Survey Methodology, June 1995
Vol. 21, No. 1, pp. 13–19
Statistics Canada

Estimation of Correlation in Randomized Response

D.R. BELLHOUSE[1]

ABSTRACT

Stanley Warner's contributions to randomized response are reviewed. Following this review, a linear model, based on random permutation models, is developed to include many known randomized response designs as special cases. Under this model optimal estimators for finite population variances and covariances are obtained within a general class of quadratic design-unbiased estimators. From these results an estimator of the finite population correlation is obtained. Three randomized response designs are examined in particular: (i) the unrelated questions model of Greenberg *et al.* (1969); (ii) the additive constants model of Pollock and Bek (1976); and (iii) the multiplicative constants model of Pollock and Bek (1976). Simple models for response bias are presented to illustrate the effect of this bias on estimation of the correlation.

KEY WORDS: Additive constants model; Linear models; Multiplicative constants model; Response bias; Unrelated question model; Variance estimation.

1. A BRIEF OVERVIEW OF WARNER'S CONTRIBUTIONS TO RANDOMIZED RESPONSE

Randomized response is a technique used to elicit responses to sensitive questions. It was developed thirty years ago by Stanley Warner (Warner 1965) to estimate a proportion under a simple random sampling design with replacement. The development was a substantial intellectual achievement requiring much originality of thought. How does one get truthful responses to sensitive questions? Warner's solution was to get the response without the interviewer knowing whether the sensitive question had actually been asked. He devised the probabilistic structure to the questioning so that an estimate of the required proportion could be obtained. In Warner's original formulation the population is divided into two mutually exclusive and exhaustive groups, A and B. It is of interest to estimate the proportion π of the population belonging to group A. To do this, a spinner is constructed with a face marked with the letters A and B. The construction is such that the spinner points to the letter A with probability p and to B with probability $1 - p$. The interviewee spins the spinner and is required only to say yes or no according to whether or not the spinner points to the interviewee's correct membership group. The with replacement design allows estimation of π by maximum likelihood.

This very original idea has received substantial attention over the past thirty years. Since Warner's original work, several randomized response techniques have been suggested for the estimation of a proportion or set of proportions as in polytomous data, or for the estimation of a population mean with continuous data. A variation on Warner's original theme is asking the sensitive question or an unrelated question with probabilities p and $1 - p$ respectively. This was originally due to Greenberg *et al.* (1969). Other variations with continuous data include adding a random variable to the response to the sensitive question or multiplying the response by a random variable. The underlying theme to any of these techniques is the masking of the original response in such a way that the sensitive information cannot be attributed to any single respondent but that information on the sensitive attribute can be extracted from the whole sample. A substantial literature, including a monograph by Chaudhuri and Mukerjee (1988), has grown up around these techniques. Nathan (1988) has provided a fairly comprehensive bibliography of this literature. Umesh and Peterson (1991) have given several detailed examples from very diverse areas of the application and applicability of the techniques of randomized response.

With several different randomized response techniques, the question arises as to how to compare the different methods. Minimization of variance cannot be the sole criterion. Each method is designed to protect the privacy of the respondent. A gain in efficiency, in terms of variance, by the choice of different values of the probabilities in the randomizing device, or by the choice of one randomized response method over another, could lead to jeopardizing the privacy of the respondents. In response to this, Leysieffer and Warner (1976) and Warner (1976) formulated natural measures of respondent jeopardy. These measures are related to the probability of the interviewer being able to infer the interviewee's response to the sensitive attribute. The theory of respondent jeopardy is reviewed in Chaudhuri and Mukerjee (1988) and some practical considerations regarding respondent jeopardy are reviewed in Umesh and Peterson (1991).

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7.

Stanley Warner made two other contributions to the literature of randomized response. The first contribution is directly related to the results obtained here. With the explosion of new ideas and new techniques in randomized response, Warner (1971) formulated a linear model which unified the theory. Most of the randomized response techniques at that time could be put in his linear model framework. The second contribution was in response to the growing use of telephone interviewing. Stem and Steinhorst (1984) described randomized response methods applicable to telephone interviewing and to mail questionnaires. Warner (1986) suggested practical natural randomizing devices, such as the serial numbers on paper money, for use in telephone interviewing.

The major topics in randomized response methodology are: the development of randomized response techniques, the comparison of these techniques through the concept of respondent jeopardy, the construction of reasonable randomizing devices, the development of a unified theory of randomized response, and the validation of randomized response techniques through field studies. Stanley Warner's contributions to randomized response touch on most of these major developments in the subject. Moreover, most of these contributions were substantial and influential. He is the originator of the technique. His original setup of a dichotomous population was quickly generalized to a polytomous one and to populations with continuous measurement. New randomized response techniques continue to be developed. Warner was at the forefront of evaluating randomized response designs through the modeling of respondent jeopardy. His work in the development of a unified linear model for randomized response designs was the foundation on which a unified theory of randomized response has been built.

2. INTRODUCTION TO ESTIMATION OF CORRELATION

Consider a finite population of size N with two measurements of interest x_j and y_j for $j = 1, \ldots, N$. It is of interest to estimate the finite population correlation

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

where $\sigma_{xy} = \sum (x_j - \bar{X})(y_j - \bar{Y})/N$ is the finite population covariance between the variables x and y and where σ_x^2 and σ_y^2 are the finite population variances of the variables x and y respectively. To estimate ρ a sample of fixed size n is chosen with probability $P(s)$ from the finite population where s denotes the set of finite population units chosen for the sample. The expectation operator with respect to the sampling design $P(s)$ is denoted by E_p. Estimators for ρ are obtained by replacing σ_x^2, σ_y^2 and σ_{xy} by their respective estimators, unbiased or biased, optimal in some sense or otherwise.

To illustrate the general results obtained here for estimation of the finite population correlation coefficient, three particular randomized response techniques will be considered:

(i) The unrelated questions model due to Greenberg *et al.* (1969). The sensitive question is asked with probability p and an unrelated question which is not sensitive is asked with probability $1 - p$. For estimation of the mean it is assumed that the finite population mean $\bar{X}$ of the unrelated question is known. For estimation of variance it is also assumed that σ_x^2 is known.

(ii) The additive constants model due to Pollock and Bek (1976). The outcome of a random variable from a known probability distribution is added to the value of the response to the sensitive question.

(iii) The multiplicative constants model due to Pollock and Bek (1976). The value of the response to the sensitive question is multiplied by the outcome of a random variable from a known probability distribution.

Edgell *et al.* (1986) have provided estimators for ρ under the unrelated questions model and the additive constants model.

Most randomized response designs that have been considered have assumed that the sampling design is simple random sampling either with or without replacement. Since the results obtained here are under a fixed size design, the simple random sampling design assumed here is without replacement.

Assume that both x and y are sensitive variables. Consequently, a randomized response technique is used to obtain information on both these variables. Let w_j and z_j, for $j \in s$ be the sampled measurements that are obtained. Let u_j and v_j for $j = 1, \ldots, N$ be the nonsensitive measurements associated with x_j and y_j respectively. Under the unrelated question model (randomized response model (i)) u_j and v_j are the responses to the unrelated questions for the j-th individual. Under the additive constants model or the multiplicative constants model (randomized response models (ii) or (iii)) u_j and v_j are the j-th outcomes of random variables from two, possibly different, known probability distributions.

3. RANDOM PERMUTATION MODELS

Several models for the finite population measurements have been put forward in the survey sampling literature. Here attention is focused on the random permutation models of Rao (1975) and Rao and Bellhouse (1978). One compelling reason for using these models is that the model parameters have a direct interpretation in the finite population of interest since model parameters in random

permutation models are also finite population parameters. In the simplest context for random permutation models it is assumed that the N-dimensional vector of finite population measurements is a random permutation of an N-dimensional vector of fixed numbers. Rao (1975) has shown how this assumption leads to a linear model. Bellhouse (1980) extended this model to randomized response designs under unequal probability sampling.

The model and associated designs applicable to unequal probability sampling are not easily applicable to estimation of variances and covariances either with or without a randomized response. Consequently, a special case of the model in Bellhouse (1980) is given here. In the model which follows there are two different expectation operators at work which together yield a composite expectation E_m. These expectation operators are: E_r, the expectation operator with respect to the randomizing device, and E_{rp}, the expectation operator with respect to the random permutation model. The composite expectations $E_m = E_{rp}E_r$ and $E = E_mE_p$. For the random permutation model we assume that the pairs (x_j, y_j), $j = 1, \ldots, N$ are a random permutation of a set of N fixed pairs of numbers, say $(p_j, q_j), j = 1, \ldots, N$. This is a special case of model (4.1) in Rao and Bellhouse (1978); the more general model in Rao and Bellhouse (1978) was used in double sampling and sampling on two occasions. The unrelated questions randomized response model (randomized response model (i)) requires an additional assumption that the quadruples $(x_j, y_j, u_j, v_j), j = 1, \ldots, N$ are a random permutation of a set of N fixed quadruples of numbers, say $(p_j, q_j, r_j, t_j), j = 1, \ldots, N$.

Assume that the randomizing device coupled with the random permutation model leads to the following linear model:

$$\begin{aligned} w_j &= \alpha_1 + \beta_1 \bar{X} + e_{1j} \\ z_j &= \alpha_2 + \beta_2 \bar{Y} + e_{2j}, \end{aligned} \tag{1}$$

for $j = 1, \ldots, N$ where $\bar{X}$ and $\bar{Y}$ are the finite population means of the x and y measurements respectively and where for $j = 1, \ldots, N$

$$E_m(e_{1j}) = E_m(e_{2j}) = 0,$$

$$E_m(e_{1j}^2) = \phi_1\sigma_x^2 + \psi_{01} + \psi_{11}\bar{X} + \psi_{21}\bar{X}^2,$$

$$E_m(e_{2j}^2) = \phi_2\sigma_y^2 + \psi_{02} + \psi_{12}\bar{Y} + \psi_{22}\bar{Y}^2,$$

$$E_m(e_{1j}\, e_{1k}) = \delta_1\sigma_x^2 + \lambda_1, E_m(e_{2j}\, e_{2k}) = \delta_2\sigma_y^2 + \lambda_2, \quad \text{for} \quad j \neq k,$$

$$E_m(e_{1j}\, e_{2j}) = \phi_3\sigma_{xy} + \psi_3, \quad \text{and}$$

$$E_m(e_{1j}\, e_{2k}) = \delta_3\sigma_{xy} + \lambda_3, \quad \text{for} \quad j \neq k. \tag{2}$$

and all other higher moments are independent of j. In the model given by (1) and (2), the α's, λ's, ϕ's, ψ's and δ's are all known constants. The finite populations variances and covariances of the sensitive questions, σ_x^2, σ_y^2 and σ_{xy} are all unknown.

For the unrelated questions model (randomized response model (i)) assume that the randomizing schemes on the two sensitive questions are independent and that sensitive question i, $i = 1, 2$, is asked with probability p_i and the associated nonsensitive questions with probability $1 - p_i$. Assume further that the sensitive questions are unrelated to the nonsensitive questions so that $\sigma_{xu} = \sigma_{yv} = \sigma_{xv} = \sigma_{yu} = 0$. This assumption is unnecessary under simple random sampling with replacement. When, in addition, a random permutation model is assumed on the quadruple (x_j, y_j, u_j, v_j) then in the model given by (1) and (2):

$$\alpha_1 = (1 - p_1)\bar{U}, \beta_1 = p_1, \alpha_2 = (1 - p_2)\bar{V}, \beta_2 = p_2,$$

$$\phi_1 = p_1, \psi_{01} = (1 - p_1)\sigma_u^2 + p_1(1 - p_1)\bar{U}^2,$$
$$\psi_{11} = -2p_1(1 - p_1)\bar{U}, \psi_{21} = p_1(1 - p_1),$$

$$\phi_2 = p_2, \psi_{02} = (1 - p_2)\sigma_v^2 + p_2(1 - p_2)\bar{V}^2,$$
$$\psi_{12} = -2p_2(1 - p_2)\bar{V}, \psi_{22} = p_2(1 - p_2),$$

$$\delta_1 = -p_1^2/(N - 1), \lambda_1 = -(1 - p_1)^2\sigma_u^2/(N - 1),$$

$$\delta_2 = -p_2^2/(N - 1), \lambda_2 = -(1 - p_2)^2\sigma_v^2/(N - 1),$$

$$\phi_3 = p_1p_2, \delta_3 = -\phi_3/(N - 1),$$
$$\psi_3 = (1 - p_1)(1 - p_2)\sigma_{uv},$$
$$\text{and} \quad \lambda_3 = -\psi_3/(N - 1). \tag{3}$$

Note that the model assumptions require that the finite population variance-covariance matrix of the nonsensitive questions is known as well as the finite population means.

For the additive constants model (randomized response model (ii)) assume that the random variables u and v that are added to the value of the responses to the two sensitive questions are independent with means μ_u and μ_v and variances σ_u^2 and σ_v^2 respectively. When the random permutation model is assumed on the pair (x_j, y_j) then in the model given by (1) and (2):

$$\alpha_1 = \mu_u, \beta_1 = 1, \alpha_2 = \mu_v, \beta_2 = 1,$$

$$\phi_1 = \phi_2 = \phi_3 = 1, \psi_{01} = \sigma_u^2, \psi_{02} = \sigma_v^2,$$

$$\delta_1 = \delta_2 = \delta_3 = -1/(N - 1),$$

$$\psi_{11} = \psi_{21} = \psi_{12} = \psi_{22} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0. \tag{4}$$

In the multiplicative constants model, two independent random variables, u and v with means μ_u and μ_v and variances σ_u^2 and σ_v^2 respectively, are multiplied respectively by the value of the response on the x-variable and the y-variable. When the random permutation model is assumed on the pair (x_j, y_j) then in the model given by (1) and (2):

$$\alpha_1 = \alpha_2 = 0, \beta_1 = \mu_u, \beta_2 = \mu_v,$$

$$\phi_1 = \mu_u^2 + \sigma_u^2, \phi_2 = \mu_v^2 + \sigma_v^2, \phi_3 = \mu_v \mu_v,$$

$$\psi_{21} = \sigma_u^2, \psi_{22} = \sigma_v^2,$$

$$\delta_1 = -\mu_u^2/(N-1), \delta_2 = -\mu_v^2/(N-1),$$
$$\delta_3 = -\mu_u\mu_v/(N-1), \text{ and}$$

$$\psi_{01} = \psi_{11} = \psi_{02} = \psi_{12} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0. \quad (5)$$

4. ESTIMATION OF VARIANCE AND COVARIANCE

Consider estimation of σ_y^2 so that the appropriate data are z_j for units $j \in s$. The general class of quadratic estimators of σ_y^2 is of the form:

$$e_{bs} = b_{s..} + \sum_{j\in s} b_{sj.} z_j + \sum_{j\in s} b_{sjj} z_j^2 + \sum\sum_{i\neq j\in s} b_{sij} z_i z_j, \quad (6)$$

where the coefficients of the z's are defined for all s, all $j \in s$ and all pairs $(i, j) \in s$.

In the context of randomized response, an estimator e_b in the class defined by (6) is design-unbiased for σ_y^2 if $E_p E_r(e_b) = \sigma_y^2$ and is *pm*-unbiased if $E(e_b) = \sigma_y^2$. Conditions under which an estimator e_b is *pm*-unbiased are obtained upon taking the expectation E of (6) under (1) and (2). On equating coefficients in $\bar{Y}^0$, $\bar{Y}^1$, $\bar{Y}^2$ and σ_y^2 four equations in four unknowns are obtained. The solution to these four equations yields the following conditions under which estimators in the class defined by (6) are *pm*-biased for σ_y^2:

$$E_p\left(\sum_{j\in s} b_{sjj}\right) = \frac{\beta_2^2}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = A_2, \quad (7)$$

$$E_p\left(\sum\sum_{i\neq j\in s} b_{sij}\right) = -\frac{\beta_2^2 + \psi_{22}}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = -(A_2 + B_2), \quad (8)$$

$$E_p\left(\sum_{j\in s} b_{sj.}\right) = \frac{(2\alpha_2\psi_{22} - \beta_2\psi_{12})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = C_2, \quad (9)$$

and

$$E_p(b_{s..}) = \frac{\lambda_2(\beta_2^2 + \psi_{22}) - (\alpha_2^2\psi_{22} - \alpha_2\beta_2\psi_{12} + \beta_2^2\psi_{02})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = D_2. \quad (10)$$

In order to obtain the optimal estimator we need to define an associated class of quadratic estimators of 0. This is given by

$$e_{cs} = c_{s..} + \sum_{j\in s} c_{sj.} z_j + \sum_{j\in s} c_{sjj} z_j^2 + \sum\sum_{i\neq j\in s} c_{sij} z_i z_j.$$

The conditions for an estimator e_c in this class to be *pm*-biased for 0 are

$$E_p(c_{s..}) = E_p\left(\sum_{j\in s} c_{sj.}\right) = E_p\left(\sum_{j\in s} c_{sjj}\right) = E_p\left(\sum\sum_{i\neq j\in s} c_{sij}\right) = 0. \quad (11)$$

Derivation of the minimum variance quadratic design-unbiased estimator of σ_y^2 follows along the same lines as that used for the finite population mean by Rao and Bellhouse (1978) for cases without randomized response and by Bellhouse (1980) for cases with randomized response. The covariance $E(e_b e_c)$ under the composite expectation is determined under the model such that only expectations of the form E_p remain to be determined. From this expression the coefficients b are set to make $E(e_b e_c) = 0$ under the conditions in (11). The values of the coefficients b are then determined from the conditions in (7) through (10). From a theorem on minimum variance unbiased estimation of Rao (1952), the resulting estimator is the optimal *pm*-unbiased estimator of σ_y^2. If there exists a design such that this estimator is also design-unbiased for σ_y^2, then by arguments similar to those given in Theorem (2.4) of Rao and Bellhouse (1978), the estimator is also the optimal design-unbiased estimator of σ_y^2. We present results for *pm*-unbiased estimators first (Theorems 1 and 2) and then present results for design-unbiased estimators under the three randomized response schemes.

Theorem 1. Under the model defined by (1) and (2) and for any design of fixed size n, the *pm*-variance of e_b, $E_{rp}[E_p E_r(e_b - \sigma_y^2)^2] = E(e_b - \sigma_y^2)^2$, is minimized for the estimator given by

$$(A_2 + B_2)s_z^2 - B_2\frac{1}{n}\sum_{j\in s} z_j^2 + C_2\bar{z} + D_2, \quad (12)$$

where $\bar{z}$ is the sample mean of the data and

$$s_z^2 = \frac{1}{n-1}\sum_{j\in s}(z_j - \bar{z})^2$$

is the sample variance of the data obtained through randomized response where A_2, B_2, C_2 and D_2 are defined in (7) through (10) respectively.

Proof. Under the model given by (1) and (2) the covariance $E(e_b e_c)$ is algebraically quite lengthy but may be expressed in the following form:

$$\boldsymbol{b}^T\boldsymbol{G}\boldsymbol{c} + H, \tag{13}$$

where $\boldsymbol{b}^T$ is the vector

$$\left[E_p(b_{s..}), E_p\left(\sum_{j\in s} b_{sj.}\right), E_p\left(\sum_{j\in s} b_{sjj}\right), E_p\left(\sum\sum_{i\neq j\in s} b_{sij}\right)\right], \tag{14}$$

and $\boldsymbol{c}^T$ is the same as (14) with the b's replaced by c's. The 4×4 matrix $\boldsymbol{G}$ in (13) contains functions of the first order moments of z_j and the second order moments of e_{2j} in (1). The expression H in (13) is a sum of terms of the form

$$\kappa \sum b_{sij}\, c_{skl}, \tag{15}$$

where the summation symbol is up to a quadruple sum, where the subscripts of b could be replaced by a dot (.) and where κ is a function of second through fourth order moments of e_{2j} in (1). Note that these moments are all independent of j. In (15) the sum is a single sum over $j \in s$ when, for example, the subscripts $i = j = k = l$ or when $i = k$ and j and l are replaced by dots. The sum is a double sum over $i \neq k \in s$ when, for example, $i \neq k$ and j and l are replaced by dots. This process continues to the quadruple sum in which $i \neq j \neq k \neq l$. From (11) $E(e_b e_c)$ reduces to 0 if $b_{s..} = h_1$, $b_{sj} = h_2$, $b_{sjj} = h_3$, and $b_{sij} = h_4$, where the h_i are constants. From (7) through (10) and the fact that the design is of fixed size we obtain

$$b_{s..} = D_2,\ b_{sj.} = C_2/n,\ b_{sij} = -\frac{A_2 + B_2}{n(n-1)},\ b_{sjj} = A_2/n,$$

so that the estimator in (12) minimizes the variance in the *pm*-unbiased class of quadratic estimators of σ_y^2. Q.E.D.

By the same arguments

$$(A_1 + B_1)s_w^2 - B_1\frac{1}{n}\sum_{j\in s} w_j^2 + C_1\bar{w} + D_1, \tag{16}$$

is the optimal *pm*-unbiased estimator for σ_x^2 where

$$A_1 = \frac{\beta_1^2}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}},$$

$$B_1 = \frac{\beta_1^2 + \psi_{21}}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}},$$

$$C_1 = \frac{(2\alpha_1\psi_{21} - \beta_2\psi_{11})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}}, \quad \text{and}$$

$$D_1 = \frac{\lambda_1(\beta_1^2 + \psi_{21}) - (\alpha_1^2\psi_{21} - \alpha_1\beta_1\psi_{11} + \beta_1^2\psi_{01})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1\psi_{21}}.$$

The same technique can be used to estimate the covariance σ_{xy}. The general class of quadratic estimators of σ_{xy} is of the form

$$e_{ds} = d_s + \sum_{j\in s} d_{1sj}\, z_j + \sum_{j\in s} d_{2sj}\, w_j + \sum\sum_{i\neq j\in s} d_{sij}\, w_i z_j,$$

where the coefficients of the w's and z's are defined for all s, all $j \in s$ and all pairs $(i, j) \in s$. The result on the covariance is stated without proof in

Theorem 2. Under the model defined by (1) and (2) and for any design of fixed size n, the *pm*-variance of e_d, $E_{rp}[E_p E_r(e_d - \sigma_{xy})^2] = E(e_d - \sigma_{xy})^2$, is minimized for the estimator given by

$$\frac{s_{wz} - (\psi_3 - \lambda_3)}{\phi_3 - \delta_3}, \tag{17}$$

where

$$s_{wz} = \frac{1}{n-1}\sum_{j\in s}(w_j - \bar{w})(z_j - \bar{z})$$

is the sample covariance between w and z.

An estimator for ρ is obtained from (12), (16) and (17). In the additive constants randomized response model (randomized response model (ii)) the estimator of ρ is given by

$$\hat{\rho}_{ac} = \frac{s_{wz}}{\sqrt{(s_w^2 - \sigma_u^2)(s_z^2 - \sigma_v^2)}}. \tag{18}$$

This is the same as the estimator obtained by Edgell *et al.* (1986). Under the multiplicative constants model (randomized response model (iii)) the estimator reduces to

$$\hat{\rho}_{mc} = \frac{s_{wz}}{\sqrt{s_w^2 - \frac{\sigma_u^2/\mu_u^2}{1 + \sigma_u^2/\mu_u^2}\frac{1}{n}\sum_{j\in s} w_j^2}\sqrt{s_z^2 - \frac{\sigma_v^2/\mu_v^2}{1 + \sigma_v^2/\mu_v^2}\frac{1}{n}\sum_{j\in s} z_j^2}}, \tag{19}$$

for $\mu_u \neq 0$ and $\mu_v \neq 0$. When $\mu_u = 0$ the coefficient of $\sum w_j^2$ is $1/n$ and when $\mu_v \neq 0$ the coefficient of $\sum z_j^2$ is $1/n$. The estimator for ρ under the unrelated questions model (randomized response model (i)) is

$$\hat{\rho}_{uq} = \frac{s_{wz} - \frac{(1-p_1)(1-p_2)}{p_1 p_2} S_{uv}}{\sqrt{\hat{S}_x^2 \hat{S}_y^2}}, \tag{20}$$

where $S_{uv} = N\sigma_{uv}/(N-1)$ and where

$$\hat{S}_x^2 = s_w^2 - (1-p_1)\frac{1}{n}\sum_{j\in s} w_j^2 + 2(1-p_1)\bar{U}\bar{w} - (1-p_1)\bar{U}^2 - (1-p_1)\sigma_u^2\left(p_1 + \frac{1-p_1}{N-1}\right)$$

and

$$\hat{S}_y^2 = s_z^2 - (1-p_2)\frac{1}{n}\sum_{j\in s} z_j^2 + 2(1-p_2)\bar{V}\bar{z} - (1-p_2)\bar{V}^2 - (1-p_2)\sigma_v^2\left(p_2 + \frac{1-p_2}{N-1}\right).$$

When $p_1 = p_2$ this may be compared to the estimator in Edgell *et al.* (1986). The resulting estimator for $\hat{\rho}_{uq}$ differs from the estimator in Edgell *et al.* (1986) who assume that $\sigma_{uv} = 0$. They also use biased estimators of σ_x^2 and σ_y^2. Edgell *et al.*'s estimator for σ_y^2 is obtained by writing the design variance of $\bar{z}$ under simple random sampling with replacement as

$$\sigma_z^2/n = \sum_{j=1}^{N} (z_j - \bar{Z})^2/(Nn). \tag{21}$$

The design variance of $\bar{z}$ under the randomizing device is

$$[p_2\sigma_y^2 + (1-p_2)\sigma_v^2 + p_2(1-p_2)(\bar{Y} - \bar{V})^2]/n. \tag{22}$$

Expression (22) is found in Greenberg *et al.* (1971). The estimator for σ_y^2 is found by equating (22) to the left hand side of (21), by substituting sample the estimator of σ_z^2 and the randomized response estimator of $\bar{Y}$ in the resulting equation, and then by solving for σ_y^2.

Each of the estimators of the finite population variances and covariance, which are the components of $\hat{\rho}$ in (18), (19) and (20), are design-unbiased under the appropriate randomized response model for any design with joint inclusion probability for units i and j given by $\pi_{ij} = n(n-1)/[N(N-1)]$. Consequently, each estimator is the optimal design-unbiased estimator for its finite population paramater counterpart. To obtain the appropriate unbiased estimators in (18), multiply the numerator and denominator each by $(N-1)/N$. The resulting numerator is design-unbiased for σ_{xy} and the expressions under the square root sign in the denominator of (18) are unbiased for σ_x^2 and σ_y^2. In (19) it is necessary to multiply the numerator and denominator by $(N-1)/[N\mu_u\mu_v]$ in order to obtain the correct form of the design-unbiased estimators. The correct estimators are obtained in (20) when the multiplier is $(N-1)/(Np_1p_2)$.

In any of the randomized response designs, the simplest estimate of the variance of $\hat{\rho}$ is the jackknife estimate of variance. Jackknife estimates of variance for $\hat{\rho}$ can be obtained from formulae (4.2.3) or (4.2.5) in Wolter (1985).

5. EFFECT OF RESPONSE BIAS

In the additive constants model, the respondent is asked to add a random variable u to x and an independent random variable v to y. Instead, the respondent may add different independent random variables, say u' and v'. The means and variances of u' and v' may differ from those of u and v. It is reasonable to assume, however, that $\sigma_{u'}^2 \geq \sigma_u^2$ and $\sigma_{v'}^2 \geq \sigma_v^2$. One example in which this situation might occur is the following. The respondent does not want to add on the outcome of a random variable near to the mean of the distribution of the random variable. In this case the distribution of response bias could be modelled by the original distribution with an interval around the mean in which any outcome from the original distribution which falls in the chosen interval is set to one of the end points of the interval. On taking separately the expectations of the numerator and the expression under each of the square root signs in the denominator of (18) the expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 + \sigma_{u'}^2 - \sigma_u^2}\sqrt{\sigma_y^2 + \sigma_{v'}^2 - \sigma_v^2}}, \tag{23}$$

is obtained. From (23) it may be noted that the response bias leads to an estimate of correlation lower than the true value.

The multiplicative constants model is the same as the additive constants model with the exception that the responses to the sensitive questions are multiplied by the random variables. As in the response bias model for additive constants, assume that u' and v' are used by the

respondent instead of u and v. Then on taking separately the expectations of the numerator and the expressions under each of the square root signs in the denominator of (19) the expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2+\frac{\sum_{j=1}^{N}x_j^2}{N\mu_{u'}^2}\,\frac{\sigma_u^2\mu_{u'}^2-\sigma_{u'}^2\mu_u^2}{\sigma_u^2+\mu_u^2}}\sqrt{\sigma_y^2+\frac{\sum_{j=1}^{N}y_j^2}{N\mu_{v'}^2}\,\frac{\sigma_v^2\mu_{v'}^2-\sigma_{v'}^2\mu_v^2}{\sigma_v^2+\mu_v^2}}}, \tag{24}$$

is obtained. If $\mu_u = \mu_{u'}$, $\mu_v = \mu_{v'}$, $\sigma_{u'}^2 \geq \sigma_u^2$ and $\sigma_{v'}^2 \geq \sigma_v^2$, as in the case of the additive constants model, then from (24) the response bias leads to an overestimate of the correlation.

In the unrelated questions model a reasonable model for response bias is to assume that the sensitive questions are answered with probability $p_1' < p_1$ and $p_2' < p_2$. In general the effect of this response bias is dependent on the relative values of the various probabilities, the means and variances of the sensitive questions, and the means and variances of the nonsensitive questions. Under simple random sampling without replacement and the response bias model, the design expectation of the numerator of (20) is given by

$$p_1'p_2'\left[S_{xy}+\frac{(1-p_1')(1-p_2')-(1-p_1)(1-p_2)}{p_1'p_2'}\,S_{uv}\right],$$

which is greater than $p_1'p_2'S_{xy}$. Likewise the design expectation of $\hat{S}_x^2$ in (20) is

$$\begin{aligned}S_x^2[p_1'^2 &+ (N-1)p_1'(p_1-p_1')/N]\\ &+ (p_1-p_1')S_u^2[p_1' - (p_1' + 2p_1 - 2)/N]\\ &+ p_1'(p_1-p_1')(\bar{X}-\bar{U})^2,\end{aligned}$$

which is greater than $p_1'^2\hat{S}_x^2$ when N is large. If $S_{uv} = 0$, then the response bias leads to an underestimate of the correlation.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Stan Warner and was written for the Stanley Warner Memorial Session at the Statistical Society of Canada meetings in Banff, Alberta, May 1994. The research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BELLHOUSE, D.R. (1980). Linear models for randomized response designs. *Journal of the American Statistical Association*, 75, 1001-1004.

CHAUDHURI, A., and MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

EDGELL, S.E., HIMMELFARB, S., and CIRA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. *Psychological Bulletin*, 100, 251-256.

GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R., and HORVITZ, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., and HORVITZ, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250.

LEYSIEFFER, F.W., and WARNER, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.

NATHAN, G. (1988). A bibliography on randomized response: 1965-1987. *Survey Methodology*, 14, 331-346.

POLLOCK, K.H., and BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884-886.

RAO, C.R. (1952). Some theorems on minimum variance unbiased estimation. *Sankhyā* (A), 12, 27-42.

RAO, J.N.K. (1975). On the foundations of survey sampling. In *A Survey of Statistical Design and Linear Models*. (Ed. J.N. Srivastava). Amsterdam: North-Holland, 489-505.

RAO, J.N.K., and BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2, 125-141.

STEM, D.E., and STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association*, 79, 555-564.

UMESH, U.N., and PETERSON, R.A. (1991). A critical evaluation of the randomized response method. *Sociological Methods and Research*, 20, 104-138.

WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

WARNER, S.L. (1976). Optimal randomized response models. *International Statistical Review*, 44, 205-212.

WARNER, S.L. (1986). The omitted digit randomized response model for telephone applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.