Ann. Inst. Statist. Math.
Vol. 54, No. 3, 659–666 (2002)

ESTIMATION OF THE SIZE AND MEAN VALUE OF A STIGMATIZED CHARACTERISTIC OF A HIDDEN GANG IN A FINITE POPULATION: A UNIFIED APPROACH

RAGHUNATH ARNAB[1] AND SARJINDER SINGH[2]*

[1]*Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban-4000, South Africa,* e-mail: arnab@pixie.udw.ac.za
[2]*Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, Canada S7N 5E6,* e-mail: sarjinder@yahoo.com

(Received May 1, 2000; revised April 25, 2001)

Abstract. We consider the problem of estimating the size and the mean value of a stigmatized quantitative characteristic of a sub-group (or hidden gang) in a finite population using a unified approach. The proposed method may be useful in estimating the proportion and the mean income of terrorists, "hijackers" or "freedom fighters" of a particular country, including those who operate across different countries.

Key words and phrases: Randomized response, sampling strategies, estimation of proportion, estimation of mean.

1. Introduction

Warner (1965) was the first to suggest an ingenious method of counteracting deficient responses to sensitive questions. His idea has spawned further research which is documented in several publications, viz. Fox and Tracy (1986), Chaudhuri and Mukerjee (1987, 1988), Sheers (1992) and Bellhouse (1995). Recent developments in randomized response methods not covered by these reviewers include Franklin (1989), Kuk (1990), Mangat and Singh (1990), Singh and Singh (1993), Mangat (1994), Kerkvliet (1994), Singh (1994), Bansal *et al.* (1994), Singh *et al.* (1994, 1996, 1998), Mangat *et al.* (1995), Chaudhuri *et al.* (1996), Chang and Liang (1996), Arnab (1996, 1998), Chadhuri *et al.* (1998), Lee and Hong (1998) and Van Der Heijden *et al.* (1998).

Singh *et al.* (1998) have considered a different problem which is found to be very useful in solving domestic problems in a particular country or it may be applied to different countries. For example, a government may be interested in estimating the average income of victims or perpetrators of domestic violence which is rife due to the increase of unscrupulous people. The real applications of randomized response survey in actual practice were listed by Kerkvliet (1994). A comparison of randomized response techniques with direct questioning has also been considered in the context of social security fraud by Van Der Heijden *et al.* (1998). Interestingly, the focus of that study was to assess the validity of the responses to sensitive questions on social security fraud, obtained by using four different methods. They compared two different varieties of randomized

*Now at School of Mathematics and Statistics, 4217 Herzberg Lab, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6.

responses with computer-assisted self-administered questionnaires (CASAQ) and direct questioning in an experimental setting. They found that the validity could be assessed because all the respondents interviewed had already been identified to committing social security fraud. They set an experiment in such a way that the interviewers did not know that the respondents had been caught for fraud, and the respondents were oblivious that the researchers had this information. Since the actual status of the respondents was known, they found it possible to compare percentages of false negatives in the four different approaches. They also used two additional questions as follows. Which are the respondents willing to admit to having committed fraud? Are the respondent characteristics that predict positive responses to the sensitive questions the same for all methods? Currently, the applications of randomized response techniques are growing rapidly although we still feel that there is a communication gap between the theoreticians of RRT and the social survey statisticians who collect information on sensitive issues.

Under simple random sampling and with replacement (SRSWR) design, Singh *et al.* (1998) addressed the following types of problems:

1. Estimation of the proportion of persons in population P having an annual income greater than or equal to \$60,000 (say, gang G_1) along with their average income.

2. Estimation of the proportion of persons having extra marital relations (say, gang G_2) in the whole population and their average income.

3. Estimation of the proportion of politically active persons in the country (say, gang G_3) and their average income [or the average number of murders committed by them].

4. Estimation of the proportion of persons involved in a particular crime (say, gang G_4) along with the average value, μ_x, of any stigmatized quantitative character (say, X) of the same gang.

Due to easy access to fast computers today, the application of SRSWR sampling in actual practice is limited. It is very convenient to analyse data with advanced techniques such as those given by Hansen and Hurwitz (1943), Horvitz and Thompson (1952) , and Rao *et al.* (1962) techniques.

In the present paper, we propose a unified method for estimating the proportion of individuals belonging to a certain sensitive group and at the same time we propose a method of estimation of the mean of a stigmatized quantitative character (such as income) of the same group. The present method of estimation can be applied to any sampling design and to wider class of estimators.

2. Setup of the problem

Let P be a finite population of size N (known) and N_G (unknown) be the total number of persons belonging to some sensitive group G. Let X_i be the value of a stigmatized quantitative variable, X say, for a person i in the sensitive group (Gang) G.

Sampling design: Independent samples s_k's $k = 1, 2$ of sizes n_k's are selected from population P following the sampling designs p_k. Let $\pi_i(k)$ and $\pi_{ij}(k)$ be the first and second order inclusion probabilities for the sampling design p_k.

Randomized Response (RR) Technique: If the respondent labelled i is selected in the sample s_k, then he has to disclose the true response X_i if he belongs to the group G; otherwise he has to produce a randomize response R_{ki} following a certain randomized design. In other words, the persons appearing in the k-th sample are provided with a randomization device R_k (say). The device may be a spinner, deck of cards or computer

grid which generates a random variable taking the value greater than or equal to Ψ_0 (say) from a given probability distribution (Normal, Weibull or Gamma etc.). Each respondent i selected in the sample is requested to draw one random number (which is obviously greater than or equal to Ψ_0) from the randomization device R_k, without showing it to the interviewer. It is to be noted that for the chosen randomization device R_k, being almost similar in range with X_i is better for collecting sample data, especially in a sensitive survey. Now he/she is requested to report the actual value of the stigmatized quantitative variable, say X_i, iff he/she belongs to gang G; otherwise he/she is requested to report the random number R_{ki}, say, as drawn. The value of Ψ_0 depends upon the problem under consideration. For example, under problem 1, the value of Ψ_1 is \$60,000; under problems 2 and 3, the value of Ψ_0 can be taken as zero or any other suitable value. The choice of Ψ_0 can be made such that the respondents' privacy will not be jeopardised if they respond honestly. It is assumed that the distribution of the randomization device, R_k, is known to the interviewer but not the number drawn by the respondent. Let θ_k and σ_k^2 denote the known mean and variance of the randomization device R_k. Suppose π denotes the proportion of persons belonging to the gang G in the population. Thus for the i-th unit (if it is included in the sample s_k), we obtain a response

$$Z_{ki} = \begin{cases} X_i, & \text{if } i \in G \\ R_{ki}, & \text{if } i \notin G \end{cases} \tag{2.1}$$

where X_i and R_{ki} denote, respectively, the actual value of the stigmatized quantitative character and the random variable drawn by the i-th respondent in the k-th sample, $k = 1, 2$.

3. Development of some useful results

Here we consider the problem of estimation of the population proportion, given by

$$\pi = \frac{N_G}{N} \tag{3.1}$$

and the population mean of the desired sub-group or hidden gang, given by

$$\mu_x = \frac{1}{N_G} \sum_{i \in G} X_i \tag{3.2}$$

where X_i is the value of the quantitative characteristic associated with the respondent. Let G be the population of the gang, $\bar{G} = P - G$ and

$$I_i = \begin{cases} 1, & \text{if } i \in G \\ 0, & \text{if } i \notin G. \end{cases}$$

The randomized response from the i-th unit for the sample s_k as given in (2.1) can therefore be written as

$$Z_{ki} = X_i I_i + (1 - I_i) R_{ki} = X_i I_i + R_{ki} I_i^{'} \tag{3.3}$$

where $I_i^{'} = 1 - I_i$. Denoting the expectation and variance as E_R, V_R respectively with respect to the randomization device, we have

$$E_R(Z_{ki}) = X_i I_i + E_R(R_{ki}) I_i^{'} = X_i I_i + \theta_k I_i^{'} = \gamma_i(k) \quad \text{(say)} \tag{3.4}$$

and

$$V_R(Z_{ki}) = I_i' V_R(R_{ki}) = I_i' \sigma_k^2. \tag{3.5}$$

Now consider the following linear homogenous unbiased estimator for $E_R(\bar{Z}_k)$, where $\bar{Z}_k = \frac{1}{N}\sum_{i=1}^{N} Z_{ki}$, based on the sampling design p_k as

$$T_k = \frac{1}{N}\sum_{i \in s_k} b_{s_k i} Z_{ki} \tag{3.6}$$

where $b_{s_k i}$'s are known constants satisfying design unbiasedness (p_k-unbiasedness) condition

$$\sum_{s_k \ni i} b_{s_k i} p_k(s_k) = 1. \tag{3.7}$$

Denoting the expectation and variance as by E_p, V_p respectively with respect to the sampling design p_k, we get the following theorems:

THEOREM 3.1. *The estimator T_k is an unbiased estimator for*

$$E_R(\bar{Z}_k) = \pi \mu_x + (1 - \pi)\theta_k. \tag{3.8}$$

PROOF. We have

$$\begin{aligned} E(T_k) &= E_p E_R(T_k) = \frac{1}{N} E_p \sum_{i \in s_k} b_{s_k i} \gamma_i = \frac{1}{N}\sum_i \gamma_i(k) \\ &= \frac{1}{N}\sum_i \{X_i I_i + I_i' \theta_k\} \\ &= \frac{1}{N}\sum_{i \in G} X_i + \frac{\theta_k}{N}\sum_i I_i' \\ &= \frac{N_G}{N}\frac{1}{N_G}\sum_{i \in G} X_i + \theta_k \frac{(N - N_G)}{N} \\ &= \pi \mu_x + (1 - \pi)\theta_k. \end{aligned}$$

Hence the theorem.

THEOREM 3.2. *The variance of the estimator T_k is given by*

$$V_k = V(T_k) = \frac{1}{N^2}\left[\sigma_k^2 \sum_i \alpha_i(k) I_i' + \left\{\sum_i \gamma_i^2(k)(a_i(k) - 1) + \sum_{i \neq j}\sum \gamma_i(k)\gamma_j(k)(\alpha_{ij}(k) - 1)\right\}\right] \tag{3.9}$$

where $\alpha_i(k) = \sum_{s_k \ni i} b_{s_k i}^2 p_k(s_k)$ and $\alpha_{ij}(k) = \sum_{s_k \ni i,j} b_{s_k i} b_{s_k j} p_k(s_k)$.

PROOF. We have

$$
\begin{aligned}
N^2V(T_k) &= N^2[E_p[V_R(T_k)] + V_p[E_R(T_k)]] \\
&= E_p\left[\sum_{i\in s_k} b_{s_k i}^2 \sigma_k^2 I_i^{'}\right] + V_p\left[\sum_{i\in s_k} b_{s_k i}\gamma_i(k)\right] \\
&= \sigma_k^2 \sum_i I_i^{'} \sum_{s_k \ni i} b_{s_k i}^2 p_k(s_k) + \sum_i \gamma_i^2(k) \sum_{s_k \ni i} b_{s_k i}^2 p_k(s_k) \\
&\quad + \sum_{i\neq j}\sum \gamma_i(k)\gamma_j(k) \sum_{s_k \ni i,j} b_{s_k i} b_{s_k j} p_k(s_k) - \left(\sum_{i=1}^{N} \gamma_i(k)\right)^2 \\
&= \sigma_k^2 \sum_i a_i(k) I_i^{'} + \left\{\sum_i \gamma_i^2(a_i(k)-1) + \sum_{i\neq j}\sum \gamma_i(k)\gamma_j(k)(a_{ij}(k)-1)\right\}.
\end{aligned}
$$

Hence the theorem.

THEOREM 3.3. *An unbiased estimator of the variance $V(T_k)$ is given by*

$$
\hat{\hat{V}}_k = \hat{V}_k + \frac{\sigma_k^2}{N}(1-\hat{\pi}(k)) \tag{3.10}
$$

where

$$
\hat{V}_k = \frac{1}{N^2}\left[\sum_{i\in s_k} \frac{Z_{ki}^2}{\pi_i(k)}(\alpha_i(k)-1) + \sum_{i\neq j\in s_k}\sum \frac{Z_{ki}Z_{kj}}{\pi_{ij}(k)}(\alpha_{ij}(k)-1)\right]
$$

and

$$
\hat{\pi}(k) = \frac{1}{N}\sum_{i\in p_k} \frac{I_i}{\pi_i(k)}.
$$

PROOF. Consider

$$
\begin{aligned}
E(\hat{V}_k) &= \frac{1}{N^2} E_p\left[\sum_{i\in s_k} \frac{\gamma_i^2(k)+\sigma_k^2 I_i^{'}}{\pi_i(k)}(\alpha_i(k)-1) + \sum_{i\neq j\in s_k}\sum \frac{\gamma_i(k)\gamma_j(k)}{\pi_{ij}(k)}(\alpha_{ij}(k)-1)\right] \\
&= \frac{1}{N^2}\left[\sum_i \gamma_i^2(k)(\alpha(k)_i - 1) + \sum_{i\neq j\in s_k}\sum \gamma(k)_i\gamma_j(k)(\alpha_{ij}(k)-1)\right. \\
&\qquad\qquad \left. + \sigma_k^2 \sum_i (\alpha_i(k)-1) I_i^{'}\right] \\
&= V(T_k) - \frac{\sigma_k^2}{N^2}\sum_i I_i^{'} = V(T_k) - \frac{\sigma_k^2(1-\pi(k))}{N}.
\end{aligned}
$$

Hence the theorem.

COROLLARY 3.1. *An alternative unbiased estimator for estimating $V(T_k)$ is given by*

$$\hat{\hat{V}}_k^* = \hat{V}_k^* + \frac{\sigma_k^2}{N}(1 - \hat{\pi}(k)) \tag{3.11}$$

where

$$\hat{V}_k^* = \frac{1}{N^2}\sum\sum_{i<j\in s_k}\left(\frac{\pi_i(k)\pi_j(k) - \pi_{ij}(k)}{\pi_{ij}(k)}\right)\left(\frac{Z_{ki}}{\pi_i(k)} - \frac{Z_{kj}}{\pi_j(k)}\right)^2 + \frac{\sigma_k^2(1 - \hat{\pi}(k))}{N},$$

assuming each $s_k (k = 1, 2)$ with positive $p(s_k)$ to have a 'common' number of distinct units.

4. Estimation of proportion

Using notations and results from the previous section, we have the following theorem.

THEOREM 4.1. *An unbiased estimator of π is given by*

$$\hat{\pi} = 1 - \frac{T_1 - T_2}{\theta_1 - \theta_2} \tag{4.1}$$

with variance

$$V(\hat{\pi}) = \frac{V(T_1) + V(T_2)}{(\theta_1 - \theta_2)^2} \tag{4.2}$$

and an unbiased estimator of $V(\hat{\pi})$ is given by

$$\hat{V}(\hat{\pi}) = \frac{\hat{\hat{V}}_1 + \hat{\hat{V}}_2}{(\theta_1 - \theta_2)^2}. \tag{4.3}$$

PROOF. Straightforward from previous section.

The estimator of proportion in (4.1) becomes non-functional if $\theta_1 - \theta_2 = 0$. It is important to keep in mind while making a pair of randomization device such that $\theta_1 - \theta_2 \neq 0$. The expression (4.2) indicates that the variance of the proportion is inversely proportional to the difference $\theta_1 - \theta_2$. Thus one should choose θ_1 and θ_2 such that the difference $\theta_1 - \theta_2$ is maximum without threatening the privacy of the respondents.

5. Estimation of mean

In this section we consider the problem of estimation of mean μ_x of the quantitative character of the sub-group G of the population. We have from the Theorem 3.1

$$E(T_k) = \pi\mu_x + (1 - \pi)\theta_k, \qquad k = 1, 2.$$

This implies that

$$\hat{\mu}_x = \frac{T_1 - (1 - \hat{\pi})\theta_1}{\hat{\pi}}$$

$$= \frac{\theta_1 - \theta_2}{(\theta_1 - \theta_2) - (T_1 - T_2)} \left\{ T_1 - \frac{(T_1 - T_2)\theta_1}{\theta_1 - \theta_2} \right\}$$
$$= \frac{T_1\theta_1 - T_1\theta_2 - T_1\theta_1 + T_2\theta_1}{(T_2 - \theta_2) - (T_1 - \theta_1)}$$
$$= \frac{T_2\theta_1 - T_1\theta_2}{(T_2 - \theta_2) - (T_1 - \theta_1)}.$$

Thus an estimator of μ_x is given by

$$\hat{\mu}_x = \frac{d_1}{d_2} \tag{5.1}$$

where $d_1 = T_2\theta_1 - T_1\theta_2$ and $d_2 = (T_2 - \theta_2) - (T_1 - \theta_1)$.

Acknowledgements

The authors are thankful to the Editor-in-Chief Professor Ryoichi Shimizu for encouragement to revise the original version of the paper according to the comments by two reviewers. The comments by the two learned reviewers have also been duly acknowledged and appreciated. The opinions and results discussed in this paper are of authors' and not necessarily of any institute or organisation.

References

Arnab, R. (1996). Randomized response trials: A unified approach for qualitative data, *Comm. Statist. Theory Methods*, **25**(6), 1173–1183.

Arnab, R. (1998). Randomized response surveys: Optimum estimation of a finite population total, *Statist. Papers*, **39**, 405–408.

Bansal, M. L., Singh, S. and Singh, R. (1994). Multi-character survey using randomized response technique, *Comm. Statist. Theory Methods*, **23**(6), 1705–1715.

Bellhouse, D. R. (1995). Estimation of correlation in randomized response, *Survey Methodology*, **21**, 13–19.

Chang, H. J. and Liang, D. H. (1996). A two stage unrelated randomized response procedure, *Austral. J. Statist.*, **38**(1), 43–52.

Chaudhuri, A. and Mukerjee, R. (1987). Randomized response techniques: A review, *Statist. Neerlandica*, **41**, 27–44.

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

Chaudhuri, A., Maiti, T. and Roy, D. (1996). A note on competing variance estimator in randomized response surveys, *Austral. J. Statist.*, **38**(1), 35–42.

Chaudhuri, A., Adhikary, A. K. and Maiti, T. (1998). A note on non-negative mean square error estimation of regression estimators in randomized response surveys, *Statist. Papers*, **39**, 409–415.

Fox, J. and Tracy, P. (1986). *Randomized Response: A Method for Sensitive Surveys*, Sage Publication, Beverly Hills.

Franklin, L. A. (1989). A comparison of estimators for randomized response sampling with continuous distributions from a dichotomous population, *Comm. Statist. Theory Methods*, **18**(2), 489–505.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations, *Ann. Math. Statist.*, **14**, 333–362.

Horvitz, D. G. and Thompson, D. J. (1952). A generalisation of sampling without replacement from a finite universe, *J. Amer. Statist. Assoc.*, **47**, 663–685.

Kerkvliet, J. (1994). Estimating a logit model with randomized data: The case of cocaine use, *Austral. J. Statist.*, **36**(1), 9–20.

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly, *Biometrika*, **77**, 436–438.

Lee, G. S. and Hong, K. H. (1998). An improved unrelated question model, *Korean Journal of Applied Statistics*, **11**, 415–421.

Mangat, N. S. (1994). An improved randomized response strategy, *J. Roy. Statist. Soc. Ser. B*, **56**, 93–95.

Mangat, N. S. and Singh, R. (1990). An alternative randomized response procedure, *Biometrika*, **77**(2), 439–442.

Mangat, N. S., Singh, R., Singh, S., Bellhouse, D. R. and Kashani, H. B. (1995). On efficiency of estimator using distinct respondents in randomized response survey, *Survey Methodology*, **21**, 21–23.

Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). A simple procedure of unequal probability sampling without replacement, *J. Roy. Statist. Soc. Ser. B*, **24**, 482–491.

Sheers, N. (1992). A review of randomized response technique, *Measurement and Evaluation in Counselling and Development*, **25**, 27–41.

Singh, S. (1994). Unrelated question randomized response sampling using continuous distributions, *J. Indian Soc. Agricultural Statist.*, **46**(3), 349–361.

Singh, S. and Singh, R. (1993). Generalized Franklin's model for randomized response sampling, *Comm. Statist. Theory Methods*, **22**(2), 741–755.

Singh, S., Mangat, N. S. and Singh, R. (1994). On estimation of mean/total of stigmatized quantitative variable, *Statistica*, **54**, 383–386.

Singh, S., Joarder, A. H. and King, M. L. (1996). Regression analysis using scrambled responses, *Austral. J. Statist.*, **38**(2), 201–211.

Singh, S., Horn, S. and Chowdhury, S. (1998). Estimation of stigmatized characteristics of a hidden gang in finite population, *Australian & New Zealand Journal of Statistics*, **40**, 291–298.

Van Der Heijden, P. G. M., Van Gils, G., Bouts, J. and Hox, J. (1998). A comparison of randomized response, CASAQ, and direct questioning, eliciting sensitive information in the context of social security fraud, *Kwantitatieve Methoden*, **59**, 15–34.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias, *J. Amer. Statist. Assoc.*, **60**, 63–69.