (tiempo, energía, estructura, economía, etc.), realizar censo poblacional, es decir, contar todo y cada uno de los elementos de la población es muy costoso y casi imposible de logar, de hecho, casi todas las inferencias se consiguen por medio del muestreo. Muestreo, es decir, la forma de obtener estimaciones confiables de los parámetros poblacionales constituye una base fundamental para las investigaciones científicas cuantitativas. En todas las situaciones en donde se pretende obtener inferencias precisas sobre los parámetros de las poblaciones, el muestreo constituye la piedra angular sobre cual depende la generación del sesgo, la precisión y la exactitud de las conclusiones. Dicho esto, el tamaño óptimo de la muestra a su vez define y determina el grado de confianza en los resultados. Por estas razones, en este trabajo nos avocamos a demostrar la forma de estimar el tamaño óptimo de la muestra para algunos casos selectos en las ciencias sociales y naturales.

Referencias

Badii, M.H. & J. Castillo. 2009. Muestreo Estadistica: Conceptos y Aplicaciones. UANL, Monterry.

Castillo Serna, L. 2011. Manual Práctico de Estadística para las Ciencias de la Salud. Trillas, México, D.F., México.

Cochran, W.G. 1977. Sampling Techniques. Wiley, New York.

Hauschke, D., V Steinijans & I. Pigot, 2007. Bioequivalence Studies in Drug Development, Methods and Applications, John Wiley & Sons Ltd., The Atrium, Southern gate, Chichester, West Sussex PO19 8SQ, England.

http://www.danielsoper.com/statcalc3/calc.aspx?id=10.

Natrella, M.G. 1963. Experimental Statistics. Handbook 91, national Bureau of Standards, Washington, D.C.

Ryan, T.P. 2013. Sample Size Determination and Power. Wiley, Hoboken, New Jersey.