InnOvaciOnes de NegOciOs 5(1): 53 - 65, 2008

Tamaño óptimo de la muestra (Optimum sample size)

Badii, M.H., J. Castillo & A. Guillen
UANL, San Nicolás, N.L, México, mhbadii@yahoo.com.mx

Key words: Bias, estimation, population, sample

Abstract. The basics of sample size estimation process are described. Assuming the normal distribution, the procedures for estimation of sample size for the mean; with and without knowledge of the population variance, and population proportion are noted. Sample size for more than one population feature is also given.

Palabras clave: Estimación, muestra, población, sesgo

Resumen. Se describen los fundamentos del proceso de la estimación del tamaño óptimo de la muestra. Suponiendo una distribución normal para una población, se notan los procedimientos de la estimación del tamaño óptimo de la muestra para la media muestral con y sin el conocimiento de la varianza poblacional. Se presenta el tamaño óptimo de la muestra con más de una característica poblacional.

Introducción

La pregunta de qué tan grande debe ser una muestra surge inmediatamente al inicio del planteamiento de cualquier encuesta o experimento (Badii et al., 2006, Badii & Castillo, 2007, Badii et al., 2007a, b). Esta es una pregunta importante y no se debe tratar a la ligera. Tomar una muestra más grande de lo necesario para obtener los resultados deseados es un desperdicio de recursos, mientras que, por otro lado, las muestras demasiado pequeñas con frecuencia dan resultados que carecen de uso práctico, y podemos fallar en la obtención de los objetivos de nuestro análisis.

Tenemos algo de error de muestreo debido a que no hemos estudiado a la población completa. Siempre que tomamos una muestra, perdemos algo de información útil con respecto a la población. Si queremos tener un alto grado de precisión, tenemos que tomar una muestra suficiente de la población para asegurarnos la obtención de la información requerida. El error de muestreo se

puede controlar si seleccionamos una muestra cuyo tamaño sea el adecuado. En general, cuenta más precisión se quiera, más grande será el tamaño de la muestra necesaria.

En este trabajo se estudia cómo determinar el tamaño de la muestra de acuerdo con la situación de cada experimento. A continuación se proporciona un método para determinar el tamaño de la muestra cuando se desea estimar la proporción de una población. Mediante extensiones directas de estos métodos, es posible determinar el tamaño necesario de las muestras para situaciones más complicadas.

Por lo tanto, el objetivo de la estimación por intervalos es el de obtener intervalos estrechos con alta confiabilidad. Si se observan los componentes de un intervalo, se ve que su dimensión está determinada por la magnitud de la cantidad: (*Coeficiente de confiabilidad*) X (*error estándar*) ya que la magnitud total del intervalo es el doble de esta cantidad. Para un determinado error estándar, el aumento de confiabilidad implica un coeficiente de contabilidad mayor, para un error estándar fijo, produce un intervalo de mayor dimensión. Por otra parte, si se fija el coeficiente de confiabilidad, la única forma de reducir la dimensión del intervalo es la reducción del error estándar. Dado que el error estándar es igual $\frac{\sigma}{\sqrt{n}}$ y σ es una constante, la única forma de obtener un error estándar menor es tomar una muestra grande. ¿Qué tan grande debe ser la muestra? Esto depende del tamaño de que es la desviación estándar de la población, así como del grado de confiabilidad y dimensión del intervalo deseados.

Supóngase que se desea obtener un intervalo que se extiende *d* unidades hacia uno y otro lado de estimador. Ello se enuncia:

***d* = (*Coeficiente de confiabilidad*) X (*error estándar*)** (1)

Si el muestreo va ser con reemplazos, a partir de una población infinita o de una que sea lo suficiente grande como para ignorar la corrección para población finita, la ecuación 1 se transforma en:

$$d = z\frac{\sigma}{\sqrt{n}} \tag{2}$$

la cual, cuando se resuelve para *n*, da.

$$n = \frac{z^2 \sigma^2}{d^2} \tag{3}$$

Cuando el muestreo se hace sin reemplazos a partir de una población finita y pequeña, se requiere de la corrección para población finita y la ecuación 3 queda de la siguiente forma:

$$d = z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{4}$$

que al resolverse para *n*, resulta :

$$n = \frac{N z^2 \sigma^2}{d^2 (N-1) + z^2 \sigma^2} \tag{5}$$

En caso de que se pueda ignorar la corrección para población finita, la ecuación 5 se reduce a la ecuación 3.

Las fórmulas para el tamaño de la muestra requieren del conocimiento de σ^2 pero, como ya se ha señalado, la varianza de la población casi siempre se desconoce. Como resultado, es necesario estimar σ^2. Las fuentes de estimación de σ^2 que se utilizan con más frecuencia son las siguientes. **1.** Se extrae una muestra piloto o preliminar de la población y se utiliza la varianza calculada a partir de esta muestra como una estimación de σ^2. Las observaciones utilizadas en la muestra piloto se toman como parte de la muestra final, de modo que *n* (el tamaño calculado de la muestra) – n_1 (el tamaño de la muestra piloto) = n_2 (el número de observaciones necesarias para satisfacer el requerimiento total del tamaño de la muestra). **2.** A partir de estudios anteriores o similares es posible obtener estimaciones de σ^2. **3.** Si se cree que la población de la cual se extrae la muestra posee una distribución aproximadamente normal, se puede aprovechar el hecho de que la amplitud es aproximadamente igual a seis desviaciones estándar y calcular σ = R/6. Este método requiere algún conocimiento acerca de los valores mínimos y máximo de la variable en la población.

Tamaño óptimo de la muestra

Ejemplo 1. Un nutriólogo del departamento de salud, al efectuar una encuesta entre una población de muchachas adolescentes con el fin de determinar su ingestión diaria promedio de proteínas, buscó el consejo de un experto en bioestadística con respecto al tamaño de la muestra que debe tomar. ¿Qué procedimiento debe seguir el experto de bioestadística para asesorar al nutriólogo? Antes de que el estadístico pueda ayudar al nutriólogo, este debe proporcionar tres elementos de información: la dimensión deseada del intervalo de confianza, el nivel de confianza deseado y la magnitud de la varianza de la población.

Solución. Supóngase que el nutriólogo requiere un intervalo con una dimensión de aproximadamente 10 unidades, es decir, la estimación se debería encontrar alrededor de las 5 unidades del valor real en ambas direcciones. Supóngase que se decide por un coeficiente de confianza de 0.95 y que con base en su experiencia previa percibe que la desviación estándar de la población es probablemente alrededor de 20 gramos. El estadístico tiene ya la información necesaria para calcular el tamaño de la muestra: z = 1.96, σ = 20, y d = 5. Supóngase que el tamaño de la población es grande, así que el estadístico puede ignorar la corrección para población finita y utilizar la ecuación 3. Con las sustituciones adecuadas, el valor de n se calcula como:

$$n = \frac{(1.96)^2 (20)^2}{(5)^2} = 61.47$$

Se recomendó que el nutriólogo tome una muestra de tamaño 62. Al calcular el tamaño de una muestra a partir de las ecuaciones 3 ó 5, el resultado se redondea al siguiente número entero mayor si los cálculos dan un número con decimales.

Tamaño de muestra para estimar una media

Suponga que una Universidad está efectuando una investigación acerca de los ingresos anuales de los estudiantes del último año de una Facultad dada. Se sabe, por la experiencia obtenida, que la desviación estándar de los ingresos anuales de la población completa (1,000 estudiantes) de los egresados es de aproximadamente $1,500. ¿Qué tan grande debe ser la muestra que la universidad debe tomar con el fin de estimar los ingresos medios

anuales de los estudiantes del último año dentro de más y menos $500 y con un nivel de confianza de 95%?

¿Exactamente qué es lo que se pide en este problema? La universidad va a tomar una muestra de un cierto tamaño, determinar la media de la muestra, y utilizarla como estimación puntual de la media de la población. Quiere tener la certeza de 95% de que el ingreso medio anual real no esté más de $500 por encima y por debajo de la estimación puntual. En resumen tenemos:

$z\sigma_{\bar{x}}$ = $500, y *z* = 1.96, podemos deducir el error estándar de la media como

$$1.96\,\sigma_{\bar{x}} = \$500$$

$$\sigma_{\bar{x}} = \$500/1.96 = \$255 = \text{error estándar de la media}$$

Utilizando la ecuación del error estándar, podemos sustituir el valor conocido de la desviación estándar de la población que es de $1,500 y el valor calculado del error estándar de $255 y despejar *n*:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\$255 = \frac{\$1500}{\sqrt{n}}$$

$$\sqrt{n} = \frac{\$1500}{\$255} = 5.882$$

n = 34.6 tamaño de muestra para la precisión especificada

Por tanto, como *n* debe ser mayor o igual a 34.6, la universidad deberá tomar una muestra de 35 estudiantes para obtener la precisión que desea en la estimación del ingreso medio anual de los estudiantes.

Tamaño de muestra para estimación de la media desconocida

La determinación del tamaño de la muestra es muy importante puesto que si tomamos una muestra muy pequeña no será significativa y si la tomamos

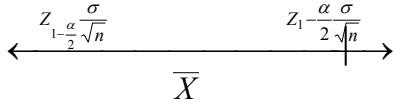
muy grande estamos desperdiciando recursos. Usaremos los intervalos de confianza para calcular tamaño de muestra; si vemos con cuidado el intervalo de confianza para la media.

$$P(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 1 - \alpha \quad (6)$$

y deseamos estrechar el intervalo, tenemos varias opciones siguientes. **1.** disminuir el nivel de confianza: 1-α. **2.** aumentar el tamaño de la muestra, lo que disminuye el error estándar, puesto que σ es fija. De estas dos opciones, la primera no es muy recomendable porque aumentamosα, el riesgo de que μ no esté en el intervalo.

Hay una consecuencia interesante que se desprende de la relación entre el error máximo de estimación (diferencia entre el estimador y el parámetro) y el riesgo (a definido anteriormente) que es la determinación del tamaño de la muestra. Observamos que la longitud o amplitud del intervalo:

$$L = 2Z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (7)$$



Donde, el error máximo de estimación es

$$E = \frac{L}{2} = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad (8)$$

Donde, podemos despejar n si conocemos el error máximo de estimación E; el riesgo a y la varianza poblacional

$$n = \left(\frac{Z_{1-\frac{\alpha}{2}} \sigma}{E} \right)^2 \quad (9)$$

Si el muestreo es sin reemplazo, introducimos el factor de corrección por población finita $\sqrt{\frac{N-n}{N-1}}$ de donde:

$$E = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \tag{10}$$

que al resolver para n, se tiene:

$$n = \frac{NZ^2_{1-\frac{\alpha}{2}}\sigma^2}{E^2(N-1) + Z^2{}_{1-\frac{\alpha}{2}}\sigma^2} \tag{11}$$

Si N es muy grande en comparación con n se puede ignorar el factor de corrección por población finita.

Tamaño de muestra para estimar una porción

Los procedimientos utilizados para determinar los tamaños de muestra para estimar una porción de población son parecidos a los que se utilizan para estimar una media de población. Suponga que deseamos encontrar a estudiantes de una universidad grande. Deseamos determinar qué porción de éstos está a favor de un nuevo sistema de evaluación. Nos gustaría contar con un tamaño de muestra que nos permita tener una certeza de 90% de que estamos estimando la verdadera porción de la población de 40,000 estudiantes que está a favor de nuevo sistema de evaluación, más menos 0.02.

De acuerdo con la tabla *z* del apéndice, el valor de *z* correspondiente a un nivel de confianza de 90%, es de 1.64 errores estándar a partir de media. Queremos que nuestra estación esté dentro de 0.02, de modo que podemos simbolizar el proceso de la siguiente manera

$$z\sigma_{\bar{p}} = 0.02$$

Y *z = 1.64*

Entonces $1.64\,\sigma_{\bar{p}} = 1.64$

Tamaño óptimo de la muestra

Si ahora sustituimos los valores que se tienen para $\sigma_{\bar{p}}$ en la parte derecha de ecuación, obtenemos:

$$1.64\sqrt{\frac{pq}{n}} = 0.02$$

$$\sqrt{\frac{pq}{n}} = 0.0122$$

$$\frac{pq}{n} = 0.00014884$$

Donde

$$n = \frac{pq}{0.00014884}$$

Para hallar *n*, todavía necesitamos una estimación de los parámetros *p* y *q* de la población.

Si tenemos una buena idea de la porción real de estudiantes que están a favor del nuevo sistema, podemos utilizar esto como nuestra mejor estimación para calcular *n*. Pero si no tenemos idea del valor de *p*, entonces nuestra mejor estrategia es determinarlo de manera tal que escogemos *n* conservadoramente. En este punto del problema, *n* es igual al producto de *p* y *q* dividido entre 0.00014884. La manera de obtener *n* más grande es generando el numerador más grande posible de esa expresión, lo cual sucede cuando elegimos *p* = 0.5 y *q* = 0.5. Entonces *n* queda como

$$n = \frac{pq}{0.00014884}$$

$$n = \frac{(0.5)(0.5)}{0.00014884} = 1{,}680 \quad \text{tamaño de muestra}$$

Como repuesta, para estar 90% seguros de que estimamos la porción real dentro de 0.02, debemos escoger una muestra aleatoria simple de 1,680 estudiantes para ser entrevistados.

En el problema que acabamos de resolver, hemos tomado un valor para *p* que representó en la estrategia más conservadora. El valor de 0.5 generó la muestra más grande posible. Pudimos hablar de otro valor de *p* si

hubiéramos sido capaces de estimar uno o si hubiésemos tenido una buena idea de su valor real. Siempre que estas dos últimas soluciones están ausentes, puede tomar el valor más conservador posible de *p,* a saber *p*=0.5.

Para ilustrar que 0.5 produce el valor más grande posible para el tamaño de la muestra, en la Tabla 1 resolvemos el problema de sistema de evaluación utilizando varios valores diferentes de *p*. Del tamaño de las muestras asociado con tales valores, se puede ver que para el intervalo de valores de *p* que va desde 0.3 a 0.7, el cambio en el tamaño de muestra correspondiente es relativamente pequeño. Por tanto, incluso si usted ya sabia que la verdadera porción de población es 0.3 y de todos utilizó 0.5, usted hubiera muestreado solamente 269 personas más (1,680 - 1,411) de lo que era realmente necesario para el grado de precisión deseado. Obviamente, adivinar valores de *p* en casos como éste no parece ser tan crítico como parecía a primera vista.

Tabla 1. Tamaño de muestra *n* asociado con diferentes valores de *p* y *q*.

Valor de *p*	**Valor de *q* = (1-p)**	***pq*/0.00014884**	**Tamaño de muestra *n***
0.2	0.8	(.2)(.8)/.00014884	1,075
0.3	0.7	(.3)(.7)/.00014884	1,411
0.4	0.6	(.4)(.6)/.00014884	1,613
0.5	0.5	(.5)(.5)/.00014884	1,680
0.6	0.4	(.6)(.4)/.00014884	1,613
0.7	0.3	(.7)(.3)/.00014884	1,411
0.8	0.2	(.8)(.2)/.00014884	1,075

Tamaño de muestra con más de una característica

En la mayoría de las encuestas se obtiene información sobre más de una característica. Un método para determinar el tamaño de muestra es especificar los márgenes de error para la característica que se considera más importante para la encuesta. Se hace primero una estimación separada del tamaño de muestra necesaria para cada una de estas características de importancia.

Cuando han sido completadas las estimaciones de características simples de n, es tiempo de hacer una apreciación de la situación. Puede suceder que los tamaños de muestra requeridos sean aproximadamente

iguales. Si la n más grande cae dentro de los límites del presupuesto existente, esta *n* es seleccionada. Más comúnmente, existe una variación suficiente entre los tamaños de muestra de tal manera que nos hace dudar al escoger la más grande, ya sea por consideraciones presupuestales o porque esto daría un estándar global de precisión sustancialmente más alto que el considerado en un principio. En este caso, el estándar de precisión deseado puede ser disminuido para ciertas características, con el fin de permitir el uso de un valor de n más pequeño. En algunos casos los tamaños de muestra *n*, requeridos para las diferentes características son tan distintos que algunos de estos pueden ser eliminados de la encuesta, puesto que con los recursos disponibles la precisión esperada para estas características es totalmente inadecuada. La dificultad puede no ser simplemente la del tamaño de la muestra. Algunas características requieren de un tipo diferente de muestreo en comparación con otras. En poblaciones que son muestreadas en forma repetida, es útil juntar la información relativa a aquellas características que pueden ser combinadas económicamente en una encuesta general y aquellas que necesitan métodos especiales. Como un ejemplo, en la Tabla 2. se presenta una clasificación.

Tabla 2. Un ejemplo de los diferentes tipos de características en encuestas regionales.

Tipo	Descripción de las características	Tipo de muestreo necesario
1	Muy extendido en toda la región ocurriendo con una frecuencia razonable en todas partes.	Una encuesta general con baja proporción de muestreo.
2	Muy extendido en toda la región pero con baja frecuencia.	Una encuesta general pero con una proporción más alta de muestreo.
3	Ocurriendo con frecuencia razonable en la mayoría de las partes de la región, pero con distribución más esporádica, estando ausente en algunas partes y muy concentrada en otras.	Un muestreo estratificado de alta intensidad en las distintas partes de la región. Algunas veces puede ser incluido en una encuesta general con muestreo adicional.
4	Distribución muy esporádica en una pequeña parte de la región.	No apropiada para una encuesta general. Requiere un muestreo acorde con su distribución.

De características en 4 tipos, sugerida por la experiencia obtenida en encuestas agrícolas regionales. Con esta clasificación, una encuesta general quiere decir que las unidades están distribuidas con bastante regularidad sobre alguna región, como por ejemplo en una encuesta simple aleatoria.

Tamaño de muestra para estimar una porción: intervalo de confianza conocido

El método para estimar el tamaño de la muestra cuando se requiere estimar la proporción de una población es esencialmente el mismo que se describió para estimar la media de una población. Se aprovecha el hecho de que la mitad del intervalo deseado d, se puede igualar al producto del coeficiente de confiabilidad y el error estándar.

Si se supone que el muestreo ha sido hecho de manera aleatoria y que existen condiciones que garanticen que la distribución de *p* sea aproximadamente normal, se obtiene la siguiente fórmula para n cuando el muestreo es con reemplazo, cuando se realiza a partir de una población infinita o cuando la población muestreada es lo suficientemente grande como para hacer innecesario el uso de la corrección para población finita.

$$n = \frac{z^2 pq}{d^2} \tag{12}$$

Si la corrección para la población infinita no puede pasarse por alto, la fórmula para *n* es.

$$n = \frac{Nz^2 pq}{d^2(N-1) + z^2 pq} \tag{13}$$

Cuando *N* es grande en comparación con *n* (es decir, $n/N \leq 0.5$) se puede pasar por alto la corrección para población finita y la ecuación 4 se reduce a la ecuación 2.

Como puede observarse, ambas fórmulas requieren que se conozca p, que es la proporción de población que posee la característica de interés. Obviamente, dado que éste es el parámetro que se desea estimar, será desconocido. Una solución para este problema consiste en tomar una muestra piloto y calcular una estimación para utilizarla en lugar de p dentro de la formula para n. Algunas veces el investigador tendrá noción de algún límite superior para *p* que podrá utilizar la fórmula. Por ejemplo, si se desea estimar la proporción de alguna población que presente una cierta condición, es posible

que se crea que la proporción real no puede ser mayor que, digamos, 0.30. Se sustituye entonces *p* por 0.30 en la fórmula para *n*. Si es imposible obtener una mejor estimación, se puede igualar *p* a 0.5 y resolver para *n*. Dado que *p* = 0.5 en la fórmula proporciona el máximo valor de n, este procedimiento dará una muestra lo suficientemente grande para alcanzar la confiabilidad y la dimensión del intervalo deseado. Sin embargo, puede ser más grande de lo necesario y resultará más costosa que si se dispusiera de una mejor estimación de *p*. Este procedimiento se debe utilizar únicamente si no se dispone de una mejor estimación de *p*.

Ejemplo 2. Se plantea realizar una encuesta para determinar que proporción de familias en cierta área carece de servicios médicos. Se cree que la proporción no puede ser mayor que 0.35. Se desea un intervalo de confianza del 95 por ciento de d = 0.05. ¿De qué tamaño se debe seleccionar la muestra de familia?

Solución: Si es posible ignorar la corrección para población finita, se tiene que:

$$n = \frac{(1.96)^2 (0.35)(0.65)}{(0.05)^2} = 349.6$$

Por lo tanto, el tamaño de la muestra es de 350.

Conclusión

Partiendo de la realidad de la escasez del los recursos (financiero, energético, temporal, material, etc.) para la investigación, se recalca la relevancia de la estimación del tamaño óptimo de la muestra. La base de la ciencia experimental es muestreo con base y rigor científico. En la obtención de cualquier tipo de la información, la colección de los datos constituye el primer paso. La subestimación o los tamaños pequeños de la muestra por debajo del tamaño óptimo, ocasiona un alto nivel del sesgo, es decir, el incremento de la distancia entre el valor esperado de la muestra y el parámetro poblacional. Por otro lado, la sobreestimación (tamaños de la muestra por encima del tamaño óptimo) no produce sesgo, más sin embargo, provoca la pérdida de los recursos que tampoco es permisible. Por tanto, el cálculo y la utilización del tamaño óptimo de la muestra es fundamentalmente crucial para tener una idea correcta

y representativa de la población bajo del estudio y que a su vez optimiza la distribución y utilización de los recursos escasos.

Referencia

Badii, M. H., A. E. Flores, R. Foroughbakhch & H. Quiróz. 2000. Fundamentos de muestreo. Pp. 129-144. En: M. H. Badii, A. E. Flores y L. J. Galán (eds.). Fundamentos y Perspectivas de Control Biológico. UANL, Monterrey.

Badii, M.H., J. Castillo & A. Wong. 2006. Diseños de distribución libre. InnOvaciOnes de NegOciOs, 3(1): 141-174.

Badii, M.H. & J. Castillo (eds.). 2007. Técnicas Cuantitativas en la Investigación. UANL, Monterrey.

Badii, M.H., R. Ramírez & J. Castillo. 2007a. Papel de estadística en la investigación científica. InnOvaciOnes de NegOciOs, 4(1): 81-114.

Badii, M.H., J. Castillo, R. Rositas & G. Ponce. 2007b. Experimental designs. Pp. 335-348. In: M.H. Badii & J. Castillo (eds.). Técnicas Cuantitativas en la Investigación. UANL, Monterrey.

Casagrande J.T., M.C. Pike & P.G. Smith. 1978. An improved approximate formula for calculating sample sizes for comparing binomial distributions. Biometrics 34:483-486.

Connett, J.E., J.A. Smith, & R.B.McHuch, 1987. Sample size and power for pair-matched case-control studies. Statist.Med. 6:53-59.

Desu, M.M. & D. Raghavasrao. 1990. Simple size Methodology. Academic press, Bostom Massachusetts, 135 pp.

Fless, J.L., A. Tytun & H.K. Ury. 1980. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 36:343-346.

Roscoe, J.T., & J.A. Byars. 1971. Sample size restraints commonly imposed on the use of the chi-square statistic. J. Amer. Statist.Assoc. 66: 755-759