REVISTA MATEMATICA de la
Universidad Complutense de Madrid
Volumen 1, números 1, 2, 3; 1988.

Estimación bajo no-respuestas aleatorias al usar diseños de probabilidades desiguales

C. N. BOUZA

ABSTRACT. Singh-Singh (1979) considered the nonresponses as generated by a mechanism which is equivalent to simple random sampling (srs). They proposed an estimator under the hypothesis associated. It is unbiased and its variance is a function of an unknown parameter. An approximate expresion of its variance is presented in this paper. The robustness of the estimator, if the nonresponse mechanism is not fixed by srs, is characterized when the correct approach is related with the need of subsampling the nonresponse stratum.

1. INTRODUCCION

Singh-Singh (1979) propusieron como estimador para el total Y de la variable $\mathscr{Y}$ al estimador del tipo Horvitz-Thompson

$$\hat{Y}_{HTR} = n \sum_{i=1}^{n_1} Y_i/\pi_i n_1 \tag{1.1}$$

siendo

a) n_1 el total de respuestas;
b) $\pi_i = nP_i$ la probabilidad de inclusión en la muestra, del individuo i-ésimo;
c) Y_i el valor de $\mathscr{Y}$ en el individuo i-ésimo.

La muestra s es seleccionada con probabilidades desiguales de la población U. Se supone que las no-respuestas son generadas por un mecanismo aleatorio d equivalente al diseño muestral *msa* sin reemplazo. Esto es, que

$$s_1 = \{i \in s\text{: } i \text{ responde en la primera visita}\}$$

1980 Mathematics Subject Classification (1985 revision): 62D05.
Editorial de la Universidad Complutense. Madrid, 1988.

es una submuestra simple aleatoria de s. El enfoque clásico, véase Cochran (1981), fija que la población está particionada en dos estratos. Uno de ellos contiene a los individuos que responderán en la primera visita (U_1) y el otro estrato (U_2) al resto. El tamaño de cada estrato U_i es N_i y se supone que las no-respuestas (*nr*) son generadas por un mecanismo relacionado con factores que dependen, en cierta forma, del valor de la variable. Es por ello que se hace necesario seleccionar una submuestra s_2' de la muestra de los no-respondientes s_2 para obtener información sobre el comportamiento de los individuos de U_2. Diversos trabajos tratan de determinar cuál debe ser el tamaño de s_2', véase Cochran (1981) y Bouza (1981) para más detalles.

En este trabajo se hace un estudio de la varianza de (1.1) bajo dos criterios de aproximación. Uno de ellos es el basado en el desarrollo en Series de Taylor (*st*) de variables del tipo n_i^{-t}, cuya esperanza, aproximadamente, es $(E(n_i))^{-t}$. El otro criterio permite el desarrollo de aproximaciones de razones de variables aleatorias mediante su expansión mediante estas series. Este fue presentado por Funatsu (1982) y su aplicación permite obtener una expresión aproximada de la varianza de (1.1) cuando $n \to \infty$. Expresiones alternativas de esa aproximación son desarrolladas utilizando el método de Evaluación Puntual Multivariada (*mep*) propuesto por Manly (1986).

Estos análisis son llevados a cabo y se elaboran criterios que permiten fijar los errores originados por la aceptación de d y el uso de (1.1) cuando realmente es d' el enfoque del submuestreo del estrato de los no-respondientes el que genera las *nr*. Los errores se caracterizan por la varianza y el sesgo. Se efectúa un estudio comparativo basado en diez poblaciones de mediano tamaño provenientes de estudios de problemas biológicos, sociológicos y psicológicos. Este fue efectuado a partir de experimentos de Montecarlo bajo diversas estructuras de *nr*.

2. EL ESTIMADOR Y SU ERROR

Cuando d genera las *nr* pueden fijarse tres etapas de aleatorización:

1. Aleatorización generada por la variable n_1.
2. Aleatorización debida a s para n_1 fijo.
3. Aleatorización para todos los valores posibles de n_1.

Por ello,

$$E_3(\hat{Y}_{HTR} \mid d) = \sum_{i=1}^{n} Y_i/\pi_i = \hat{Y}_{HT}$$

y la insesgadez de (1.1) para Y se demuestra aplicando la relación

$$E(\hat{Y}_{HTR} \mid d) = E_1 E_2 E_3(\hat{Y}_{HTR} \mid d)$$

donde E_t es la esperanza condicionada en la etapa de aleatorización t.

Si el enfoque de los subestratos es el adecuado, cuando fijamos n_1 se tiene que s_1 es una muestra de U_1 y

$$E_2E_3(\hat{Y}_{HTR}\mid d')=E_2\left(n\sum_{i=1}^{n_1}Y_i/\pi_i n_1\right)=(n/n_1)E_2(\hat{Y}_{HT1})=nY_1/n_1$$

siendo $\hat{Y}_{HT1}$ el estimador de Horvitz Thompson para la población U_1.

Utilizando la aproximación

$$n_i^{-t}\doteq(E(n_i))^{-t}$$

se tiene que

$$E(\hat{Y}_{HTR}\mid d')\doteq Y_1/W_1=Y'$$

por lo que (1.1) posee sesgo

$$B_T\doteq(W_2/W_1)Y_1-Y_2$$

Si el mep es utilizado para obtener la aproximación tenemos que

$$E(\hat{Y}_{HTR}\mid d')\doteq nY_1(nW_1-W_2)-1=nY_1/(F_1-W_2)-1=Y''$$

al hacer $F_i=nW_i$, para $i=1,2$. Entonces el sesgo está dado por

$$B_S\doteq(W_2-F_1)Y_2/(F_1-W_2)=-Y_2$$

Note que, en general, $|B_S|>|B_T|>0$.

Para deducir la varianza de (1.1) es conveniente utilizar la relación

$$V(\hat{Y}_{HTR}\mid P)=E_1E_2V_3(\hat{Y}_{HTR}\mid P)+E_1V_2E_3(\hat{Y}_{HTR}\mid P)+V_1E_2E_3(\hat{Y}_{HTR}\mid P)\quad(2.1)$$

Cuando $P=d$ y $n_2=n-n_1$

$$E_2V_3(\hat{Y}_{HTR}\mid d)=E_2\left[(1-f)\left(\sum_{i=1}^{n}z_i^2-n\bar{z}^2\right)/(n_1(n-1))\right]=$$

$$=nn_2\left[\sum_{j=1}^{N}(Y_j^2/\pi_j)-(V(\hat{Y}_{HT})+Y^2)/n\right]/n_1(n-1)$$

denotando $z_i=Y_i/\pi_i$ y $f=n_1/n$. Por otra parte, como $E_3(\hat{Y}_{HTR}\mid d)=\hat{Y}_{HT}$ se tiene que

$$V_1E_2E_3(\hat{Y}_{HTR}\mid d)=0$$

$$E_1V_2V_3(\hat{Y}_{HTR}\mid d)=V(\hat{Y}_{HT})$$

por lo que la varianza está dada por

$$V(\hat{Y}_{HTR}|d)=\left[\sum_{j=1}^{N}(Y_j^2/\pi_j)-\sum_{j=1}^{N}\sum_{t\neq j=1}^{N}Y_jY_t\pi_{jt}/\pi_j\pi_t(n-1)\right]E(n_2/n_1)+$$
$$+V(\hat{Y}_{HT})=V_0E(n_2/n_1)+V(\hat{Y}_{HT});\ n_2=n-n_1 \tag{2.2}$$

Este resultado fue obtenido por Singh-Singh (1979) desarrollando su estimación a partir de una expresión equivalente.

Para obtener una expresión aproximada de la varianza utilizaremos los resultados de Funatsu (1982). En nuestro problema $F_i>0$, $i=1,2$; $P(n_1\in]L_1,\ L_2[)=1$ para todo $Lj\geqslant 0$, $j=1,2$. Por tanto, se cumplen las hipótesis fijadas para que

$$\frac{n_2}{n_1}\doteq\frac{F_2}{F_1}[(tF_2+n_2-F_2t)/tF_2]\,[(tF_1+n_1-F_1t)/tF_1]^{-1}$$

admita un desarrollo en Series de Taylor, pues $t>L_1/2F_1>0{,}5$.

Haciendo $]L_1,\ L_2[=]0,\ n[$ podemos tomar $t=1$ y

$$\frac{n_2}{n_1}\doteq\frac{F_2}{F_1}\left[\sum_{i=0}^{\infty}(-1)^i((n_1-F_1)/F_1)^i+((n_2-F_2)/F_2)(n_1-F_1)/F_1)^i\right]$$

En el tratamiento de los estimadores de razón es usual que no sean significativos los términos de orden mayor que dos. Aceptando esta hipótesis la esperanza aproximada de n_2/n_1 es

$$E(n_2/n_1)\doteq(F_2/F_1)[1+(V(n_1)/F_1^2)-(\text{Cov}\ (n_1,\ n_2)/F_1F_2)]$$

Entonces, si n es suficientemente grande

$$V(\hat{Y}_{HTR}|d)\doteq V(\hat{Y}_{HT})+V_0(1+2/n)=V_T \tag{2.3}$$

Manly (1986) generalizó el método de evaluación puntual al caso multivariado proponiendo como aproximación para la esperanza de cualquier función $h\ (n_1,\ ...,\ n_k)$ a

$$E(h(n_1,\ ...,\ n_k))\doteq\sum_{j=1}^{k}(h_{j+}+h_{j-})/2k$$

donde h_{j+} es la función evaluada en el punto

$$(E(n_1),\ ...,\ E(n_{j-1}),\ E(n_j)+V(n_j))^{1/2},\ E(n_{j+1}),\ ...,\ E(n_k))$$

y h_{j-} tiene una definición similar al sustituir el término j-ésimo por $E(n_j)-(V(n_j))^{1/2}$. Este criterio de aproximación lo denominamos mep y su uso fija que

$$E(n_2/n_1)\doteq W_2(2F_1-1)/2W_1(F_1-W_2)$$

Sustituyendo este resultado en (2.2) tenemos que

$$V(\hat{Y}_{HTR}|d) \doteq V(\hat{Y}_{HT}) + V_0(W_2(2F_1 - 1)/2W_1(F_1 - W_2)) = V_S \qquad (2.4)$$

Para estudiar el comportamiento de estas aproximaciones es necesario evaluar (2.3) y (2.4) fijando valores de n, W_1 y W_2.

Si $P = d'$ entonces $V_3(\hat{Y}_{HTR}|d') = 0$, mientras que

$$E_1 V_2 E_3(\hat{Y}_{HTR}|d') = V(\hat{Y}_{HT1}) E_1(n/n_1)^2$$

y

$$V_1 E_2 E_3(\hat{Y}_{HTR}|d') = Y_1^2 V_1(n/n_1)$$

Las aproximaciones por Series de Taylor son, respectivamente,

$$V(\hat{Y}_{HT1}) E_1(n/n_1)^2 \doteq n^2 V(\hat{Y}_{HT1})/F_1^2 = W_1^{-2} V(\hat{Y}_{HT1})$$

como

$$V_1(n/n_1) \doteq n^2 E_1((F_1 - n_1)/n_1 F_1)^2$$

se tiene nuevamente que hallar la esperanza de una razón de variables aleatorias y como se satisface que

$$E_1(F_1 - n_1) \neq 0 \text{ y } F_1 \in]0, N[$$

se pueden aplicar los resultados de Funatsu (1982) obteniéndose

$$V_1(n/n_1) \doteq (W_2/(F_1 + F_1 W_2)F_1)(1 + (2F_1(3 - W_2^3) - F_1(W_2 + 1)^2 + 6F_1)/(F_1 + F_1 W_2^2)^2 + (F_1^2(7 + W_2(n+1))/(F_1^2 + F_1 W_2)F_1 W_2) = g(n, W_1, W_2)$$

de ahí que

$$V(\hat{Y}_{HTR}|d') \doteq W_1^2 V(\hat{Y}_{HT1}) + Y_1^2 n^2 g(n, W_1, W_2) = V'_T \qquad (2.5)$$

Para aplicar el mep es necesario utilizar la aproximación para la varianza de $h(n_1, ..., n_k)$. Esta es aproximada por

$$V(h(n_1, ..., n_k)) \doteq \sum_{j=1}^{k} (h_{j+} - h_{j-})^2/4$$

véase Manly (1986).

Entonces el desarrollo de los componentes de (2.1) determina que

$$V(\hat{Y}_{HT1})E_1(n/n_1)^2 \doteq V(\hat{Y}_{HT1})(1-F_2^2)/(W_1^2(1-F_2)^2(1+F_2)^2)$$

y la varianza es aproximada por

$$V(\hat{Y}_{HTR}|d') \doteq V(\hat{Y}_{HT1})((1-F_2^2)/(W_1^2(1-F_2)^2) + Y_1^2F_2/W_1(F_1-W_2)^2 = V_S' \quad (2.6)$$

al utilizar el mep pues

$$Y_1^2V_1(n/n_1) \doteq Y_1^2F_2/W_1(F_1-W_2)^2$$

3. EXPERIMENTO DE MONTECARLO

Las expresiones obtenidas de la varianza y el sesgo de (1.1) bajo los dos modelos y los dos métodos de aproximación no son de fácil comparación analítica. Para evaluar su comportamiento se tomaron diez poblaciones y se generaron distintas estructuras de *nr* aleatoriamente fijando que $W_1 > 0$, lo que es usual, y tomando un 10 % de cada población estudiada.

Como parámetro de comparación utilizaremos la media de los valores de los coeficientes de variación y la de la razón de los errores cuadráticos medios respecto al total poblacional. Se simularon 150 encuestas, cuyos resultados se presentan en la Tabla 1.

Tabla 1
Resultados de 150 encuestas con estructura de no respuesta simulada

Población	*Porcentaje de Variación Promedio*					
	Coeficiente de variación				*Variación en ECM*	
	d		*d'*		*d vs d'*	
	T	*S*	*T*	*S*	*T*	*S*
1.....	5,8	7,5	0,1	4,5	97,3	97,4
2.....	4,9	6,6	2,7	0,06	83,3	82,2
3.....	3,3	4,2	9,8	0,9	97,6	97,1
4.....	1,5	1,9	0,3	0,2	97,0	97,2
5.....	40,8	52,5	0,2	0,1	52,8	97,2
6.....	86,5	110,2	148,3	121,1	51,4	30,9
7.....	1,8	2,8	0,7	0,2	96,9	97,1
8.....	1,5	1,9	0,2	0,3	79,0	30,7
9.....	32,0	40,5	6,8	8,7	97,0	97,2
10.....	8,3	10,7	4,0	2,7	96,5	97,2

Es de apuntar que el coeficiente $(2F_1-1)/2W_1(F_1-W_2)$ fue como promedio, 21,22 veces mayor que $(1+2/n)$. Esto se refleja en la Tabla 1 al ver que

$\bar{C}\bar{V}_S$ es siempre mayor que $\bar{C}\bar{V}_T$. De ahí la recomendabilidad de la aproximación por Series de Taylor aunque la ganancia en precisión, porcentualmente sea poco notable con respecto al mep. Es de apuntar que al no ser d el modelo adecuado el uso de (1.1) determina que la varianza explique sólo un pequeño porcentaje del error de estimación. Por eso la estimación de ella no brinda, generalmente, una correcta visión del error que se comete.

Reconocimientos: Agradezco las sugerencias de un revisor anónimo, las que permitieron introducir cambios que mejoraron el presente trabajo.

Referencias

BOUZA, C. N.: Sobre el problema de la fracción de submuestreo para el caso de las no respuestas. *Trab. de Estad. y de Inv. Operativa.* Vol. 32, 30-36, 1981.

COCHRAN, W. G.: *Técnicas de muestreo.* Ed. Espasa. México, 1981.

MANLY, B. F. J.: The point evaluation method for approximating functions means, variances and covariances. *Biometrical J.* Vol. 28, 949-956, 1986.

FUNATSU, Y.: A method of deriving valid approximate expressions for the bias in ratio estimation. *J. Stat. Planning and Inference.* Vol. 6, 216-225, 1982.

SINGH, S. and SINGH, R.: On random non response in unequal probability sampling. *Sankhya. Serie C.* 41, 127-137, 1979.

Departamento de Matemática Aplicada
Facultad de Matemática y Cibernética
Universidad de La Habana

Recibido: 18 de febrero de 1988
Revisado: 8 de julio de 1988