ESTIMACION DE LOS ERRORES MUESTRALES MEDIANTE EL METODO DE LOS CONGLOMERADOS ULTIMOS

Félix Aparicio Pérez
Centro de Investigaciones Sociológicas

RESUMEN. En este artículo se explica el modelo de los Conglomerados Ultimos para la estimación de los errores muestrales en encuestas basadas en muestras aleatorias. Se incide en la importancia que tiene el conocimiento de estos errores muestrales, así como en la imposibilidad de estimarlos *a priori,* siendo necesario estimarlos *a posteriori,* una vez realizada la encuesta y basándose en la información recogida en ella. Finalmente, se incluye un ejemplo práctico de estimación de los errores muestrales aplicando el modelo de los Conglomerados Ultimos a una de las encuestas realizadas por el CIS.

1. ERRORES EN ENCUESTAS Y CENSOS

En toda encuesta realizada sobre una determinada población, es sabido que las estimaciones obtenidas para las variables en estudio (renta, porcentaje de adeptos a una ideología, etc.) no se corresponden exactamente con los valores reales. Incluso en una encuesta efectuada sobre toda la población en estudio (censo) ocurre este fenómeno.

Es más, un mismo tipo de encuesta que se repita en ocasiones diferentes, pero bajo las mismas condiciones, dará resultados distintos. Las causas de esto son de muy diversa índole, pero podríamos clasificarlas básicamente en dos tipos:

1) Errores de muestreo.
2) Errores ajenos al muestreo.

Los errores ajenos al muestreo son los que influyen incluso en los censos. Se llaman así porque no se deben a la utilización de una muestra, sino a otras causas, como pueden serlo errores accidentales, mala actuación de los agentes encuestadores, defectos de procedimiento, etc.

Los errores de muestreo se deben a la aleatoriedad de la muestra, es decir, a que, por su misma naturaleza, las estimaciones obtenidas a partir de una muestra aleatoria de la población son variables aleatorias y fluctúan de vez en vez, según las leyes del azar. Esto hace que, aun en el caso de que no existieran los errores ajenos al muestreo, si realizáramos distintas encuestas con diseños muestrales semejantes se obtendrían resultados diferentes.

Lo interesante es intentar mantener ambos tipos de errores dentro de unos márgenes aceptables, para que podamos tener confianza en nuestras estimaciones.

Pero para ello será necesario, primero, que conozcamos los errores que tenemos.

En este artículo nos centraremos en los errores muestrales, dejando a un lado los errores ajenos al muestreo. Sólo diré sobre estos últimos que su estimación requiere, en general, repetir parte de las entrevistas por parte de agentes o inspectores especializados, mientras que, para estimar los errores muestrales, tan sólo hace falta tener las entrevistas grabadas en soporte apto para su utilización en ordenador.

2. ERRORES MUESTRALES

Volviendo a los errores muestrales, diré que los libros de muestreo probabilístico afirman que, si la muestra es grande, son pequeños y, además, dan fórmulas para calcularlos. En la sección siguiente damos algunas de ellas como ejemplo.

Estas fórmulas han sido ampliamente utilizadas por los centros e institutos de estudios de mercado y opinión y aparecen en las fichas técnicas que éstos adjuntan a sus estudios y que, frecuentemente, se pueden encontrar en la prensa y revistas cuando en ellas se publica alguna encuesta.

Sin embargo, los errores muestrales que se obtienen aplicando las fórmulas a que hemos hecho referencia no son correctos en la práctica, debido a que, en la realidad, no suelen cumplirse algunas de las hipótesis que harían

que fuesen correctos. Estas hipótesis que no suelen cumplirse son las siguientes:

1) El modelo de muestreo empleado en la mayoría de las ocasiones es bastante más complicado que aquellos modelos para los que, en los libros, se dan las fórmulas de errores muestrales.

2) Los estimadores no siempre cumplen las hipótesis de normalidad que se les supone en las fórmulas. A este respecto, debe resaltarse que la normalidad no sólo depende del tamaño de la muestra (aplicación del Teorema Central del Límite), sino que el hecho de que existan elementos en la muestras que tengan valores mucho más grandes o pequeños que la mayoría en las variables estimadas, puede ser causa de que no haya normalidad (Cochran, pp. 70-71).

3) Estas fórmulas no tienen en cuenta la utilización de filtros en las preguntas que puedan dar lugar a que sólo respondan a algunas preguntas determinadas subpoblaciones.

4) Las fórmulas no nos proporcionan el error muestral relativo, sino sólo el absoluto. Sobre este punto volveremos en las secciones 3 y 5.

Cochran (pp. 72-73) dice que en algunos casos en que se ha investigado la relación entre errores muestrales teóricos y reales se ha encontrado que los reales son hasta cuatro veces superiores a los teóricos. Esto no quiere decir que siempre vaya a suceder algo semejante, pero nos da idea del elevado nivel de incertidumbre que implica la estimación de errores muestrales *a priori,* mediante las fórmulas de los libros de muestreo.

Por otra parte, existen métodos que permiten estimar los verdaderos errores muestrales en que se incurre al efectuar una encuesta aleatoria. A ellos nos referiremos en la sección cuarta.

Antes, en la sección tercera, daremos algunos conceptos y fórmulas necesarios para comprender el resto del artículo.

3. MUESTRAS ALEATORIAS Y CONCEPTOS ASOCIADOS

Se llama población a un conjunto de objetos (personas, hogares, empresas, etc.) de los que deseamos obtener información.

Se llama muestra a un subconjunto de la población, al cual vamos a investigar y, como consecuencia de esta investigación, esperamos inferir características de la población (ej.: encuestamos a 1.000 personas y, a partir de ellas, queremos saber la opinión de los españoles sobre un tema).

Existen muchas formas de crear muestras de una población; por ejemplo,

un investigador médico puede escoger una muestra de conejillos de indias alargando el brazo y tomando aquellos que están más cerca de él, o un empleado administrativo puede tomar una muestra de expedientes escogiendo los 50 primeros expedientes de un fichero.

Sin embargo, nosotros estamos interesados en un tipo muy particular de muestras. Estas son las muestras aleatorias, es decir, aquellas muestras en que la elección de los objetos que forman parte de la muestra se hace al azar. Este tipo de muestreo tiene algunas ventajas. Una es que permite emplear la teoría de probabilidades a la hora de obtener conclusiones, cuantificando así los errores en que se incurre y estimando la precisión con que se trabaja (éste es precisamente el objeto del presente artículo). Otra gran ventaja es que, al dejar en manos del azar la elección de las unidades encuestadas, elimina vicios que pueden producir errores (ejemplos corrientes de estos vicios son entrevistar a las personas que vivan cerca del entrevistador o a las que se ofrezcan a ser entrevistadas).

Un estimador es una función de la información obtenida a partir de la muestra que, normalmente, utilizamos para obtener conclusiones sobre la población. Por ejemplo, en un modelo de muestreo aleatorio elemental, si 500 entrevistados de 1.000 afirman que poseen televisor, deduciremos que el 50 por 100 de la población tiene televisor. El estimador aquí es la proporción muestral, o sea el cociente 500/1.000 estima a la proporción poblacional de personas que tienen televisor.

Como la muestra es aleatoria, los estimadores son variables aleatorias y están, por lo tanto, sujetos a fluctuación. Así, en el ejemplo que acabamos de poner, si realizamos varias encuestas, cada una de ellas a 1.000 personas, para intentar averiguar el porcentaje de personas de la población que tienen televisor, en una encuesta podríamos obtener que 500 entrevistados lo tienen, en otra que son 480 los que lo tienen, en otra que son 510 los que lo tienen, etc. Vemos cómo llegamos a distintas estimaciones con cada una de las muestras. Ahora bien, todas las estimaciones son parecidas; están, si el proceso se ha realizado correctamente, en torno al verdadero valor (desconocido en general). Bien, pues se llama error muestral a la desviación típica del estimador. La desviación típica es una medida de la dispersión; por ello, el error muestral pretende informarnos de si podemos o no esperar pequeñas fluctuaciones entre los estimadores que obtenemos y los verdaderos valores que estimamos (desconocidos, en general). Por tanto, un estimador con un error muestral grande nos servirá de muy poco, puesto que no tendremos confianza de que deba estar cerca del valor verdadero a estimar.

Se trata, pues, de conocer las desviaciones típicas de los estimadores empleados.

Sea N el tamaño de la población (el número de individuos que la componen) y sea n el tamaño de la muestra.

Llamamos proporción poblacional P a la fracción de la población o tanto por uno de individuos que poseen un determinado atributo (personas que tienen televisor o partidarios de una idea, por ejemplo). Se llama proporción muestral a la fracción o tanto por uno de individuos de la muestra que poseen ese mismo atributo. Llamamos P a la proporción poblacional y p a la muestral. P será el número de individuos de la población que poseen el atributo dividido por N y p será el número de individuos de la muestra que poseen el atributo dividido por n.

El estimador más usual de P es p, como decíamos antes; por eso, escribimos a veces $\hat{P} = p$.

Para el modelo de muestreo más sencillo, que es el aleatorio simple, los libros de muestreo dicen que el error muestral de p, $D(p)$, es:

$$D(p) = \left[\frac{N - n}{N - 1} \cdot \frac{P \cdot (1 - P)}{n} \right]^{1/2} \qquad [3.1]$$

Como P es desconocido *a priori*, se suele tomar el caso más desfavorable en [3.1], que es $P = 1/2$ (o sea, el máximo de [3.1] se alcanza en $P = 1/2$, como demuestro en el Apéndice 1); queda, pues, una cota superior teórica del error muestral:

$$D(p)^{\sup} = \left[\frac{N - n}{N - 1} \cdot \frac{1}{4 \cdot n} \right]^{1/2} \qquad [3.2]$$

Esta fórmula [3.2] es correcta, pero, desgraciadamente, ha sido y es mal interpretada y utilizada con una enorme frecuencia. En efecto, [3.2] nos da (suponiendo que se cumplan las hipótesis de normalidad y que el muestreo sea aleatorio simple de poblaciones finitas) la máxima desviación típica que puede tener el estimador de P, *pero a nosotros lo que nos importa en realidad no es esta desviación típica, sino ella dividida por P, es decir, el error relativo, no el absoluto.* Por ejemplo, si P vale 0,3 y $D(p)$ vale 0,2, la estimación es mucho peor que si P vale 0,7 y $D(p)$ sigue valiendo 0,2, puesto que una desviación típica de 0,2 sobre una magnitud pequeña, como 0,3 es mucho más importante que sobre una magnitud grande como 0,7.

Para evitar este problema se trabaja con el coeficiente de variación, que es la desviación típica dividida por la media (y multiplicada por 100, para expresarlo como tanto por ciento); así, para el ejemplo que acabamos de poner, el coeficiente de variación de la estimación de $P = 0{,}3$ es $\frac{0{,}2}{0{,}3} \cdot 100 = 66{,}66$, mientras que, para la estimación de $P = 0{,}7$ es mucho menor; en concreto, es $\frac{0{,}2}{0{,}7} \cdot 100 = 28{,}57$.

Es decir, el error relativo, expresado como coeficiente de variación, en tanto por ciento es:

$$\text{C. V.}(p) = \frac{\left[\frac{N-n}{N-1} \cdot \frac{P \cdot (1-P)}{n}\right]^{1/2}}{P} \cdot 100 \qquad [3.3]$$

Como P no suele ser conocido, el coeficiente de variación se estima por la misma expresión [3.3], sustituyendo P por p.

Bien, pues [3.2] nos dice que el error muestral *absoluto* mínimo es el dado allí, pero a nosotros *nos interesa,* como hemos dicho antes, *el error muestral relativo,* es decir, [3.3], y no es ya cierto que el error relativo mínimo se alcance en $P = 1/2$, sino que crece cuando P decrece y tiende a infinito cuando P tiende a cero (como es lógico, si tenemos un atributo que poseen muy pocas personas de la población, tendremos mucho error al estimarlo a partir de la muestra, dado que habrá poquísimas personas de la muestra que lo tengan). En el Apéndice 1 demuestro matemáticamente que el error relativo crece cuando P decrece y que tiende a infinito cuando P tiende a cero.

Como antes de realizar las entrevistas no conocemos P, ni su estimador p, no podemos dar, mediante una fórmula parecida a [3.1], una estimación *a priori* del error relativo, ni podemos dar una fórmula análoga a [3.2], pues sabemos (Apéndice 1) que el error relativo máximo no existe, sino que tiende a infinito cuando P tiende a cero. Podríamos como mucho dar una tabla de posibles errores relativos para distintos valores de P, 0,1, 0,2, ..., 0,9, por ejemplo, sustituyendo cada uno de ellos en [3.3]. Pero aun así, no conseguiríamos evitar los otros problemas de la estimación *a priori* de los errores muestrales a que hacíamos referencia en la sección anterior.

En la sección 5, donde se da un ejemplo de aplicación del modelo de los Conglomerados Ultimos, incidimos más sobre esto.

Afortunadamente, existen métodos para, basándose en la información recogida en la encuesta, estimar los verdaderos errores muestrales.

Los dos métodos más empleados son el método de los Conglomerados Ultimos, que es el objeto de este artículo, y el método de las Pseudorreiteraciones con Semimuestras (véase Sánchez-Crespo, secciones 11.5 y 11.6). He escogido para este artículo el método de los Conglomerados Ultimos debido a su mayor simplicidad.

4. EL MODELO DE LOS CONGLOMERADOS ULTIMOS

4.1. *Descripción general*

En primer lugar, explicaremos el fundamento del modelo para, después, aplicarlo al caso de un muestreo estratificado.

Se denomina conglomerado último al conjunto de individuos de la muestra que pertenecen a una misma unidad primaria (la definición de unidad primaria depende del modelo de muestreo; más adelante veremos un ejemplo), independientemente de que se realicen una o varias etapas dentro de cada unidad primaria. El submuestreo dentro de cada unidad primaria ha de ser independiente del efectuado en las demás.

El método de los conglomerados últimos sólo requiere que haya dos o más unidades primarias en la muestra y es muy adecuado, por su simplicidad, cuando no se necesitan estimaciones separadas de las contribuciones de las distintas etapas del muestreo a los errores muestrales.

En el caso de muestreo con reposición las fórmulas dadas por el modelo son insesgadas. Si el muestreo es, como suele serlo en la práctica, sin reposición, las fórmulas son sesgadas. Ahora bien, si el tamaño poblacional es grande (lo es en la práctica en la mayoría de los casos), los sesgos son muy reducidos.

Sea θ el parámetro poblacional en estudio (θ puede ser una proporción poblacional, P, por ejemplo, u otro tipo de parámetro). Estimamos θ por

$$\hat{\theta} = \sum_{i=1}^{m} \frac{\hat{\theta}_i}{m} \qquad [4.1]$$

Donde $\hat{\theta}_i$, $i = 1, \ldots, m$ es el estimador de θ en la unidad primaria i-ésima (suponemos que existen m de estas unidades primarias en la muestra).

Su varianza es:

$$V(\hat{\theta}) = V\left(\frac{1}{m} \cdot \sum_{i=1}^{m} \hat{\theta}_i\right) = \frac{E(\hat{\theta}_i - \theta)^2}{m} =$$

$$= \frac{1}{m} \frac{\sum_{i=1}^{N} (\hat{\theta}_i - \theta)^2}{N} \qquad [4.2]$$

Un estimador insesgado suyo será (si el muestreo es con reposición):

$$\hat{V}(\hat{\theta}) = \frac{\sum_{i=1}^{m} (\hat{\theta}_i - \theta)^2}{m \cdot (m-1)} \qquad [4.3]$$

Como decíamos antes, para muestreo sin reposición estas fórmulas tendrán pequeños sesgos.

El error muestral estimado será, pues, la raíz cuadrada de [4.3], dado que [4.3] es la varianza estimada y nosotros queremos la desviación típica.

Vimos en la sección anterior que una forma de normalizar los errores muestrales es expresarlos en la forma de coeficiente de variación, es decir, dividirlos por el valor estimado y multiplicarlos por 100 (esto último sólo para expresarlos en tanto por ciento).

Así, pues, llamando $\hat{D}(\hat{\theta})$ a la raíz cuadrada de [4.3], definimos:

$$\hat{C.V.}(\hat{\theta}) = \frac{\hat{D}(\hat{\theta})}{\hat{\theta}} \cdot 100 \qquad [4.4]$$

En lo sucesivo, supondremos que tenemos un modelo de muestreo estratificado, en el cual los estratos son divisiones territoriales (clásicamente, intersección de región o provincia y tamaño de hábitat). Esta hipótesis no resta generalidad al modelo de los Conglomerados Ultimos; es más, todo el modelo está dado en [4.1], [4.3] y [4.4]. Lo que vamos a ver en el resto del artículo es cómo se aplica el modelo en la práctica a un esquema de muestreo estratificado del tipo que acabamos de describir.

Si tomamos como marco para realizar la muestra el Censo Electoral, podemos definir como unidades primarias las secciones electorales, con lo cual un conglomerado último serían los individuos entrevistados dentro de la misma sección electoral.

Supongamos que deseamos calcular los errores muestrales en que se incurre al estimar un parámetro poblacional dentro de un ámbito territorial cualquiera.

Reduciremos el problema a efectuar la estimación dentro de cada estrato perteneciente a ese ámbito territorial.

Como caso particular, si deseamos estimar el error muestral en un municipio, tomamos todo restringido a ese municipio (en el modelo de muestreo propuesto, un estrato consta de uno o varios municipios).

4.2. *Estimación de errores muestrales en un estrato*

Sea el estrato h, llamamos $\hat{\theta}^h$ a la estimación del parámetro poblacional en estudio dentro del estrato considerado.

Supongamos que existen n^h conglomerados últimos (secciones electorales, en nuestro caso), dentro del estrato h.

Sea $\hat{\theta}_i^h$ la estimación del parámetro poblacional en el conglomerado último i-ésimo del estrato h, $i = 1, \ldots, n^h$.

Será, por [4.1]:

$$\hat{\theta}^h = (1/n^h) \cdot \sum_{i=1}^{n_h} \hat{\theta}_i^h \qquad [4.5]$$

y también, por [4.3]:

$$\hat{V}(\hat{\theta}^h) = \sum_{i=1}^{n_h} \frac{(\hat{\theta}_i^h - \hat{\theta}^h)^2}{n_h \cdot (n_h - 1)} \qquad [4.6]$$

Por tanto, el error muestral, expresado en forma de coeficiente de variación, será, de [4.5], [4.6] y [4.4], llamando $\hat{D}(\hat{\theta}^h)$ a la raíz cuadrada de [4.6]:

$$\hat{C.V.}(\hat{\theta}^h) = \frac{\hat{D}(\hat{\theta}^h)}{\hat{\theta}^h} \cdot 100 \qquad [4.7]$$

4.3. *Estimación de los errores muestrales en un territorio*

Sea un determinado territorio. Si queremos estimar los errores muestrales en este territorio debe de cumplirse una de las siguientes situaciones:

Situación 1: El territorio en cuestión es la unión de uno o varios estratos.

Situación 2: El territorio está contenido completamente dentro de un estrato.

La situación 1 es la más corriente; son casos particulares suyos la estimación de errores muestrales en una comunidad autónoma, en el total nacional, en una provincia (sólo si el diseño muestral lo permite en el sentido de que los estratos estén definidos a partir de las provincias o de las regiones).

La situación 2 sólo se va a presentar en la práctica si se quieren estimar los errores muestrales en un muncipio concreto que no sea el único de su estrato. Este caso es el más sencillo, basta con aplicar las fórmulas [4.5] a [4.7] sin extender los sumatorios a todo el estrato, sino tan sólo al municipio o territorio en cuestión. Supongamos, pues, que estamos en la situación 1, es decir, el territorio es unión de uno o varios estratos.

En este caso, si el territorio es exactamente un estrato, basta con emplear las fórmulas [4.5] a [4.7]; supondremos, pues, que el territorio es unión de más de un estrato.

Sean L de estos estratos los que componen el territorio, entonces será, por la teoría del muestreo estratificado (llamando $\hat{\theta}$ al estimador del parámetro en el territorio en estudio):

$$\hat{\theta} = \sum_{h=1}^{L} W_h \cdot \hat{\theta}^h \qquad [4.8]$$

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^{L} W_h^2 \cdot \hat{V}(\hat{\theta}^h) \qquad [4.9]$$

y, llamando $\hat{D}(\hat{\theta})$ a la raíz cuadrada de [4.9],

$$\hat{C.V.}(\hat{\theta}) = \frac{\hat{D}(\hat{\theta})}{\hat{\theta}} \cdot 100 \qquad [4.10]$$

Donde es

$$W_h = \frac{N_h}{N} \qquad [4.11]$$

N_h es la población del estrato h, y N es la población del territorio en que se desea calcular el error muestral, es decir, W_h es la fracción de población que representa el estrato h respecto del territorio en que se desean estimar los errores muestrales.

Por tanto, para estimar los errores muestrales en este territorio basta con calcular [4.5] y [4.6] para cada estrato y, a partir de ellos y de [4.11], calcular [4.8], [4.9] y [4.10].

Con esto, queda resuelto el problema, al menos en teoría. En la práctica puede ser necesario matizar algunos puntos.

A veces los W_h no son conocidos; esto suele suceder cuando la muestra se diseña para que el proceso se realice con pesos en vez de con factores de elevación. En este caso, si la muestra es autoponderada no hay ningún problema: la fracción de población que hay en cada estrato coincide con la muestral, o sea:

$$W_h = w_h = \frac{n_h}{n} \qquad [4.12]$$

Donde n_h es el tamaño de la muestra en el estrato h y n es el tamaño de la muestra en todo el territorio (no confundir n_h con n^h de [4.5]).

El problema se plantea cuando la muestra no es autoponderada. En este caso, debemos obtener los W_h aplicando [4.11], o recurrir a los pesos.

El camino más sencillo es obtener los W_h; no obstante, haremos referencia al empleo de los pesos.

Se define el peso del estrato h como:

$$P_h = \frac{n \cdot N_h}{N \cdot n_h} \qquad [4.13]$$

Por tanto, conocido P_h, n_h y n, es sencillo despejar de [4.13], teniendo en cuenta [4.11]:

$$W_h = P_h \cdot \frac{n_h}{n} \qquad [4.14]$$

Ahora bien, debe tenerse en cuenta que los P_h de [4.13] son los pesos referidos al territorio en el cual se desean estimar los errores mustrales; por tanto, si, por ejemplo, se estiman los errores muestrales para cada región, *hay que emplear los pesos en esa región,* no en el total nacional.

Las variables numéricas se tratarán en la forma habitual, mientras que las variables categóricas darán lugar al cálculo de errores muestrales para cada categoría de cada variable, excepto las categorías correspondientes al No Sabe, No Contesta y otras semejantes.

Se puede plantear el problema en que, en algunos estratos de la muestra, sólo habrá una unidad primaria y no se podrá, por tanto, estimar en estos estratos los errores muestrales debido a que en el denominador de [4.6] aparecerá un cero.

La solución a este problema es «fundir» estratos, es decir, considerar dos o más estratos como si fueran uno sólo. Con esto, se podrán estimar los errores muestrales.

Naturalmente, los W_h de los estratos fundidos son la suma de los de los estratos que componen la fusión.

De todas formas, en las muestras más usuales no se suelen cargar estratos concretos, sino regiones o provincias. Esto facilita mucho las cosas, dado que, en este caso, dentro de cada región o provincia la muestra es autoponderada y, por tanto, no hay que tener en cuenta peso alguno ni conocer los W_h para obtener las estimaciones de errores muestrales en los estratos de cada región o provincia ni para estimarlos a nivel de esa región o provincia (pero sí para estimarlos a nivel nacional).

5. UN EJEMPLO PRACTICO

A continuación, veremos un ejemplo de aplicación del Modelo de los Conglomerados Ultimos a una encuesta concreta.

La encuesta escogida es el estudio 1750 del CIS.

Se trata de un estudio postelectoral de las elecciones autonómicas catalanas de 1988. En este estudio se realizaron 2.897 entrevistas en toda Cataluña, de las cuales 1.199 se efectuaron en la provincia de Barcelona, 599 en la de Gerona, 499 en la de Lérida y 600 en la de Tarragona.

La muestra no es autoponderada por provincias, aunque sí lo es por estratos dentro de cada provincia. Los estratos son intersección de provincia y de tamaño de hábitat.

Como en todas las encuestas del CIS, la elección de los individuos a entrevistar, en la última fase del muestreo, es decir, dentro de cada sección electoral, se hace cumplimentando unas cuotas de sexo y edad. Esto podría dar lugar a pensar que no es aplicable el modelo de los Conglomerados Ultimos, que se basa en la hipótesis de que tenemos una muestra aleatoria.

Considero que esto no es así, debido a las siguientes causas:

1) El modelo de los Conglomerados Ultimos, que es el que utilizamos en este trabajo para estimar los errores muestrales es correcto si, subjetivamente, cada entrevistador se comporta en la elección de los individuos en forma aleatoria.

2) Aun si el entrevistador comete sesgos, pero éstos tienden a compensarse en cierta forma, la Ley de los Grandes Números haría que el método de estimación siguiera siendo válido en forma aproximada.

Además, como ya dije al principio, no pretendo en este artículo estimar todos los errores de la encuesta, sino sólo la variabilidad de las estimaciones.

Se escogió como pregunta a la que aplicar el método el partido político al cual afirmaban haber votado los encuestados en las elecciones autonómicas catalanas de 1988. Esta es una pregunta filtrada. Se eliminan los casos en que el entrevistado afirma no haber votado o no dice el partido al que votó. En concreto, 1.152 casos son no válidos y 1.747 válidos.

Por tanto, para cada categoría de las respuestas y para cada provincia, así como para el total de Cataluña, se obtuvo un error muestral estimado (previamente se obtuvieron los errores muestrales por estrato, pero aquí no los reflejamos).

El proceso se realizó mediante programación a medida dentro del paquete estadístico SAS, y los resultados fueron:

TABLA 1

Errores muestrales, estimados por el modelo de los Conglomerados Ultimos, del partido al que los entrevistados recuerdan haber votado, expresado como coeficiente de variación, en tanto por ciento (C. V.) y como desviación típica (σ)

	Cataluña	*Barcelona*	*Gerona*	*Lérida*	*Tarragona*	
AP	14,7	18,0	68,0	32,0	27,6	C. V.
	0,005	0,006	0,004	0,014	0,014	σ
CDS	15,7	19,5	40,7	37,4	29,7	C. V.
	0,005	0,006	0,006	0,013	0,014	σ
CiU	*4,0*	*5,3*	*4,5*	*7,5*	*6,7*	C. V.
	0,018	0,023	0,028	0,038	0,034	σ
ERC	16,3	22,6	34,8	21,0	28,3	C. V.
	0,007	0,009	0,013	0,020	0,013	σ
IC/PSUC	10,3	11,7	27,3	28,4	23,1	C. V.
	0,010	0,013	0,014	0,017	0,015	σ
PSC-PSOE	*5,6*	*6,6*	10,4	14,1	11,0	C. V.
	0,018	0,022	0,027	0,033	0,032	σ

Veamos ahora cómo se interpreta la tabla 1.

Para ello, primero debemos seguir algún criterio con el cual decidir cuándo es aceptable un coeficiente de variación.

Si seguimos el criterio del Instituto Nacional de Estadística en algunas de sus publicaciones, tenemos que resulta inaceptable un coeficiente de variación de 10 o superior (o sea, si la desviación típica es igual o superior al 10 por 100 de la media estimada, lo cual parece razonable, porque, de esta forma, suponiendo normalidad en el estimador, sólo nos desviamos más de un 19,6 por 100 de la media en un 5 por 100 de las estimaciones que hagamos).

Por tanto, mirando en la tabla los coeficientes de variación mayores y menores de 10 (en *cursiva,* los menores de 10), vemos que no podemos fiarnos de las estimaciones de recuerdo de voto de AP, el CDS, ERC ni del PSUC, y sí podemos fiarnos (en cuanto a errores muestrales se refiere) de las de CiU. Asimismo, se ve en la tabla que podemos fiarnos también de las del PSOE a nivel de toda Cataluña y de la provincia de Barcelona.

Como también se puede ver mirando a la tabla, parece ser algo más fiable la estimación de CiU en Gerona que en Barcelona, a pesar de que hay más muestra en Barcelona que en Gerona. Esto puede deberse a dos motivos. El primero es que hay más voto a CiU en Gerona que en Barce-

lona, y en el segundo es que, con toda probabilidad, el electorado de CiU en Gerona es más homogéneo que el de Barcelona.

Vemos, pues, que no es sólo el tamaño muestral el que influye en los errores muestrales, sino también la homogeneidad de los encuestados respecto a cada pregunta realizada, el hecho de que la pregunta esté filtrada y no se realice a todos los encuestados, así como otros factores difíciles de precisar en general, más específicos de cada pregunta y encuesta. Pero, eso sí, el modelo de los Conglomerados Ultimos nos estima correctamente los errores muestrales, independientemente de todos estos factores, mientras que la estimación *a priori* de errores muestrales, por las fórmulas del tipo de [3.1], [3.2] y [3.3] no tiene en cuenta más que el tamaño de la muestra.

Como es lógico —y puede comprobarse en la tabla—, los errores muestrales son menores a nivel regional que a nivel provincial, pues el tamaño muestral de la región es la suma de los de las provincias. De la misma forma, los errores muestrales de los estratos son mayores que los de las provincias, aunque no hemos reflejado aquí la tabla de errores muestrales por estratos.

En la tabla 2 damos los tamaños poblacionales y muestrales de la encuesta, así como la estimación de los párametros en estudio.

TABLA 2

Tamaños poblacionales, muestrales teóricos y muestrales reales (eliminando los individuos que no contestan correctamente), y estimadores de los parámetros en estudio de Cataluña y sus cuatro provincias

	Cataluña	*Barcelona*	*Gerona*	*Lérida*	*Tarragona*
N_h	5.978.638	4.614.364	488.342	352.049	523.883
n_h teóricos	2.897	1.199	599	499	600
n_h reales	1.747	745	358	303	341
$\hat{\theta}$ AP	0,0334	0,0338	0,0054	0,0427	0,0504
n_h reales	56	27	2	10	17
$\hat{\theta}$ CDS	0,0314	0,0310	0,0152	0,0342	0,0481
n_h reales	55	25	6	10	14
$\hat{\theta}$ CiU	0,4672	0,4423	0,6329	0,5120	0,5027
n_h reales	911	337	235	171	168
$\hat{\theta}$ ERC	0,0427	0,0389	0,0362	0,0942	0,0471
n_h reales	90	29	13	29	19
$\hat{\theta}$ PSUC	0,0974	0,1088	0,0502	0,0614	0,0644
n_h reales	140	83	17	17	23
$\hat{\theta}$ PSOE	0,3195	0,3360	0,2602	0,2334	0,2873
n_h reales	495	244	85	66	100

Comparando las tablas 1 y 2, se ve que los coeficientes de variación mayores de 10 van asociados siempre a tamaños muestrales menores de 150 entrevistas. Esto sugiere que, en otras preguntas e investigaciones semejantes a la presente, no se utilicen datos basados en menos de este número de entrevistas. Incluso una cota más razonable parece ser la de las 200 entrevistas, pues los coeficientes de variación próximos a 10 ya son dudosos. No obstante, la última palabra la tiene siempre la estimación *a posteriori* de los errores muestrales.

En la siguiente tabla 3 se expresan los errores muestrales absolutos máximos que podrían esperarse *a priori* del estudio (aplicando [3.2]); utilizamos los n_h teóricos (parte A de la tabla 3) y los n_h reales (parte B de la tabla 3). [3.2] supone que el muestreo es aleatorio simple de poblaciones finitas (cosa falsa, pues es estratificado, con subestratificación en algunas ciudades y empleo de conglomerados). Esta misma hipótesis se emplea en las tablas restantes de esta sección.

TABLA 3

Errores muestrales absolutos (expresados en desviación típica) que cabría esperar del estudio si el muestreo fuera aleatorio simple de poblaciones finitas
(La parte A de esta tabla se puede obtener *a priori*)

	Cataluña	*Barcelona*	*Gerona*	*Lérida*	*Tarragona*	
A · n_h teóricos	0,0093	0,0144	0,0204	0,0224	0,0204	σ
B · n_h reales	0,0120	0,0183	0,0264	0,0287	0,0271	σ

Como es obvio, estos errores son los máximos que da [3.2] para cualquier valor de P (que coinciden, como dijimos, con los de $P = 1/2$).

Comparando las tablas 1 y 3, vemos que los errores teóricos de la tabla 3 son superados por la realidad o, mejor dicho, por la estimación de la realidad de la tabla 1, en algunos casos. Por ejemplo, CiU en Barcelona tiene en la tabla 1 un σ de 0,023, mientras que el máximo teórico de la tabla 3 es $\sigma = 0{,}0144$, para la parte A de la tabla y es $\sigma = 0{,}0183$ para la parte B de la tabla.

En resumen, los errores muestrales del PSOE y de CiU quedan subestimados en la tabla 3 en todos los territorios, cuando se supone que la tabla 3 nos da cotas superiores de los mismos y, además, al ser el modelo de muestreo mejor que el aleatorio simple, la cota superior debería cumplirse con

más razón. Vemos que, incluso la parte B de la tabla 3 sigue subestimando los errores muestrales, a pesar de haber empleado los n_h reales (desconocidos *a priori,* o sea, esta parte de la tabla 3 no se podría haber efectuado *a priori*). Este experimento viene a confirmar la necesidad de estimar los errores muestrales *a posteriori* mediante el modelo de los Conglomerados Ultimos u otro semejante.

Vemos que el porcentaje máximo en que la parte A de la tabla 3 subestima a los errores muestrales obtenidos en la tabla 1 es del 93 por 100, en el caso de las estimaciones de CiU y del PSOE para toda Cataluña. En cuanto a la parte B de la tabla 3 (no calculable *a priori*), el porcentaje de subestimación máximo es del 50 por 100, también para CiU y el PSOE en toda Cataluña.

Respecto de la parte A de la tabla 3, que es la que se puede obtener *a priori,* CASI SE DOBLAN, pues, en la práctica, los errores muestrales teóricos.

En la tabla 4 damos los errores muestrales absolutos y relativos, obtenidos mediante [3.1] y [3.3], pero con los p obtenidos tras procesar la encuesta (desconocidos *a priori,* por tanto, es imposible crear esta tabla 4 *a priori*).

TABLA 4

Errores muestrales absolutos (A) y relativos (B) utilizando [3.1] y [3.3], respectivamente, para los valores estimados de P (desconocidos «a priori») y para los n_h reales (también desconocidos «a priori»)

		Cataluña	*Barcelona*	*Gerona*	*Lérida*	*Tarragona*	
AP	A	0,0043	0,0066	0,0039	0,0116	0,0118	σ
	B	12,9	19,6	71,9	27,2	23,5	C. V.
CDS	A	0,0042	0,0064	0,0065	0,0104	0,116	σ
	B	13,3	20,5	42,6	30,5	24,1	C. V.
CiU	A	0,0119	0,0182	0,0255	0,0287	0,0271	σ
	B	2,6	4,1	4,0	5,6	5,4	C. V.
ERC	A	0,0048	0,0071	0,0099	0,0168	0,0115	σ
	B	11,3	18,2	27,3	17,8	24,4	C. V.
PSUC	A	0,0071	0,0114	0,0115	0,0138	0,0133	σ
	B	7,3	10,5	23,0	22,5	20,6	C. V.
PSOE	A	0,0112	0,0173	0,0232	0,0243	0,0245	σ
	B	3,5	5,1	8,9	10,4	8,5	C. V.

Comparando las tablas 1 y 4 vemos que, aun a pesar de haber empleado en la tabla 4 los P estimados y los n_h reales, que son todos ellos valores desconocidos *a priori,* estamos en la tabla 4 infravalorando los errores muestrales, tanto absolutos como relativos; en concreto, de todos los errores relativos de la tabla 4, sólo tres no están infravalorados (AP en Barcelona y el CDS en Barcelona y en Tarragona).

Esta prueba creo que es concluyente de que no podemos fiarnos de [3.1] ni de [3.3], ni siquiera utilizando los P y los n_h correctos (que, insisto, son desconocidos *a priori*).

Finalmente, a modo de ejemplo, en la tabla 5 doy los errores muestrales estimados *a priori* mediante [3.1] y [3.3] para diferentes hipótesis de valores de P. Esta tabla da unas estimaciones, en general, incorrectas de los errores muestrales, pero menos malas que [3.2], que tan sólo proporciona una cota teórica del error muestral absoluto. Aquí, para cada P, tenemos una estimación de los errores muestrales absolutos y relativos.

Creo que los investigadores que, a pesar de lo reflejado en este trabajo, no utilicen el modelo de los Conglomerados Ultimos u otro semejante para estimar los errores muestrales, sea por falta de medios o por otras causas, al menos deberían crear *a priori* una tabla semejante a la tabla 5 que les diera una idea de la variación de los errores relativos teóricos para los distintos valores de P.

TABLA 5

Errores muestrales absolutos (A) y relativos (B) teóricos obtenidos por [3.1] y [3.3] para distintos valores de P. Se utilizan también los n_h teóricos
(Esta tabla se puede obtener *a priori*)

		Cataluña	*Barcelona*	*Gerona*	*Lérida*	*Tarragona*
$P = 0{,}01$	A	0,0018	0,0029	0,0041	0,0045	0,0041
	B	18,5	28,7	40,7	44,5	40,6
$P = 0{,}05$	A	0,0040	0,0063	0,0089	0,0097	0,0089
	B	8,1	12,6	17,8	19,5	17,8
$P = 0{,}1$	A	0,0056	0,0087	0,0123	0,0134	0,0122
	B	5,6	8,7	12,3	13,4	12,2
$P = 0{,}2$	A	0,0074	0,0116	0,0163	0,0179	0,0163
	B	3,7	5,8	8,2	8,9	8,2
$P = 0{,}3$	A	0,0085	0,0132	0,0187	0,0205	0,0187
	B	2,8	4,4	6,2	6,8	6,2
$P = 0{,}4$	A	0,0091	0,0141	0,0200	0,0219	0,0200
	B	2,3	3,5	5,0	5,5	5,0
$P = 0{,}5$	A	0,0093	0,0144	0,0204	0,0224	0,0204
	B	1,9	2,9	4,1	4,5	4,1

		Cataluña	*Barcelona*	*Gerona*	*Lérida*	*Tarragona*
$P = 0,6$	A	0,0091	0,0141	0,0200	0,0219	0,0200
	B	1,5	2,4	3,3	3,7	3,3
$P = 0,7$	A	0,0085	0,0132	0,0187	0,0205	0,0187
	B	1,2	1,9	2,7	2,9	2,7
$P = 0,8$	A	0,0074	0,0116	0,0163	0,0179	0,0163
	B	0,9	1,4	2,0	2,2	2,0
$P = 0,9$	A	0,0056	0,0087	0,0123	0,0134	0,0122
	B	0,6	1,0	1,4	1,5	1,4
$P = 0,95$	A	0,0040	0,0063	0,0089	0,0097	0,0089
	B	0,4	0,7	0,9	1,0	0,9
$P = 0,99$	A	0,0018	0,0029	0,0041	0,0045	0,0041
	B	0,2	0,3	0,4	0,4	0,4

De la tabla 5 se puede comprobar cómo los errores muestrales absolutos teóricos son máximos en $P = 0,5$ (como dijimos en la sección 3), pero los relativos crecen al decrecer 0 (como también dijimos en la sección 3). Sabemos también que tienden a infinito al tender P a 0.

A pesar de que sabemos que estamos infravalorando los verdaderos errores muestrales, esta tabla 5 nos dice, por ejemplo, que no podemos fiarnos de ninguna estimación de un parámetro P que valga 0,01; también nos dice que no podemos fiarnos, a nivel provincial, de ninguna estimación de un $P = 0,05$ (coeficientes de variación mayores de 10 en la tabla 5). Supone, pues, una notable mejora sobre [3.2], como ya hemos dicho antes.

6. CONCLUSION

Considero que las tablas de errores muestrales como la tabla 1 de la sección 5 son un instrumento esencial para poder obtener conclusiones de una encuesta, pues nos dicen de qué datos no podemos fiarnos y, por tanto, si obtenemos conclusiones a partir de ellos, esto es bajo nuestra entera responsabilidad.

Finalmente, decir que el hecho de que las estimaciones de los errores muestrales sean pequeñas no garantiza la exactitud de nuestras conclusiones, pues todavía existen los errores ajenos al muestreo, que pueden producir sesgos notables (desviaciones sistemáticas de los verdaderos valores que queremos estimar). Ahora bien, los errores muestrales constituyen un buen test, es decir, si son grandes sabemos que no podemos tener confianza en las estimaciones (aunque si son pequeños, aun así no podemos tener plena confianza en ellas. Para tenerla, habría que estimar los errores ajenos al muestreo).

APENDICE 1

Demostraciones

En primer lugar, veamos que [3.2] es cierto, o sea, que el máximo de [3.1], para P entre 0 y 1 se alcanza en $P = 1/2$.

En efecto, para $P = 0$ y para $P = 1$, [3.1] vale 0 y es positiva entre 0 y 1; basta, pues, con derivar [3.1] respecto de P e igualar a cero la derivada. Tenemos (llamando C a la constante, para N y n fijos, que aparece en [3.1]):

$$C \cdot \frac{1 - 2 \cdot P}{2 : \sqrt{P \cdot (1 - P)}} = 0, \text{ luego } P = 1/2, \text{ c.q.d.}$$

Para comprobar que es máximo basta con hallar la segunda derivada y ver que es menor que cero en $P = 1/2$.

Ahora veremos que [3.3] tiende a infinito cuando P tiende a cero. En efecto, N y n son constantes para cada muestra dada; por tanto, los ignoramos, queda:

$$\lim_{P \to 0} \frac{\sqrt{P \cdot (1 - P)}}{P} = \overset{\text{L'Hopital}}{\lim_{P \to 0} \frac{1 - 2 \cdot P}{\sqrt{P \cdot (1 - P)} \cdot 2}} = \infty \text{ c.q.d.}$$

Finalmente, veremos que el coeficiente de variación, según [3.3] es función monótona decreciente de P, para $P \in [0,1]$.

Basta con derivar en [3.3], tenemos, llamando C a la misma constante que un poco más arriba que [3.3] queda:

$$C \cdot 100 \cdot \frac{\sqrt{P \cdot (1 - P)}}{P}, \text{ si derivamos respecto de } P, \text{ es:}$$

$$\frac{-1}{2 \cdot P^2 \cdot \sqrt{P \cdot (1 - P)}} < 0 \text{ si } P \text{ está entre cero y uno.}$$

Luego [3.3] decrece entre 0 y 1.

BIBLIOGRAFIA

COCHRAN, W. G. (1974): *Técnicas de Muestreo,* Compañía Editorial Continental, S. A., 4.ª reimpresión, México.

SÁNCHEZ-CRESPO, J. L. (1980): *Curso Intensivo de Muestreo en Poblaciones Finitas,* INE, 2.ª ed., Madrid.

CRITICA DE LIBROS