Nociones Introductorias de Muestreo Estadístico

(Introductory Notions of Statistical Sampling)

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

UANL, San Nicolás, N.L., México & UAAAN, Saltillo Cuah. Mexico

Abstract. Fundamentals of statistical sampling are described. True notions of bias, accuracy as well as the essential fearutres of a good estimador are noted. Probabilistic and non probabilistic sampling is defined. Relevance of bionomics, confidence degree, dispersion and the statistical model of choosing for sampling as prerequisites of pilot sampling are emphasized. Optimal sample sizes for population with distinct features are defined and estimated.

Keywords. Dispersion, estimation, probabilistic sampling

Resumen. Se describen e explican las nociones básicas del muestreo estadístico. Se definen el significado del sesgo, exactitud, precisión, y rasgos esenciales de un buen estimador generado por el muestreo. Se notan el muestreo probabilístico y no-probabilístico con sus rasgos asociados. Se enfatizan la relevancia de la bionomía, el grado de confiabilidad, la distribución espacial y el modelo estadístico apropiado como requisitos dentro del muestreo piloto. Se define y se estima el tamaño óptimo de la muestra para poblaciones con características diferentes.

Palabras claves. Dispersión, estimación, muestreo probabilístico

Introducción

Para comprender el muestreo, es necesario familiarizarse con los conceptos siguientes (Badii & McMurtry, 1990, Badii & Castillo, 2007, Badii et al., 1994a, 1994b, 1995a, 1995b, 1996a, 1996b, 1998, 2000, 2006, 2007a, 2007b, Southwood, 1966, 1978, Kogan & Herzog, 1980, Taylor, 1984). El muestreo es una manera de obtener información deseada de una población basada en ciertos criterios. El investigador realiza muestreo porque no se puede medir la totalidad de la población (censo), debido a que los recursos financieros, laborales, temporales, estructurales, etc., son limitados. Hay dos propósitos para el muestreo.

1. **Objetivos generales:** Determinar los parámetros poblacionales tales como densidad, porcentaje de mortalidad, tasa de reproducción, etc. Por tanto, el objetivo general nos permite conocer la población mediante sus parámetros.

2. **Objetivos específicos:** Estimar la dinámica poblacional, detectar y cuantificar los factores claves, es decir, aquellos factores que ocasionan los cambios poblacionales, y medir la diferencia entre la intensidad de estos factores.

Los investigadores deben armonizar sus criterios acerca de las siguientes relacionadas con el muestreo.

Variable. A la característica mesurable de una unidad experimental se le denomina la *variable*.
Parámetro. El *parámetro* es la variable innata de la población.
Estimación o estadística. A la variable muestral se da el nombre de la *estimación* o la *estadística* de la muestra.
Sesgo. La diferencia entre el parámetro y la estimación tiene el nombre del *sesgo*.
Precisión. La *precisión* es la medida de la variabilidad entre diferentes estimaciones. La precisión se mide por alguna medida de error, como por ejemplo, la varianza, el error estándar de la media, coeficiente de variación o la variación relativa de varias muestras repetitivas.
Precisión relativa Neta (PRN). Cuando la selección de dos o más esquemas de muestreo depende del costo (tiempo, dinero, esfuerzo, etc.), la precisión a parte de la variabilidad, también involucra el costo asociado. En relación a la precisión, se denomina un programa de muestreo más eficiente (confiable) si produce resultados menos variables o más confiables por unidad de costo. La precisión relativa neta (PRN) se calcula mediante: PRN = $100/(VR*C_M)$, donde, VR = variación relativa = Error estándar de la media como una fracción de la media = EE_m/m; y C_M = costo relativo de muestreo.
Exactitud. La *exactitud* es la medida de la distribución de los datos muestreales con respecto al parámetro poblacional. La exactitud se mide vía cuadrado media del término de error o residual.
Estimador. Una expresión matemática que al sustituir los datos de la muestra genere la estimación muestreal.
Rasgo de un buen estimador. Un *estimador* bueno es aquel que es preciso, exacto, sin sesgo, con una distribución conocida y finalmente robusto a la violación de los supuestos de los modelos.
Eficiencia. Una estimación *eficiente* es aquella que produce resultados con mínima varianza por unidad de costo, o con máxima media por unidad de costo.
Consistencia. Se le llama *consistente* a una estimación, si la proporción de los valores esperados de la muestra dentro de una cantidad pequeña fija del parámetro se acerca a 100%, a medida que el tamaño de la muestra se incrementa.
Suficiencia. Se le denomina a una estimación *suficiente*, si se captura, independientemente del tamaño de la muestra, toda la información de la población que está contenida en las observaciones muestreales.
Intervalo de confianza (I.C.). Un intervalo que abarca la estimación (estadística) de la muestra y especifica que el parámetro poblacional se ubica dentro de este intervalo con una probabilidad especifica.

Antes del inicio de un programa formal del muestreo el investigador debe realizar un premuestreo (muestreo piloto o muestreo preliminar) para poder contar con las siguientes informaciones necesarias (Southwood, 1978):

1. **Bionomía**; es decir, el ciclo (historia) de vida, fenología y toda la información que refleje la información acerca de la población.

2. **Grado de confiabilidad** para determinar el tamaño de la muestra de tal manera que se optimice la relación entre el costo de los recursos para el muestreo y la información que se genera por el muestreo

3. **Dispersión espacial** de los elementos, es la forma en que los elementos se arreglan en el espacio basado de la interacción evolutiva de los factores internos (biología y etología) y externos (dispersión de los recursos y la heterogeneidad ambiental) para el uso óptimo de los recursos.

4. **Modelo de análisis**; esto es, determinar el tipo de información o datos que se van a colectar y medir el ajuste de los supuestos del modelo.

Definición estadística de muestreo

En términos estadísticos, se deben conocer los siguientes factores (Cochran, 1977). La *población* que es una colección total de observaciones en una determinada escala espacio-temporal, de las cuales se desea hacer inferencia; el *elemento* es una observación sobre la cual se hace el muestreo; la *Unidad Muestreal* (UM), se trata de un conjunto de elementos de una población y es en realidad la composición física o la magnitud de la muestra; el *cuadro* representa la lista total de las UM's; y la *muestra* que es una selección de UM's de un cuadro.

Debido a la importancia de la UM (Unidad Muestreal) es relevante considerar los siguientes criterios propuestos por Morris (1955). **A.** Todas las unidades muestreales deben tener la misma probabilidad de ser seleccionadas. **b.** La unidad muestreal debe ser estable, de lo contrario, es necesario medir el grado de cambio de forma sencilla y continua. **c.** Es importante poder convertir las unidades muestreales en unidades absolutas. **d.** El tamaño de la UM debe ser adecuado para que permita un balance razonable entre el costo y el beneficio de la muestra. **e.** La estructura de las UM's debe ser de manera más sencilla.

La colecta de cada individuo requiere gasto de tiempo, energía, etc., en otras palabras el muestreo cuesta. Tomando en cuenta la noción del costo del muestreo, la pregunta principal sería ¿qué cantidad de información desea obtener el investigador? Es obvio que el grado de información adquirida depende de qué tan grande va a ser el tamaño de la muestra, ya que a mayor tamaño de la muestra, mayor será la cantidad de información obtenida. Sin embargo, el exceso en el tamaño de la muestra indica el desperdicio del recurso, en otras palabras debe existir un tamaño óptimo de la muestra que se debe cuantificar para obtener información con un óptimo nivel de precisión y un adecuado uso de los recursos. Además del tamaño óptimo de la muestra, también el tipo de diseño muestreal controla la cantidad de la información que se puede adquirir.

Muestreo probabilístico y no-probabilístico

Hay que distinguir entre el muestreo probabilístico y muestreo no-probabilístico.
En el muestreo probabilístico se diseña un plan de muestreo en donde:

1. Cada elemento tiene una probabilidad conocida de estar incluida en la muestra.

2. La selección de los datos de la muestra están basada en un proceso aleatorio consistente con estas probabilidades.

3. Las estimaciones también están basadas en estas probabilidades.

La probabilidad de la selección no necesariamente debe ser igual para todos los elementos de la población y solo basta conocer esta probabilidad. El muestreo probabilístico provee formulas para estimar el error estándar, el sesgo, la precisión, la exactitud y los límites del intervalo de confianza.

El muestreo probabilístico no es la única manera de seleccionar datos de la población. La alternativa denominada el muestreo no-probabilístico consiste en conseguir información sobre la población por medio de preguntar de las personas que conocen la población y puedan apuntalar a los elementos típicos y por ende, limitar la muestra a estos elementos o los elementos que son accesibles.

En realidad, cuando la población es muy variable; tiene alto nivel de varianza, y la muestra por cualquier razón, debe ser de tamaño pequeño, el muestreo no-probabilístico es frecuentemente la forma más apropiada de seleccionar datos de la población. Sin embargo, en el muestreo no-probabilístico no se puede estimar el sesgo, la precisión, etc., es decir, las ecuaciones para estimar el error estándar de la media, la proporción total e intervalos de confianza no se aplican para el muestro no-probabilístico. Por tanto, uno debe utilizar el muestreo probabilístico al menos que: **a.** el muestreo no sea factible, y **b.** el muestreo sea demasiado costoso.

Tamaño óptimo de la muestra

El tamaño óptimo de la muestra, es decir, aquel tamaño que permite un balance adecuado entre el costo del muestreo y la precisión obtenida, y además evita la sobreestimación (sobre-gasto de recursos) o subestimación (precisión no adecuada) depende de tres factores (Kogan & Herzog, 1980).

1. La cantidad de recursos disponible; es obvio que sin recurso simplemente no se puede hacer nada.
2. El grado de confiabilidad; se le puede definir de dos formas:

   a. En términos de error estándar (*EE*) como una fracción de la media (*m*) y se denomina *D*, es decir, $D = EE/m$. Para fines de investigación se selecciona la D igual a 10%, y para la aplicación hasta 25% (Southwood, 1978).

   b. En términos probabilísticos: es necesario escoger un límite sobre el error de estimación y denominarlo "*L*"; es decir, la diferencia entre el parámetro poblacional (μ) y la estimación del muestreo (*m*) debe ser menor que este límite de error. Por tanto, error de estimación = |μ-m|<L, donde "|" significa tomar el valor absoluto. Hay que designar una probabilidad (1-a) que determine la proporción del tiempo que el error de estimación es menor que el límite *L*, es decir, *p* [(error de estimación)<*L*] = (1-*a*).

3. El tipo de dispersión espacial o la forma que los elementos, las observaciones, las mediciones o los individuos se colocan en el espacio, es decir, se agrupan o se distancian el uno del otro. El tipo de dispersión espacial es resultado de dos factores:

a. Factores intrínsecos, como biología y el comportamiento de los elementos.
b. Factores extrínsecos como la distribución de los recursos y la heterogeneidad del medio ambiente, estos dos factores (internos e externos) interactúan entre sí y el resultado es una adaptación evolutiva para optimizar el uso de los recursos vitales como alimento, espacio o refugio, pareja etc. Cabe aclarar que la técnica de muestreo por el hombre, los herbívoros o los depredadores también afecta la estimación del patrón de dispersión espacial.

Debido a que el tamaño óptimo de la muestra apoya de manera fundamental a la representatividad de la población, en continuación, presentamos una descripción detallada del concepto del tamaño óptimo de la muestra.

Fundamentos

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, 2007b). Esta es una pregunta seria e 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 la precisión, la exactitud para los usos prácticos, y consecuentemente, se falla en la obtención de los objetivos del análisis.

Hay que recalcar que el muestreo se debe a la lógica inductiva, es decir, ir de lo particular a lo general para poder predecir un sistema total en base a información específica. Por consiguiente, el muestreo jamás es libre de error, en otras palabras, siempre tenemos algo de error de muestreo debido a que no hemos estudiado 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 suficientemente grande 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, cuando más precisión se quiera, más grande será el tamaño de la muestra necesaria.

En esta sección 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.

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***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 confiabilidad mayor, para un error estándar fijo, y esto 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 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:

$$d = (\textit{Coeficiente de confiabilidad}) * (\textit{error estándar})$$

Si el muestreo se realiza 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 anterior se transforma en:

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

La cual, cuando se resuelve para *n*, da la siguiente ecuación:

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

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 arriba queda de la siguiente forma:

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

que al resolverse para *n*, resulta :

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

En caso de que se pueda ignorar la corrección para población finita, la ecuación anterior se reduce a la ecuación $n = \frac{z^2\sigma^2}{d^2}$.

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 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.

2. 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, donde, R = rango. Este método requiere algún conocimiento acerca del rango, es decir, los valores mínimos y máximo de la variable en la población.

Ejemplo

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 estadística con respecto al tamaño de la muestra que debe tomar.

¿Qué procedimiento debe seguir el experto de estadística 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 cada dirección. Supóngase que se decide por un coeficiente de confianza de 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 a base de la ecuación $n = \frac{z^2\sigma^2}{d^2}$ y con los datos siguientes: $z = 1.96$, $\sigma = 20$, 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 arriba mencionada. 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 para tal efecto, el resultado se redondea al siguiente número entero mayor si los cálculos dan un número con decimales.

Tamaño de la 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 determinada. 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 = 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*:

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

225 = 1500/√*n*, por tanto, *n* = 34.6

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.

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.

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

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 aumentamos el valore de α, 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 (definido anteriormente) que es la determinación del tamaño de la muestra. Observamos que la longitud o amplitud del intervalo:

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

$$Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \qquad\qquad Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$\overline{X}$$

Donde, el error máximo de estimación es

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

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

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

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

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

que al resolver para n, se tiene:

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

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

Tamaño de la muestra para estimar una proporción

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 analizar a estudiantes de una universidad grande. Queremos 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*, 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 estimación esté dentro de 0.02, de modo que podemos simbolizar el proceso de la siguiente manera

$$z\sigma_{\bar{p}} = 0.02$$
$$z = 1.64$$
$$\text{Entonces } 1.64\,\sigma_{\bar{p}} = 0.02$$

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

$$1.64\sqrt{\frac{pq}{n}} = 0.02$$
$$\sqrt{\frac{pq}{n}} = 0.0122$$
$$d^2/Z^2 = 0.02^2/1.96^2 = 0.00014884$$

$$pq/n = d^2/Z^2$$
$$n = z^2\,pq\,/\,d^2$$
$$n = \frac{pq}{0.00014884}$$

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:

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

$$n = \frac{(0.5)(0.5)}{0.00014884} = 1{,}680$$

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 modos utilizó 0.5, usted hubiera muestreado solamente 269 personas más (1,680 - 1,411 = 269) 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*.

p	***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 óptimo de la muestra para múltiples rasgos

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.

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

Tipo	**Descripción de los rasgos**	**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, ausente en algunas partes y muy concentrada en otras.	Un muestreo estratificado de alta intensidad en las distintas partes de la región. A veces puede ser incluido en una encuesta general con muestreo adicional.
4	Distribución muy esporádica en una pequeña parte de la región.	No apropiada una encuesta general. Requiere un muestreo acorde con su distribución.

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.

Hay 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.

Muestreo de una proporción

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 tomado 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.

$$n = z^2 pq / d^2$$

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

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

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 $d = z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ se reduce a la ecuación $d = z\frac{\sigma}{\sqrt{n}}$ o $n = z^2 \sigma^2 / d^2$.

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 fórmula para *n*. Algunas veces el investigador tendrá noción de algún límite superior para *p* que podrá utilizar en 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, 0.3. Se sustituye entonces p por 0.3 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 $p = 0.5$, 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 costoso 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

Se plantea realizar una encuesta para determinar qué 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 y un margen de error $d = 0.05$. ¿De qué tamaño se debe seleccionar la muestra de las familias?

Solución

Si es posible ignorar la corrección para población finita, se utiliza la ecuación siguiente: $n = z^2\ pq\ /\ d^2$, y como resultado el tamaño óptimo de la muestra será: $n = (1.96)^2(0.35)(0.65)/(0.05)^2 = 349.6$ o sea 350 familias.

De manera resumida para los datos con la distribución normal, se utilizan las siguientes ecuaciones para estimar el tamaño óptimo de la muestra (Tabla 3).

Tabla 3. Ecuaciones para estimar el tamaño óptimo de la muestra (n_{opt}).

Tipo de datos	**Población infinita**	**Población finita**
Datos continuos	$n_{opt} = Z^2V/E^2$	$n_{opt} = NV/[(\{N\text{-}1\}E^2/Z^2) + V]$
Proporciones	$n_{opt} = Z^2pq/E^2$	$n_{opt} = Npq/[(\{N\text{-}1\}E^2/Z^2) + pq]$

n_{opt} = Tamaño óptimo de la muestra
V = Varianza de la muestra
Z = 1.96 para I.C. igual a 95%
$E = d$ = Margen del error
P = Proporción de ocurrencia del evento
$q = 1 - p$ = Proporción de no ocurrencia del evento

Conclusiones

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 el muestreo lo cual se toma de un universo con fundamento 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.

Referencias

Badii, M.H. & J.A. McMurtry. 1990. Field experiments on predation, dispersión, regulation and population changes. Publ. Biol. 4(1-2): 43-48.

Badii, M.H., A.E. Flores, S. Flores & S. Varela. 1994a. Statistical description of population distribution and fluctuation of citrus rust mite (Acari: Eriophyidae) on orange fruit in Nuevo Leon, Mexico. Biotam, 6(1): 1-8.

Badii, M.H., A.E. Flores, S. Flores & S. Varela. 1994b. Comparative estimation of distribution statistics of citrus rust mite (Acari: Eriophyidae) on leaves of three different orange orchards in Nuevo León, Mexico. Biotam, 6(1): 9-16.

Badii, M.H., A.E. Flores, R. Foroughbakhch & H. Quiroz. 1995a. Análisis conceptual de muestreo. Pp. 123-136. En: W. Rosa (ed.). VI Curso Nacional de Control Biológico. SMCB, Tapachula , Chiapas. México.

Badii, M.H., A.E. Flores, R. Torres & H. Quiroz. 1995b. Muestreo y evaluación económica de las plagas. Pp. 1-13. En: H. Fuente (ed.). Curso Internacional sobre Manejo de Huertas de Cítricos. SAGAR, INIFAP, CIRNE., General Allende, N. L., México.

Badii, M.H., A.E. Flores, S. Varela, S. Flores & R. Foroughbakhch. 1996a. Dispersion indices of citrus rust mite (Acari: Eriophyidae) on orange in Tamaulipas, Mexico. Pp. 17-20. En: R. Mitchel, D. Horn, G. R. Needham y W. C. Welbourn (eds.). Acarology IX: Volume 1, Proceedings. Section I: Behavior and Physiological Ecology. Ohio Biological Survey, Columbus, Ohio.

Badii, M.H., A.E. Flore, R. Foroughbakhch, H. Quiróz & R. Torres.1996b. Ecología de manejo integrado de plagas (mip) con observaciones sobre control microbiano de insectos. Pp. 21-49. En: L.J. Galan Wong, C. Rodreigguez-Padilla y H. Luna-Olvera (eds.). Avances Recientes en la Biotecnología en *Bacillus Thuringiensis*. Ciencia, Universitaría no. 2. UANL.

Badii, M.H., A.E. Flores, S. Flores & R. Foroughbakhch. 1998. Population dynamics of citrus mites in northeasterm Mexico. Pp. 275-280. En: G. Needham, R. Mitchell, D. Horn and W. C. Welbourn (eds.). Acaroloy IX: Vol. 2, Symposia. Ohio Biological Survey, Columbus, Ohio.

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., J. Castillo, J. Landeros & K. Cortez. 2007a. Papel de la estadística en la investigación científica. InnOvaciOnes de NegOciOs. 4(1): 107-145.
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.
Badii, M.H. & J. Castillo. 2009. Muestreo Estadístico: Conceptos y Aplicaciones. UANL, 250 pp.
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. Adol. Statist. Assoc. 66: 755-759.

***Acerca de los Autores**

El Dr. Mohammad Badii es Profesor e Investigador de la Universidad Autónoma de Nuevo León. San Nicolás, N.L., México, 66450. mhbadiiz@gmail.com
La Dra. Amalia Guillén es Profesora-Investigadora, Egresada de FACPYA-UANL. Monterrey, NL.
El Dr. E. Cerna es Profesor e Investigador en UAAAN, Saltillo, Coah.
El Dr. J. Valenzuela es Profesor e Investigador en UAAAN, Saltillo, Coah.