Metodología de Encuestas
Vol 3, Núm 1, 2001, 65-73

APLICACIÓN DEL MUESTREO SISTEMÁTICO EN EL DISEÑO DE ENCUESTAS

Cristina Fernández Alvaro
Begoña Salamanca Miño
Fernando López Blázquez
Universidad de Sevilla

RESUMEN

La obtención de muestras utilizando el muestreo sistemático es algo que se ha llevado a cabo desde principios del siglo XX, debido a la simplicidad de su manejo. Como inconveniente, los diseños sistemáticos clásicos no son estimables, por lo que no podemos suministrar estimadores directos del error de muestreo cometido.

Este inconveniente ha tratado de salvarse a través de distintas variantes sobre la versión más simple del muestreo sistemático. En este trabajo, proponemos otra variante, en la que se hace aleatorio el paso del muestreo sistemático, siguiendo una determinada variable, lo que nos conduce a un diseño equivalente al muestreo aleatorio simple y, por tanto, estimable. Conseguimos así un procedimiento secuencial de obtención de muestras de gran utilidad en poblaciones que tienen una cierta estructura, y fácilmente implementable en la realización de encuestas.

Palabras clave: muestreo sistemático, hipergeométrica negativa, muestreo aleatorio simple.

Introducción

La forma en la que se lleva a cabo el muestreo es determinante para los resultados que posteriormente se van a obtener de la investigación. El hecho de que existan muchos esquemas de muestreo se debe a que se desea reducir el error que se comete al dar estimaciones sobre los parámetros, es decir, el error inherente de muestreo. Pero esta abundancia en el campo de la teoría no viene acompañada por un abanico equivalente de posibilidades de aplicación. Son muy distintas la visión teórica de un diseño muestral y su posterior realización práctica. Así, en general, podemos decir que todos los diseños muestrales tienen dificultades de implementación. De hecho, la realización en la práctica de un esquema muestral u otro, cuando se está condicionado por el coste económico y el tiempo, va a depender de las buenas propiedades que presente uno sobre el otro y, sobre todo, de las posibilidades que tengamos de llevar uno de ellos a cabo según el marco disponible de la población.

El muestreo aleatorio simple es el diseño muestral más empleado, fundamentalmente cuando se verifica la propiedad de que la información que proporcionan los individuos es equivalente. Pero también presenta dificultades de implementación a nivel práctico (véase Cochran, 1974; e Iachan,1982).

Supongamos, por ejemplo, que se muestrea entre los alumnos de una facultad y, dado que se cuenta con el marco en los registros de secretaría, se puede llevar a cabo un muestreo aleatorio simple, asignando a cada individuo un número y realizando un sorteo. El problema, en este caso, lo constituye buscar al alumno, especialmente cuando esa unidad presenta dificultades para su localización (cambio de dirección, abandono de la carrera,...).

En diseños más complicados, los problemas son mayores.

El muestreo aleatorio simple tiene una característica importante respecto al error de muestreo y es el hecho de que es controlable a través del tamaño de la muestra. Pero cuando los marcos de la población no están perfectamente definidos, los problemas del muestreo aleatorio simple, y en general de cualquier otro esquema de muestreo, son mayores. Aparece, con ello, toda la problemática de los marcos imperfectos.

Por esta razón, desde el principio de la Teoría del Muestreo, se han venido utilizando los muestreos secuenciales, que no tienen toda su base en el marco y que trabajan más directamente con los elementos de la población con los que se va "encontrando" , siguiendo determinados criterios. Así, los primeros estudios en teoría de muestras se hicieron por esquemas sistemáticos (véase Buckland,1950). Entre los procesos secuenciales más conocidos:

- Bowley, en su encuesta de 1912, según Bellhouse (1988), recurrió a la selección sistemática de uno de cada diez edificios.
- En las primeras aplicaciones de las muestras, se empleó el muestreo sistemático para obtener muestras sobre el uso de la tierra y estudios de bosques.

- También se obtuvo información suplementaria al censo de 1940 en EEUU a partir de muestras sistemáticas del mismo de tamaño: uno de cada veinte personas.

Según la regla que se utilice para obtener la muestra en el muestreo secuencial, se conseguirá que este muestreo sea equivalente a algún determinado diseño muestral. Nuestro objetivo será conseguir un muestreo aleatorio simple, ya que ello facilitaría todos los cálculos posteriores.

El muestreo sistemático clásico, al contrario de otros diseños más complejos, tiene una fácil implementación. Basta tener la lista de la población y recorrerla una vez para obtener la muestra, por lo que en situaciones en las que contemos con la población una sola vez (encuestas a la salida de cines, teatros...) es de gran utilidad.

Es, por tanto, el tipo de muestreo secuencial más sencillo, ya que permite obtener con mucha facilidad la muestra de tamaño dado. Así por ejemplo, si se dispone de una lista telefónica, el muestreo sistemático permite obtener la muestra recorriendo la población y escogiendo un elemento de cada n/N (si se desean muestras de tamaño n de una población de tamaño N)

El muestreo sistemático clásico presenta, sin embargo, un inconveniente importante: se trata de un diseño *no estimable*, en el sentido de que no se pueden dar estimadores del error de muestreo cometido, ya que no todo par de elementos puede aparecer en la muestra (véase Hedayat y Sinha, 1991). Esta circunstancia provoca que las probabilidades de inclusión de segundo orden sean nulas y no se pueda determinar directamente la medida del error.

Para evitarlo se siguen tres caminos:

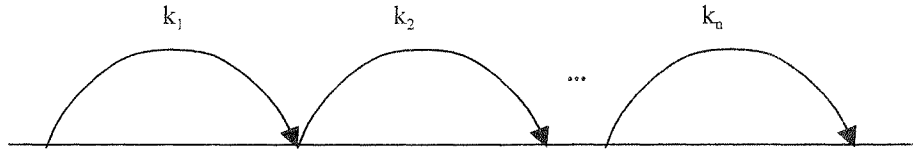
- El primero implica la utilización de los métodos de estimación del error muestral por replicación, lo que suele ocasionar un aumento en el tamaño muestral.
- Un segundo camino supone que la población U investigada es una parte de una población más general sobre la que se suponen ciertas hipótesis. Estos modelos de superpoblación permiten estimar el error de muestreo utilizando las propiedades muestrales que se derivan de la población general. Así, si la superpoblación es completamente aleatoria, como es el caso de poblaciones recogidas en listas alfabéticas de personas, se puede estimar el error de muestreo usando las fórmulas del muestreo aleatorio simple.
- La tercera vía trata de solucionar el problema de que todo par de elementos no puede pertenecer a la muestra. Para ello hace variable el tamaño del paso del muestreo sistemático. En esta tercera dirección se incluye el esquema de muestreo que proponemos. Introducimos un diseño muestral sistemático en el que el paso del diseño es una variable aleatoria que evoluciona de un paso a otro según una cierta ley. Con ello conseguimos que el diseño sistemático sea estimable, al permitir que la probabilidad de inclusión de cualquier par de elementos de la población en la muestra sea siempre positiva. Además, el diseño resultante debe no solo permitir la extracción de la muestra de un modo fácil, sino también determinar los errores de muestreo. Para ello, se considera una determinada variable aleatoria que define el salto, lo que nos conduce a que el diseño final sea

precisamente el muestreo aleatorio simple. Gracias a esta estrategia, podemos recurrir a las fórmulas de este diseño para la determinación de los estimadores y los errores de muestreo.

Esquema de muestreo

Como se ha comentado en el muestreo sistemático clásico, al imponer un salto determinístico no es posible determinar los errores de muestreo. Para salvar este problema vamos a considerar el salto aleatorio, pero además buscaremos que el diseño resultante sea equivalente a un diseño tan conocido como es el muestreo aleatorio simple.

El paso se lleva a cabo a través de la realización de una variable aleatoria y, por tanto, de comportamiento imprevisible a priori, por lo que cualquier par de elementos tiene probabilidad no nula de ser seleccionado:



La variable dependerá de los elementos que ya se hayan recorrido y de los que aún quedan por recorrer.

Para conseguir los dos objetivos que hemos comentado, ser estimable y equivalente a un muestreo aleatorio simple, trabajamos con la distribución hipergeométrica negativa: $HN(n,N\text{-}n,l)$ definida de la siguiente forma: si tenemos una urna con bolas blancas y negras, esta ley nos indica el número de extracciones sin reposición que tenemos que hacer hasta obtener l bolas negras. En nuestro caso l=1, por lo que la distribución se representa como:

$$HN(n,N\text{-}n)$$

Tratamos de relacionar una extracción sucesiva de bolas de una urna con el muestreo. Para ello, se asocian todos los elementos de esa urna con los elementos de la población, siendo la muestra los elementos que corresponden a las bolas negras. De este modo, el número de bolas totales representa el total de la población N, el número de bolas negras el tamaño de la muestra n y el número de bolas blancas será el resto.

En cada paso utilizaremos una ley hipergeométrica negativa de parámetros: el número de bolas blancas y negras que queden hasta ese momento. Es decir, el paso r-ésimo, K_r, seguirá:

$$HN\left(n-r+1, N-n-\sum_{i=1}^{r-1} k_i + r - 1\right)$$

siendo:

N: el número total de individuos de la población.

n: el número de elementos de la muestra.

k_i: la realización de la variable K_i, con $1 \leq i \leq r\text{-}1$ y $k_0 = 1$.

El esquema de muestreo es el siguiente:

Sea $U = \{e_1,...,e_N\}$ la población en la que vamos a obtener una muestra de tamaño n.

Paso 1: Obtenemos un valor aleatorio de la distribución $HN(n,N\text{-}n)$, que llamamos k_1, lo que nos dice que el elemento e_{k_1} es el primer elemento seleccionado para la muestra.

Paso 2: Obtenemos un valor aleatorio de una $HN(n,N\text{-}n\text{-}k_1+1)$ que llamamos k_2, lo que nos dice que el elemento $e_{k_1+k_2}$ es el segundo elemento seleccionado para la muestra.

.....

Paso r: Obtenemos un valor aleatorio de una distribución

$$HN\left(n-r+1, N-n-\sum_{i=1}^{r-1} k_i + r - 1\right)$$

que llamamos k_r lo que nos dice que el elemento $e_{k_1+\cdots+k_r}$ es el r-ésimo elemento seleccionado para la muestra.

Seguiremos $r = r+1,...,n$.

Cuando $r=n$ tendremos los n elementos seleccionados y una muestra con los siguientes elementos de la población:

$$(e_{k_1}, e_{k_1+k_2}, \ldots, e_{k_1+k_2+\cdots+k_n})$$

La demostración de que este procedimiento es equivalente a un muestreo aleatorio simple puede verse en Fernández y Salamanca (2001).

En una situación real, se debe proporcionar al entrevistador una lista con las realizaciones de las n variables aleatorias que hemos planteado. Esta lista le indica al encuestador cuántas personas (caso de que sea una encuesta personal) debe dejar pasar hasta preguntar a una (en términos del esquema de urnas: cuántas bolas blancas salen antes de una negra).

Imaginemos que estamos preguntando en una fila. Siguiendo este esquema preguntaríamos al individuo k_1-ésimo. A partir de él y dando el salto de k_2 indicado en la lista (realización de la variable k_2), seleccionaríamos al individuo k_1+k_2-ésimo. Así sucesivamente hasta que tuviéramos la muestra completa de n individuos.

Ejemplo A

Supongamos que se quiere conocer la opinión del alumnado sobre una asignatura en la que está matriculado. Disponemos de las actas con los nombres de los alumnos ordenados alfabéticamente. Imaginemos que se cuenta con 300 alumnos y que se ha decidido obtener una muestra de 10. Para realizar la selección, se recorre la lista una sola vez, aplicando el muestreo sistemático de paso aleatorio que hemos propuesto.

En primer lugar, obtenemos una realización de una $HN(10,290)$. El resultado es, por ejemplo, 67. Seleccionaríamos por tanto al individuo 67 de la lista.

El próximo paso se lleva a cabo a partir de otra hipergeométrica negativa, pero con distintos parámetros. Concretamente se tratará de $HN(9,224)$, pues 224=290-67+1. La realización de esta variable resulta ser 66. Luego, seleccionaremos al individuo 67+66=133 de la lista.

Así sucesivamente, iremos obteniendo las distintas realizaciones de las variables aleatorias, concretamente:

Paso 3: Realización de $HN(8,159)$: 4. Seleccionaremos al individuo 137.
Paso 4: Realización de $HN(7,156)$: 4. Seleccionaremos al individuo 141.
Paso 5: Realización de $HN(6,153)$:14. Seleccionaremos al individuo 155.
Paso 6: Realización de $HN(5,140)$:20. Seleccionaremos al individuo 175.
Paso 7: Realización de $HN(4,121)$:10. Seleccionaremos al individuo 185.
Paso 8: Realización de $HN(3,112)$:10. Seleccionaremos al individuo 195.
Paso 9: Realización de $HN(2,103)$:46. Seleccionaremos al individuo 241.
Paso 10: Realización de $HN(1,58)$:50. Seleccionaremos al individuo 291.

Tenemos, por tanto, los individuos que componen la muestra. Como este esquema es equivalente al muestreo aleatorio simple, podremos calcular directamente los errores de muestreo.

Otra variable que puede considerarse a la hora de llevar a cabo el muestreo, especialmente en poblaciones no estáticas (como son personas en un determinado lugar), es la forma en que estas personas se "ordenan" secuencialmente para ser seleccionadas. Si se cuenta con una única fila no hay ningún problema, pues puede considerarse el orden más natural: según se encuentran en la cola. Pero si no es así, es necesario decidir cuál va a ser ese orden (contar de derecha a izquierda, por filas,...) Ésta es una cuestión muy importante, ya que la elección depende precisamente de ese orden.

Sabiendo de antemano el orden en el que se escoge a los encuestados y conociendo la duración de cada entrevista, podemos prever cuántos encuestadores se

necesitan para realizar la investigación o en qué momentos es necesario un segundo encuestador para escoger al individuo seleccionado mientras el primero lleva a cabo la entrevista.

Existe otro procedimiento secuencial que consigue un diseño equivalente al muestreo aleatorio simple, dado por Fan, Muller y Rezucha en 1962 (véase Fernández y Mayor, 1995). No obstante, este procedimiento implica hacer N realizaciones aleatorias en lugar de n, como se ha planteado en el esquema que se presenta en este trabajo.

Comportamiento asintótico

Cuando N sea muy grande y la fracción de muestreo $f = n/N$ pequeña, las variables aleatorias, K_i, hipergeométricas negativas, que determinarán el tamaño del paso en el muestreo sistemático de paso aleatorio, pueden aproximarse por una misma ley, una geométrica con parámetro la fracción de muestreo, $K \approx Ge(f)$.

La ley geométrica es fácil de generar:

Generar $U \in u(0,1)$,

$$y = \text{int}\left(\frac{\ln u}{\ln(1-\text{f})}\right),$$

entonces $y \approx Ge(f)$. Con lo que, para la elección de la muestra de tamaño n, se hace repetir n veces el procedimiento, y se van acumulando los números enteros obtenidos.

El tamaño esperado del recorrido de la muestra es N unidades ($=nN/n$), por lo que la muestra podría ser de tamaño menor que n. Para evitar tal problema, se parte de muestras de tamaño fijo, al considerar la población de un modo circular $\{e_1, e_2, ..., e_N, e_1, e_2, ...\}$ y se comienza el muestreo sistemático con paso generado por la ley geométrica $Ge(f)$ desde una unidad $e_{(1)}$ escogida al azar, entre todas las unidades de U.

Ejemplo B

Supongamos que el objetivo es muestrear a la población que asiste a un espectáculo a la salida del mismo, para preguntar su opinión sobre la obra. Suponemos que esa población sale ordenada en una fila, o bien se ha determinado el orden previamente.

Para la simulación, realizada con el paquete informático Maple V, se considera una población de 3000 personas, ordenadas a la salida del espectáculo, y se decide un tamaño de muestra de 20 personas. Al considerar un comportamiento asintótico, el paso seguirá una ley $Ge(1/150)$ Se calcula un número aleatorio en el intervalo (1, 3.000), para aleatorizar la primera persona a partir de la cual vamos a contar, y consideramos teóricamente la secuencia de personas cíclica. Según el primer resultado,

se ha obtenido que comenzamos por la persona 595, con lo que se debe empezar a contar a partir de ese número de persona.

Las aplicaciones informáticas nos permiten generar directamente una distribución geométrica de la que obtenemos 20 realizaciones:

120, 332, 144, 364, 111, 88, 596, 79, 741, 25, 320, 61,
10, 19, 127, 24, 120, 12, 66, 171.

En el procedimiento, se acumulan estas realizaciones a partir del número de persona inicial. Si se sobrepasa el número total de personas, 3.000, se continúa el proceso por la unidad 1, como si ésta ocupara el orden 3001.

Procediendo de tal forma, se obtiene la secuencia del orden de las personas que hay que entrevistar:

1	715	*11*	515
2	1.047	*12*	576
3	1.191	*13*	586
4	1.555	*14*	605
5	1.666	*15*	732
6	1.754	*16*	756
7	2.350	*17*	876
8	2.429	*18*	888
9	170	*19*	954
10	195	*20*	1.125

Una vez realizada esta labor, es necesario ordenar la muestra, con el objetivo de trabajar con la población una sola vez, cuando vayan saliendo del espectáculo:

1	170	*11*	888
2	195	*12*	954
3	515	*13*	1.047
4	576	*14*	1.125
5	586	*15*	1.191
6	605	*16*	1.555
7	715	*17*	1.666
8	732	*18*	1.754
9	756	*19*	2.350
10	876	*20*	2.429

Conclusiones

Desde que Bowley llevó a cabo su estudio sobre viviendas, se comprobó que el muestreo sistemático clásico es fácil de llevar a la práctica, ya que se trataba de buscar la vivienda seleccionada en la calle, en contraposición con la búsqueda de un determinado número de vivienda, que puede que no esté numerada correctamente o que haya desaparecido. Este problema no lo presenta el diseño sistemático. Pero este diseño no está libre de inconvenientes, como hemos comentado en el trabajo, siendo el principal problema la estimación del error de muestreo.

El método que nosotros proponemos es más complicado que el sistemático clásico, ya que es necesario especificar el número de saltos que debe hacerse para tener la muestra, mientras que en el modelo clásico hay más automatismo en el salto. Pero, una vez que se han realizado las variables aleatorias y tenemos los saltos, se indican al entrevistador en una lista, por lo que éste no debe más que seguir ese esquema contando los elementos en función de lo que se le ha proporcionado.

Es un proceso más fácil en la calle que el muestreo aleatorio simple, ya que éste fija de antemano los elementos con los que se va a encontrar el entrevistador, obviando su existencia, mientras que el sistemático no determina tanto el elemento como la forma de llegar a él.

Referencias

Bellhouse, D.R. (1988) Systematic sampling. En Krishnariah PR y Rao Eds, Sampling, *Handbook of Statistics*. North-Holland: Sampling Vol 6, 125-145.

Buckland, W.R. (1950). A review of the literature of systematic sampling. Journal of. Royal Statistic Society. B13, 208-215.

Cochran, W.G. (1974) *Técnicas de muestreo*. México: CECSA.

Fan C.T., Muller, M.E., Rezucha, I. (1962) *Development of sampling plans by using sequential (item by item) selection techniques and digital computers*. J. Amer. Statit. Assoc. 57, 387-402.

Fernández, F.R., Mayor, J.A. (1995) *Muestreo en poblaciones finitas: Curso básico*. Barcelona: EUB.

Fernández, C, Salamanca, B. (2001) Un muestreo sistemático de paso aleatorio. Boletín de la SEIO. Marzo 2001.

Hedayat, A.S. y Sinha, B. (1991). *Design and Inference en Finite Population Sampling*. New York: Wiley

Iachan, R. (1982). Systematic Sampling: a critical survey. International Statistic Review, 50, 213-303.