

Las encuestas por muestreo: uso y abuso

Juan José Obagi Araújo*

Resumen: Con base en la teoría del muestreo se analizan los ejercicios de encuestas de opinión. Se presentan algunos casos que evidencian el mal uso y la desinformación hacia la cual conducen las encuestas de opinión mal fundamentadas estadísticamente.

Abstract: In this paper the author presents the statistical basis of the sample theory. Some examples of misuse and lack of information originated in poorly supported opinion polls are shown.

1. Introducción

El primer semestre de 1998 se presenta como uno de los períodos más prolijos en encuestas en nuestro país, debido a la realización de tres elecciones, eventos que motivan tanto a candidatos como a las entidades dedicadas a este tipo de trabajos. Por esta razón, es oportuno y conveniente hacer algunas reflexiones sobre el tema, a la luz de la teoría de muestreo.

Algunas acepciones para el término Encuesta, son:

- Investigación (del latín *aquaesita*)
- Indagación (utilizado en Psicología)
- Pesquisa(en el ámbito judicial)

Es decir, dependiendo del contexto se tiene una acepción diferente. Si nos circunscribimos al contexto estadístico, una encuesta es una herramienta por medio de la cual se solicita y obtiene información de fuentes *previamente definidas*. En este sentido, no cabe la espontaneidad; si las fuentes son tomadas espontáneamente, por iniciativa de quien solicita o de quien suministra la información, se tiene un simple "sondeo", cuyos resultados deben ser usados con

* Ingeniero Industrial de la Universidad Tecnológica de Pereira, MS. en Estadística de George Washington University. Profesor Asociado del Departamento de Procesos Productivos de la Pontificia Universidad Javeriana.

cautela. El término *fuentes previamente definidas*, significa que la encuesta corresponde a una investigación estadística formal, esto es, a un censo o a una investigación por muestreo. En el primer caso, se investiga la totalidad de fuentes posibles; en el segundo caso, el procedimiento de selección elegido define las fuentes a las cuales se aplicará la encuesta.

Es en este último caso donde interesa detenerse, pues las encuestas vía censo son viables cuando se trata de construir un marco de información sobre un tema particular, cuya vigencia puede oscilar entre 5 y 10 años, aspecto que, sumado a razones de tiempo y costo, determina la preferencia hacia las investigaciones por muestreo.

¿Y qué es una investigación por muestreo? Es un estudio sobre un tema específico, en el cual las fuentes de información son seleccionadas como resultado de la aplicación de un método de muestreo, con el propósito de inferir información acerca de la población o conjunto al cual pertenecen dichas fuentes. Por tanto, es inevitable hablar del término *muestra*, esencia de una investigación por muestreo.

Una muestra es simplemente una parte de una población, seleccionada para realizar una investigación por muestreo. El tipo de selección determina el tipo de muestra. Así, una muestra puede ser: a) Probabilística, si la selección de las fuentes es aleatoria; b) No-Probabilística, si las fuentes fueron seleccionadas usando procedimientos no-aleatorios, es decir, por conveniencia o necesidad.

Una muestra probabilística permite producir estimaciones de valores poblacionales (parámetros) cuya confiabilidad y error pueden ser medidos, mediante procedimientos matemáticos. Una muestra no-probabilística no lo permite.

Ahora bien, la pregunta favorita cuando se necesita realizar una investigación por muestreo es: “*Tengo una población de tal tamaño, ¿cuántas encuestas debo hacer para que mi muestra sea representativa?*”.

Al respecto, una muestra probabilística no siempre es representativa; **la aleatoriedad no garantiza la representatividad de una muestra**. Este último aspecto tiene que ver con la distribución de la misma sobre la población estudiada. De otra parte, antes de preocuparse por el tamaño de la muestra, es preciso establecer el *estadístico* sobre el cual se basará el cálculo correspondiente, aspecto ignorado u omitido por quienes tratan el tema a la ligera.

También, ya realizada la investigación, se pregunta: “*Pero si la muestra es bastante grande, ¿por qué no es representativa?*”.

El tamaño de la muestra, es decir, el número de unidades o elementos seleccionados para ser investigados, es otro ingrediente en la investigación por muestreo. **El tamaño de muestra per se tampoco garantiza su representatividad**: muestras consideradas grandes pueden ser mucho menos representativas que otras de menor tamaño.

En cuanto a la determinación del tamaño de muestra, éste debe obedecer a un cálculo matemático si se trata de una muestra probabilística. Obviarlo, sería contradictorio. Surge entonces el interrogante: ¿De qué depende el tamaño de una muestra aleatoria?

El tamaño de una muestra aleatoria depende **esencialmente**, de:

1. La **variabilidad**, medida por medio de la varianza, en términos absolutos o relativos, del estadístico (variable) fundamental del estudio;
2. La **confiabilidad**, dada por la probabilidad de que el *parámetro* o valor poblacional que se estima, se encuentre dentro de un determinado intervalo;
3. El **error**, con el cual se espera estimar el valor poblacional de interés.

Los dos primeros factores intervienen de manera directa y el tercero de manera inversa, lo cual significa que el tamaño de la muestra se incrementa cuando crecen la confiabilidad y/o la variabilidad y cuando disminuye el error. (Ecuación (1)).

Pero la formulación matemática para su cálculo depende del estadístico escogido y del método de muestreo con el cual se trabaje. Si se utiliza **Muestreo Aleatorio Simple**, el método más elemental y sencillo, el cual garantiza que cada unidad de la población tenga la misma probabilidad de ser seleccionada, la expresión cuando se trata de estimar el parámetro **media poblacional**, es:

$$(1) \quad n_x = \frac{k^2 N \sigma^2}{k^2 \sigma^2 + N e^2}$$

donde:

- n_x : Tamaño de muestra necesario para estimar la media poblacional.
- k : Múltiplo del error estándar, para alcanzar la confiabilidad deseada.
- σ^2 : Varianza de la característica de interés en la población investigada.
- e : Error absoluto con el cual se requiere producir la estimación de la media.
- N : Tamaño de la población investigada.

Para apreciar la influencia del *estadístico* de interés, a continuación se presenta la formulación para estimar el parámetro **proporción poblacional**.

$$(2) \quad n_p = \frac{k^2 N P Q}{k^2 P Q + N e^2}$$

donde:

- P : Proporción de la característica de interés en la población.
- $Q = 1 - P$

Como P no se conoce, en su lugar debe utilizarse una estimación preliminar obtenida por medio de una muestra piloto o de información secundaria.

¿Y cómo influye el tamaño de la población? O en términos probabilísticos, ¿cómo afecta la fracción de muestreo (n/N) el error de muestreo en la estimación?

La respuesta, algunas veces sorprendente para estudiantes e investigadores es que, para poblaciones grandes relativas a la muestra, el tamaño absoluto de la muestra (n) y no la fracción de muestreo, la cual incluye N , tamaño poblacional, determina la precisión de la estimación realizada a partir de la muestra.

Lo anterior es consecuencia del hecho de que, cuando N es grande relativo a n , el factor de corrección para poblaciones finitas $(N - n) / (N - 1)$, se aproxima a uno. Entonces, la definición matemática de la varianza del estadístico **media muestral**, se simplifica sin la presencia del tamaño N de la población, a saber:

$$(3) \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} * \frac{N - n}{N - 1}; \text{ Si } N \gg n \Rightarrow \frac{N - n}{N - 1} \cong 1; \Rightarrow \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

En resumen, la población tiene influencia en la determinación del tamaño de muestra, solamente si se considera finita y “pequeña”; las poblaciones consideradas “grandes” se asumen infinitas para propósitos de cálculo, lo cual se traduce en simplificación matemática, por la no presencia del tamaño de la población. Ahora bien, es pertinente aclarar los términos “grande” y “pequeño”, en este caso. No existe una cifra que opere como lindero. Sin embargo, poblaciones cuyo tamaño esté por encima de 100,000 unidades, pueden ser consideradas grandes para todo propósito práctico. Por ejemplo, si el tamaño de muestra se calcula sobre poblaciones urbanas de grandes ciudades (Santa Fe de Bogotá, Medellín, Cali, Barranquilla, Bucaramanga, Cartagena, Manizales, Pereira) puede omitirse el tamaño poblacional.

Ahora bien, de la misma manera como el tamaño de muestra se fundamenta en un estadístico, asimismo los errores muestrales se refieren a estadísticos, y concretamente están ligados a las estimaciones o valores obtenidos de la investigación muestral. Es decir, el error de muestreo pertenece al estimador de un parámetro y no a la muestra como un todo. Por esta razón, es incorrecto mostrar como parte de la ficha técnica de una investigación por muestreo (no encuesta, como periódicamente se informa) un porcentaje de error para la muestra, omitiendo el estimador o estadístico al cual hace referencia dicho error. En otras palabras, la muestra per se no tiene error pero sí ocasiona que las estimaciones obtenidas a partir de ella lleven consigo un error muestral. Este aspecto facilita la manipulación de los resultados de una investigación por muestreo, porque presenta a la muestra “con error irrisorio” cuando dicho error pertenece a una de las tantas estimaciones posibles resultantes de esa investigación. Este tema, desde luego, merece una mayor dedicación; no obstante, el siguiente ejemplo ayuda a aclarar la anterior afirmación:

Se desea indagar la preferencia en Santa Fe de Bogotá por determinado candidato, quien parece tener un 20% de la población votante a su favor. El equipo asesor del candidato solicita diseñar la muestra con un error no mayor de 1 punto porcentual respecto a la proporción verdadera (parámetro) a favor del candidato, y con una confianza de 95% de que esa proporción será estimada dentro de ese error. ¿Cuál será el tamaño de muestra necesario?

El *estadístico* de interés, en este caso, es la **Proporción Muestral**, estimador de la Proporción Poblacional. Su varianza está dada por (pq/n) .

Dado que la población votante de Santa Fe de Bogotá es, para propósitos muestrales, una población grande, puede omitirse el factor de corrección para poblaciones finitas, lo cual simplifica el cálculo, y la fórmula (2) se convierte en:

$$(4) \quad n_p = \frac{k^2 PQ}{e^2}$$

Cuyo resultado es $n= 6,147$

Este tamaño de muestra garantizará, con una confianza de 95%, que las estimaciones de todas aquellas proporciones poblacionales mayores o iguales a 20%, tendrán un error no mayor de 1 punto porcentual. Por tanto, proporciones inferiores al 20%, serán estimadas con un error mayor al estipulado. Este aspecto no se menciona en una ficha técnica, lo cual de alguna manera confunde o desinforma al lector o usuario no versado en la materia, y peor aún, facilita la manipulación de los resultados. En gracia de discusión, y para que el lector saque sus conclusiones, es de observar que la mayoría de estudios de opinión en el mercado de encuestas tienen un tamaño de muestra mucho menor que el resultado anterior.

Veamos un segundo ejemplo: En uno de los diarios capitalinos, se publicó el resultado de un estudio de opinión, realizado el 02/18/98, una de cuyas preguntas era: "*¿Cree usted que Samper le va a cumplir a Santa Fe de Bogotá con la financiación del Metro?*". La ficha técnica mostraba un tamaño de muestra de 437 unidades, una confiabilidad de 95% y un error máximo permisible de 4.7%.

La aplicación de la fórmula (4), revela que con esa confiabilidad y ese error, una muestra de tamaño 437, sólo permitiría producir estimaciones confiables para proporciones poblacionales no menores de 80%, lo cual quiere decir que las estimaciones presentadas en esa fecha (SI: 72.19% y NO: 22.96%) presentan un error mucho mayor que el anunciado, o visto de otra manera, la confiabilidad para ese mismo error es significativamente menor.

Ojalá estas cortas reflexiones acerca de los fundamentos de la teoría del muestreo les permitan a estudiantes, investigadores e interesados en el tema evaluar la validez de los resultados de las encuestas de opinión que con tanta frecuencia se difunden en el país.