TRABAJOS DE ESTADISTICA Y
DE INVESTIGACION OPERATIVA
Vol. 32, Núm. 2, 1981, pp. 30 a 36

SOBRE EL PROBLEMA DE LA FRACCION DE SUBMUESTREO PARA EL CASO DE LAS NO RESPUESTAS

C. N. Bouza
Universidad de La Habana

1. Introducción

Usualmente se utiliza el muestreo simple aleatorio sin reemplazo como diseño muestral para obtener una muestra s de una población finita U. La estimación de la media poblacional $\bar{X}$ es obtenible al evaluar la la característica de interés X en los individuos muestreados y calcular

$$\bar{x} = \sum_{u \in s} X(u)/n$$

en la práctica no se obtiene respuesta de todos los muestreados.

Esto implica que U está particionado en dos estratos dados por

$$U_1 = (u \in U: \text{ responden en la primera visita})$$

$$U_2 = (u \in U: \text{ no responden en la primera visita})$$

Ellos se asocian a las ponderaciones usuales $W_i = N_i/N$, $i = 1, 2$, en

las que N_i es el número de individuos en U_i y N es el tamaño de U.

El muestrista al construir su estrategia supone que $N_2 = 0$. Al evidenciarse que esto no es cierto, algunas transformaciones deben realizarse para estimar $\bar{X}$. Una solución lógica es extraer una submuestra s_2' de s_2 al definir

$$s_i = (u \in s : u \in U_i), \qquad i = 1, 2$$

El problema que se debe solucionar ahora es el de determinar una fracción de submuestreo adecuada para fijar n_2' como el tamaño de s_2'.

Suponiendo que todos los individuos reentrevistados dan su respuesta un estimador insesgado de $\bar{X}$, ver Cochran (1963), es

$$\bar{x}' = w_1 \bar{x}_1 + w_2 \bar{x}_2' \tag{1}$$

donde $w_i = n_i/n$, $i = 1, 2$, es la ponderación dada la muestra observada al tomar n_i como el tamaño de la muestra s_i y n como el de s. Las medias muestrales envueltas en esta fórmula están dadas por

$$\bar{x}_1 = \sum_{u \in s_1} X(u)/n_1$$

$$\bar{x}_2' = \sum_{u \in s_2'} X(u)/n_2'$$

La varianza de (1) al fijar s, tiene la expresión

$$V = V(\bar{x}' : s) = ((N - n)/nN)\, S^2 + w_2^2\, ((n_2 - n_2')/n_2'\, n_2)\, S_2^2 \tag{2}$$

en la que

$$S_2^2 = \sum_{u \in U_2} (X(u) - \bar{X}_2)^2/(N_2 - 1)$$

siendo $\bar{X}_2$ la media poblacional del estrato U_2.

Una función de costo para esta estrategia es

$$C = c_0\, n + c_1\, n + c_2\, n_2' \tag{3}$$

Es de destacar que la varianza y el costo expresados por (2) y (3) respectivamente son variables aleatorias. De ahí la necesidad de utilizar sus esperanzas para evaluar el comportamiento de la precisión y el costo de una cierta regla de submuestreo.

Para obtener n_2' se utiliza una regla de submuestreo. Las más conocidas son las de Hansen-Hurwitz (1946) y la de Srinath (1971). Ambas dependen de un parámetro arbitrariamente fijado por el estadístico. En este trabajo se propone una regla que no requiere de la definición de parámetros. En ella n_2' depende solo de los resultados de la muestra observada. La comparación de las tres reglas es desarrollada, utilizando los errores y costos esperados asociados a ellas.

2. Reglas de Hansen-Hurwitz y de Srinath

La varianza de $\bar{x}'$, dada por (2), depende de las variables aleatorias n_2 y n_2'. Por su parte la función de costo (3) está supeditada a n_2'. De ahí que sea necesario utilizar las esperanzas de ambas funciones para evaluar la precisión y el costo de una regla de submuestreo.

Hansen y Hurwitz (1946) propusieron utilizar como regla para obtener el tamaño de la submuestra s_2', la relación siguiente

$$n_2' = n_2/K \tag{4}$$

el muestrista, debe fijar el valor del parámetro K como mayor que uno.

Sustituyendo la relación anterior en (2) se obtiene que la varianza condicionada a s es

$$V = ((N-n)/nN)\, S^2 + (n_2\ (K-1)/n^2)\, S_2^2$$

cuya esperanza

$$E\ (V) = ((N-n)/nN)\, S^2 + (W_2\ (K-1)/n)\, S_2^2 = V_H \tag{5}$$

sigue dependiendo del parámetro K.

Analizando el promedio teórico del costo dado por la función dada por (3) al sustituir, previamente la expresión (4) se tiene que

$$E(C) = c_0 + c_1 n + n W_2/K = C_H \tag{6}$$

la que es también función de K.

La regla de Srinath (1971) establece que el tamaño de la submuestra es calculable mediante

$$n_2' = n_2^2/(Hn + n_2) \tag{7}$$

En este caso el parámetro H tiene como intervalo de definición $]0, \infty[$. Su uso determina que la varianza tenga la forma

$$V = ((N - n)/nN)\, S^2 + H S_2^2/n = V_S \tag{8}$$

Note que al afijar el tamaño de s_2' mediante esta regla la aleatoriedad de la varianza desaparece al depender sólo del parámetro H y no de n_2. Sin embargo, para obtener el costo esperado es necesario que se cumpla la relación

$$E(n_2 - E(n_2))^t = 0$$

para todo $t \geqslant 2$. En este caso

$$E(C) \doteq c_0 n + c_1 n + c_2 n W_2/(H + W_2) = C_S \tag{9}$$

es la esperanza del costo.

3. Regla propuesta

Como se ha visto en las dos reglas anteriores los parámetros K y H juegan un papel importante en la determinación de n_2' y en la de los costos y variancias promedio.

Una regla en la que se puede obviar esta situación, debe basarse en

los resultados de la muestra s. Se propone utilizar

$$n_2' = n_2 \, w_2 \tag{10}$$

donde w_2 es la ponderación de s_2 en la muestra observada definida en la sección 1.

Utilizando esta relación, se obtiene que el coeficiente de S_2^2 en (2) es de la forma

$$w_2^2 \, (n_2 - n_2')/n_2' \, n_2 = (n - n_2)/n^2 = n_1/n^2$$

por lo que la varianza esperada bajo esta regla es

$$V_B = ((N-n)/nN) \, S^2 + W_1 \, S_2^2/n \tag{11}$$

La obtención del costo esperado depende de $E \, (n_2^2)$ pues

$$C = c_0 + c_1 n + c_2 n_2^2/n$$

es la expresión de la función de costo bajo esta regla.

Dado el uso de muestreo simple aleatorio sin reemplazo n_2 es una variable con distribución hipergeométrica. Usando este hecho y la relación

$$E \, (n_2^2) = V \, (n_2) + (E \, (n_2))^2$$

se obtiene que

$$E \, (C) = c_0 + c_1 n + c_2 \, ((N - n)/n \, (N - 1) \, W_1 \, W_2 + (nW_2)^2)$$

Cuando U es suficientemente grande para aceptar que $N - n \doteq N - 1$ esta expresión se simplifica un poco, quedando

$$C_B \doteq c_0 + c_1 n + c_2 W_2 \, (W_1 + n^2 W_2)/n \tag{12}$$

Al aceptar esta aproximación también se tiene que $(N - n)/N \doteq 1$, por lo que las varianzas esperadas también se simplifican al hacerlo su

primer término.

Como se nota, las funciones evaluativas de la precisión y el costo del muestreo asociadas a esta regla, no dependen de constantes inducidas por el estadístico.

4. Comparación de las reglas

Para hacer una valoración de las distintas reglas, se compararán los comportamientos de sus errores y costos esperados.

Tomando en cuenta la precisión de las reglas de Hansen-Hurwitz y de Srinath, se preferirá la primera si

$$K < (nH/W_2) + 1$$

Como K es un parámetro mayor que uno, esta relación no es satisfecha. Al compararla con la regla propuesta, la de Hansen-Hurwitz será preferida cuando

$$K < (W_1/W_2 n) + 1$$

lo que tampoco se cumple. De ahí que las reglas de Srinath y la desarrollada en este trabajo sean más precisas. Al comparar éstas, el criterio de preferencia está dado porque

$$H < W_1$$

observando (7) y recordando que $H > 0$, se tiene que la regla de Srinath es más precisa si se toma una submuestra grande.

Analizando el costo, se tiene que es más barato utilizar la regla de Hansen-Hurwitz que la de Srinath si se cumple la desigualdad

$$H > K - W_2$$

Utilizando el criterio que determina la preferencia de esta última regla con respecto a la propuesta al utilizar el error esperado, se tiene que debe satisfacerse

$$H < W_1 < K - W_2$$

Esto implica que su aplicación determina mayores errores que la regla propuesta, si es más barata que la de Hansen-Hurwitz. Por otra parte, si

$$1 < K\,((W_1/N^2) + W_2)$$

esta regla es más barata que la desarrollada en este trabajo pero esto conspira contra su precisión, pues K debe ser grande. Una comparación similar con la de Srinath arroja que deberá preferirse ésta cuando

$$1 > ((H/W_2) + 1)\ \ ((W_1/n^2 W_2) + 1)$$

la que no se satisface ni en el caso extremo $W_1 = H = 0$.

Puede notarse la existencia de dificultades diversas al buscar normas para fijar H o K para garantizar estimaciones precisas y bajos costos. La regla propuesta da una solución a este problema obteniendo aceptables niveles para ambos criterios.

BIBLIOGRAFIA

COCHRAN W. G.: (1963). Sampling Techniques. J. Wiley, N. York.

HANSEN M. H. and Hurwitz W. N.: (1946). The problem of non response in sample survey. J. American Stat. Ass. 41, 517-29.

SRINATH K. P.: (1971). Multiphase sampling in non response problems. J. American Stat. Ass. 66, 583-86.