UN MÉTODO PARA LA DETERMINACIÓN DEL TAMAÑO DE MUESTRA EN ENCUESTAS SOBRE POBLACIONES FINITAS

TOMÁS GARZA H. y JOSÉ A. CORONEL
El Colegio de México

INTRODUCCIÓN

UNO DE LOS PROBLEMAS PRINCIPALES en el diseño de muestras para investigación en poblaciones finitas es la determinación del tamaño de muestra —es decir, el número de unidades que debe seleccionarse— para que ésta alcance un nivel de "representatividad" deseado. El concepto mismo de "representatividad" tiene acepciones diferentes, aunque para el estadístico profesional no existe realmente problema en este sentido cuando el propósito de la investigación por muestreo se refiere a una sola característica, cualitativa o cuantitativa, de los elementos de la población: en este caso, la teoría de las distribuciones en el muestreo arroja suficiente luz sobre la situación como para no dejar dudas al respecto. Así, por ejemplo, si se tratara de estudiar, digamos, la estructura de edades en una población dada, el diseño de la muestra tendría que hacerse a partir de una consideración de la variabilidad de la característica "edad" en los elementos de la población, y el tamaño de muestra se determinaría, en esencia, en términos de la variabilidad esperada de dicha característica dentro de la muestra.

Sin embargo, una investigación por muestreo se realiza generalmente para estudiar de manera simultánea varias características en los elementos de la población, y entonces la idea de representatividad de la muestra es más difícil de definir, ya que el comportamiento particular (o "marginal", para usar la terminología convencional) de cada característica suele ser diferente al de las demás, de manera que tendría que atenderse al mismo tiempo a todas ellas para fijar el tamaño adecuado de la muestra. En una encuesta en que se pretendiera estudiar a la vez, por ejemplo, la edad, el ingreso personal y el tamaño de la familia, se obtendrían diferentes tamaños de muestra según que ésta se diseñara atendiendo a la distribución, en la población, de las edades, los ingresos o el número de miembros en la familia, y evidentemente la solución no es tomar el tamaño de muestra más desfavorable, según puede concluirse de una breve reflexión sobre el concepto de distribución conjunta de las características consideradas.

Desde un punto de vista general, el problema es muy complicado, y su solución requeriría de una teoría estadística de distribuciones en varias dimensiones, que desgraciadamente no existe aún más que para

casos particulares de escaso interés en las ciencias sociales. En esta nota enfocaremos el problema en forma simplificada y propondremos una solución de índole práctica que ha sido utilizada en diversos estudios hechos en El Colegio de México.

I. Descripción del problema

Supondremos que se trata de obtener, mediante una muestra extraída al azar (esto se refiere a la naturaleza probabilística de la selección, y no al diseño mismo, que puede ser estratificado, polietápico, etc.), información respecto de k características definidas en cada uno de los elementos de una población de tamaño T. Con cada uno de dichos elementos está asociado, entonces, un vector $x^r = (x_1^r, x_2^r, \ldots, x_k^r)$, donde la componente j-ésima, es decir, x_j^r, es el valor (o atributo) que toma la característica j-ésima en el elemento r-ésimo de la población.

Ordinariamente está uno interesado en hacer estimaciones, a partir de la muestra, sobre los valores totales $\sum_r x_j^r$, o los promedios $\sum_r x_j^r/T$, de la j-ésima característica, si ésta es de naturaleza numérica, o bien sobre las distribuciones (ya sean en proporciones o en totales) en cada una de las categorías posibles, cuando se trata de un carácter cualitativo. En estudios de índole económica o sociológica es frecuente, además, intentar alguna estimación sobre las relaciones que guardan entre sí las diversas variables definidas en la población; así, por ejemplo, podría interesar el comportamiento de la variable "ingreso" en relación con el de la variable "edad", o alguna otra combinación análoga entre las diversas variables.

Limitaremos nuestra atención al caso en que el conjunto de valores que puede tomar cada característica numérica, así como el número de categorías de cada carácter cualitativo, son finitos, y designaremos con m_j el número de valores o categorías, según el caso, correspondiente a la característica j-ésima. De acuerdo con lo anterior, puede considerarse que existe una clasificación en $M = m_1 m_2 \ldots m_k$ clases, o celdas, de manera que cada elemento de la población pertenece a exactamente una de dichas clases (aunque, desde luego, habrá casi siempre un cierto número de celdas vacías). Para ilustrar este concepto, volvamos al ejemplo mencionado antes, en que se consideraban 3 características (esto es, $k = 3$), a saber, edad, ingreso y tamaño de familia. Si suponemos, por simplicidad, que se toman en cuenta sólo las edades de 21 a 65, en grupos de 5, tenemos $m_1 = 9$; si tomamos 5 clases de ingresos, resultará $m_2 = 5$, y, finalmente, de 1 a 14 miembros en la familia, habrá $m_3 = 14$; entonces, $M = 9 \times 5 \times 14 = 630$. Cada elemento de la población pertenece así a una de las 630 celdas que se han definido, aun cuando no todas ellas necesariamente contienen a algún miembro de la población.

Supongamos ahora que una investigación por muestreo esté enfocada al estudio de las k variables definidas en cada elemento, y que, como es lo usual, se intente no solamente el análisis por separado, o sea, marginal, de cada una de ellas, sino que interese también obtener información sobre el comportamiento conjunto de dos, o más

de las k variables. Entonces, es importante garantizar que, al efectuarse la selección de la muestra, cada una de las celdas que sean de significación para el análisis quede adecuadamente representada en ella, esto es, que contenga una proporción adecuada de elementos respecto de los que dicha celda tiene en la población total.

Veremos a continuación el modelo probabilístico aplicable a esta situación, a fin de poder plantear formalmente el problema.

II. Modelo probabilístico para el problema

Podemos considerar, sin pérdida de generalidad, que la muestra se va a extraer sin reemplazo y mediante un esquema de aleatoriedad completa, es decir, asignando a cada elemento la misma probabilidad de selección. Podemos numerar arbitrariamente las M celdas en que se ha clasificado a los elementos de la población, del 1 al M, y definir las probabilidades $p_i = prob$ (un elemento extraído al azar pertenece a la i-ésima celda), para $i = 1, 2, \ldots, M$. Bajo las suposiciones hechas arriba, cada p_i es simplemente la proporción de la población que pertenece a la celda respectiva a la población total, de modo que, si designamos con A_i estos valores, tenemos $p_i = A_i / \Sigma A_i$, y claramente $\Sigma A_i = T$, el total de elementos en la población.

Supongamos que se extrae una muestra de tamaño N y que el número de elementos que caen en la i-ésima celda es n_i, para $i = 1, 2, \ldots, M$, de manera que, por hipótesis, $\Sigma n_i = N$. Evidentemente las n_i son variables aleatorias, y es un resultado conocido de la teoría de la probabilidad que la distribución conjunta de dichas variables es la llamada hipergeométrica, esto es, se tiene que

$$prob\ (n_1 = c_1,\ n_2 = c_2, \ldots, n_M = c_M) = \frac{\binom{A_1}{c_1}\binom{A_2}{c_2}\cdots\binom{A_M}{c_M}}{\binom{A_1 + A_2 + \ldots + A_M}{N}},$$

donde $c_1, c_2, \ldots, c_M$ son constantes enteras no negativas que satisfacen $c_1 + c_2 + \ldots + c_M = N$.

Podemos ahora establecer un conjunto de condiciones que debe cumplir cada uno de los números n_i que resultan en cada celda al extraer la muestra. Estas condiciones las fija, por supuesto, el investigador en términos de las necesidades de información que requiera en cada celda, y generalmente pueden establecerse mediante la colección de desigualdades

$$n_1 \geq d_1,\ n_2 \geq d_2, \ldots, n_M \geq d_M,$$

donde, de nuevo, las d_i son constantes no negativas arbitrarias. En otras palabras, se está pidiendo que el total que aparece, dentro de la muestra, en la celda i-ésima, sea no menor que un cierto número d_i que puede ser fijado de modo arbitrario. Naturalmente, la naturaleza aleatoria del experimento no permite asegurar que se cumplirá lo an-

terior, pero se puede, en cambio, prescribir una probabilidad mínima de que suceda, y la magnitud de dicha probabilidad, que en general dependerá del tamaño de la muestra (y, por supuesto, de las d_i), será una medida de la bondad esperada de la muestra para los fines del investigador.

En las condiciones anteriores, el problema puede plantearse de la siguiente manera: determinar un número N tal que se satisfaga la desigualdad

$$prob\ (n_1 \geq d_1, n_2 \geq d_2, \ldots, n_M \geq d_M) \geq q, \qquad (1)$$

donde q es un número positivo y menor que 1, que se fija arbitrariamente en atención a las necesidades de la investigación: un valor pequeño de q dará lugar a una N relativamente pequeña, pero que quizá no garantizará debidamente que las celdas queden representadas en la muestra, en tanto que una q demasiado grande, si bien tal vez satisfará esta última condición, llevará con seguridad a un tamaño de muestra excesivamente alto.

En todo caso, el problema es esencialmente de cálculo, pues la desigualdad (1) es equivalente a

$$\sum_{\substack{c_i \geq d_i \\ i=1,\ldots,M}} \frac{\binom{A_1}{c_1}\binom{A_2}{c_2}\cdots\binom{A_M}{c_M}}{\binom{A_1 + A_2 + \ldots + A_M}{N}} \geq q, \qquad (2)$$

de donde habría que determinar la (mínima) N que la satisface. No es exagerado afirmar que la única forma viable de abordar este problema es a través de aproximaciones sucesivas, probando diferentes valores de N, pero aun esto presenta serias dificultades en la evaluación numérica de la suma múltiple que aparece en (2), como puede verificarse con sólo intentar el desarrollo de un procedimiento de cálculo para efectuar las operaciones indicadas.

En la siguiente sección presentaremos un método aproximado para evaluar la suma que aparece en (2) de manera eficiente.

III. Método propuesto para la solución

Existe la posibilidad de simplificar notablemente el esfuerzo de cálculo en el problema que hemos planteado recurriendo a las llamadas *desigualdades de Bonferroni*,[1] combinadas con resultados posteriores obtenidos por Mallows.[2] Describiremos a continuación en qué consisten dichos resultados, y su aplicación a nuestro problema.

[1] Véase, por ejemplo, W. Feller, *Introduction to Probability Theory*, Vol. 1, Nueva York, John Wiley & Sons, 1957, p. 100.

[2] C. L. Mallows, "An inequality involving multinomial coefficients", *Biometrika*, 55, 1968, pp. 422-424.

Designemos con E_i el evento $n_i \geq d_i$, y con F_i su complemento, para $i = 1, 2, \ldots, M$. Entonces, se cumple que

$$prob\,(F_1 \cup F_2 \cup \cdots \cup F_M) \leq \sum_{i=1}^{M} prob\,(F_i), \tag{3}$$

donde el símbolo $\cup$ denota la operación de unión entre dos eventos. Éste es un resultado elemental (conocido como desigualdad de Boole) en la teoría de la probabilidad, y puede establecerse fácilmente por inducción matemática.

Por otra parte, extendiendo las desigualdades de Bonferroni (de las cuales la fórmula (3) puede verse como un caso particular), Mallows (*op. cit.*) obtuvo que

$$prob\,(E_1 \cap E_2 \cap \ldots \cap E_M) \leq \prod_{i=1}^{M} prob\,(E_i), \tag{4}$$

donde $\cap$ denota, como es usual, la operación de intersección entre dos eventos, y Π es el símbolo convencional para indicar el producto de los factores indicados.

Ahora bien, aplicando una de las conocidas leyes de de Morgan al lado izquierdo de (3) obtenemos

$$prob\,(F_1 \cup F_2 \cup \cdots \cup F_M) = 1 - prob\,(F_1 \cap E_2 \cap \ldots \cap E_M),$$

de donde resulta que

$$prob\,(E_1 \cap E_2 \cap \ldots \cap E_M) \geq 1 - \sum_{i=1}^{M} prob\,(F_i),$$

y combinando este último resultado con (4) tenemos, finalmente,

$$1 - \sum_{i=1}^{M} prob\,(F_i) \leq prob\,(E_1 \cap E_2 \cap \ldots \cap E_M) \leq \prod_{i=1}^{M} prob\,(E_i). \tag{5}$$

Notemos que la expresión $prob\,(E_i \cap E_2 \cap \ldots \cap E_M)$ es justamente igual a la suma múltiple que aparece en (2), y el resultado (5) nos da una cota inferior y otra superior para dicha suma, y ambas cotas aparecen sólo en términos de las probabilidades marginales correspondientes $prob\,(E_i)$ y $prob\,(F_i)$, que son, por supuesto, complemento una de la otra, y fáciles de calcular. En efecto, tenemos

$$prob\,(E_i) = prob\,(n_i \geq d_i)$$

$$= \sum_{j=d_i}^{n} \frac{\binom{A_i}{j}\binom{T-A_i}{n-j}}{\binom{T}{n}}, \tag{6}$$

y esta suma presenta problemas de poca importancia en comparación con la que aparece en (2). Por otra parte, cuando las A_i son grandes puede verificarse que cada uno de los términos hipergeométricos es virtualmente igual al correspondiente en la distribución binomial de parámetros $(n, A_i/T)$, y puede entonces usarse esta última o bien la aproximación normal respectiva para evaluar la suma sin pérdida apreciable de precisión en los cálculos.

En la siguiente sección aplicaremos estos resultados a un ejemplo concreto para ilustrar su utilización en la práctica.

IV. Un ejemplo de la aplicación del método

Considérese el problema de extraer una muestra aleatoria de la población del Distrito Federal, con el propósito de tener representadas las siguientes características simultáneamente: sexo, edad, sector de ocupación, condición migratoria y tamaño de la localidad de residencia (dentro del Distrito Federal hay localidades urbanas y rurales). Si tomamos 2 categorías de sexo, 9 grupos de edad, 5 sectores de ocupación, 2 condiciones migratorias (esto es, migrantes y no migrantes) y 3 tamaños de localidad de residencia (a saber, 1-2 500, 2 501-20 000, y 20 000 en adelante), tenemos un total de 540 celdas, y se trata, entonces, de determinar un tamaño de muestra que garantice una representatividad adecuada de las mismas dentro de la muestra.

Puede observarse, *a priori*, que un problema de esta índole va a requerir de una muestra muy grande, en virtud del alto nivel de detalle que se estipula. Debe aclararse que el diseño que se propone como ejemplo fue originalmente planteado para extraer la muestra directamente de la información censal respectiva, de modo que el tamaño total de la población T es de alrededor de 7 millones. En este caso, la utilización de equipo moderno de computación permite almacenar y manejar volúmenes considerables de información, de manera que el tamaño de la muestra puede ser muy grande sin que ello ocasione dificultades para la investigación.

El marco muestral utilizado para el ejemplo fue la información del Censo de Población de 1960, de donde se obtuvieron estimaciones para los valores A_i correspondientes a cada celda, y con base en ellos se efectuaron los cálculos para determinar el tamaño de muestra que sería necesario extraer del Censo de 1970 a fin de lograr la representación deseada en cada una de las 540 celdas definidas anteriormente.

Las sumas (6) se calcularon utilizando la aproximación normal a la distribución hipergeométrica, de manera que tomamos

$$\text{prob}\,(E_i) \doteq \frac{1}{\sqrt{2\pi}} \int_{a_i}^{\infty} e^{-y^2/2}\, dy\,, \tag{7}$$

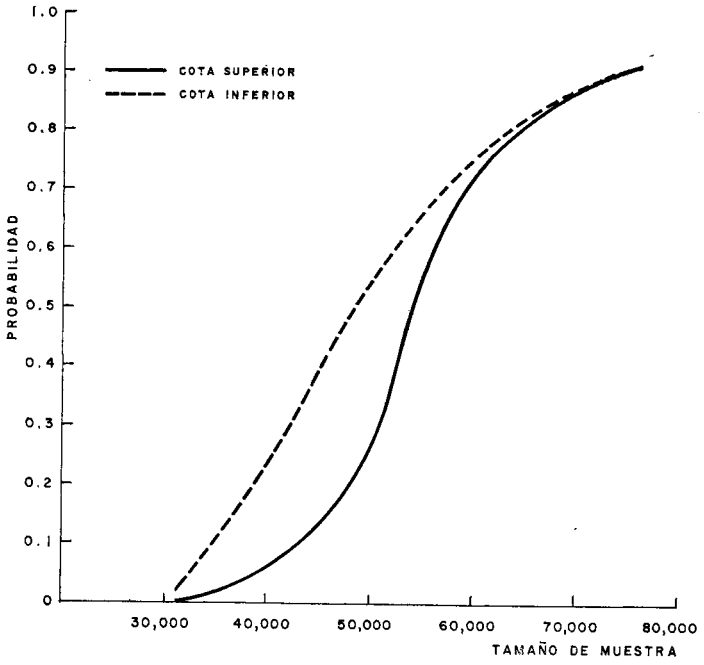
donde $a_i = (d_i - N A_i/T - 0.5)/\sqrt{(NA_i/T)\,(1 - N A_i/T)}$, y se consideraron diversas posibilidades para las d_i, de las cuales mencionamos dos a continuación:

a) Se requiere que la muestra contenga, en la celda *i*-ésima, el mínimo de las cantidades (50, $0.02 \times A_i$), en el caso en que dicha celda contenga al menos 300 individuos en el Censo de 1960. Si la celda tenía entonces menos de 300, no se le impone condición alguna.

b) Se requiere lo mismo que en *a)* para las celdas que en 1960 tenían al menos 300, y además el 8 % del total si éste era inferior a 300 en 1960.

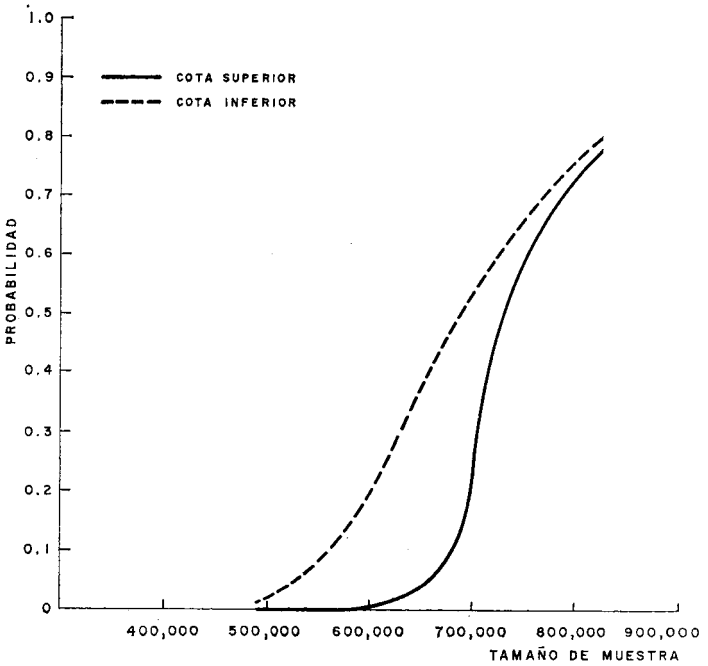
Notamos que la condición *b)* es mucho más restrictiva que la *a)*, y esto se refleja, como era de esperar, en el tamaño de la muestra.

Los cálculos necesarios se efectuaron en la computadora B-5500 del Centro de Cálculo de la Universidad Nacional Autónoma de México, y los resultados se presentan en las gráficas 1 y 2, donde aparecen los tamaños de muestra correspondientes a diferentes probabilidades de garantizar las condiciones *a)* y *b)*, respectivamente. Es interesante observar la eficiencia de las cotas propuestas en este trabajo, pues a partir de aproximadamente 0.8 se ve que de hecho coinciden en los dos casos presentados. Como dato complementario puede mencionarse que los cálculos para construir cada una de las gráficas requirieron de un promedio de 7 segundos en el referido equipo de cómputo.



Gráfica 1

COTAS SUPERIOR E INFERIOR A LA PROBABILIDAD DE SATISFACER LOS REQUISITOS *a)*



Gráfica 2

COTAS SUPERIOR E INFERIOR A LA PROBABILIDAD DE SATISFACER LOS REQUISITOS *b)*