Las encuestas por muestreo: uso y abuso (II)

*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 autor presents the statistical basis of the sample theory. Some examples of misuse and lack of information originated in poorly supported opinion polls are shown.*

En el primer artículo sobre este tema aparecido en el número anterior, se mencionó el concepto de error de muestreo y se expresó que merecía una mayor dedicación:

> *"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 ...." [Obagi, 1998: 71]*

Existen dos clases de errores en las investigaciones estadísticas, según su naturaleza. Son el *error de muestreo* o *error muestral* y el *error no-muestral.* El segundo, está presente en cualquier tipo de investigación incluyendo las investigaciones por muestreo, mientras el primero, *error muestral*, es propio de estas últimas. Más claramente, una investigación censal presentará error no-muestral, pero no podrá presentar error de muestreo. Una investigación por muestreo presentará ambos tipos de errores.

¿En qué consisten el error no-muestral y el error de muestreo?

El error no-muestral es el error producido por causas ajenas al muestreo. Estas causas son múltiples y van desde la misma planeación de la investigación hasta el procesamiento de la información. Una de las etapas donde más se introduce este error es en la recolección de datos. Particularmente, el instrumento de recolección (cuestionario o formulario) es una de las mayores causas de error. Un cuestionario mal confeccionado por su cantidad de preguntas, por la redacción de ellas, por el enlace entre éstas, es causa de error no-muestral. Aún, si el formula-

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

rio estuviese correctamente diseñado, la respuesta del informante así como la manera de registrarla, son también causa de error no-muestral. A manera de ilustración, supóngase una investigación de naturaleza económica en la cual las fuentes de información, industrias, comercios, servicios, etc., suministren datos no consecuentes con su nivel de actividad económica, bien porque no cuenten con registros adecuados o actualizados, o bien porque deliberadamente sobrestimen sus costos o subestimen sus ingresos, al asumir erróneamente que dicha investigación pueda estar relacionada en alguna forma con aspectos tributarios. Así, el error no-muestral puede ser debido a **errores de respuesta** (su origen radica básicamente en el suministro de los datos por las fuentes de información y/o en el diligenciamiento del formulario por parte de los recolectores o entrevistadores) y a **errores de no-respuesta** (formulario inadecuado, digitación, procesamiento, etc.).

El problema entonces es cómo llegar a estimaciones muestrales razonables, en presencia de dichos errores. Para tratar de aclarar el asunto, tómese un modelo muy simple. Supóngase que el gerente de una empresa lanza un dado para determinar aleatoriamente la cifra que él va a suministrar en una encuesta. Si la cifra verdadera fuera **y**, la cifra suministrada sería **y+e**, donde el valor esperado y la varianza del error **e**, fuesen respectivamente, $E[\mathbf{e}] = 0$ y $V(\mathbf{e}) = 10$. Si se toma una muestra de dos empresas, el valor esperado de la media muestral será $E[\mathbf{y}] + E[\mathbf{e}] = \mu$, la media poblacional. Pero la varianza del estimador (media muestral) será $V(\mathbf{y}) + V(\mathbf{e}) = V(\mathbf{y}) + 10/2$, la cual es más grande que la varianza obtenida cuando no están presentes errores de respuesta, es decir, errores no-muestrales (Este último cálculo supone independencia estadística entre **y** y **e**; de lo contrario, habría necesidad de introducir el término de covarianza entre ellas). En este caso, la media muestral ($\overline{\mathbf{Y}}$) continúa siendo un *estimador insesgado* (aquellos estimadores para los cuales *el valor esperado de su distribución muestral, es igual al valor verdadero por estimar*) de la media poblacional (μ), pero su varianza se ha incrementado. Sin embargo, si los errores de respuesta son deliberados (no aleatorios), $E[\boldsymbol{e}]$ es diferente de cero y la estimación estará sujeta a sesgo, el cual se denomina ***sesgo de respuesta***. Aún más, ordinariamente habrá alguna correlación entre ***y*** y ***e***, desvirtuando el supuesto anterior de independencia estadística. Por ejemplo, grandes empresas subestimarán y pequeñas empresas sobreestimarán determinadas características (variables), o viceversa. De otra parte, los datos recolectados de una misma empresa pueden diferir de entrevistador a entrevistador, debido a la gran diversidad de personalidades y caprichos. ¿Cómo corregir esta situación? Existe una posibilidad (más no la única) de hacerlo.

Es posible recoger la información de la totalidad de la muestra por correo, lo cual significa una reducción considerable en los costos de recolección. Después se selecciona una muestra aleatoria de un tamaño muchísimo menor y se envía a los mejores entrevistadores a esas fuentes de información. Ellos tendrían acceso a los registros verdaderos, a partir de los cuales transcribirían la información correcta. Esta técnica, llamada ***doble muestreo***, se utiliza entonces para elaborar una estimación más precisa del valor real desconocido. El estimador utilizado es denominado un ***estimador de diferencia***, de la forma $\boldsymbol{y}\text{-}\boldsymbol{x}+\boldsymbol{X}$. Este estimador está sujeto a un ***sesgo de respuesta*** mucho menor.

Podría haber otro tipo de sesgo presente en los resultados. Sucede cuando algunas de las unidades en la muestra no responden o no suministran información por completo. Este sesgo requiere de otro procedimiento para su disminución.

En el procesamiento de la información, es posible introducir error en la digitación de los datos y si no se dispone de un buen programa de corrección de inconsistencias, el error puede incrementarse en esta etapa. Aquí, el uso de lectoras ópticas permite mejorar los resultados, disminuyendo el error pero sin eliminarlo del todo.

El sesgo se refiere entonces a los errores sistemáticos que afectan a cualquier investigación. Conceptualmente, los sesgos se distinguen de los ***errores variables***, los cuales se suponen aleatorios. Estos últimos pueden ser de muestreo o no muestrales. Generalmente, los errores de muestreo constituyen la mayor parte de los errores variables de una investigación por muestreo y los sesgos provienen fundamentalmente de fuentes ajenas al muestreo.

En la teoría del muestreo hay una formulación ampliamente aceptada, la cual une el error variable y el sesgo en lo que se conoce como ***error total***. Se habla de éste como la **raíz cuadrada positiva del error cuadrático medio (RECM)** y suele remplazar al **error estándar (EE)** —el cual se *define como la desviación estándar de la distribución de muestreo del estimador en cuestión*—, mientras que el **error cuadrático medio (ECM)** sustituye a la varianza. Esta relación entre el error variable de muestreo y el sesgo, es

$$(1) \qquad E[\bar{y} - \mu]^2 = E\,[\,\bar{y} - E[\bar{y}]]^2 + [E[\bar{y}] - \mu]^2$$

El valor esperado se toma sobre la distribución de todos los valores posibles del estimador media muestral ($\overline{Y}$). Este valor medio puede ser igual al valor poblacional (μ), o puede no serlo. La diferencia entre los dos es el llamado sesgo de muestreo. Se dice entonces que un diseño de muestra, para estimar una media poblacional, es insesgado si $E[\overline{Y}] = \mu$. Nótese que ésta no es propiedad de una sola muestra, sino de toda la distribución de muestreo, y no pertenece sólo al procedimiento de selección o al de estimación, sino conjuntamente a ambos. Al respecto, es pertinente señalar que no todos los estimadores usados en una investigación por muestreo son insesgados. La media y la varianza obtenidas a partir de muestreo irrestricto aleatorio son insesgados; pero muchos otros estimadores no lo son, por ejemplo, la desviación estándar de una muestra irrestricta aleatoria o la media de una muestra sistemática, entre otros. Sin embargo, en todos los casos el sesgo de muestreo se vuelve despreciable cuando se incrementa el tamaño de la muestra.

Con relación a la ecuación (1), las desviaciones cuadráticas medias de los resultados posibles de la muestra con respecto al **valor verdadero** (μ) se analizan mediante sus dos componentes: la desviación cuadrática media de los errores variables alrededor del valor promedio E[y] y el cuadrado de la desviación de ese promedio con respecto al valor verdadero μ.

Por tanto, el error total o la raíz del error cuadrático medio, es

$$(2) \qquad \text{Error total} = (EV^2 + \text{sesgo}^2)^{1/2},$$

donde :

$$(3) \qquad EV^2 = E[\bar{y} - E[\bar{y}]]^2$$

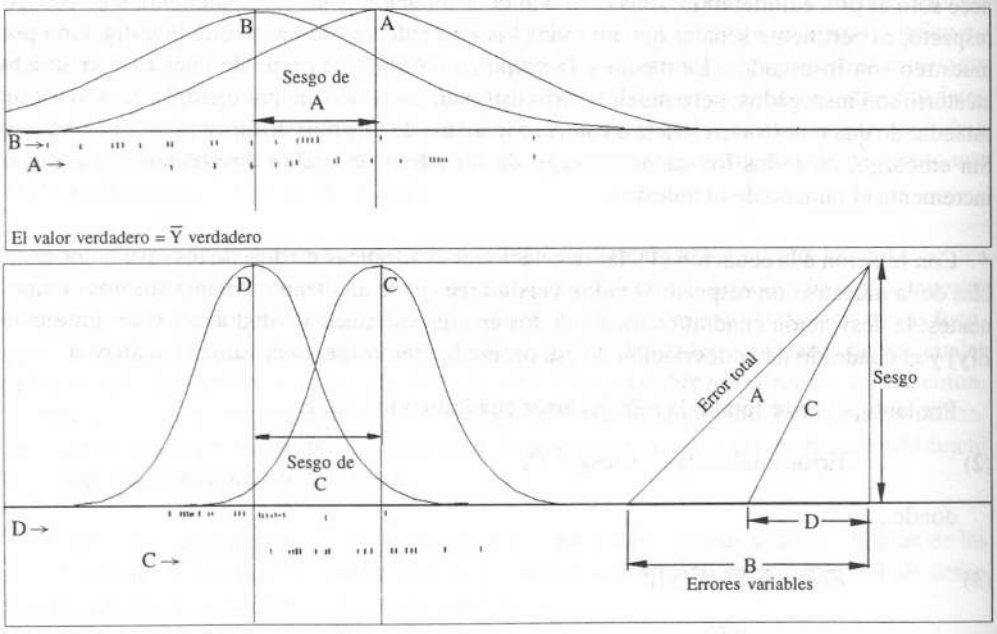
$$(4) \qquad \text{sesgo} = E[\overline{Y}] - \mu$$

Cuando la única causa de los errores variables, **EV**, radica en los errores de muestreo, el valor de $\mathbf{EV^2}$ es igual a la varianza de muestreo.

Los términos **exactitud** y **precisión** se usan para separar los efectos del sesgo. En general, la precisión se refiere a los errores variables pequeños y excluye los efectos del sesgo. La exactitud se refiere a los errores totales pequeños e incluye los efectos del sesgo. Así, en un diseño muestral preciso deben presentarse errores variables pequeños; pero un diseño exacto, además de ser preciso, ha de tener cero sesgo ó muy cercano a cero. En otras palabras, un diseño muestral con sesgo grande será preciso si sus errores variables son pequeños, aunque no será exacto. Estos conceptos pueden relacionarse de manera burda con los de *confiabilidad* y *validez* en algunas ciencias, especialmente la psicología; la confiabilidad se refiere sobre todo a la precisión de las medidas y la validez a la ausencia de sesgo en las medidas.

Como ilustración considérese un diseño tipo **C** en el cual se utiliza un cuestionario enviado por correo, con gran precisión debido a que proviene de una muestra grande, pero con poca exactitud por la debilidad de los métodos de respuesta, deficiencias de marco de muestreo, una gran cantidad de no-respuesta ó una combinación de estos factores, los cuales causan un sesgo grande. Represéntese ahora con **A** un diseño que utiliza el mismo cuestionario por correo pero de tamaño reducido. Otro diseño **B** exhibe los resultados insesgados de los buenos métodos de muestreo, pero hay imprecisión debido al tamaño pequeño de la muestra [Obagi, 1998: 71] lo cual se debe a pocas entrevistas o a muchas entrevistas en pocos conglomerados. Y por último, el diseño **D** muestra los efectos de los métodos insesgados con la alta precisión obtenida de un tamaño de muestra grande.

Figura 1. La relación del sesgo con el error variable.



Los errores totales, los cuales incluyen a los variables, se refieren a resultados promedio, es decir, a resultados esperados y habituales, no a los resultados posibles de una muestra individual. Por ejemplo, en la Figura 1 se aprecia que los resultados de algunas de las muestras de **D**, el mejor diseño, están más lejos de **μ** que algunas muestras de **A**, el peor diseño.[Kish, 1978] Curiosamente, A tiene más muestras próximas a **μ** que **C**, siendo éste mejor; pero **A** también tiene muchas más muestras que se alejan más de ese valor que las de **C**. Las curvas normales representan las distribuciones de probabilidad (distribuciones muestrales) de los estimadores en cada uno de los diseños. Los errores variables corresponden a las desviaciones estándar de las distribuciones y los sesgos a la distancia de los centros de las curvas al valor verdadero **μ**. Nótese que en los diseños **A** y **C** hay sesgos grandes, mientras que **B** y **D** no exhiben sesgo alguno. De otra parte, los diseños **C** y **D** son más precisos, porque sus errores estándar son más pequeños que los de los diseños **A** y **B.** El único exacto es el diseño **D**, porque su error variable y su sesgo son pequeños. Obsérvense también las pequeñas rayas al pie de las cuatro distribuciones, las cuales indican la forma como están concentrados los datos con respecto al valor verdadero. Así, mientras los de **B** y **D** se concentran alrededor de **μ**, los de **A** y **C** están cargados hacia su derecha. Ahora, con relación a los triángulos, en los cuales las alturas representan sesgo, las bases los errores variables y las hipotenusas los errores totales, los diseños **B** y **D** presentan errores totales iguales a sus errores variables debido a la eliminación del sesgo en cada uno. Si bien el error total de **C** (hipotenusa) es grande, a pesar de tener error variable más pequeño, el diseño **A** muestra el error total más grande, lo cual lo hace el peor de los cuatro. Los dos diseños **C** y **D** son precisos (en comparación con **A** y **B**), pero sólo **D** es exacto.

Entonces, los sesgos tienen efectos importantes en los errores totales de una encuesta; un diseño preciso será muy inexacto si tiene sesgo grande. Por eso, la medición de error con base sólo en el error estándar, subestima el error total de la encuesta, y en consecuencia los intervalos basados en errores estándar, para estimar el valor verdadero, resultan en errores mayores de los que se sospecha.

El error de muestreo o error muestral es el ingrediente principal de los errores variables de una investigación por muestreo, y por tanto, del error total; es el error producido como consecuencia de no investigar la totalidad de la población objeto de estudio, sino una parte de ella. Se refiere a un estimador (variable aleatoria) de un valor verdadero desconocido, el cual se trata de estimar a partir de la muestra. Por ejemplo, es posible hablar del error de muestreo para la media muestral, la proporción muestral, etc. Es incorrecto por eso, hablar del error muestral de una encuesta sin especificar el estimador al cual hace referencia. Más aún, *diferentes niveles de un mismo estimador corresponden a diferentes errores muestrales en una misma encuesta*. Su enunciado matemático es el del error estándar, es decir, la raíz cuadrada positiva de la varianza de la distribución muestral del estimador en cuestión, a saber:

$$(5) \qquad \sigma_{\bar{y}} = \{E[\bar{y} - \mu]^2\}^{1/2} = \left[\frac{E[y - \mu]^2}{n}\right]^{1/2} = \sigma_y / \sqrt{n}$$

para el caso de la media muestral de una **muestra irrestricta aleatoria (m.i.a.)** con remplazo o de poblaciones grandes, y

$$(6) \qquad \sigma_{\hat{p}} = \sqrt{pq}/\sqrt{n}$$

para la proporción muestral de una **(m.i.a.)** con remplazo o de poblaciones grandes.

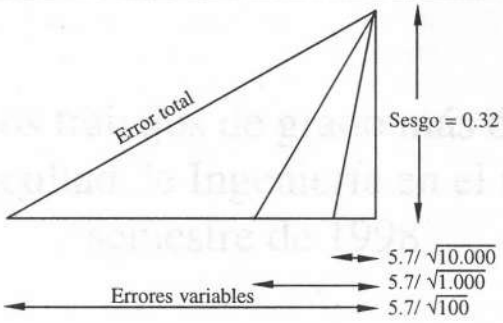
Recuérdese que para poblaciones pequeñas o muestreo sin remplazo, las dos fórmulas anteriores deben ajustarse por el factor de corrección para población finita, (N-n) / (N-1).

Este es el error del cual se habla en las fichas técnicas de las encuestas que ordinariamente se publican, pero su presentación es incorrecta pues se le asigna a la muestra cuando debe imputársele al estimador, y específicamente, al nivel del estimador sobre el cual se fundamentó el diseño de la muestra. Si adicional a ésto se considera la influencia de los errores no-muestrales, y particularmente del sesgo, el error presentado comúnmente no sólo es una subestimación del error total, sino que facilita su manipulación con fines periodísticos o de satisfacción del cliente de turno. Un ejemplo reciente fueron las encuestas preelectorales de mayo y junio de 1998 en Colombia, las cuales mostraron resultados totalmente contrarios a la realidad, en la primera vuelta, y luego, para la segunda vuelta, se utilizó como tabla de salvación la figura del "empate técnico", que permitía pronosticar a la fija el triunfo de cualquiera de los candidatos. Si se consideran los errores descritos, y no sólo el de un determinado nivel del estimador utilizado, es altamente improbable que ocurra esa figura. Este caso es un ejemplo de lo que puede ocurrir cuando el interés científico se sacrifica ante el interés económico, político ó de otra clase.

El siguiente ejemplo puede contribuir a aclarar la influencia de los errores muestrales y no-muestrales en el error total de una investigación por muestreo. Se trata de una investigación sobre el avalúo promedio de inmuebles para vivienda. Una muestra de propietarios entrevistados arrojó un avalúo promedio de 92 millones de pesos, con una desviación estándar de 57 millones de pesos. Posteriormente, se contrató con peritos o profesionales en finca raíz el avalúo de una muestra de los inmuebles en referencia. Estas medidas permitieron una estimación del sesgo de los propietarios al informar su avalúo, considerando como valor verdadero, el suministrado por los peritos. Entonces, el sesgo se define como la diferencia entre el avalúo reportado por los propietarios y el realizado por los peritos. El sesgo promedio resultó positivo y equivalente a 3.2 millones de pesos, es decir, el 3.5% del avalúo promedio de los propietarios.

Para percibir el efecto del sesgo en el error total hay que considerarlo conjuntamente con el error estándar de la media muestral, el estimador fundamental del estudio. En unidades de 10^7, el error total de una m.i.a. de tamaño **n**, es $(5.7^2/n)^{1/2}$, sin el sesgo. El sesgo incrementa este valor a $\{(5.7^2/n)+0.32^2\}^{1/2}$. La Figura 2 permite ver que si el sesgo es pequeño, hay un efecto *moderado* en el error total de una muestra de tamaño n=100, pero se vuelve *dominante* para n=1,000 y *abrumador* para n=10,000.[Kish, 1978] Obsérvese que para disminuir el error total a la mitad, se requiere incrementar el tamaño de muestra aproximadamente 100 veces. El último renglón muestra a $n' = 5.7^2/\{(5.7^2/n)+0.32^2\}$, el cual es el número de observaciones para lograr un estimador insesgado con el mismo error total que el del estimador no-insesgado, con sesgo promedio de $\$0.32 * 10^7$. Esta es una medida del efecto del sesgo en el error total de diversos tamaños muestrales.

Figura 2. Ejemplo del efecto de un sesgo constante en diversos tamaños muestrales



	n = 100	n = 1.000	n = 10.000
$\sqrt{5.7^2/n}$	0.57	0.18	0.06
$\sqrt{5.7^2/n + 0.32^2}$	0.65	0.37	0.32
$\sqrt{5.7^2/(5.7^2/n + 0.32^2)}$	76	240	308

(sesgo = 0.32 y s = 5.7)

Referencias

Hansen, M.H., Hurwitz, W.N., Madow, W.G. *Sample survey methods and theory.* New York: John Wiley and sons, 1963.

Hansen, M.H., Hurwitz, W.N., Bershad, M. *Measurement errors in censuses and surveys.* En: Bulletin of the International Statistical Institute, 38(2), 1961, 359-374.

Kish, L. *Sampling survey.* New York: John Wiley and sons, 1978.

Obagi, J.J. *Las encuestas por muestreo: uso y abuso.* En: *Ingeniería y Universidad,* 2,(1), 1998, 67-71.

Raj, D. *On the relative accuracy of some sampling techniques.* En: Journal of the American Statistical Association, 53, 1958, 98-101.