Statistics and Applications {ISSN 2452-7395 (online)}
Volume 15 Nos. 1&2, 2017 (New Series), pp 161-166

Detection of Hidden Gangs in the Finite Population Using Randomized Response Sampling

Sarjinder Singh and Polly D. Allred
Texas A&M University—Kingsville, USA

Received: January 16, 2017; Revised: February 03, 2017; Accepted: February 07, 2017

Abstract

This paper considers the problem of the estimation of sizes of hidden gangs in a finite population, and the estimation of mean values of stigmatized quantitative characteristics of those gangs. An estimator of the correlation coefficient between two sensitive quantitative characteristics of a particular hidden gang has been suggested. The bias and variance expressions of the proposed estimators for estimating the population means of two sensitive quantitative characteristics of a particular gang are derived to the first order of approximation.

Key words: Hidden gangs, Randomized Response Technique, Estimation of mean and correlation coefficient.

1. Introduction

Singh et al. (1998) introduced the concept of hidden gangs in finite populations and restricted the problem to the estimation of the mean of only one stigmatizing characteristic of a single gang. A "hidden gang" is simply a group of persons engaged in a secret activity (for example, extramarital affairs).

The estimation of the correlation between two sensitive quantitative characteristics of a single gang is also possible. Consider estimating the proportionate sizes in the finite population Ω, or averages of stigmatizing quantitative characteristics of a hidden gang G. People may be unwilling to admit membership in gang G due to fear of retribution or social embarrassment. Likewise, gang members may be reluctant to divulge values of sensitive characteristics.

Example: Say, gang G represents the group of people in the overall population who are having extramarital affairs. It may be interesting to:

- estimate the proportion of persons having extramarital relations in the whole population;
- estimate the income X of persons in gang G; and
- estimate the number of murders Y committed by persons in gang G;
- estimate the correlation ρ_{xy} between income X and number of murders Y in gang G.

In general, it may be interesting to estimate the proportion of persons involved in a particular gang G, along with mean values, μ_x and μ_y, and correlation coefficient ρ_{xy}, of any two stigmatized quantitative characteristics, X and Y, of the same gang.

Corresponding Author: Sarjinder Singh
E-mail: sarjinder.singh@tamuk.edu

2. Single Gang Problem with Two Sensitive Quantitative Characteristics

Select two independent samples, S_1 and S_2, of sizes n_1 and n_2, respectively, from the population $\Omega = \{w_1, w_2, \ldots, w_N\}$ of N units, using simple random sampling with replacement (SRSWR). Let $\pi = N_G/N$ be the proportion of persons involved in gang G. Let X and Y be two quantitative sensitive characteristics of the hidden gang G.

Each respondent in each of the first sample S_1 and the second sample S_2 is asked about membership in gang G using the pioneer Warner (1965) model wherein each respondent is provided with a randomization device bearing two questions:

(i) "Are you a member of hidden gang G?" with probability P;
(ii) "Are you not a member of hidden gang G?" with probability (1-P). (2.1)

A random variable t_{ij} for $i = 1, 2$, and $j = 1,\ldots,n_1$ or $j = 1,\ldots,n_2$ is defined such that $t_{ij} = 1$ if the i[th] respondent reports "yes" and $t_{ij} = 0$, otherwise. Following Warner (1965), an estimator of π based on the first sample information is given by

$$\hat{\pi}_1 = \frac{\frac{1}{n_1}\sum_{j=1}^{n_1} t_{1j} - (1-P)}{2P-1}. \tag{2.2}$$

Similarly, another estimator of ρ based on the second sample information is given by

$$\hat{\pi}_2 = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} t_{2j} - (1-P)}{2P-1}. \tag{2.3}$$

Theorem 2.1: An unbiased estimator of π is given by

$$\hat{\pi} = \frac{1}{2}\left[\hat{\pi}_1 + \hat{\pi}_2\right]. \tag{2.4}$$

Proof: The theorem follows easily from $E(\hat{\pi}_i) = \pi$ for $i = 1, 2$.

Each respondent selected in the first sample is also requested to use one of two randomization devices, R_1 and R_2, according to the following rule. If the j^{th} respondent belongs to the gang G, then a scrambled response $X_j R_1$ is reported, otherwise the scrambled response $Y_j R_2$ is reported. The expected value, $E(R_k) = \theta_k$, and variance, $V(R_k) = \gamma_k^2$, $k = 1,2$, for the randomization devices are assumed known. The distributions of the randomization devices R_1 and R_2 can be adjusted so that the ranges of the scrambled responses are similar in order to guarantee the respondents' protection of privacy. Thus in the first sample, the distribution of responses will be

$$Z_{1j} = \begin{cases} X_j R_1 \text{ with probability } \pi \\ Y_j R_2 \text{ with probability } (1-\pi) \end{cases}, \tag{2.5}$$

and the expected value of Z_{1j} for $j = 1,\ldots,n_1$ is given by

$$E(Z_{1j}) = \pi\mu_x\theta_1 + (1-\pi)\mu_y\theta_2. \tag{2.6}$$

Each respondent selected in the second independent sample is also requested to use one of two randomization devices R_3 and R_4, according to the following rule. If the j^{th} respondent belongs to the gang G, then a scrambled response $X_j R_3$ is reported, otherwise the scrambled response $Y_j R_4$ is reported. The expected value, $E(R_k) = \theta_k$, and the variance, $V(R_k) = \gamma_k^2$, $k = 3, 4$, for the randomization devices are assumed known. The distributions of the randomization devices R_3 and R_4 can be adjusted so that the ranges of the scrambled responses is similar in order to guarantee the respondents' protection of privacy. Thus in the second sample, the distribution of responses will be

$$Z_{2j} = \begin{cases} X_j R_3 \text{ with probability } \pi \\ Y_j R_4 \text{ with probability} (1-\pi) \end{cases}, \tag{2.7}$$

and the expected value of Z_{2j} for $j = 1, ..., n_2$ is given by

$$E(Z_{2j}) = \pi \mu_x \theta_3 + (1-\pi)\mu_y \theta_4 . \tag{2.8}$$

Theorem 2.2: An estimator for the mean μ_x of the sensitive quantitative characteristic X of the hidden gang G is given by

$$\hat{\mu}_x = \frac{1}{\hat{\pi}(\theta_1\theta_4 - \theta_2\theta_3)} \left[\theta_4 \bar{Z}_1 - \theta_2 \bar{Z}_2\right], \tag{2.9}$$

and an estimator for the mean μ_y of the sensitive quantitative characteristic Y of the hidden gang G is given by

$$\hat{\mu}_y = \frac{1}{(1-\hat{\pi})(\theta_1\theta_4 - \theta_2\theta_3)} \left[\theta_1 \bar{Z}_2 - \theta_3 \bar{Z}_1\right] \tag{2.10}$$

where $\bar{Z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$, for $i = 1, 2$.

Corollary 2.1: Construct the randomization devices R_i, $i = 1, 2, 3, 4$, such that $\theta_1\theta_4 \neq \theta_2\theta_3$.

Corollary 2.2: The following results are useful in obtaining the bias and variance expressions of the estimators $\hat{\mu}_x$ and $\hat{\mu}_y$ by the ratio method of estimation.

$$V(\bar{Z}_1) = \frac{1}{n_1}\left[\pi\left\{\sigma_x^2\left(\gamma_1^2 + \theta_1^2\right) + \mu_x^2\gamma_1^2\right\} + (1-\pi)\left\{\sigma_y^2\left(\gamma_2^2 + \theta_2^2\right) + \mu_y^2\gamma_2^2\right\} + \pi(1-\pi)\left(\mu_x\theta_1 - \mu_y\theta_2\right)^2\right] = \frac{\sigma_{Z_1}^2}{n_1}, \tag{2.11}$$

$$V(\bar{Z}_2) = \frac{1}{n_2}\left[\pi\left\{\sigma_x^2\left(\gamma_3^2 + \theta_3^2\right) + \mu_x^2\gamma_3^2\right\} + (1-\pi)\left\{\sigma_y^2\left(\gamma_4^2 + \theta_4^2\right) + \mu_y^2\gamma_4^2\right\} + \pi(1-\pi)\left(\mu_x\theta_3 - \mu_y\theta_4\right)^2\right] = \frac{\sigma_{Z_2}^2}{n_2}, \tag{2.12}$$

$$Cov\left(\bar{Z}_1, \bar{Z}_2\right) = 0, \tag{2.13}$$

$$V(\hat{\pi}) = \frac{1}{4}\left[V(\hat{\pi}_1) + V(\hat{\pi}_2)\right] = \frac{1}{4}\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left\{\pi(1-\pi) + \frac{P(1-P)}{(2P-1)^2}\right\}\right], \tag{2.14}$$

$$Cov\left(\bar{Z}_1, \hat{\pi}\right) = Cov\left(\bar{Z}_1, \frac{1}{2}\hat{\pi}_1\right) = \frac{\left(\theta_1\mu_x - \theta_2\mu_y\right)}{2n_1}\left\{\pi(1-\pi) + \frac{P(1-P)}{(2P-1)^2}\right\}, \tag{2.15}$$

and

$$Cov\left(\bar{Z}_2, \hat{\pi}\right) = Cov\left(\bar{Z}_2, \frac{1}{2}\hat{\pi}_2\right) = \frac{\left(\theta_3\mu_x - \theta_4\mu_y\right)}{2n_2}\left\{\pi(1-\pi) + \frac{P(1-P)}{(2P-1)^2}\right\}. \tag{2.16}$$

Theorem 2.4: The variances and covariance of the estimators $\hat{\mu}_x$ and $\hat{\mu}_y$ to the first order of approximation, are given by

$$V(\hat{\mu}_x) \approx \theta_4^2 V(\bar{Z}_1) + \theta_2^2 V(\bar{Z}_2) + \mu_x^2 V(\hat{\pi}) - 2\mu_x Cov\{\theta_4\bar{Z}_1 - \theta_2\bar{Z}_2, \hat{\pi}\}, \tag{2.17}$$

$$V(\hat{\mu}_y) \approx \theta_1^2 V(\bar{Z}_2) + \theta_3^2 V(\bar{Z}_1) + \mu_y^2 V(\hat{\pi}) + 2\mu_y Cov\{\theta_1\bar{Z}_2 - \theta_3\bar{Z}_1, \hat{\pi}\}, \tag{2.18}$$

and

$$Cov(\hat{\mu}_x, \hat{\mu}_y) \approx \frac{(\theta_1\theta_4 + \theta_2\theta_3)E(\bar{Z}_1)E(\bar{Z}_2) - \theta_1\theta_2\left\{V(\bar{Z}_1) + \left(E(\bar{Z}_1)\right)^2\right\} - \theta_3\theta_4\left\{V(\bar{Z}_2) + \left(E(\bar{Z}_2)\right)^2\right\}}{\pi(1-\pi)(\theta_1\theta_4 - \theta_2\theta_3)^2} - \mu_x\mu_y. \tag{2.19}$$

Theorem 2.5: The optimum values of n_i can be obtained by minimizing the determinant of the variance-covariance matrix

$$D = \begin{vmatrix} V(\hat{\mu}_x) & Cov(\hat{\mu}_x, \hat{\mu}_y) \\ Cov(\hat{\mu}_x, \hat{\mu}_y) & V(\hat{\mu}_y) \end{vmatrix}, \tag{2.20}$$

subject to the condition that $n_1 + n_2 = n$. The determinant (2.20) can be written as

$$D = \begin{vmatrix} \frac{V_1}{n_1} + \frac{V_2}{n_2} & \frac{C_1}{n_1} + \frac{C_2}{n_2} \\ \frac{C_1}{n_1} + \frac{C_2}{n_2} & \frac{V_3}{n_1} + \frac{V_4}{n_2} \end{vmatrix}, \tag{2.21}$$

where the V_i and C_i are the coefficients of n_i^{-1}, for $i = 1,2$, in (2.17), (2.18) and (2.19).

It is difficult to develop a theoretical formula for optimum values of n_i, but a computer grid search method can be developed to find such optimum values. The optimum values of n_i will depend upon the population parameters similar to those reported by Singh et al. (1998) and Singh et al. (1996).

3. Estimation of Correlation Coefficient: Single Gang Problem

Samples, S_1 and S_2, of $n_1 + n_2 = n$ units have been selected from the population. The j^{th} respondent in each sample has been requested to report two responses Z_{1j} and Z_{2j} using the randomization devices defined as above. Note that two independent samples of different sizes cannot be used to estimate the correlation coefficient. Each respondent (from either sample) has also been requested to use the same Warner (1965) device as defined in (2.1) above, so an estimator of π is also given by

$$\hat{\pi}^* = \frac{\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} t_{ij} - (1-P)}{2P-1}. \tag{3.1}$$

In order to estimate the correlation ρ_{xy} between the two sensitive quantitative characteristics X and Y of the hidden gang G, the parameters of the four randomization devices R_k, $k = 1,2,3,4$, must have been adjusted so that

$$\frac{\gamma_1^2 + \theta_1^2}{\gamma_2^2 + \theta_2^2} \neq \frac{\gamma_3^2 + \theta_3^2}{\gamma_4^2 + \theta_4^2}. \tag{3.2}$$

The correlation coefficient between the two sensitive quantitative characteristics X and Y of the gang G is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{3.3}$$

where $\sigma_{xy} = \dfrac{\sigma_{Z_1Z_2} - \pi^2\theta_1\theta_3\sigma_x^2 - (1-\pi)^2\theta_2\theta_4\sigma_y^2}{\pi(1-\pi)(\theta_1\theta_4 + \theta_2\theta_3)}$. (3.4)

$$\sigma_x^2 = \frac{\psi_1\left(\gamma_4^2 + \theta_4^2\right) - \psi_2\left(\gamma_2^2 + \theta_2^2\right)}{\pi\left[\left(\gamma_1^2 + \theta_1^2\right)\left(\gamma_4^2 + \theta_4^2\right) - \left(\gamma_2^2 + \theta_2^2\right)\left(\gamma_3^2 + \theta_3^2\right)\right]}. \tag{3.5}$$

$$\sigma_y^2 = \frac{\psi_2\left(\gamma_1^2 + \theta_1^2\right) - \psi_1\left(\gamma_3^2 + \theta_3^2\right)}{(1-\pi)\left[\left(\gamma_1^2 + \theta_1^2\right)\left(\gamma_4^2 + \theta_4^2\right) - \left(\gamma_2^2 + \theta_2^2\right)\left(\gamma_3^2 + \theta_3^2\right)\right]}. \tag{3.6}$$

and where

$$\psi_1 = \sigma^2{}_{Z_1} - \pi(1-\pi)\left(\mu_x\theta_1 - \mu_y\theta_2\right)^2 - \pi\gamma_1^2\mu_x^2 - (1-\pi)\gamma_2^2\mu_y^2, \tag{3.7}$$

and $\psi_2 = \sigma^2{}_{Z_2} - \pi(1-\pi)\left(\mu_x\theta_3 - \mu_y\theta_4\right)^2 - \pi\gamma_3^2\mu_x^2 - (1-\pi)\gamma_4^2\mu_y^2$. (3.8)

Theorem 3.1: An estimator for the correlation coefficient ρ_{xy} between two sensitive quantitative characteristics X and Y of a particular hidden gang G is given by

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.9}$$

where $s_{xy} = \dfrac{s_{Z_1Z_2} - \left(\hat{\pi}^*\right)^2\theta_1\theta_3 s_x^2 - \left(1-\hat{\pi}^*\right)^2\theta_2\theta_4 s_y^2}{\hat{\pi}^*\left(1-\hat{\pi}^*\right)(\theta_1\theta_4 + \theta_2\theta_3)}$, (3.10)

$$s_x^2 = \frac{\hat{\psi}_1\left(\gamma_4^2+\theta_4^2\right)-\hat{\psi}_2\left(\gamma_2^2+\theta_2^2\right)}{\hat{\pi}^*\left[\left(\gamma_1^2+\theta_1^2\right)\left(\gamma_4^2+\theta_4^2\right)-\left(\gamma_2^2+\theta_2^2\right)\left(\gamma_3^2+\theta_3^2\right)\right]}, \tag{3.11}$$

and $s_y^2 = \frac{\hat{\psi}_2\left(\gamma_1^2+\theta_1^2\right)-\hat{\psi}_1\left(\gamma_3^2+\theta_3^2\right)}{\left(1-\hat{\pi}^*\right)\left[\left(\gamma_1^2+\theta_1^2\right)\left(\gamma_4^2+\theta_4^2\right)-\left(\gamma_2^2+\theta_2^2\right)\left(\gamma_3^2+\theta_3^2\right)\right]}$; (3.12)

and where $\hat{\psi}_1 = s_{Z_1}^2 - \hat{\pi}^*\left(1-\hat{\pi}^*\right)\left(\hat{\mu}_x\theta_1-\hat{\mu}_y\theta_2\right)^2 - \hat{\pi}^*\gamma_1^2\hat{\mu}_x^2 - \left(1-\hat{\pi}^*\right)\gamma_2^2\hat{\mu}_y^2$, (3.13)

and $\hat{\psi}_2 = s_{Z_2}^2 - \hat{\pi}^*\left(1-\hat{\pi}^*\right)\left(\hat{\mu}_x\theta_3-\hat{\mu}_y\theta_4\right)^2 - \hat{\pi}^*\gamma_3^2\hat{\mu}_x^2 - \left(1-\hat{\pi}^*\right)\gamma_4^2\hat{\mu}_y^2$. (3.14)

Also, $s_{Z_1Z_2} = (n-1)^{-1}\sum_{j=1}^{n/2}\left(Z_{1j}-\bar{Z}_1\right)\left(Z_{2j}-\bar{Z}_2\right)$, (3.15)

$$s_{Z_1}^2 = (n-1)^{-1}\sum_{j=1}^{n/2}\left(Z_{1j}-\bar{Z}_1\right)^2, \tag{3.16}$$

and $s_{Z_2}^2 = (n-1)^{-1}\sum_{j=1}^{n/2}\left(Z_{2j}-\bar{Z}_2\right)^2$, (3.17)

are the estimators of $\sigma_{Z_1Z_2}$, $\sigma_{Z_1}^2$, and $\sigma_{Z_2}^2$, respectively .

The estimator of the correlation coefficient defined at (3.9) differs from those developed by Bellhouse (1995) and Singh (1991, 2016) because, here, the correlation between two sensitive quantitative characteristics of a particular hidden gang is estimated, whereas the previous techniques estimated the correlation between two sensitive quantitative characteristics of the whole population.

Acknowledgements
The authors are thankful to Dr. Hukum Chandra and a referee for comments on the original version of the manuscript.

References

Bellhouse, D.R. (1995). Estimation of correlation in randomized response. *Survey Methodology*, **21(1)**, 13-19.

Singh, S. (1991). *On Improved Strategies in Survey Sampling*. (Unpublished Dissertation). Punjab Agricultural University, Ludhiana, India.

Singh, S. (2016). On the estimation of correlation coefficient using scrambled responses. Handbook of Statistics 34 (Edited by Prof. A. Chaudhuri, Prof. T.C. Christofides and Prof. C.R. Rao), ELSEVIER, pp 43-89.

Singh, S., Horn, S. and Chowdhury, S. (1998). Estimation of stigmatized characteristics of a hidden gang in a finite population. *Australian & New Zealand Journal of Statistics*, **40(3)**, 291-297.

Singh, S., Singh, R. and Mangat, N.S. (1996). Estimation of mean of a stigmatized quantitative variable for a sub-group of the population. *Metron*, **54 (3-4)**, 83-91.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60(309)**, 63-69.