MUESTREO DE RESPUESTAS ALEATORIZADAS EN POBLACIONES FINITAS: UN ENFOQUE UNIFICADOR

RANDOMIZED RESPONSE SAMPLING IN FINITE POPULATIONS: A UNIFYING APPROACH

Víctor **Soberanis-Cruz**[1], Gustavo **Ramírez-Valverde**[2*], Sergio **Pérez-Elizalde**[2] y Félix **González-Cossio**[2]

[1]Universidad de Quintana Roo. Colonia del Bosque. Chetumal. 77000. Quintana Roo, México (vsobera@correo.uqroo.mx). [2]Campus Montecillo. Colegio de Postgraduados. 56230. Montecillo, Estado de México (gramirez@colpos.mx)

Resumen

La técnica de respuesta aleatorizada (RA) introducida por Warner (1965), fue diseñada para disminuir la no respuesta sobre aspectos sensibles y para proteger la confidencialidad del entrevistado. En este trabajo se propone un nuevo enfoque que utiliza la información contenida en la correlación entre la variable de interés y una variable inocua *W*, enfoque que denominaremos modelo C. Así mismo se propone, bajo un muestreo de poblaciones finitas y en el marco de la teoría de los estimadores-π (Särndal *et al*., 1992; Cassel *et al*., 1977), la unificación de varios de los modelos propuestos de RA en uno más general (modelo G). Además, se obtienen las varianzas de los distintos modelos y se observa que, bajo ciertas restricciones, el estimador del modelo C es más eficiente.

Palabras clave: Estimadores-π, pregunta sensitiva, respuesta aleatorizada, técnica de Warner.

Abstract

The randomized response technique (RR), introduced by Warner (1965) was designed to avoid non-answers to questions about sensitive issues and protect the privacy of the interviewee. In this paper, a new approach is proposed using information contained in the correlation between the variable of interest and an innocuous variable *W*; this approach is called here, the C model. Likewise it is proposed, under a finite populations sampling scheme, and in the framework theory of π estimators (Särndal *et al*., 1992; Cassel *et al*., 1977) the unification of several of the RR models in a more general one (G model). Furthermore, the variances of the different models are obtained and, under certain restrictions, the C model estimator is more efficient.

Key words: π-estimators, sensitive question, randomized response, Warner's technique.

Introducción

En estudios de muestreo por encuestas el interés frecuentemente se centra en aspectos sensibles o confidenciales para las personas entrevistadas, tales como uso de drogas, evasión de impuestos, preferencias sexuales, honestidad en exámenes, opinión respecto a autoridades, etcétera. Por tal motivo, algunos entrevistados se niegan a responder (fenómeno de no-respuesta) la pregunta con la que se pretende obtener información sobre el aspecto sensible, o proporcionan respuestas falsas; en cualquiera de los dos casos las estimaciones son sesgadas.

La técnica de respuesta aleatorizada (RA), introducida por Warner (1965), propone una solución para la protección de la confidencialidad del entrevistado, y consiste en la utilización de un mecanismo aleatorio (MA) por medio del cual se selecciona una de dos preguntas: ¿pertenece al grupo con la característica A? o ¿pertenece al grupo que no tiene la característica A?, donde A es la característica sensible de interés.

*Autor responsable ❖ Author for correspondence.
Recibido: Septiembre, 2006. Aprobado: Marzo, 2008.
Publicado como ARTÍCULO en Agrociencia 42: 537-549. 2008.

Introduction

In survey studies, interest is frequently focused on issues that are sensitive or confidential for the interviewees, such as use of drugs, tax evasion, sexual preferences, honesty in tests, opinions on authorities, etc. For this reason, some interviewees refuse to respond (no response phenomenon) to the questions designed to obtain information on a sensitive aspect, or they give false answers. In either case the estimations are biased.

The random response technique (RR), introduced by Warner (1965), proposes a solution to protect the privacy of the interviewee consisting of using a random mechanism by which one of two questions is selected: do you belong to the group with characteristic A? or do you belong to the group that does not have characteristic A?, in which A is the sensitive characteristic of interest. The interviewee will answer yes or no, and the interviewer has no possibility of knowing which question was answered by the interviewer, thus protecting its privacy.

The RR technique has encouraged a series of approaches, among which the following models are

El entrevistado contestará sí o no y el entrevistador no tiene la posibilidad de saber qué pregunta contestó el entrevistado, protegiendo así la confidencialidad del mismo.

La técnica RA ha propiciado que se generen una serie de enfoques, entre los que destacan los siguientes modelos: a) el W (Warner, 1965), b) el U con pregunta inocua *W* no relacionada (Greenberg *et al*., 1969), c) el C, d) el H (Horvitz *et al*., 1976), e) el D (Devore, 1977) y, f) el M (Mangat y Singh, 1990), cada modelo se describe a continuación.

El modelo U (Greenberg *et al*., 1969) es de respuesta aleatorizada con preguntas no relacionadas. Al igual que el modelo W tiene un mecanismo aleatorio que selecciona una de dos preguntas, pero mientras una pregunta corresponde al aspecto sensible, ¿perteneces al grupo con la característica A?, la segunda pregunta no tiene que ver con el aspecto sensible; es sobre algún otro aspecto inocuo *W*; ésto es, no afecta la sensibilidad del entrevistado. Por ejemplo, si la primera pregunta es ¿evade usted impuestos?, la segunda pregunta podría ser ¿le gusta el cine? La comparación de los modelos W y U se ha hecho en el marco de poblaciones infinitas (Moors, 1971), resultando el modelo U más eficiente que el W.

Horvitz *et al*. (1976) proponen el modelo H, que permite una mayor protección del anonimato del entrevistado sin utilizar la pregunta complementaria; cada elemento de la muestra responde aleatoriamente una de tres proposiciones: (1) la pregunta sensitiva, (2) una instrucción que dice sí y (3) una instrucción que dice no, a ser escogidas con probabilidades p_1, p_2 y p_3, con $p_1 + p_2 + p_3 = 1$.

En el modelo M el mecanismo aleatorio proporciona n respuestas independientes con dos componentes aleatorias. El modelo D es análogo al U, con una diferencia básica: la pertenencia al grupo inocuo *W* se establece con probabilidad uno.

Chaudhuri y Mukerjee (1988) presentan una buena reseña sobre los trabajos pioneros en respuestas aleatorizadas. Algunos trabajos más recientes son los de Lakshmi y Raghavarao (1992); Mangat *et al*. (1993); Chua y Tsui (2000); Padmawar y Vijayan (2000); y Chaudhuri (2001). Un enfoque bayesiano al modelo de Warner puede verse en Winkler y Franklin (1979) y Bar-Lev *et al*. (2003).

En este trabajo se propone un nuevo esquema (modelo C) que permite que la pregunta inocua del modelo U esté correlacionada con la variable sensitiva *Y*, pero que no afecta la sensibilidad del individuo, manteniéndose así la confidencialidad del entrevistado. En este nuevo enfoque se aprovecha la información contenida en la correlación de la variable sensible con la variable inocua para tener una

outstanding: a) the W model (Warner, 1965), b) the U model with an innocuous unrelated question W (Greenberg *et al*., 1969), c) the C model, d) the H model (Horvitz *et al*., 1976), e) the D model (Devore, 1977), and f) the M model (Mangat and Singh, 1990). Each of these is described below.

The U model (Greenberg *et al*., 1969) is a randomized response approach with unrelated questions. Like the W model, it has a random mechanism that selects one of two questions, but while one question refers to a sensitive subject (Do you belong to the group with characteristic A?), the second question has nothing to do with the sensitive subject, but is about some other innocuous aspect *W*; that is, it does not affect the interviewee's sensitivity. For example, if the first question is "Do you evade taxes?", the second question could be "Do you like the movies?" The W and U models were compared within the framework of infinite populations (Moors, 1971), and the outcome revealed that the U model was more efficient than the W.

Horvitz *et al*. (1975) propose the H model, which allows for greater protection of the interviewee's anonymity, without the use of the complementary question. Each element of the sample responds randomly to one of three propositions: (1) the sensitive question, (2) an instruction that says "yes", and (3) an instruction that says "no", to be chosen with probabilities of p_1, p_2 and p_3, with $p_1 + p_2 + p_3 = 1$.

In the M model, the random mechanism provides n independent responses with two random components. The D model is analogous to U, with a basic difference: belonging to the innocuous group *W* is established with probability one.

Chaudhuri and Mukerjee (1988) present a good review of the pioneer work in randomized responses. Some more recent studies are those of Lakshmi and Raghavarao (1992), Mangat *et al*. (1993), Chua and Tsui (2000), Padmawar and Vijayan (2000), and Chaudhuri (2001). A Bayesian approach to the Warner model can be seen in Winkler and Franklin (1979) and Bar-Lev *et al*. (2003).

This paper proposes a new scheme (C model) in which the innocuous question of the U model is correlated with the sensitive variable Y, but does not affect the individual's sensitivity, thus maintaining the privacy of the interviewee. This new scheme uses information contained in the correlation between the sensitive and the innocuous variable to attain a better estimation in terms of bias and variance, under a finite populations approach. Also, it is proposed that these schemes be unified into a G randomized response model, such that the W, U, C, H, D, and M models be particular cases. The variances of the different models

mejor estimación en términos de sesgo y varianza, bajo un esquema de muestreo de poblaciones finitas. Asimismo, se propone unificar estos esquemas en un modelo G de respuesta aleatorizada, tal que los modelos W, U, C, H, D, y M sean casos particulares. Se obtienen las varianzas de los distintos modelos y se estudia por simulación la dispersión del estimador; y resulta que el modelo C es más eficiente.

Materiales y Métodos

Población bajo estudio

Se considera una población finita $U=\{1,2,\ldots,N\}$. Se definen las subpoblaciones: $Y_i=\{k\in U: y_k=i\}$, $i=0,1$; y W_i, $i=0,1$. El tamaño de la población N se supone conocido. El tamaño de la muestra se denota por n, el cual no necesariamente es fijo. Sea y la variable dicotómica que denota la pertenencia de un individuo al grupo con la característica sensible de interés, con y_k el valor de y para el k-ésimo elemento de la población. Así, y_k es desconocida pero no aleatoria. Además, y_k=1 si el k-ésimo individuo tiene la característica sensitiva A y y_k=0 en caso contrario. Lo que se desea estimar es $t_A=\sum_U y_k$, el total de los individuos en la población con la característica sensible A.

Procedimiento de muestreo

Para el modelo general (G) el procedimiento de muestreo es:

Etapa 1 (selección de la muestra). Se toma una muestra de tamaño n de acuerdo con el diseño de muestreo $p(s)$ con probabilidades positivas de inclusión π_k y π_{kl}, donde

$$\pi_k=\Pr\{S\ni k\}=\sum_{s\ni k}p(s)\ \text{ y }\ \pi_{kl}=\Pr(S\ni k\ \&\ l)=\sum_{s\ni k\&l}p(s)$$

Para cada elemento k en la muestra S se tiene $I_k=1$ si $k\in S$, $I_k=0$ de otra forma; nótese que $I_k(S)$ es función de la variable aleatoria S. Además,

$$\Delta_{kl}=Cov(I_k,I_l)=E_p(I_kI_l)-E_p(I_k)E_p(I_l)$$
$$=\pi_{kl}-\pi_k\pi_l\leq \text{ cuando }\ 0k\neq l\ \text{ y }\ \Delta_{kk}=\pi_k(1-\pi_k)$$
$$\pi_{kk}=\pi_k\breve{y}_k=\frac{y_k}{\pi_k},\ \breve{\Delta}_{kl}=\frac{\Delta_{kl}}{\pi_{kl}}$$

Etapa 2 (recopilación de la información). Las entrevistas se realizan a los individuos en la muestra de acuerdo al MA definido por el modelo de respuesta aleatorizada empleado. El MA induce para cada $k\in S$ una variable aleatoria Z_k, tal que la combinación lineal $\hat{Z}_k=aZ_k+b_k$ es una estimación insesgada de y_k, donde a y b_k son constantes conocidas que dependen del MA; por tanto, $E_{MA}(\hat{Z}_k)=y_k$, $V_{MA}(\hat{Z}_k)=a^2V_{MA}(Z_k)$ $E_{MA}(b_k)=b_k\ \forall k\in S$ y el cálculo de $V_{MA}(Z_k)$ también depende de MA.

Materials and Methods

Population under study

A finite population $U=\{1,2,\ldots,N\}$ is considered. The following subpopulations are defined: $Y_i=\{k\in U: y_k=i\}$, $i=0,1$; and W_i, $i=0,1$. It is assumed that the size of population N is known. Sample size is denoted by n, which is not necessarily fixed. Let y be the dichotomic variable that refers to the individual's belonging to the group with the sensitive characteristic of interest, with y_k the value of y for the k^{th} element of the population. Thus, y_k is unknown but not random. Also, y_k=1 if the kth individual has the sensitive characteristic A, and y_k=0 in the opposite case. What is to be estimated is $t_A=\sum_U y_k$, the total number of individuals of the population with the sensitive characteristic A.

Sampling procedure

For the general model (G), the sampling procedure is:

Step 1 (sample selection). A sample size of size n is taken in accordance with the sampling design $p(s)$ with positive probabilities of inclusion π_k and π_{kl}, where

$$\pi_k=\Pr\{S\ni k\}=\sum_{s\ni k}p(s)\ \text{ and }\ \pi_{kl}=\Pr(S\ni k\ \&\ l)=\sum_{s\ni k\&l}p(s)$$

For each element k in sample S, $I_k=1$ if $k\in S$; I_k=0, otherwise. Note that $I_k(S)$ is a function of the random variable S. Also,

$$\Delta_{kl}=Cov(I_k,I_l)=E_p(I_kI_l)-E_p(I_k)E_p(I_l)$$
$$=\pi_{kl}-\pi_k\pi_l\leq \text{ when }\ 0k\neq l\ \text{ y }\ \Delta_{kk}=\pi_k(1-\pi_k)$$
$$\pi_{kk}=\pi_k\breve{y}_k=\frac{y_k}{\pi_k},\ \breve{\Delta}_{kl}=\frac{\Delta_{kl}}{\pi_{kl}}$$

Step 2 (information gathering). The interviews of individuals in the sample are conducted in accordance with the MA defined by the randomized response model used. For each $k\in S$, the MA induces a random variable Z_k, so that the linear combination $\hat{Z}_k=aZ_k+b_k$ is an unbiased estimation of y_k, where a and b_k are known constants that depend on the MA; therefore, $E_{MA}(\hat{Z}_k)=y_k$, $V_{MA}(\hat{Z}_k)=a^2V_{MA}(Z_k)$ $E_{MA}(b_k)=b_k\ \forall k\in S$, and the calculation of $V_{MA}(Z_k)$ also depends on MA.

Unified approach (G model)

In a similar way to the generalization developed by Chaudhuri and Mukerjee (1988), a unified approach of the considered models is proposed. In addition, it is proposed that the information contained in the correlation between the innocuous and the sensitive variable be used to generate model C.

Enfoque unificador (modelo G)

De manera semejante a la generalización desarrollada por Chaudhuri y Mukerjee (1988), se propone un enfoque unificador de los modelos considerados. Además, se propone aprovechar la información contenida en la correlación entre la variable inocua y la sensitiva, generando el modelo C.

El estimador general que se propone para la estimación del total $t_A = \sum_U y_k$ es:

$$\hat{t}_{AG,\pi} = \sum_S \frac{\hat{Z}_k}{\pi_k} \tag{1.1}$$

Se tiene que

$$\begin{aligned} E\left(\hat{t}_{AG,\pi}\right) &= E_p E_{MA}\left(\sum_S \frac{\hat{Z}_k}{\pi_k}\right) = E_p\left(\sum_S \frac{E_{MA}\left(\hat{Z}_k\right)}{\pi_k}\right) \\ &= E_p\left(\sum_S \frac{y_k}{\pi_k}\right) = \sum_U y_k = t_A, \end{aligned} \tag{1.2}$$

por lo que el estimador es insesgado bajo el diseño de muestreo $p(s)$. También

$$\begin{aligned} V\left(\hat{t}_{AG,\pi}\right) &= E_p V_{MA}\left(\sum_S \frac{\hat{Z}_k}{\pi_k}\right) + V_p E_{MA}\left(\sum_S \frac{\hat{Z}_k}{\pi_k}\right) \\ &= E_p\left(\sum_S \frac{V_{MA}\left(\hat{Z}_k\right)}{\pi_k^2}\right) + V_p\left(\sum_S \frac{E_{MA}\left(\hat{Z}_k\right)}{\pi_k}\right) \\ &= a^2 E_p\left(\sum_S \frac{V_{MA}\left(Z_k\right)}{\pi_k^2}\right) + V_p\left(\sum_S \frac{y_k}{\pi_k}\right) \\ &= V_p\left(\sum_S \frac{y_k}{\pi_k}\right) + a^2 E_p\left(\sum_S \frac{V_{MA}\left(Z_k\right)}{\pi_k^2}\right) \\ &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + a^2 E_p\left(\sum_S \frac{V_{MA}\left(Z_k\right)}{\pi_k^2}\right) \end{aligned} \tag{1.3}$$

La varianza para cualquier modelo que cumpla con las condiciones del modelo G está formada por dos términos: el primero depende del diseño de muestreo $p(s)$ y los valores y_k, esta parte es común a todos los modelos y será denotada por V_G; el otro término depende del mecanismo aleatorio empleado. Por tanto, para comparar la varianza de los distintos modelos es suficiente la comparación del segundo término de la varianza.

A continuación se muestra que los modelos W, U, C, H, D y M, son casos particulares del modelo G, con las constantes presentadas en el Cuadro 1.

Modelo de Warner (W): preguntas complementarias

Para el mecanismo aleatorio del modelo de Warner se tiene

$$Z_k = \begin{cases} y_k & \text{con probabilidad } p \\ 1 - y_k & \text{con probabilidad } 1-p \end{cases} \tag{1.4}$$

The general estimator proposed for estimating total $t_A = \sum_U y_k$ is

$$\hat{t}_{AG,\pi} = \sum_S \frac{\hat{Z}_k}{\pi_k} \tag{1.1}$$

We have that

$$\begin{aligned} E\left(\hat{t}_{AG,\pi}\right) &= E_p E_{MA}\left(\sum_S \frac{\hat{Z}_k}{\pi_k}\right) = E_p\left(\sum_S \frac{E_{MA}\left(\hat{Z}_k\right)}{\pi_k}\right) \\ &= E_p\left(\sum_S \frac{y_k}{\pi_k}\right) = \sum_U y_k = t_A, \end{aligned} \tag{1.2}$$

so that the estimator is unbiased under the sampling design $p(s)$. Also,

$$\begin{aligned} V\left(\hat{t}_{AG,\pi}\right) &= E_p V_{MA}\left(\sum_S \frac{\hat{Z}_k}{\pi_k}\right) + V_p E_{MA}\left(\sum_S \frac{\hat{Z}_k}{\pi_k}\right) \\ &= E_p\left(\sum_S \frac{V_{MA}\left(\hat{Z}_k\right)}{\pi_k^2}\right) + V_p\left(\sum_S \frac{E_{MA}\left(\hat{Z}_k\right)}{\pi_k}\right) \\ &= a^2 E_p\left(\sum_S \frac{V_{MA}\left(Z_k\right)}{\pi_k^2}\right) + V_p\left(\sum_S \frac{y_k}{\pi_k}\right) \\ &= V_p\left(\sum_S \frac{y_k}{\pi_k}\right) + a^2 E_p\left(\sum_S \frac{V_{MA}\left(Z_k\right)}{\pi_k^2}\right) \\ &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + a^2 E_p\left(\sum_S \frac{V_{MA}\left(Z_k\right)}{\pi_k^2}\right) \end{aligned} \tag{1.3}$$

The variance for any model that satisfies the conditions of the G model is composed of two terms. The first depends on the sampling design $p(s)$ and the y_k values. This part is common to all of the models and will be denoted V_G. The other term depends on

Cuadro 1. Resumen de las constantes asociadas a los distintos modelos considerados.

Table 1. Summary of the constants associated with the different models considered.

Modelo	a	b
W	$\frac{1}{2p-1}$	$-\frac{1-p}{2p-1}$
H	$\frac{1}{p_1}$	$-\frac{(1-p)}{p}$
U, C	$\frac{1}{p}$	$-\frac{(1-p)w_k}{p}$
D	$\frac{1}{p}$	$-\frac{(1-p)}{p}$
M	$\frac{1}{t+(1-t)(2p-1)}$	$-\frac{(1-t)(1-p)}{t+(1-t)(2p-1)}$

para cada $k \in S$, y

$$\begin{aligned}\hat{Z}_k &= \frac{Z_k - (1-p)}{2p-1} \\ &= aZ_k + b_k\end{aligned} \tag{1.5}$$

con $a = \frac{1}{2p-1}$ y $b_k = \frac{1-p}{2p-1}$; de modo que

$$E_{MA}(Z_k) = (2p-1)y_k + (1-p),$$
$$V_{MA}(Z_k) = p(1-p),$$

y

$$E_{MA}(\hat{Z}_k) = y_k, k \in S; V_{MA}(\hat{Z}_k) = a^2 V_{MA}(Z_k) \tag{1.6}$$

De (1.5) y (1.6) se puede ver que el modelo W es un caso particular del modelo G. Si $\hat{t}_{A1,\pi}$ es el estimador del total $t_A = \sum_U y_k$ en el modelo de Warner, de (1.1) obtenemos

$$\begin{aligned}\hat{t}_{A1,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \frac{1}{2p-1}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{1}{\pi_k}\right]\end{aligned} \tag{1.7}$$

De (1.3) y (1.6)

$$\begin{aligned}V(\hat{t}_{A1,\pi}) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{p(1-p)}{(2p-1)^2}\left(\sum_U \frac{1}{\pi_k}\right) \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + V_0\left(\sum_U \frac{1}{\pi_k}\right)\end{aligned} \tag{1.8}$$

donde $V_0 = \frac{p(1-p)}{(2p-1)^2}$. La selección de p debe hacerse de tal modo que el entrevistado esté convencido de la protección de su anonimato. Obviamente, la selección $p=1/2$ es la más convincente; no obstante, es inadmisible ya que para este valor no está definido el estimador.

Modelo U

El modelo U propuesto por Greenberg *et al*. (1969) utiliza el siguiente mecanismo aleatorio:

$$Z_k = \begin{cases} y_k & \text{con probabilidad } p \\ w_k & \text{con probabilidad } 1-p \end{cases} \tag{2.1}$$

la esperanza es:

$$E_{MA}(Z_k) = y_k p + w_k(1-p)$$

y

$$\begin{aligned}\hat{Z}_k &= \frac{Z_k - (1-p)w_k}{p} \\ &= aZ_k + b_k\end{aligned} \tag{2.2}$$

the random mechanism used. Therefore, to compare the variance of the different models, it suffices to compare the second term of the variance.

Below, it is shown that the W, U, C, H, D and M models are particular cases of the G model, with the constants presented in Table 1.

Warner model (W): complementary questions

For the random mechanism of the Warner model we have

$$Z_k = \begin{cases} y_k & \text{with likelihood } p \\ 1-y_k & \text{with likelihood } 1-p \end{cases} \tag{1.4}$$

For each $k \in S$, and

$$\begin{aligned}\hat{Z}_k &= \frac{Z_k - (1-p)}{2p-1} \\ &= aZ_k + b_k\end{aligned} \tag{1.5}$$

with $a = \frac{1}{2p-1}$ and $b_k = \frac{1-p}{2p-1}$, so that

$$E_{MA}(Z_k) = (2p-1)y_k + (1-p),$$
$$V_{MA}(Z_k) = p(1-p),$$

and

$$E_{MA}(\hat{Z}_k) = y_k, k \in S; V_{MA}(\hat{Z}_k) = a^2 V_{MA}(Z_k) \tag{1.6}$$

From (1.5) and (1.6), it can be seen that the W model is a particular case of the G model. If $\hat{t}_{A1,\pi}$ is the estimator of the total $t_A = \sum_U y_k$ in the Warner model, from (1.1) we obtain

$$\begin{aligned}\hat{t}_{A1,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \frac{1}{2p-1}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{1}{\pi_k}\right]\end{aligned} \tag{1.7}$$

From (1.3) and (1.6)

$$\begin{aligned}V(\hat{t}_{A1,\pi}) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{p(1-p)}{(2p-1)^2}\left(\sum_U \frac{1}{\pi_k}\right) \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + V_0\left(\sum_U \frac{1}{\pi_k}\right)\end{aligned} \tag{1.8}$$

where $V_0 = \frac{p(1-p)}{(2p-1)^2}$. Selection of p should be done in such a way as that the interviewee is convinced of the protection of his privacy. Obviously, the selection $p=1/2$ is the most convincing. Nevertheless, it is inadmissible since the estimator for this value is not defined.

con $a=\frac{1}{p}$ y $b_k=-\frac{1-p}{p}w_k$. Observe que aunque b_k depende de k, es constante respecto al MA, lo cual implica que el modelo U es un caso particular del modelo G. Además, se tiene que

$$\begin{aligned} V_{MA}(Z_k) &= E_{MA}\left(Z_k^2\right)-E_{MA}^2(Z_k) \\ &= E_{MA}(Z_k)-E_{MA}^2(Z_k) \\ &= py_k+(1-p)w_k-\left[py_k+(1-p)w_k\right]^2 \\ &= p(1-p)(y_k-w_k)^2 \end{aligned} \tag{2.3}$$

Denotando con $\hat{t}_{A2,\pi}$ al estimador del total $t_A=\sum_U y_k$, de (2.2) y de (1.1) se tiene,

$$\begin{aligned} \hat{t}_{A3,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \sum_S\left[\frac{Z_k-(1-p)w_k}{p}\right]/\pi_k \\ &= \frac{1}{p}\left[\sum_S\frac{Z_k}{\pi_k}-(1-p)\sum_S\frac{w_k}{\pi_k}\right], \end{aligned} \tag{2.4}$$

De (2.3) y de la ultima igualdad en (1.3) se tiene que

$$\begin{aligned} V\left(\hat{t}_{A2,\pi}\right) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}+\left(\frac{1}{p}\right)^2 p(1-p)E_p\left(\sum_S\frac{(y_k-w_k)^2}{\pi_k^2}\right) \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}+\frac{1-p}{p}\sum_U\frac{(y_k-w_k)^2}{\pi_k} \end{aligned} \tag{2.5}$$

En la expresión (2.5), si y y W están correlacionadas se reduce la varianza del estimador.

Modelo H

El modelo H es una alternativa al esquema de Warner que da mayor protección al entrevistado, y consiste en que cada elemento de la muestra selecciona aleatoriamente una de tres proposiciones: (1) la sensitiva Q_A, (2) una instrucción que dice si y (3) una que dice no, con probabilidades p_1, p_2, p_3 y $p_1+p_2+p_3=1$.

Para el modelo H se imponen las siguientes restricciones:

$$\begin{aligned} &1>p=p_1>1/2, 0<1-p_1-p_2\le p_2<1 \\ &-p_1 \Leftrightarrow 1-p_1-2p_2\le 0, p_2<1-p_1 \end{aligned} \tag{3.1}$$

Se tiene

$$Z_k=\begin{cases} y_k & \text{con probabilidad } p_1, \\ 1 & \text{con probabilidad } p_2, \\ 0 & \text{con probabilidad } 1-p_1-p_2, \end{cases} \tag{3.1a}$$

donde

The U model

The U model proposed by Greenberg *et al*. (1969) uses the following random mechanism:

$$Z_k=\begin{cases} y_k & \text{with likelihood } p \\ w_k & \text{with likelihood } 1-p \end{cases} \tag{2.1}$$

The expectation is:

$$E_{MA}(Z_k)=y_kp+w_k(1-p)$$

$$\begin{aligned} \hat{Z}_k &= \frac{Z_k-(1-p)w_k}{p} \\ &= aZ_k+b_k \end{aligned} \tag{2.2}$$

and,

with $a=\frac{1}{p}$ y $b_k=-\frac{1-p}{p}w_k$. Note that, although b_k depends on k, it is constant with respect to MA, which implies that the U model is a particular case of the G. model. Also, we have

$$\begin{aligned} V_{MA}(Z_k) &= E_{MA}\left(Z_k^2\right)-E_{MA}^2(Z_k) \\ &= E_{MA}(Z_k)-E_{MA}^2(Z_k) \\ &= py_k+(1-p)w_k-\left[py_k+(1-p)w_k\right]^2 \\ &= p(1-p)(y_k-w_k)^2 \end{aligned} \tag{2.3}$$

With $\hat{t}_{A2,\pi}$ denoting the estimator of the total $t_A=\sum_U y_k$ of (2.2) and from (1.1) we have

$$\begin{aligned} \hat{t}_{A3,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \sum_S\left[\frac{Z_k-(1-p)w_k}{p}\right]/\pi_k \\ &= \frac{1}{p}\left[\sum_S\frac{Z_k}{\pi_k}-(1-p)\sum_S\frac{w_k}{\pi_k}\right], \end{aligned} \tag{2.4}$$

From (2.3) and from the last equality in (1.3) we have

$$\begin{aligned} V\left(\hat{t}_{A2,\pi}\right) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}+\left(\frac{1}{p}\right)^2 p(1-p)E_p\left(\sum_S\frac{(y_k-w_k)^2}{\pi_k^2}\right) \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}+\frac{1-p}{p}\sum_U\frac{(y_k-w_k)^2}{\pi_k} \end{aligned} \tag{2.5}$$

In the expression (2.5), if y and W are correlated, the variance of the estimator is reduced.

The H model

The H model is an alternative to the Warner scheme and provides greater protection for the interviewee. It consists in

$$E_{MA}(Z_k) = y_k p_1 + p_2 \tag{3.2}$$

y

$$\begin{aligned}\hat{Z}_k &= \frac{Z_k - p_2}{p_1} = \frac{1}{p_1} Z_k - \frac{p_2}{p_1} \\ &= aZ_k + b_k\end{aligned} \tag{3.2a}$$

con $a = \dfrac{1}{p_1}$ y

$$b_k = -\frac{p_2}{p_1} \Rightarrow E_{MA}(\hat{Z}_k) = y_k \;\; y \;\; V_{MA}(\hat{Z}_k) = a^2 V_{MA}(Z_k) \tag{3.3}$$

De (3.2a) y (3.3) se puede ver que el modelo H es un caso particular del modelo G. Si $\hat{t}_{A2,\pi}$ es el estimador del total $t_A = \sum_U y_k$ en el modelo H, de (1.1) se tiene

$$\begin{aligned}\hat{t}_{A3,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \frac{1}{p_1}\left[\sum_S \frac{Z_k}{\pi_k} - p_2 \sum_S \frac{1}{\pi_k}\right]\end{aligned} \tag{3.4}$$

De (1.3) y de (3.3) se obtiene

$$\begin{aligned}V(\hat{t}_{A3,\pi}) &= \sum\sum_S \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \left(\frac{1}{p_1}\right)^2 \\ &E_p\left\{\sum_S \frac{1}{\pi_k^2}\left[p_1(1-p_1-2p_2)y_k + p_2(1-p_2)\right]\right\} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p_1-2p_2}{p_1^2}\sum_U \frac{y_k}{\pi_k} + \frac{p_2(1-p_2)}{p_1^2}\sum_U \frac{1}{\pi_k}\end{aligned} \tag{3.5}$$

Considerando a $V(\hat{t}_{A3,\pi})$ como función de p_2 únicamente tomando una p_1 fija, se puede ver que $V(\hat{t}_{A3,\pi})$ es creciente en p_2, y como $\dfrac{1-p_1}{2} \le p_2 < 1 - p_1$, entonces $V(\hat{t}_{A3,\pi})$ es mínima cuando $p_2 = \dfrac{1-p_1}{2}$ y (3.5) se reduce a

$$V(\hat{t}_{A3,\pi}) = \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1}{4}\frac{1-p_1^2}{p_1^2}\left(\sum_U \frac{1}{\pi_k}\right) \tag{3.6}$$

Modelo D

Esta propuesta es análoga al modelo U, con una diferencia básica, la pertenencia al grupo inocuo *W* es con probabilidad uno. Por ejemplo, la pregunta inocua *W* puede ser: ¿está usted vivo? Para este modelo tenemos

$$Z_k = \begin{cases} y_k & \text{con probabilidad } p \\ 1 & \text{con probabilidad } 1-p. \end{cases} \tag{4.1}$$

that each element of the sample randomly selects one of three propositions: (1) the sensitive one Q_A, (2) an instruction that says "yes", and (3) one that says "no", with probabilities of p_1, p_2, p_3, and $p_1 + p_2 + p_3 = 1$.

For the H model the following restrictions are imposed:

$$\begin{aligned}1 > p = p_1 > 1/2, 0 < 1 - p_1 - p_2 \le p_2 < 1 \\ -p_1 \Leftrightarrow 1 - p_1 - 2p_2 \le 0, p_2 < 1 - p_1\end{aligned} \tag{3.1}$$

We have

$$Z_k = \begin{cases} y_k & \text{with probability } p_1, \\ 1 & \text{with probability } p_2, \\ 0 & \text{with probability } 1-p_1-p_2, \end{cases} \tag{3.1a}$$

where

$$E_{MA}(Z_k) = y_k p_1 + p_2 \tag{3.2}$$

and

$$\begin{aligned}\hat{Z}_k &= \frac{Z_k - p_2}{p_1} = \frac{1}{p_1} Z_k - \frac{p_2}{p_1} \\ &= aZ_k + b_k\end{aligned} \tag{3.2a}$$

With $a = \dfrac{1}{p_1}$ and

$$b_k = -\frac{p_2}{p_1} \Rightarrow E_{MA}(\hat{Z}_k) = y_k \;\; y \;\; V_{MA}(\hat{Z}_k) = a^2 V_{MA}(Z_k) \tag{3.3}$$

From (3.2a) and (3.3) it can be seen that the H model is a particular case of the G model. If $\hat{t}_{A2,\pi}$ is the estimator of the total $t_A = \sum_U y_k$ in the H model, from (1.1) we have

$$\begin{aligned}\hat{t}_{A3,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \frac{1}{p_1}\left[\sum_S \frac{Z_k}{\pi_k} - p_2 \sum_S \frac{1}{\pi_k}\right]\end{aligned} \tag{3.4}$$

From (1.3) and (3.3) we obtain

$$\begin{aligned}V(\hat{t}_{A3,\pi}) &= \sum\sum_S \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \left(\frac{1}{p_1}\right)^2 \\ &E_p\left\{\sum_S \frac{1}{\pi_k^2}\left[p_1(1-p_1-2p_2)y_k + p_2(1-p_2)\right]\right\} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p_1-2p_2}{p_1^2}\sum_U \frac{y_k}{\pi_k} + \frac{p_2(1-p_2)}{p_1^2}\sum_U \frac{1}{\pi_k}\end{aligned} \tag{3.5}$$

Considering $V(\hat{t}_{A3,\pi})$ a function of p_2, taking only a fixed p_1, it can be seen that $V(\hat{t}_{A3,\pi})$ is increasing in p_2, and since

De modo que

$$E_{MA}(Z_k) = y_k p + (1-p)$$

y

$$\begin{aligned}\hat{Z}_k &= \frac{Z_k - (1-p)}{p} \\ &= aZ_k + b_k,\end{aligned} \tag{4.2}$$

con $a = \frac{1}{p}$ y $b_k = -\frac{1-p}{p}$. Denotando por $\hat{t}_{A4,\pi}$ al estimador del total $t_A = \sum_U y_k$ en el modelo D, de (1.1) se tiene

$$\begin{aligned}\hat{t}_{A4,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \frac{1}{p}\sum_S \frac{Z_k}{\pi_k} - \frac{1-p}{p}\sum_S \frac{1}{\pi_k}\end{aligned} \tag{4.3}$$

Para obtener la varianza del estimador correspondiente al modelo D primero vemos que

$$\begin{aligned}V_{MA}(Z_k) &= E_{MA}(Z_k^2) - \left[E_{MA}(Z_k)\right]^2 \\ &= E_{MA}(Z_k) - \left[E_{MA}(Z_k)\right]^2 \\ &= py_k + (1-p) - \left[py_k + (1-p)\right]^2 \\ &= p(1-p)(1-y_k)\end{aligned}$$

Así

$$\begin{aligned}E_p\left(\sum_S \frac{V_{MA}(Z_k)}{\pi_k^2}\right) &= E_p\left(\sum_S \frac{p(1-p)(1-y_k)}{\pi_k^2}\right) \\ &= p(1-p)\sum_U \frac{1-y_k}{\pi_k},\end{aligned}$$

finalmente se obtiene:

$$\begin{aligned}V(\hat{t}_{A4,\pi}) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1}{a^2}p(1-p)\sum_U \frac{1-y_k}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_U \frac{1-y_k}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_{Y_0} \frac{1}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_{Y_0\cap W_0} \frac{1}{\pi_k} + \frac{1-p}{p}\sum_{Y_0\cap W_1} \frac{1}{\pi_k}.\end{aligned} \tag{4.4}$$

Posteriormente se verá que la expresión anterior puede utilizarse para comparar la varianza del modelo D con la del modelo C.

Modelo M

En el modelo M propuesto por Mangat y Singh (1990) el MA proporciona n respuestas independientes con dos componentes

$\frac{1-p_1}{2} \le p_2 < 1-p_1$, then $V(\hat{t}_{A3,\pi})$ is minimum when $p_2 = \frac{1-p_1}{2}$ and (3.5) becomes

$$V(\hat{t}_{A3,\pi}) = \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1}{4}\frac{1-p_1^2}{p_1^2}\left(\sum_U \frac{1}{\pi_k}\right) \tag{3.6}$$

The D model

This proposal is analogous to the U model with a substantial difference: it belongs to the innocuous W group with a probability of one. For example, the innocuous question W may be "are you alive?" For this model we have

$$Z_k = \begin{cases} y_k & \text{with likelihood } p \\ 1 & \text{with likelihood } 1-p. \end{cases} \tag{4.1}$$

so that

$$E_{MA}(Z_k) = y_k p + (1-p)$$

and

$$\begin{aligned}\hat{Z}_k &= \frac{Z_k - (1-p)}{p} \\ &= aZ_k + b_k,\end{aligned} \tag{4.2}$$

with $a = \frac{1}{p}$ and $b_k = -\frac{1-p}{p}$. With $\hat{t}_{A4,\pi}$ denoting the estimator of the $t_A = \sum_U y_k$ in the D model, from (1.1) we have

$$\begin{aligned}\hat{t}_{A4,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= \frac{1}{p}\sum_S \frac{Z_k}{\pi_k} - \frac{1-p}{p}\sum_S \frac{1}{\pi_k}\end{aligned} \tag{4.3}$$

To obtain the variance of the estimator corresponding to the D model, first we see that

$$\begin{aligned}V_{MA}(Z_k) &= E_{MA}(Z_k^2) - \left[E_{MA}(Z_k)\right]^2 \\ &= E_{MA}(Z_k) - \left[E_{MA}(Z_k)\right]^2 \\ &= py_k + (1-p) - \left[py_k + (1-p)\right]^2 \\ &= p(1-p)(1-y_k)\end{aligned}$$

Thus,

$$\begin{aligned}E_p\left(\sum_S \frac{V_{MA}(Z_k)}{\pi_k^2}\right) &= E_p\left(\sum_S \frac{p(1-p)(1-y_k)}{\pi_k^2}\right) \\ &= p(1-p)\sum_U \frac{1-y_k}{\pi_k},\end{aligned}$$

aleatorias. La primera componente consta de dos proposiciones, seleccionadas con probabilidades t y $1-t$, respectivamente: (1) pertenezco al grupo A, y (2) ir a la segunda componente. La segunda componente también consta de dos proposiciones seleccionadas con probabilidades p y $1-p$: (1) pertenezco al grupo A, y (2) pertenezco al grupo $\overline{A}$. Se tiene entonces

$$Z_k = \begin{cases} y_k & \text{con probabilidad } T \\ Py_k + (1-P)(1-y_k) & \text{con probabilidad } 1-T \end{cases} \tag{5.1}$$

$$\begin{aligned} E_{MA}(Z_k) &= y_k t + (1-t)\left[py_k + (1-p)(1-y_k)\right] \\ &= \left[t + (1-t)(2p-1)\right]y_k + (1-t)(1-p) \end{aligned}$$

$$\begin{aligned} \hat{Z}_k &= \frac{Z_k - (1-t)(1-p)}{\left[t + (1-t)(2p-1)\right]} \\ &= \frac{1}{t + (1-t)(2p-1)} Z_k - \frac{(1-t)(1-p)}{t + (1-t)(2p-1)} \\ &= aZ_k + b_k, \end{aligned} \tag{5.2}$$

con $a = \dfrac{1}{t + (1-t)(2p-1)}$ y $b_k = -\dfrac{(1-t)(1-p)}{t + (1-t)(2p-1)}$. De modo que

$$\begin{aligned} \hat{t}_{A5,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= a\sum_S \frac{Z_k}{\pi_k} + b\sum_S \frac{1}{\pi_k} \\ &= \frac{1}{t + (1-t)(2p-1)} \sum_S \frac{Z_k}{\pi_k} - \frac{(1-t)(1-p)}{t + (1-t)(2p-1)} \sum_S \frac{1}{\pi_k} \end{aligned} \tag{5.3}$$

Para obtener la varianza del estimador para el modelo M, se define $\alpha = \dfrac{1}{a}$ y $\beta = (1-p)(1-t)$, con lo que se obtiene

$$\begin{aligned} V_{MA}(Z_k) &= E\left(Z_k^2\right) - \left[E(Z_k)\right]^2 \\ &= \alpha y_k + \beta - (\alpha y_k + \beta)^2 \\ &= \alpha(1 - \alpha - 2\beta) y_k + \beta(1-\beta) \\ &= \beta(1-\beta), \end{aligned}$$

ya que $1 - \alpha - 2\beta = 0$; de modo que:

$$\begin{aligned} V\left(\hat{t}_{A5,\pi}\right) &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + a^2 E_p\left(\sum_S \frac{V_{MA}(Z_k)}{\pi_k^2}\right) \\ &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \frac{1}{\alpha^2}\beta(1-\beta)\sum_U \frac{1}{\pi_k} \\ &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \frac{\beta(1-\beta)}{(1-2\beta)^2}\sum_U \frac{1}{\pi_k} \end{aligned} \tag{5.4}$$

Nótese que (5.4) es finita sí y sólo sí $\beta \neq 1/2$, lo cual es cierto si $p > 1/2$ y $t > 1/2$.

finally, we obtain

$$\begin{aligned} V\left(\hat{t}_{A4,\pi}\right) &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \frac{1}{a^2} p(1-p) \sum_U \frac{1-y_k}{\pi_k} \\ &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \frac{1-p}{p} \sum_U \frac{1-y_k}{\pi_k} \\ &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \frac{1-p}{p} \sum_{Y_0} \frac{1}{\pi_k} \\ &= \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \frac{1-p}{p} \sum_{Y_0 \cap W_0} \frac{1}{\pi_k} + \frac{1-p}{p} \sum_{Y_0 \cap W_1} \frac{1}{\pi_k}. \end{aligned} \tag{4.4}$$

Below, we will see that the above expression can be used to compare the D model with that of the C model.

The M model

In the M model proposed by Mangat and Singh (1990) the MA provides n independent responses with two random components. The first component comprises two propositions, selected with probabilities t and $1-t$, respectively: (1) I belong to group A, and (2) go to the second component. The second component also comprises two propositions selected with probabilities p and $1-p$: (1) I belong to group A, and (2) I belong to group $\overline{A}$. We thus have

$$Z_k = \begin{cases} y_k & \text{with likelihood } T \\ Py_k + (1-P)(1-y_k) & \text{with likelihood } 1-T \end{cases} \tag{5.1}$$

$$\begin{aligned} E_{MA}(Z_k) &= y_k t + (1-t)\left[py_k + (1-p)(1-y_k)\right] \\ &= \left[t + (1-t)(2p-1)\right]y_k + (1-t)(1-p) \end{aligned}$$

$$\begin{aligned} \hat{Z}_k &= \frac{Z_k - (1-t)(1-p)}{\left[t + (1-t)(2p-1)\right]} \\ &= \frac{1}{t + (1-t)(2p-1)} Z_k - \frac{(1-t)(1-p)}{t + (1-t)(2p-1)} \\ &= aZ_k + b_k, \end{aligned} \tag{5.2}$$

with $a = \dfrac{1}{t + (1-t)(2p-1)}$ and $b_k = -\dfrac{(1-t)(1-p)}{t + (1-t)(2p-1)}$. So that

$$\begin{aligned} \hat{t}_{A5,\pi} &= \sum_S \frac{\hat{Z}_k}{\pi_k} \\ &= a\sum_S \frac{Z_k}{\pi_k} + b\sum_S \frac{1}{\pi_k} \\ &= \frac{1}{t + (1-t)(2p-1)} \sum_S \frac{Z_k}{\pi_k} - \frac{(1-t)(1-p)}{t + (1-t)(2p-1)} \sum_S \frac{1}{\pi_k} \end{aligned} \tag{5.3}$$

To obtain the variance of the estimator for the M model, and $\alpha = \dfrac{1}{a}$ y $\beta = (1-p)(1-t)$ are defined. With these we obtain

Modelo C

Una forma de mejorar la precisión de un estimador es introducir información auxiliar correlacionada con la variable de interés. A diferencia del modelo U, que considera la introducción de una variable inocua no relacionada con la variable sensitiva *y*, en el modelo C la variable inocua *W* está correlacionada con *y*, pero manteniendo la confidencialidad del entrevistado; el procedimiento de muestreo es exactamente como en el modelo U.

Un ejemplo donde se puede aplicar este modelo C es cuando se desea estimar el número total de empresas que evaden impuestos y se utiliza como variable inocua el tamaño de la empresa. Se espera que exista una asociación entre la variable sensible (evadir impuestos) y la variable inocua (tamaño de la empresa), asumiendo que existe un censo de las empresas y su tamaño.

Si se denota por $\hat{t}_{A6,\pi}$ al estimador del total $t_A = \sum_U y_k$ en el modelo C, se tiene que

$$\hat{t}_{A6,\pi} = \frac{1}{p}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{w_k}{\pi_k}\right] \tag{6.1}$$

y

$$\begin{aligned} V\left(\hat{t}_{A6,\pi}\right) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_U \frac{\left(w_k - y_k\right)^2}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_{W_0}\frac{y_k}{\pi_k} + \frac{1-p}{p}\sum_{W_1}\frac{1-y_k}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_{W_0\cap Y_1}\frac{1}{\pi_k} + \frac{1-p}{p}\sum_{W_1\cap Y_0}\frac{1}{\pi_k}. \end{aligned} \tag{6.2}$$

La expresión 6.2 muestra que la varianza del estimador $\hat{t}_{A6,\pi}$ decrece al aumentar la correlación de *W* con *y*. Asimismo, la última igualdad en (6.2) abre la posibilidad de comparar la varianza del estimador de este modelo C con la varianza de los estimadores correspondientes a otros modelos; en particular, con la del D dada por (4.4). En el Cuadro 2 se resumen los resultados.

Estudio de simulación

Se realizó un estudio de simulación para destacar el efecto de la correlación entre la variable *W* y la variable sensitiva *y* en la reducción de la varianza del estimador.

La comparación entre los modelos estudiados se hizo considerando el esquema de muestreo aleatorio simple (MAS). Así, tenemos $p(s) = p(S = s) = \binom{N}{n}^{-1}$ $\forall s \in S$, donde *S* es la colección de todas las posibles muestras; las probabilidades de inclusión son

$$\pi_k = \frac{n}{N} = f; \pi_{kl} = \frac{n(n-1)}{N(N-1)}, k \neq l; \pi_{kk} = \pi_k;$$

$$\Delta_{kl} = -\frac{f(1-f)}{N-1} \text{ para } k \neq l, \Delta_{kk} = f(1-f)$$

$$\begin{aligned} V_{MA}(Z_k) &= E\left(Z_k^2\right) - \left[E(Z_k)\right]^2 \\ &= \alpha y_k + \beta - \left(\alpha y_k + \beta\right)^2 \\ &= \alpha(1-\alpha-2\beta)y_k + \beta(1-\beta) \\ &= \beta(1-\beta), \end{aligned}$$

since $1-\alpha-2\beta = 0$, and so

$$\begin{aligned} V\left(\hat{t}_{A5,\pi}\right) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + a^2 E_p\left(\sum_S \frac{V_{MA}(Z_k)}{\pi_k^2}\right) \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1}{\alpha^2}\beta(1-\beta)\sum_U \frac{1}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{\beta(1-\beta)}{(1-2\beta)^2}\sum_U \frac{1}{\pi_k} \end{aligned} \tag{5.4}$$

Note that (5.4) is finite if and only if $\beta \neq 1/2$. This is true if $p > 1/2$ and $t > 1/2$.

The C model

One way to improve precision of an estimator is to introduce auxiliary information correlated with the variable of interest. Unlike the U model, which considers the introduction of an innocuous variable related to a sensitive variable *y*, in the C model the innocuous variable *W* is correlated with *y*, but maintains the interviewee's privacy. The sampling procedure is exactly like in the U model.

An example of where the C model can be applied is when total number of enterprises that evade taxes is to be estimated, and size of the enterprise is used as the innocuous variable. It is expected that there would be an association between the sensitive variable (tax evasion) and the innocuous variable (size of the enterprise), assuming that there is a census of enterprises and their size.

If the estimator of the total $t_A = \sum_U y_k$ in the C model is denoted $\hat{t}_{A6,\pi}$, we have

$$\hat{t}_{A6,\pi} = \frac{1}{p}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{w_k}{\pi_k}\right] \tag{6.1}$$

and

$$\begin{aligned} V\left(\hat{t}_{A6,\pi}\right) &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_U \frac{\left(w_k - y_k\right)^2}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_{W_0}\frac{y_k}{\pi_k} + \frac{1-p}{p}\sum_{W_1}\frac{1-y_k}{\pi_k} \\ &= \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{1-p}{p}\sum_{W_0\cap Y_1}\frac{1}{\pi_k} + \frac{1-p}{p}\sum_{W_1\cap Y_0}\frac{1}{\pi_k}. \end{aligned} \tag{6.2}$$

Expression 6.2 shows that the variance of the $\hat{t}_{A6,\pi}$ decreases when the correlation between *W* and *y* increases. Also, the last equality in (6.2) opens up the possibility of comparing the variance

Cuadro 2. Ecuaciones para los estimadores y sus varianzas.
Table 2. Equations for the estimators and their variances.

Modelo	Estimador	Varianza del estimador
W	$\frac{1}{2p-1}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{1}{\pi_k}\right]$	$\sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + V_0\left(\sum_U \frac{1}{\pi_k}\right)$
U, C	$\frac{1}{p}\left[\sum_S \frac{Z_k}{\pi_k} - (1-p)\sum_S \frac{w_k}{\pi_k}\right]$	$V_G + \frac{1-p}{p}\sum_U \frac{(w_k - y_k)^2}{\pi_k}$
H	$\frac{1}{p_1}\left[\sum_S \frac{Z_k}{\pi_k} - p_2\sum_S \frac{1}{\pi_k}\right]$	$V_G + \frac{1}{4}\frac{1-p_1^2}{p_1^2}\left(\sum_U \frac{1}{\pi_k}\right)$
D	$\frac{1}{p}\sum_S \frac{Z_k}{\pi_k} - \frac{1-p}{p}\sum_S \frac{1}{\pi_k}$	$V_G + \frac{1-p}{p}\left(\sum_U \frac{1-y_k}{\pi_k}\right)$
M	$\frac{1}{t+(1-t)(2p-1)}\sum_S \frac{Z_k}{\pi_k} - \frac{(1-t)(1-p)}{t+(1-t)(2p-1)}\sum_S \frac{1}{\pi_k}$	$\sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \frac{\beta(1-\beta)}{(1-2\beta)^2}\sum_U \frac{1}{\pi_k}$

Para cada modelo la simulación se realizó para los siguientes parámetros: una población de $N=1000$ individuos, de los cuales $A=702$ tendrán la característica sensible. El tamaño de muestra es $n=100$.

En el estudio de simulación se tiene fija la variable de interés (y) en la población, y se construyeron 100 vectores, cada uno representando un variable inocua W, de tal manera que las correlaciones de W con la variable sensible, van creciendo desde -1 hasta 1, esto permite ver la eficiencia del modelo C; esto es, permite ver la relación inversa que existe entre la varianza del estimador con la correlación de W con y.

Para el modelo H se imponen las restricciones:

$$1 > p = p_1 > 1/2, 0 < 1 - p_1 - p_2 \le p_2 < 1$$
$$-p_1 \Leftrightarrow 1 - p_1 - 2p_2 \le 0, p_2 < 1 - p_1$$

Los valores considerados para los parámetros son $p=p_1=t=0.70$ de donde el valor óptimo para p_2 en el modelo H, de acuerdo con (3.5), es $p_2 = \frac{1-p_1}{2} = 0.15$. Para todos los modelos se realizó una simulación de Monte Carlo con J=1000 iteraciones. Las expresiones de los estimadores para cada modelo son:

$$\hat{t}_{A1,\pi} = \frac{1}{(2p-1)}\left[\frac{1}{f}\sum_S Z_k - N(1-p)\right]$$
$$\hat{t}_{A2,\pi} = \frac{1}{p}\frac{1}{f}\left[\sum_S Z_k - (1-p)\sum_S w_k\right]$$
$$\hat{t}_{A3,\pi} = \frac{1}{p_1}\left[\frac{1}{f}\sum_S Z_k - Np_2\right]$$
$$\hat{t}_{A4,\pi} = \frac{1}{p}\frac{1}{f}\left[\sum_S Z_k - (1-p)n\right]$$
$$\hat{t}_{A5,\pi} = \frac{1}{f}\frac{1}{t+(1-t)(2p-1)}\left[\sum_S Z_k - (1-t)(1-p)n\right]$$

of the estimator of this model C with the variance of the estimators corresponding to other models, particularly that of the D model given by (4.4). The results are summarized in Table 2.

Simulation study

A simulation study was conducted to highlight the effect of the correlation between the W variable and the sensitive variable and on the reduction of the variance of the estimator.

Comparison of the models studied was done considering the simple random sampling (SRS) scheme. Thus we have $p(s) = p(S = s) = \binom{N}{n}^{-1}$ $\forall s \in S$, where S is the collection of all possible samples; the probability of inclusion is

$$\pi_k = \frac{n}{N} = f; \pi_{kl} = \frac{n(n-1)}{N(N-1)}, k \neq l; \pi_{kk} = \pi_k;$$
$$\Delta_{kl} = -\frac{f(1-f)}{N-1} \text{ for } k \neq l, \Delta_{kk} = f(1-f)$$

For each model, simulation was done for the following parameters: a population of $N=1000$ individuals, of which $A=702$ will have the sensitive characteristic. Sample size is $n=100$.

In the simulation study the variable of interest (y) is fixed in the population, and 100 vectors were constructed, each representing an innocuous variable W, such that the correlations between W and the sensitive variable increase from -1 to 1. This shows how efficient the C model is; that is, it exhibits the inverse relationship between the variance of the estimator and the correlation of W and y.

For the H model, the following restrictions are imposed:

$$1 > p = p_1 > 1/2, 0 < 1 - p_1 - p_2 \le p_2 < 1$$
$$-p_1 \Leftrightarrow 1 - p_1 - 2p_2 \le 0, p_2 < 1 - p_1$$

Resultados y Discusión

En el Cuadro 3 se presenta la media y las desviación estándar de las estimaciones simuladas. La media de las estimaciones es cercana al valor verdadero, lo que confirma el insesgamiento de los estimadores.

Se puede observar que con el mismo valor de p, en el mecanismo aleatorio, el modelo menos eficiente en cuanto a la varianza es el modelo Warner. Las varianzas estimadas disminuyen drásticamente al utilizar los modelos H, D, M o C.

La desviación estándar del estimador del total en el modelo C depende del valor de la correlación con la pregunta inocua: si la correlación es negativa, los modelos H, D y M resultan con menor desviación estándar; sin embargo, cuando la correlación es positiva la desviación estándar de los estimadores en el modelo C disminuye y es menor que la correspondiente a los otros modelos.

En la Figura 1 se puede observar que la desviación estándar sigue una relación inversa con la correlación entre la variable sensible y y la pregunta inocua W. La relación entre la correlación y la desviación estándar del estimador en el modelo C (ecuación 6.2) se preserva en otros escenarios de simulación con distintos tamaños de muestra y de población, así como con diferentes parámetros del MA.

Conclusiones

El modelo G generaliza el de respuesta aleatorizada. Los modelos considerados en este trabajo pueden ser vistos como casos particulares. De acuerdo con la expresión de la varianza del estimador del total en el modelo G, ésta se puede dividir en dos partes: la primera, común a todos los modelos RA y la segunda,

Cuadro 3. Resultados de la simulación N=1000, t_A=702, n=100, J=1000.
Table 3. Results of simulation N=1000, t_A=702, n=100, J=1000.

Modelo	$cor(y,W)$	Media	Desviación estándar
W		706.50	122.56
H		704.90	64.87
D		699.63	57.66
M		703.56	65.29
C	0.88	703.69	46.03
C	0.67	701.10	50.17
C	0.46	703.10	52.74
C	0.27	704.76	59.49
C	0.09	701.89	61.67
C	0.09	702.66	66.36
C	0.27	703.54	69.03
C	0.47	700.11	71.57
C	0.66	704.90	75.29
C	0.88	705.24	80.41

The values considered for the parameters are $p=p_1=t=0.70$, from which we obtain the optimum value for p_2 in the H model, according to (3.5) is $p_2 = \frac{1-p_1}{2} = 0.15$. For all of the models, a Monte Carlos simulation was conducted with J=1000 iterations. The expressions of the estimators for each model are the following:

$$\hat{t}_{A1,\pi} = \frac{1}{(2p-1)}\left[\frac{1}{f}\sum_S Z_k - N(1-p)\right]$$

$$\hat{t}_{A2,\pi} = \frac{1}{p}\frac{1}{f}\left[\sum_S Z_k - (1-p)\sum_S w_k\right]$$

$$\hat{t}_{A3,\pi} = \frac{1}{p_1}\left[\frac{1}{f}\sum_S Z_k - Np_2\right]$$

$$\hat{t}_{A4,\pi} = \frac{1}{p}\frac{1}{f}\left[\sum_S Z_k - (1-p)n\right]$$

$$\hat{t}_{A5,\pi} = \frac{1}{f}\frac{1}{t+(1-t)(2p-1)}\left[\sum_S Z_k - (1-t)(1-p)n\right]$$

Results and Discussion

Table 3 presents the mean and standard deviation of the simulated estimations. The mean of the estimations is close to the true value, confirming the unbiasedness of the estimators.

It can be observed that with the same p value in the random mechanism, the least efficient model, in terms of variance, is the Warner model. The estimated variances decrease drastically when H, D, M or C models are used.

The standard deviation of the estimator of the total in the C model depends on the value of its correlation with the innocuous question: if correlation is negative, the H, D, and M models have a smaller standard deviation. However, when the correlation is positive, the standard deviation of the C model estimators decreases and is lower than that of the other models.

In Figure 1, it can be observed that the standard deviation is inversely related to the correlation between the sensitive variable and the innocuous question (W). The relationship between correlation and the standard deviation of the estimator in the C model (Equation 6.2) is conserved in the other simulation scenarios that have different sample and population sizes, as well as with different MA parameters.

Conclusions

The G model generalizes randomized response sampling. The models considered in this paper can be seen as particular cases. In keeping with the expression of the variance of the estimator of the total in the G model, this can be divided into two parts: the first,

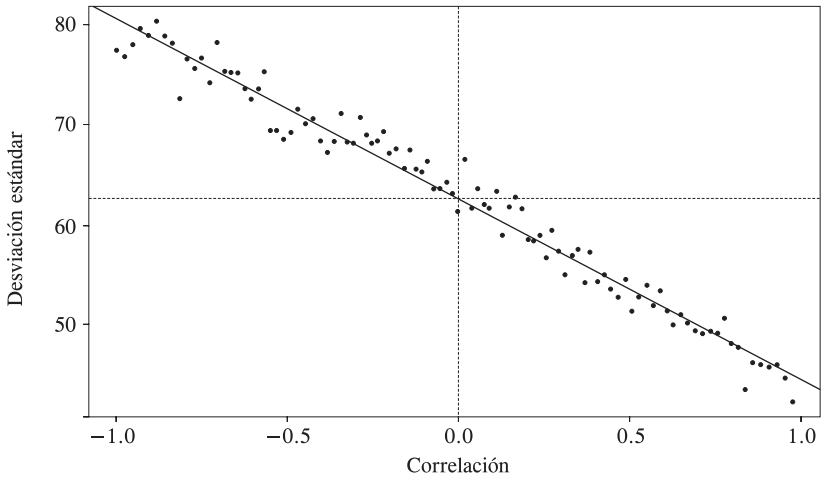

Figura 1. Relación entre cor(y,W) y la desviación estándar del estimador.
Figure 1. Relationship between cor(y,W) and the standard deviation of the estimator.

que depende del MA. Por lo anterior, la comparación de la varianza entre dos modelos sólo depende de la segunda componente de la varianza.

La expresión (6.2) indica la existencia de una relación inversa entre la varianza del estimador del total y la correlación entre las variables sensible e inocua en el modelo C. En particular, cuando la correlación es uno la precisión de estimador es máxima, ya que el segundo término de la expresión de la varianza se anula.

El estudio de simulación muestra que el estimador $\hat{t}_{A6,\pi}$ del modelo C tiene menor varianza que los demás. Ésto es, se logra una reducción muy importante en la varianza del estimador $\hat{t}_{A2,\pi}$ al considerar una variable W inocua, pero altamente correlacionada con la variable sensible y.

Literatura Citada

Bar-Lev, S. K., E. Bobovich, and B. Boukai. 2003. A Common conjugate prior structure for several randomized response models. Test 12: 101-113.

Cassel, C. M., J. K. Wretman, and C. E. Särndal. 1977. Foundations of Inference in Survey Sampling. J. Wiley. New York. 192 p.

Chaudhuri, A. 2001. Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. Journal of Statistical Planning and Inference 94: 37-42.

Chaudhuri, A., and R. Mukerjee. 1988. Randomized Response Theory and Technique. Marcel Dekker. New York. 162 p.

Chua, T. C., and A. K. Tsui. 2000. Procuring honest responses indirectly. Journal of Statistical Planning and Inference 90: 107-116.

Devore, J. L. 1977. A note on the randomized response technique. Communications in Statistics Theory and Methods 6: 1525-1529.

Greenberg, B. G., A. A. Abul-ela, W. R. Simmons, and D. C. Horvitz. 1969. The unrelated question RR model: theoretical framework. JASA 64: 520-539.

Horvitz, D. C., B. G. Greenberg, and J. R. Abernathy. 1976. Randomized response. A data gathering device for sensitive questions. International Statistical Review 44: 181-196.

common to all of the RR models, and the second, depending on the MA. Thus, comparison between the variances of two models depends only on the second component of variance.

Expression (6.2) indicates the existence of an inverse relationship between the variance of the estimator of the total and the correlation between the sensitive and the innocuous variables in the C model. Particularly, when the correlation is one, precision of the estimator is maximum since the second term of the expression of the variance is nullified.

The simulation study shows that the estimator $\hat{t}_{A6,\pi}$ of the C model has smaller variance than the other models. That is, a major reduction in the variance of the estimator $\hat{t}_{A2,\pi}$ is achieved when an innocuous variable W is highly correlated with the sensitive variable y.

—End of the English version—

Lakshmi, D. V., and D. Raghavarao. 1992. A test for detecting untruthful answering in randomized response procedures. Journal of Statistical Planning and Inference 31: 387-390.

Mangat, N. S., and R. Singh. 1990. An alternative randomized response procedure. Biometrika 77: 439-442.

Mangat, N. S., R. Singh, S. Singh and B. Singh. 1993. On Moors' randomized response model. Biometrical Journal 35: 727-732.

Moors, J. J. 1971. Optimization of the unrelated question in RR model. JASA 66: 627-629.

Padmawar, V. R., and K. Vijayan. 2000. Randomized response revisited. Journal of Statistical Planning and Inference 90: 293-304.

Särndal, C. E., B. Swensson, and J. Wretman. 1992. Model Assisted Survey Sampling. Springer Verlag. New York. 694 p.

Warner, S. L. 1965. Randomized response: A survey technique for eliminating evasive answer bias. JASA 60: 63-69.

Winkler, R. L., and L. A. Franklin, 1979. Warner's randomized response model: A Bayesian approach. JASA 74: 207-214.