ESTADISTICA ESPAÑOLA
Núm. 96, 1982, págs. 51 a 62

El diseño de respuesta aleatorizada de Warner: Un modelo de superpoblación

por J. BASULTO
Departamento de Econometría y Estadística
Universidad de Sevilla

RESUMEN

Se propone un modelo de superpoblación, en el que se ha considerado el Modelo de Respuesta Aleatorizada de Warner, para estimar la proporción de un carácter dicotómico que es sensible para los individuos de una población finita y homogénea.

Palabras clave: Población finita, Modelo de Respuesta Aleatorizada de Warner, Modelos de Superpoblación, Inferencia Bayesiana.

1. INTRODUCCION

Las encuestas dirigidas a poblaciones de individuos, que buscan medir, a partir de preguntas directas, características sensibles (aborto, fraude, drogas, etc.), suelen tropezar con dificultades tales como que los individuos no contesten o que intencionadamente falsifiquen sus respuestas.

Un intento de evitar tales dificultades consiste en utilizar diseños basados en el concepto de «Respuesta aleatorizada». Tales diseños hacen uso de un instrumento aleatorio en el proceso de realización de la pregunta. Así, el entrevistado debe contestar a una pregunta que ha sido seleccionada aleatoriamente de un conjunto de preguntas,

desconociendo el entrevistador la pregunta seleccionada. De esta manera, el entrevistado puede contestar honestamente sin revelar totalmente la información sensible.

Por otra parte, al conocerse por parte del investigador las probabilidades del instrumento aleatorio, se pueden estimar aspectos cuantitativos de características sensibles, para una cierta población de individuos, a partir de las respuestas proporcionadas por una muestra de individuos. La pérdida de información que se produce por la introducción del instrumento aleatorio puede ser menos importante que las dificultades señaladas más arriba.

Han sido desarrollados varios diseños de respuesta aleatorizada; un resumen puede verse en el trabajo de Horvitz, Greenberg y Albernathy (1976). En nuestro trabajo nos limitamos al diseño original de Warner (1965).

El análisis inferencial del diseño de Warner ha sido desarrollado por procedimientos clásicos y bayesianos. El trabajo de Winkler y cols. (1979) aborda el problema desde un punto de vista bayesiano, pero considera que la población es infinita.

El objetivo de este trabajo es utilizar la aproximación de los «Modelos de Superpoblación» al diseño de Warner. Estos modelos proporcionan una metodología fructífera para abordar los problemas de diseño, inferencia y robustez en poblaciones finitas (Basulto y Murgui, 1982).

El trabajo está estructurado en secciones; así, en la sección 2 introducimos el Modelo de Superpoblación que describe la población finita bajo estudio; en la sección 3 se obtiene la distribución que nos permite hacer estimaciones del parámetro poblacional de interés; en la sección 4 se analiza dicha distribución, haciendo comparaciones con los resultados obtenidos por procedimientos clásicos y bayesianos. Por último, en la sección 5, se resume y discute el modelo propuesto.

2. MODELO DE SUPERPOBLACION PARA EL DISEÑO DE WARNER

Vamos a considerar una población U de tamaño N, cuyas unidades están identificadas por los enteros 1, 2, ..., N. A cada unidad i, de la población, asociamos un vector aleatorio $T_i = (X_i, Z_i)$, siendo X_i y Z_i variables aleatorias dicotómicas que toman valores uno y cero. La distribución de probabilidad del vector T_i viene dada por

$$f(T_i/\theta) = f(X_i, Z_i/\theta) = f(X_i/Z_i, \theta) \cdot f(Z_i) \qquad [2.1]$$

donde

$$Pr(Z_i = 1) = p, \qquad p \neq 0{,}5$$

p es conocida

y

$$Pr(X_i = 1/Z_i, \theta) = \begin{cases} \theta & \text{si } Z_i = 1 \\ 1 - \theta & \text{si } Z_i = 0 \end{cases}$$

La cantidad θ es una variable aleatoria, con recorrido en el intervalo [0.1], que tiene una distribución Beta con parámetros a y b.

La variable aleatoria Z_i selecciona, para el individuo i, la pregunta sensible con probabilidad p; por ejemplo: ¿Es usted defraudador?, y con probabilidad $1 - p$ la pregunta sensible complementaria; por ejemplo: ¿Es usted no defraudador?

La variable aleatoria X_i, observable, toma el valor uno en el caso de respuesta afirmativa y el valor cero en el caso de respuesta negativa.

Es fácil ver que

$$Pr(Z_i = 1/\theta) = \theta p + (1 - \theta) \cdot (1 - p) = \lambda \qquad [2.2.]$$

siendo λ la probabilidad de que un individuo conteste «sí», independientemente de la pregunta seleccionada.

Por último, vamos a suponer que el conjunto de vectores aleatorios $T = (T_1, T_2, \ldots, T_N)$ son independientes, condicionados a la variable aleatoria θ, e igualmente distribuidos. Es decir:

$$f(T/\theta) = \prod_{i=1}^{N} f(T_i/\theta) \qquad [2.3]$$

Sea, ahora, la variable aleatoria $Y_i = X_i Z_i + (1 - X_i) \cdot (1 - Z_i)$. Es fácil ver que Y_i es una variable dicotómica, tomando el valor uno si $X_i = 1$ y $Z_i = 1$ (el individuo contestan «sí» a la pregunta sensible) o $X_i = 0$ y $Z_i = 0$ (el individuo contesta «no» a la pregunta sensible complementaria); en los demás casos, la variable Y_i toma el valor cero.

Estamos interesados en estimar las funciones paramétricas $\Psi_1 = \Sigma Y_i$, total de individuos que contestan «sí» a la pregunta sensible, y $\Psi_2 = \Sigma Y_i/N$, la proporción de individuos que contestan «sí» a la pregunta sensible.

3. ESTIMACION DE ψ_1 Y ψ_2 A PARTIR DE UNA MUESTRA

Sin pérdida de generalidad, tomaremos los n primeros individuos de la población U. Sea $s = \{x_1, x_2, \ldots, x_n\}$ los valores de las variables aleatorias X_i, $i = 1, 2, \ldots, n$ en los

n primeros individuos. Nuestro objetivo, en esta sección, es determinar la distribución condicionada

$$f(\psi_1/x_1, x_2, \ldots, x_n; a, b)$$

es decir, la distribución del «número de individuos en la población que contestan sí a la pregunta sensible» habiendo observado una muestra s.

Notemos que la distribución [3.1] puede escribirse como

$$f(\psi_1/s; a, b) = \int_0^1 f(\psi_1/s; \theta) \cdot f(\theta/s; a, b)\, d\theta$$

donde las distribuciones, dentro del signo de integración, son, respectivamente, la distribución de ψ_1 condicionada a s y a θ, y la distribución de θ condicionada a s.

El siguiente lema nos será útil para obtener la primera distribución condicionada que aparece bajo el signo de integración.

Lema 3.1

Sea K_i una variable aleatoria binomial con parámetros (θ_i, n_i), $i = 1, 2, 3$. Supongamos que las variables aleatorias K_1, K_2 y K_3 son independientes, entonces la distribución de probabilidad de la variable aleatoria $K = K_1 + K_2 + K_3$ viene dado por la expresión siguiente

$$f(K) = \sum_{a_1}^{a_2} \sum_{b_1}^{b_2} \binom{n_1}{K-(K_2+K_3)} \binom{n_2}{K_2} \binom{n_3}{K_3} g_1(\theta_1) \cdot g_2(\theta_2) \cdot g_3(\theta_3) \quad [3.1]$$

donde

$$\begin{aligned} a_1 &= \max(0, n_2 - (n - K)) \\ a_2 &= \min(K, n_2) \\ b_1 &= \max(0, (K - K_2) - n_1) \\ b_2 &= \min(n_3, (K - K_2)) \end{aligned}$$

también

$$\begin{aligned} g_1(\theta_1) &= \theta_1^{(K-K_2-K_3)} \cdot (1 - \theta_1)^{(n_1-K+K_1+K_2)} \\ g_j(\theta_i) &= (\theta_i^{(K_i)} \cdot (1 - \theta_i)^{(n_i-K_i)}, \qquad i = 2, 3;\ j = 2, 3 \end{aligned}$$

donde $\max(c, d)$ y $\min(c, d)$ significa el máximo y mínimo, respectivamente, de los número c y d. Además, el primer sumando es para la variable K_2 y el segundo para K_3.

Demostración: Puede verse en el apéndice

El siguiente lema proporciona la distribución de ψ_1 condicionada a s y θ.

Lema 3.2

La variable aleatoria ψ_1, condicionada a la muestra s y θ, es una suma de tres variables binomiales independientes.

Demostración

La cantidad ψ_1 se puede descomponer en ψ^1, ψ^2 y ψ^3, siendo

$$\psi^1 = \sum_{i=1+n}^{N} Y_i, \quad \psi^2 = \sum_{i=1}^{n} Y_i \cdot X_i \quad \text{y} \quad \psi^3 = \sum_{i=1}^{n} Y_i \cdot (1 - X_i)$$

notemos que $\psi^2 + \psi^3 = \sum_{i=1}^{n} Y_i$ es independiente de ψ^1, condicionado a s y θ, porque

$$f(Y_1, Y_2, \ldots, Y_n, Y_{n+1}, \ldots, Y_N/x_1, x_2, \ldots, x_n; \theta) =$$
$$= f(Y_1, \ldots, Y_n/x_1, x_2, \ldots, x_n; \theta) \cdot f(Y_{n+1}, \ldots, Y_N/\theta)$$

Igualmente ψ^2 y ψ^3 son independientes, condicionados a s y θ, porque

$$f(Y_1, Y_2, \ldots, Y_n/x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(Y_i/x_i; \theta)$$

luego, a partir de las dos últimas expresiones vemos que ψ^1, ψ^2 y ψ^3 son independientes. Además, por la definición de dichas variables aleatorias, es fácil ver que tienen distribuciones binomiales con parámetros $(\theta, N - n)$, $(\theta p^0\lambda, m)$ y $(\theta(1 - p)/(1 - \lambda), n - m)$, respectivamente. La cantidad m es el número de individuos que contestan «sí» entre los n individuos entrevistados.

A partir de los lemas 3.1 y 3.2 se deduce que la distribución de ψ_1, condicionada a s y θ, es de la forma [3.1], donde después de la serie de simplificaciones queda la expresión siguiente:

$$f(\psi_1/s; \theta) = \theta^{\psi_1} \cdot (1 - \theta)^{N-\psi_1}(1/\lambda) \cdot (1/(1 - \lambda))^{n-m} \cdot g_4(\psi_1) \qquad [3.2]$$

con

$$g_4(\psi_1) = \sum_{a_1}^{a_2} \sum_{b_1}^{b_2} \binom{N - n}{\psi_1 - \psi^2 - \psi^3} \binom{m}{\psi^2} \binom{n - m}{\psi^3} p^{(n-\psi^2-m-\psi^3)}(1 - p)^{m+\psi^3-\psi^2}$$

donde

$$a_1 = \max(0,\ m - (N - \psi_1))$$
$$a_2 = \min(\psi_1,\ m)$$
$$b_1 = \max(0,\ \psi_1 - \psi^2 - N + n)$$
$$b_2 = \min(n - m,\ \psi_1 - \psi^2)$$

Respecto a la distribución condicionada de la variable θ, resulta después de aplicar el teorema de Bayes que

$$f(\theta/s;\ a,\ b) \propto \lambda^m \cdot (1 - \lambda)^{n-m} \cdot \theta^{a-1} \cdot (1 - \theta)^{b-1} \qquad [3.3]$$

donde el símbolo $\propto$ significa que la primera parte es proporcional a la segunda parte.

Luego de [3.2] y [3.3.] resulta, después de una serie de simplificaciones, que

$$f(\psi_1/s;\ a, b) \propto g_4(\psi_1) \cdot \beta b(\psi_1/n,\ a,\ b) \binom{N}{\psi_1} \qquad [3.4]$$

donde $\beta b(\varphi_1/N,\ a,\ b)$ es una distribución Betabinomial.

La distribución [3.4] nos permite hacer estimaciones puntuales, una vez especifiquemos la función de pérdida, sobre ψ_1 (y también sobre ψ_2) o realizar estimaciones mediante intervalos.

4. ANALISIS DE LA DISTRIBUCION $f(\psi_1/s;\ a,\ b)$

En esta sección proporcionamos aproximaciones del valor medio y la varianza de la distribución [3.4].

En primer lugar vamos a calcular una aproximación del valor medio, $E(\psi_1/s;\ a, b)$, de la distribución [3.4].

De la sección 3, lema 3.2, resulta que la esperanza condicionada de ψ_1 es

$$E(\psi_1/s;\ \theta) = (N - n) \cdot \theta + m \cdot (\theta \cdot p)/\lambda + (n - m) \cdot \theta(1 - p)/(1 - \lambda)$$

Consideremos ahora la función

$$h(\theta) = m \cdot \theta p/\lambda + (n - m) \cdot \theta(1 - p)/(1 - \lambda)$$

Desarrollando en serie de Taylor la función $h(\cdot)$ alrededor del valor $E(\theta/s;\ a,\ b)$ se obtiene, después de una serie de cálculos, que

$$E(\psi_1/s;\ a,\ b) \doteq (N - n) \cdot E(\theta/s;\ a,\ b) + h(E(\theta/s;\ a,\ b)) + \\ + \operatorname{var}(\theta/s;\ a,\ b) \cdot h''(E(\theta/s;\ a,\ b))/2 \qquad [4.1]$$

donde $E(\theta/s; a, b)$ y $\text{var}(\theta/s; a, b)$ son la media y la varianza, respectivamente, de la distribución $f(\theta/s; a, b)$. La función h'' es la derivada segunda de la función h; puede verse que

$$h''(\theta) = 2 \cdot p \cdot (1 - p) \cdot (2p - 1)\{(n - m)/(1 - \hat{\lambda})^3 - m/\hat{\lambda}^3\}$$

El símbolo $\doteq$ significa «aproximadamente».

A continuación consideramos una aproximación debida a Winkler y cols. (1979) para calcular la media $E(\theta/s; a, b)$ y la varianza $\text{var}(\theta/s; a, b)$.

Sean los valores n^* y m^* soluciones del sistema

$$m^*/n^* = \{m/n - (1 - p)\}/(2p - 1) = \overset{*}{\theta}$$
$$n^* = n \cdot (2p - 1)\hat{\theta} \cdot (1 - \hat{\theta})/(\hat{\lambda} \cdot (1 - \hat{\lambda}))$$

donde $\hat{\lambda} = m/n$. Suponemos que $p > 0{,}5$ y que $1 - p < m/n < p$. Entonces

$$E(\theta/s; a, b) \doteq (a + m^*)/(a + b + n^*)$$

y

$$\text{var}(\theta/s; a, b) \doteq (a + m^*) \cdot (b + n^* - m^*)/(a + b + n^*) \cdot (a + b + n^* + 1)$$

La sustitución de estas últimas aproximaciones en la fórmula [4.1] permite obtener una aproximación del valor medio $E(\psi_1/s; a, b)$.

Si estamos interesados en estimar el parámetro ψ_2, la proporción de individuos que contestan «sí» a la pregunta sensible en toda la población, entonces una aproximación de la cantidad $E(\psi_2/s; a, b)$ es

$$E(\psi_2/s; a, b) = (1 - n/N) \cdot E(\theta/s; a, b) + 0\left(\frac{1}{n}\right) \qquad [4.2]$$

que se aproxima a $E(\theta/s; a, b)$ para N suficientemente grande. Precisamente este es el estimador propuesto por Winkler y cols. (1979) en el caso de una población infinita.

También, si en [4.2] consideramos que a y b son suficientemente pequeños, entonces

$$E(\psi_2/s; a, b) \doteq \hat{\theta}$$

que es precisamente el estimador propuesto por Warner (1965).

Para obtener, ahora, una aproximación de la varianza, $\text{var}(\psi_1/s; a, b)$, de la distribución [3.4], hacemos uso de la siguiente igualdad

$$\text{var}(\psi_1/s; a, b) = E(\psi_1^2/s; a, b) - \{E(\psi_1/s; a, b)\}^2$$

siendo

$$E(\psi_1^2/s;\, a,\, b) = (N - n) \cdot E(\theta(1 - \theta)/s;\, a,\, b)) + (N - n)^2 \cdot E(\psi_1^2/s;\, a,\, b) + \\ + E(g(\theta)/s;\, a,\, b)$$

donde

$$g(\theta) = mp(1 - p)\theta(1 - \theta)/\lambda + (mp\theta/\lambda)^2 + (n - m)p(1 - p)\theta/(1 - \lambda)^2 + \\ + \{(n - m)(1 - p)\theta/(1 - \lambda)\}^2$$

Desarrollando en serie de Taylor la función $g(\cdot)$ alrededor del valor $E(\theta/s;\, a,\, b)$ se obtiene

$$E(g(\theta)/s;\, a,\, b) \doteq g(E(\theta)/s;\, a,\, b) + \text{var}\,(\theta/s;\, a,\, b)g''(E(\theta/s;\, a,\, b))$$

De todo lo anterior puede verse que

$$\text{var}\,(\psi_1/s;\, a,\, b) = (N - n)^2\, \text{var}\,(\theta/s;\, a,\, b) + 0(N - n)$$

y para el parámetro ψ_2 se obtiene

$$\text{var}\,(\psi_2/s;\, a,\, b) = (1 - n/N)^2\, \text{var}\,(\theta/s;\, a,\, b) + 0((N - n)/N)$$

luego para N suficientemente grande respecto de n obtenemos el estimador propuesto por Winkler y cols. (1979).

5. RESUMEN Y DISCUSION

En este trabajo hemos estudiado el modelo de Warner desde el punto de vista de un Modelo de Superpoblación; así se supone que la población finita considerada es una realización del modelo

$$f(T/a,\, b) = \int_0^1 \prod_{i=1}^{N} f(T_i/\theta) \cdot f(\theta/a,\, b)d\theta \qquad [5.1]$$

donde $f(\theta/a,\, b)$ es una distribución Beta.

El modelo que el vector de variables aleatorias $Y = (Y_1, Y_2, \ldots, Y_n)$, no observables, se distribuye como

$$f(Y/a,\, b) = \int_0^1 \prod_{i=1}^{N} f(Y_i/\theta) \cdot f(\theta/a,\, b)d\theta \qquad [5.2]$$

donde las variables Y_i, $i = 1, 2, \ldots, N$, son condicionadas a θ, independientes e idénticamente distribuidas.

El modelo [5.2] significa que las variables $Y_1, Y_2, \ldots, Y_N$ son intercambiables, y de ahí que la población bajo estudio sea considerada como una población homogénea; no siendo relevantes los identificadores de las unidades de la población, en el caso de estimar funciones simétricas del vector Y, por ejemplo, ψ_1 y ψ_2 (Murgui, 1982).

Todo esto último conduce a que, una vez fijado el tamaño muestral n, la selección de una muestra de tamaño n se reduzca a tomar cualquier subconjunto de tamaño n de U. (Suponiendo un coste constante para cada unidad.)

También, al variar los parámetros a y b, el modelo describe distintos niveles de incertidumbre sobre las cantidades ψ_1 y ψ_2 bajo estudio.

A partir de observar una muestra $s = \{x_1, x_2, \ldots, x_n\}$ hemos derivado de [5.1] la distribución predictiva [3.4], proporcionando aproximaciones de su valor medio y varianza.

Por fin, el trabajo puede generalizarse si se considera otros diseños de respuestas aleatorizadas.

APENDICE

Vamos la demostración del lema 3.1.

La distribución conjunta de las variables aleatorias $\mathbf{K}_i$, $i = 1, 2, 3$, es

$$f(\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3) = \prod_{i=1}^{3} \binom{n_i}{\mathbf{K}_i} \theta_i^{k_i}(1 - \theta_i)^{n_i - \mathbf{K}_i}$$

Haciendo la transformación

$$\begin{aligned} \mathbf{K} &= \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_3 \\ \mathbf{X} &= \mathbf{K}_2 \\ \mathbf{Y} &= \mathbf{K}_3 \end{aligned}$$

se obtiene la distribución conjunta $f(\mathbf{K}, \mathbf{X}, \mathbf{Y})$.

Si queremos obtener la distribución marginal $f(\mathbf{K}, \mathbf{X})$ debe observarse que

$$0 \leq \mathbf{Y} \leq n_3, \quad \mathbf{Y} = (\mathbf{K} - \mathbf{X}) - \mathbf{K}_1 \quad \text{y} \quad 0 \leq \mathbf{K}_1 \leq n_1$$

viendo que si $(K - X) \geq n_1$ entonces $Y \geq 0$ ó $Y \geq (K - X) - n_1$, y si $(K - X) < n_1$ entonces $Y \geq 0$ ó $Y \geq (K - X) - n_1$. Luego

$$f(K, X) = \sum_{b_1}^{b_2} f(K, X, Y)$$

donde b_1 y b_2 fueron ya definidas en la sección 3.

Si, por fin, se quiere obtener la distribución $f(K)$ debe observarse que

$$0 \leq X \leq n_2 \qquad 0 \leq (K - X) \leq n - n_2$$

siendo $n = n_1 + n_2 + n_3$, viendo que si $K \geq n_2$ entonces $0 \leq X \leq n_2$, y si $K < n_2$ entonces $0 \leq X \leq K$. Luego

$$f(K) = \sum_{a_1}^{a_2} f(K, X)$$

donde a_1 y a_2 fueron ya definidas en la sección 3.

BIBLIOGRAFIA

BASULTO, J., y MUNGUI, S.: *Diseño, inferencia y robustez en poblaciones finitas*. Universidad de Sevilla. No publicado. 1982.

HORVITZ, D. G.; GREENBERG, B. G., y ABERNATHY, J. R.: «Randomized Response: A Data-Gathering Device for Sensitive Questions». *International Statistical Review*, 44, 181-196, 1976.

MURGUI, S.: *Diseño e inferencia en poblaciones finitas: Modelos de superpoblación*. Tesis doctoral. Universidad de Valencia. 1982.

WARNER, S. L.: «Randomized Response: A Survey Technique for Eliminating Evasive Anwer Bias». *Journal of the American Statistical Association*, 60, 63-69, 1965.

WINKLER, R. L., y FRANKLIN, L. A.: «Warner's Randomized Response Model: A Bayesian Approach». *Journal of the American Association*, 74, 207-214, 1979.

SUMMARY

The objective of this article is to present a Superpopulation model to estimate the proportion of a sensitive dicotomic character in a finite and

homogeneus population. We focus on the original randomized response model proposed by Warner.

Key words: Finite Population; Warner's randomized Response model; Superpopulation model; Bayesian Inference.

AMS, 1980. Subject classification: Primary 62D05 & Secondary 62F15.