Estadística Española · Volumen 57, número 186 / 2015, pp. 29-37

Estimación insesgada objetiva para no respuesta*

Mariano Ruiz Espejo
Departamento de Matemáticas Fundamentales
UNED-Madrid

Resumen

En este artículo consideramos un estimador de Bouza-Herrera (2013) en el caso de no respuesta. Justificamos que este estimador es condicionalmente sesgado para estimar la media poblacional. Obtenemos su sesgo y su varianza condicionales. Proponemos otro estimador insesgado en las mismas condiciones, calculamos su varianza y damos un estimador insesgado de la varianza del estimador propuesto. Comentamos el interés científico de los resultados en las encuestas.

Palabras clave: Estimación insesgada objetiva, Media poblacional, No respuesta.

Clasificación AMS: 62D05.

Objective unbiased estimation for nonresponse

Abstract

In this article, we consider an estimator by Bouza-Herrera (2013) in the case of nonresponse. We justify that this estimator is conditionally biased for estimating the population mean. We obtain its conditional bias and variance. We propose other unbiased estimator in the same conditions, calculate its variance and give an unbiased variance estimator of the proposed estimator. We comment the scientific interest of the results in surveys.

Keywords: Objective unbiased estimation, Population mean, Nonresponse.

AMS classification: 62D05.

* Reconozco los comentarios anónimos del evaluador del artículo que han mejorado la presentación final del mismo.

1. Introducción

El problema de no respuesta fue tratado inicialmente por Hansen y Hurwitz (1946). Estos autores dieron una solución al problema que se presenta cuando algunas de las personas encuestadas no responden el dato de la variable de interés que tratamos de estudiar por muestreo. La idea básicamente consiste en que la población queda dividida en dos estratos, uno de respuesta y otro de no respuesta. Al hacer las preguntas a los encuestados sabremos si responden o si no responden, y por tanto sabremos a qué estrato pertenece cada unidad encuestada. Consecuentemente, conoceremos los tamaños relativos de los estratos en la muestra aleatoria seleccionada, y así se construye un estimador insesgado de la media poblacional basado en la información de los tamaños relativos de respuesta y no respuesta de la muestra, de las respuestas de la muestra, y de información adicional de una submuestra de la muestra del estrato de no respuesta que no informaron su dato de la variable de interés pero que puede ser obtenido por medios más cuidadosos en una segunda fase. Así, y como refiere Cochran (1977), se resolvió el problema de la estimación insesgada de media poblacional.

Sin embargo el problema de la estimación insesgada de la varianza de este estimador insesgado de Hansen y Hurwitz (1946) ha sido resuelto recientemente por Ruiz Espejo (2011, 2013b) en el caso de usar diseño de muestreo aleatorio simple con reemplazamiento, y por Thompson (2012) en el caso de usar diseño de muestreo aleatorio simple sin reemplazamiento.

El hecho de poder estimar sin sesgo la varianza de un estimador insesgado confiere la posibilidad de estimar sin sesgo su error cuadrático medio, e incluso la de dar estimadores por intervalo aproximados del parámetro media poblacional, y de contrastar hipótesis sobre el verdadero valor de la media poblacional que tratamos de estimar. Para entender con mayor detalle estas afirmaciones sugiero los libros de Ruiz Espejo (2013a, 2014).

Otra forma de abordar la resolución del problema de no respuesta ha sido propuesta por Bouza-Herrera (2013, p. 61). Consideramos que seleccionamos una muestra aleatoria simple con reemplazamiento de tamaño *n* de una población finita de tamaño *N*. Nuestro objetivo es estimar la media poblacional de una variable de interés *y*, de modo que si la unidad de la población finita numerada es la *i*-ésima ($i = 1, 2, \dots, N$), la variable de interés tiene en dicha unidad el dato y_i. Entonces la media poblacional es

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

El estimador usual de este parámetro en muestreo aleatorio simple con reemplazamiento es la media muestral, es decir la media aritmética de la variable de interés de las unidades seleccionadas en la muestra de tamaño *n*, es decir

$$\bar{y}_s = \frac{1}{n}\sum_{j=1}^{n} y_{i_j},$$

donde el subíndice i_j indica la unidad de la población finita (unidad identificada comprendida entre los números enteros 1 y *N*) seleccionada en la *j*-ésima selección en la muestra aleatoria simple con reemplazamiento *s* de tamaño *n*, es decir, $j = 1, 2, \dots, n$. Esta media muestral es insesgada para estimar la media poblacional con el diseño de muestreo aleatorio simple con reemplazamiento de tamaño fijo *n* (Ruiz Espejo, 2013a).

Para resolver el problema de estimación insesgada de la media poblacional con no respuesta, Bouza-Herrera (2013) supone que se dispone de una media muestral de una muestra piloto en la que no se presenta la no respuesta. Esto puede ocurrir cuando se procura obtener la muestra con cuidado esmero, similar por ejemplo a cuando se selecciona la submuestra del estrato de no respuesta en el estimador propuesto por Hansen y Hurwitz (1946). Llamemos $\bar{y}_{ps}$ a la media muestral de la muestra piloto obtenida por muestreo aleatorio simple con reemplazamiento de tamaño fijo *m*.

Al obtener la muestra aleatoria simple con reemplazamiento de tamaño *n* en el estudio y exponerla a la posible no respuesta, sabemos que la media muestral de las respuestas tiene por esperanza matemática la "media del estrato de respuesta" que denotamos por $\bar{y}_1$ y que en general será distinta de la media poblacional $\bar{y}$. Por tanto, la media muestral de las respuestas tiene un sesgo conocido igual a $W_2(\bar{y}_1 - \bar{y}_2)$ donde W_2 es el tamaño relativo del estrato de no respuesta, e $\bar{y}_2$ es la media del estrato de no respuesta.

Para eliminar este sesgo, Bouza-Herrera (2013), admite que de las *n* unidades seleccionadas en la muestra hay *k* respuestas, las de las unidades $i_1, i_2, \dots, i_k$, de las que se obtiene respuesta. También hay otras *n - k* unidades que no responden y a las que podría sustituir el verdadero valor de su variable de interés por la media muestral piloto de tamaño *m* que es independiente de las demás selecciones de muestra. De este modo el estimador propuesto por este autor en su monografía es

$$t = \frac{1}{n}\sum_{j=1}^{n} y_{i_j}^{*},$$

donde $y_{i_j}^{*} = y_{i_j}$ si hay respuesta en la unidad i_j , y también $y_{i_j}^{*} = \bar{y}_{ps}$ es decir se iguala a la media de la muestra piloto de tamaño *m* cuando la unidad i_j no responde.

2. Sesgo del estimador

Como las unidades que responden en la muestra tienen la esperanza matemática en su estrato de respuesta de su variable de interés, $\bar{y}_1$, al ser todas unidades que responden. En cambio la esperanza matemática de la media muestral de la muestra piloto coincide con la media poblacional $\bar{y}$. Entonces, la esperanza condicional (a haberse obtenido *k* respuestas en la muestra de tamaño fijo *n*) es

$$E(t|k) = \frac{1}{n}\left[\sum_{j=1}^{k} E\left(y_{i_j}\right) + (n-k)E\left(\bar{y}_{ps}\right)\right]$$

$$= \frac{1}{n}[k\bar{y}_1 + (n-k)\bar{y}]$$

$$= \bar{y} + \frac{k}{n}(\bar{y}_1 - \bar{y}).$$

Por tanto, el sesgo condicional de este estimador *t* para estimar la media poblacional $\bar{y}$ es

$$B(t|k) = E(t|k) - \bar{y} = \frac{k}{n}(\bar{y}_1 - \bar{y}).$$

Y, como consecuencia, el sesgo incondicional del estimador *t* es

$$B(t) = E[B(t|k)] = W_1(\bar{y}_1 - \bar{y}) = W_1W_2(\bar{y}_1 - \bar{y}_2).$$

Este sesgo tiene por causa haber usado la 'media muestral piloto' en lugar de una 'media muestral en el segundo estrato' dentro del estimador propuesto por Bouza-Herrera (2013).

3. Varianza condicional del estimador

La varianza condicional del estimador *t* puede obtenerse de modo similar a como hemos razonado para obtener su esperanza matemática. Ahora,

$$V(t|k) = \frac{kV(y) + (n-k)^2V\left(\bar{y}_{ps}\right)}{n^2}$$

$$= \frac{k\sigma_1^2 + (n-k)^2\frac{\sigma^2}{m}}{n^2}.$$

Hemos denotado aquí por *y* a la variable de interés en el estrato de respuesta, que tiene por varianza la varianza de dicho estrato, que denotamos σ_1^2. También hemos denotado por σ^2 a la varianza de la variable de interés en la población finita. Por tanto el error cuadrático medio condicional del estimador *t* de Bouza-Herrera es

$$ECM(t|k) = V(t|k) + [B(t|k)]^2$$

$$= \frac{k\sigma_1^2 + (n-k)^2\frac{\sigma^2}{m}}{n^2} + \left[\frac{k}{n}(\bar{y}_1 - \bar{y})\right]^2.$$

4. Corrección del estimador

Proponemos a continuación un estimador *t** que corrige el sesgo del estimador de Bouza-Herrera (2013). Concretamente, sea el estimador

$$t^* = t - \hat{B}(t|k) = t - \frac{k}{n}(\bar{y}_{1s} - \bar{y}_{ps}),$$

en donde hemos denotado por $\bar{y}_{1s}$ a la media muestral de las unidades que responden en la muestra del estudio, es decir

$$\bar{y}_{1s} = \frac{1}{k}\sum_{j=1}^{k} y_{i_j}.$$

Este estimador *t** sería condicionalmente insesgado para estimar la media poblacional, ya que las esperanzas matemáticas de $\bar{y}_{1s}$ e $\bar{y}_{ps}$ son respectivamente $\bar{y}_1$ e $\bar{y}$. Pero fácilmente observamos que el estimador insesgado de la media poblacional se reduciría a la media muestral de la muestra piloto de tamaño fijo m, $t^* = \bar{y}_{ps}$, lo que hace inaprovechable las *k* respuestas obtenidas en el estudio a efecto de proponer un estimador insesgado de la media poblacional por este método.

Así pues, otro estimador posible que aproveche al mismo tiempo la información de la muestra piloto y de las respuestas del estudio, sería el estimador siguiente

$$t^{**} = g\bar{y}_{ps} + (1-g)\bar{y}_{nr}$$

en donde g es una constante por determinar, $\bar{y}_{ps}$ la media muestral de la muestra piloto, e $\bar{y}_{nr}$ es el estimador tradicional para no respuesta de Hansen y Hurwitz (1946)

$$\bar{y}_{nr} = w_1\bar{y}_{1s} + w_2\bar{y}_{(2)s}$$

donde $w_1 = n_1/n$ y $w_2 = n_2/n$ son los tamaños relativos de los estratos estimados en el estudio, siendo $n_1 = k$ y $n_2 = n - k$. Además, $\bar{y}_{1s}$ es la media muestral de las respuestas del estudio, e $\bar{y}_{(2)s}$ es la media muestral de las respuestas obtenidas de una submuestra de tamaño fijo $n_{(2)}$ de la muestra de no respuestas del estudio que tuvo un tamaño muestral $n_2 = n - k$.

El estimador t^{**} es insesgado para estimar la media poblacional $\bar{y}$, puesto que tanto $\bar{y}_{ps}$ como $\bar{y}_{nr}$ son insesgados para estimar $\bar{y}$.

5. Varianza del estimador

Su varianza es

$$V(t^{**}) = g^2 V\left(\bar{y}_{ps}\right) + (1-g)^2 V(\bar{y}_{nr})$$
$$= g^2 \frac{\sigma^2}{m} + (1-g)^2 \left[\frac{\sigma^2}{n} + \frac{(n-1)W_2^2\sigma_2^2}{nn_{(2)}}\right],$$

debido a Ruiz Espejo (2013a, p. 178) y a que las muestras piloto y de estudio son independientes. Hemos denotado por σ_2^2 a la varianza de la variable *y* en el 'estrato de no respuesta' cuyo tamaño relativo es W_2. Esta varianza se minimiza de la ecuación

$$\frac{dV(t^{**})}{dg} = 0.$$

Es decir, cuando

$$g = \frac{\frac{\sigma^2}{n} + \frac{(n-1)W_2^2\sigma_2^2}{nn_{(2)}}}{\frac{\sigma^2}{m} + \frac{\sigma^2}{n} + \frac{(n-1)W_2^2\sigma_2^2}{nn_{(2)}}}.$$

Este valor de g es en general una constante desconocida perteneciente al intervalo abierto (0, 1). Además se comprueba que

$$\frac{d^2V(t^{**})}{dg^2} = 2\left[\frac{\sigma^2}{m} + \frac{\sigma^2}{n} + \frac{(n-1)W_2^2\sigma_2^2}{nn_{(2)}}\right] > 0,$$

por lo que el valor de g obtenido anteriormente es el mínimo de la función $V(t^{**})$. Un valor de referencia para esta constante, que aunque no fuera mínimo sí sería práctico, es

$$g = \frac{m}{m+n}$$

pues atribuye a cada estimador $\bar{y}_{ps}$ e $\bar{y}_{nr}$ un peso proporcional al tamaño muestral fijo de partida en cada caso, es decir proporcional a *m* y a *n*.

6. Estimación insesgada de la varianza

Un estimador insesgado de la varianza del estimador t^{**} puede proponerse a partir de los estimadores insesgados de la varianza de los estimadores $\bar{y}_{ps}$ e $\bar{y}_{nr}$. Concretamente del modo siguiente

$$\hat{V}(t^{**}) = g^2\hat{V}(\bar{y}_{ps}) + (1-g)^2\hat{V}(\bar{y}_{nr}).$$

Aquí,

$$\hat{V}(\bar{y}_{ps}) = \frac{s_{ps}^2}{m},$$

donde s_{ps}^2 es la cuasivarianza muestral de la muestra piloto de tamaño fijo *m*, y (Ruiz Espejo, 2013a, p. 179)

$$\hat{V}(\bar{y}_{nr}) = w_1\frac{s_1^2}{n} + \widehat{\sigma_2^2}\left(\frac{w_2^2}{n_{(2)}} - \frac{w_2}{nn_{(2)}} + \frac{w_2}{n}\right) +$$

$$\frac{w_1w_2}{n-1}\left[\bar{y}_{1s}^2 - \frac{s_1^2}{n_1} + \bar{y}_{(2)s}^2 - \hat{V}(\bar{y}_{(2)s}) - 2\bar{y}_{1s}\bar{y}_{(2)s}\right],$$

donde

$$\widehat{\sigma_2^2} = \widehat{s_2^2} = \frac{n_2}{n_2-1}\widehat{\sigma_{(2)}^2} = \frac{n_2}{n_2-1}s_{(2)}^2,$$

siendo s_1^2 la cuasivarianza muestral de tamaño fijo n_1 en el primer estrato de respuesta del estudio, y $s_{(2)}^2$ la cuasivarianza muestral de tamaño $n_{(2)}$ en la submuestra en el segundo estrato o de no respuesta en el primer intento. Como

$$V(\bar{y}_{(2)s}) = V_1E_2E_3(\bar{y}_{(2)s}) + E_1V_2E_3(\bar{y}_{(2)s}) + E_1E_2V_3(\bar{y}_{(2)s}) =$$

$$V_1(\bar{y}_2) + E_1V_2(\bar{y}_{2s}) + E_1E_2\left(\frac{\sigma_{(2)}^2}{n_{(2)}}\right) = E_1\left(\frac{\sigma_2^2}{n_2}\right) + E_1E_2\left(\frac{\sigma_{(2)}^2}{n_{(2)}}\right),$$

entonces un estimador insesgado de esta varianza es

$$\hat{V}(\bar{y}_{(2)s}) = \frac{\widehat{\sigma_2^2}}{n_2} + \frac{s_{(2)}^2}{n_{(2)}} = s_{(2)}^2\left[\frac{1}{n_2-1} + \frac{1}{n_{(2)}}\right].$$

7. Conclusión

Con todo lo anterior hemos corregido el estimador de Bouza-Herrera (2013) para asegurar su insesgación en las condiciones generales en que se estudia la no respuesta en encuestas, y hemos calculado el error cuadrático medio condicional de su estimador. Además hemos hecho posible que su idea sea práctica en este contexto proponiendo una clase de estimadores que aprovechan en parte su idea y dando las condiciones prácticas para su uso como es proporcionar un estimador insesgado de su varianza, lo que completa su estudio a efectos de la inferencia basada en el uso de una muestra piloto (independiente al estudio) en la que no ha habido falta de respuesta.

Para concluir podemos afirmar que este es un método objetivo de estimación insesgada de la media poblacional y de estimación insesgada de su varianza en presencia de no respuesta alternativo al propuesto por Hansen y Hurwitz (1946) que ha sido perfeccionado con el estudio de su varianza y su estimación insesgada (Ruiz Espejo, 2013b).

Desde un punto de vista de la práctica de la estadística oficial, la realización de la muestra piloto no siempre se hace debido a las limitaciones en los recursos o de tiempo. Cuando se realiza la muestra piloto en la estadística oficial, la muestra es demasiado pequeña e incluso se selecciona por métodos no probabilísticos pues la intención del ensayo no suele ser hacer inferencias con precisión y por sistema, sino que va dirigida a testar la utilidad de los instrumentos de medida (sobre todo los cuestionarios) y la organización de los métodos de recogida de la información.

Que no haya falta de respuesta en el ensayo piloto es algo que no se suele cumplir ni siquiera aproximadamente en la práctica oficial. También, por motivo de falta de recursos y de plazos temporales, no se suele pretender en la estadística oficial obtener respuestas en el segundo estrato, lo que limita la utilidad práctica tanto del estimador de Hansen y Hurwitz (1946), como los de Ruiz Espejo (2011), Thompson (2012), Bouza-Herrera (2013) y el propuesto en este artículo. Pero desde un punto de vista de su interés científico sí es útil, al menos potencialmente, en estudios sociológicos, e incluso psicológicos y biomédicos en los que no se presenten restricciones drásticas en recursos o de plazos de tiempo.

En la práctica, aunque hemos llamado muestra piloto a la muestra en la que no se presenta no respuesta, ésta puede ser obtenida con los mismos protocolos con que se obtiene la submuestra en el estrato de no respuesta en el estudio, y no es necesario que se obtenga con anterioridad al estudio, sino que puede ser simultánea en el tiempo, anterior o posterior al estudio.

Otras alternativas como los métodos de calibrado con información auxiliar para tratar la no respuesta tienen el inconveniente de que no dan lugar por lo general a estimaciones insesgadas ni para la media poblacional ni para el error cuadrático medio del estimador, lo que hace que la inferencia no sea objetiva sino solo aproximada bajo ciertas hipótesis no ‘seguras o absolutamente ciertas’. Esta objeción no se da en los estimadores de Ruiz Espejo (2011), Thompson (2012) y el propuesto en este trabajo.

Referencias

BOUZA-HERRERA, CARLOS N. (2013). «Handling Missing Data in Ranked Set Sampling». Heidelberg. Springer.

COCHRAN, WILLIAM G. (1977). «Sampling Techniques», 3ª edición. Nueva York. Wiley.

HANSEN, MORRIS H. Y HURWITZ, WILLIAM N. (1946). «The problem of nonresponse in sample surveys». *Journal of the American Statistical Association* 41, 517-529.

RUIZ ESPEJO, MARIANO (2011). «An objective solution to the problem of unbiased estimation with nonresponse». *Statistical Reports* 13, 1-2.

RUIZ ESPEJO, MARIANO (2013a). «Exactitud de la Inferencia en Poblaciones Finitas», 1ª edición. Madrid. Bubok.

RUIZ ESPEJO, MARIANO (2013b). «Objective unbiased variance estimation with nonresponse: a review». *Statistical Reports* 18, 1-10.

RUIZ ESPEJO, MARIANO (2014). «Fundamentos de la Inferencia Estadística Objetiva», 3ª edición. Raleigh, NC. Lulu Press.

THOMPSON, STEVEN K. (2012). «Sampling», 3ª edición. Hoboken, NJ. Wiley.