Metodología de Encuestas
Vol 6, Núm 1, 2004, 53-59
ISSN: 1575-7803

VISIÓN ALGEBRAICA UNIFICADA PARA EL CÁLCULO DEL TAMAÑO DE LA MUESTRA

Viso M. Ares
S.I.P.I.E.

RESUMEN

La decisión sobre el tamaño de una muestra está en función de multitud de aspectos, principalmente del modelo de selección de la muestra y de la función poblacional a estimar. La combinación de los factores implicados genera una gran cantidad de expresiones de cálculo específicas. Esta situación lleva consigo dos consecuencias indeseables. Por un lado, dificulta la automatización del proceso en la elaboración de algoritmos de cálculo. Por otro, desorienta a los lectores e investigadores que buscan una expresión idónea. En el presente trabajo se sugiere la inclusión de variables globales en las fórmulas para el cálculo del tamaño de la muestra. Con ello, se generan dos consecuencias positivas. Por un lado, se reduce sensiblemente el número de expresiones de cálculo requeridas. Por otro, se facilita la obtención de una visión global, general o unificada en el problema del cálculo de un tamaño para la muestra

Palabras clave: *tamaño de muestra, modelos de muestreo, objetivos de inferencia.*

Introducción

Entre los procedimientos de extracción de muestras aleatorias, pueden considerarse cuatro modelos básicos: aleatorio simple, estratificado, de conglomerados monoetápico y de conglomerados con submuestreo. A su vez, las estimaciones más comunes se realizan para proporciones y medias, lo que supone la existencia de 4 x 2 = 8 combinaciones comunes. Los objetivos de inferencia implican bien a problemas de una sola variable o estableciendo diferencias (en grupos relacionados o independientes) (8 x 3 = 24). Además, las expresiones se duplican cuando se considera la existencia o no de una hipótesis alternativa en las pruebas de hipótesis (24 x 2 = 48). Por último, las situaciones de aplicación requieren distinguir entre poblaciones finitas e infinitas (48 x 2 = 96). De esta forma, considerando únicamente los contextos más usuales, debería generarse cerca de un centenar de expresiones de cálculo para las estimaciones de las varianzas o para la decisión sobre el tamaño de la muestra.

Esta variedad de situaciones complica enormemente el establecimiento de algoritmos de cálculo automático y, sobretodo, la comunicación de expresiones en manuales y la visión de patrones comunes.

La inclusión de las variables globales A, B y C

Una posible solución es la identificación de entidades comunes a grupos amplios de expresiones de cálculo y su inclusión simbólica en las fórmulas. Con tal objetivo, se sugiere aquí el establecimiento de las siguientes variables globales:

A. Medida de variación en la población (variación global o entre-grupos o dentro-grupos)
B. Medida de probabilidad o riesgos asociados a la estimación y/o decisión (distancia estandarizada de un valor o una comparación, comparación entre distancias estandarizadas
C. Medida de error en la estimación o en la decisión (error de precisión o, en el caso de hipótesis alternativa, tamaño de efecto)

Medidas básicas

Para la construcción de las expresiones, vamos a utilizar los siguientes elementos, todos referidos a la población:

σ^2	varianza de la característica.
σ_i^2	varianza de la característica en el grupo *i*.
σ, σ_i	desviación tipo global o en el grupo *i*, respectivamente.
$\rho_{1,2}$	correlación lineal simple de Pearson entre los grupos 1 y 2.

w_i ponderación del grupo *i*, de tal forma que $\sum w_i = 1$.

Z_λ medida estandarizada de la probabilidad o riesgo de errar en la estimación o en la prueba de decisión, según la distribución muestral teórica que corresponda. Para el riesgo de primera especie, $\lambda = \alpha$; para el de segunda especie, $\lambda = \beta$.

μ_j medida poblacional de la característica (sea media o proporción, en cuyo caso se considera el modelo de codificación dummy) para el supuesto de la hipótesis *j*. En el caso de hipótesis nula, *j* = 0; para la hipótesis alternativa, se recurre a *j* = 1.

e_p error de precisión, es decir, cota máxima para el error de estimación o distancia entre el valor del estimador y el parámetro poblacional. En otros términos, radio del intervalo de estimación.

H_j referencia a la hipótesis nula (*j* = 0) o alternativa (*j* = 1).

γ^2 medida estandarizada del efecto, según Cohen (1988).

Variación: A.

En los problemas unitarios (estimación de un valor poblacional), la medida de variación se corresponde con la varianza de la variable en la población:

$$A = \sigma^2$$

En los problemas relacionales, hay que distinguir si los grupos de comparación están o no relacionados entre sí. En el caso de grupos independientes, la medida de variación surge de ponderar las varianzas de los diferentes grupos:

$$A = \frac{w_2\sigma_1^2 + w_1\sigma_2^2}{w_1 w_2}$$

En la situación particular de igual peso, la expresión anterior pasa a ser

$$A = 2\left(\sigma_1^2 + \sigma_2^2\right)$$

En el caso de grupos relacionados, la medida de variación debe considerar la correlación o covariación, de tal forma que:

$$A = \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{1,2}$$

Si ocurre que la covariación es nula, la expresión anterior pasa a ser:

$$A = \sigma_1^2 + \sigma_2^2$$

Riesgo: B.

La medida de probabilidad o riesgo es representada por las distancias estandarizadas según los modelos de probabilidad asociados al ejercicio de la inferencia. Si utilizamos indistintamente el subíndice α para señalar tanto pruebas de una como de dos colas, la medida de probabilidad para un problema unitario será:

$$B = Z_{\alpha}^{2}$$

Mientras que, en el caso de un problema relacional, la expresión deberá ser:

$$B = \left(Z_{\alpha} + Z_{\beta}\right)^{2}$$

Error: C.

La medida de error depende de si la inferencia adquiere la forma de una estimación o a una decisión según el modelo de la prueba de significación de la hipótesis nula, sin consideración de la potencia de la prueba, por lo que esta medida se circunscribe al error de precisión:

$$C = e_{p}^{2}$$

Si se trata de una decisión y existe consideración de una hipótesis alternativa y el correspondiente riesgo (β), la expresión del error debe considerar si estamos o no ante un problema unitario. En el primer caso:

$$C = \left(\mu_1 - \mu_0\right)^2$$

Siendo preferible utilizar medidas estandarizadas del efecto (Cohen, 1988), la expresión anterior quedaría como:

$$C = \gamma^2 \sigma^2$$

En el caso de una comparación entre grupos, con hipótesis alternativa, la expresión será:

$$C = \left[\left(\mu_1 - \mu_0\right)_{H_1} - \left(\mu_1 - \mu_0\right)_{H_0}\right]^2$$

que, utilizando la medida estandarizada del riesgo, quedará como:

$$C = \gamma^2 \left(w_1 \sigma_1^2 + w_2 \sigma_2^2\right)$$

Visión general de las variables globales

La figura 1 muestra los valores que definen a estas tres variables, en función del conjunto de factores mencionados en los puntos anteriores.

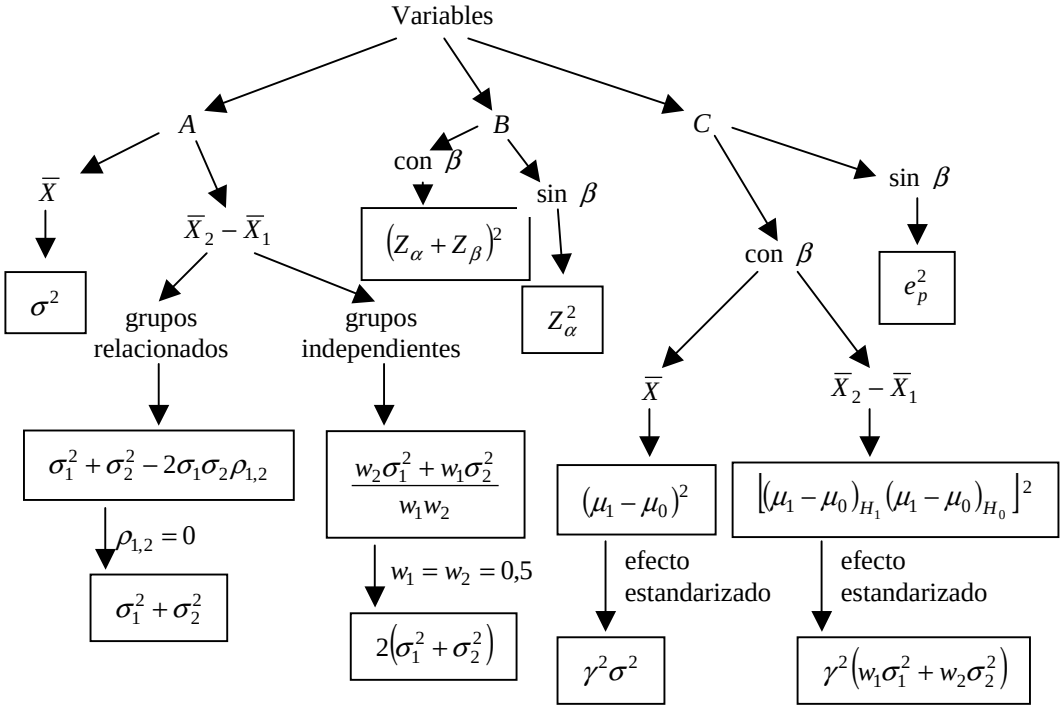

Figura 1: *expresiones para las variables A, B y C.*

Expresiones para el cálculo del tamaño de la muestra

En lo que sigue, se muestran las expresiones de cálculo adecuadas a los diferentes modelos de muestreo. Estas expresiones han sido deducidas de las correspondientes para el error típico que se encuentran en los textos al uso sobre teoría del muestreo, como por ejemplo Sukhatme (1951), Hansen y otros (1953), Kish (1965), Raj (1968), Hájek (1981) o Yates (1981). Las expresiones sobre el tamaño del efecto para los contextos de inferencia aquí mencionados, se encuentran en Cohen (1988).

En la exposición de la expresiones se ha recurrido a los siguientes elementos:

N	tamaño de la población en número de unidades elementales.
N_i	tamaño del estrato *i*.
N_C	tamaño de la población en número de conglomerados.
n	tamaño de la muestra en número de unidades elementales.
n_C	tamaño de la muestra en número de conglomerados.

w_i peso del estrato *i* en la población.

A_i valor de la variable *A* en el estrato *i*.

G tamaño medio de conglomerado en la población (N / N_C)

g tamaño medio de conglomerado en la muestra (n / n_C)

δ correlación u homogeneidad entre conglomerados.

σ_e^2 variación entre conglomerados. $\sigma_e^2 = A\dfrac{\delta(G-1)+1}{G}$. Si $G \to \infty$, $\sigma_e^2 = \sigma^2\delta$

σ_d^2 variación dentro conglomerados. $\sigma_d^2 = A\sigma_e^2$

Modelo aleatorio simple

$$n = \frac{N}{\dfrac{C(N-1)}{AB}+1} \qquad \text{para } N \to \infty \qquad n = \frac{AB}{C}$$

Modelo estratificado

$$n = \frac{\left(\sum N_i\sqrt{A_i}\right)^2}{\dfrac{CN^2}{B}+\sum N_i A_i} \qquad \text{para } N \to \infty \qquad n = \frac{B\left(\sum w_i\sqrt{A_i}\right)}{C}$$

Modelo de conglomerados monoetápico

$$n = \frac{N_C}{\dfrac{C(N_C-1)}{AB}+\dfrac{1}{G}} \qquad \text{para } N_C \to \infty \qquad n = \frac{AB}{C}\left[\delta(G-1)+1\right]$$

Modelo de conglomerados con submuestro

$$n = \frac{G\sigma_d^2 n_C}{\sigma_d^2 + (G-1)\left[n_C\dfrac{C}{B} - \sigma_e^2\dfrac{N_C-n_C}{N_C-1}\right]}$$

para $N_C \to \infty$ $\qquad n = \dfrac{G\sigma_d^2 n_C}{\sigma_d^2 + (G-1)\left[n_C\dfrac{C}{B} - \sigma_e^2\right]}$

para $G \to \infty$ $$n = \frac{\sigma_d^2}{n_C \frac{C}{B} - \sigma_e^2 \frac{N_C - n_C}{N_C - 1}}$$

para G, $N_C \to \infty$ $$n = \frac{\sigma_d^2}{n_C \frac{C}{B} - \sigma_e^2}$$

Referencias

Cohen, J. (1988). *Statistical power for the behavioral sciences.*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Hàjek, J. (1981). *Sampling from a finite population*. Nueva York: Marcel Dekker.

Hansen, M. H.; Hurwitz, W. N. y Madow, W. G. (1953). *Sample survey. Methods and theory. Vol 1: Methods and applications. Serie Wiley Classics Library*. Nueva York: John Wiley & Sons.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Raj, D. (1968). *Sampling theory*. Nueva York: McGraw-Hill.

Sukhatme, P.V. (1953). *Sampling theory with applications*. Iowa: Iowa State College Press.

Yates, F. (1981). *Sampling methods for censures and surveys*. High Wycombe, England: Charles Griffin.