METRON (2013) 71:3–8
DOI 10.1007/s40300-013-0002-3

A new randomized response device for sensitive characteristics: an application of the negative hypergeometic distribution

Sarjinder Singh · Stephen A. Sedory

Received: 10 October 2011 / Accepted: 8 August 2012 / Published online: 28 May 2013

Abstract In this paper, a new randomized response device is proposed based on the use of negative hypergeometric distribution while estimating the proportion of persons possessing a sensitive characteristic in a population. Situations where the proposed randomization device can perform better than the Warner (J Am Stat Assoc 60, 63–69, 1965), the Kuk (Biomerika 77(2), 436–438, 1990) and the Mangat (J R Statist Soc B 56, 93–95, 1994) estimators are simulated and discussed.

Keywords Sensitive variables · Estimation of proportion · Relative efficiency

1 Introduction

Warner [5] proposed an interviewing technique, called Randomized Response, to protect an interviewee's privacy and to reduce a major source of bias (evasive answers or refusing to respond) in estimating the prevalence of sensitive characteristics in surveys of human populations. Warner [5] designed a randomization device, for example a spinner or a deck of cards that consists of two mutually exclusive outcomes. In the case of cards, each card has one of the following statements: (i) I possess attribute A; (ii) I do not possess attribute A. The maximum likelihood estimator of π, the proportion of respondents in the population possessing the attribute A, is given by:

$$\hat{\pi}_w = \frac{(n_1/n) - (1 - P)}{2P - 1} \tag{1.1}$$

where n_1 is the number of individuals responding "*yes*", n is the number of respondents selected by a simple random and with replacement sample (SRSWR), and P is the proportion of cards bearing the statement, "I possess an attribute A." The variance of $\hat{\pi}_w$ is given by:

S. Singh (✉) · S. A. Sedory
Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX 78363, USA
e-mail: sarjinder@yahoo.com

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2} \tag{1.2}$$

Kuk [1] introduced an ingenious randomized response model in which respondents belonging to a sensitive group A are instructed to use a deck of cards having the proportion θ_1 of cards with the statement, "I belong to group A" and if respondents belong to non-sensitive group A^c then they are instructed to use a different deck of cards having the proportion θ_2 of cards with the statement, "I do not belong to group A". Again let π be the true proportion of persons belonging to the sensitive group A. Then, the probability of a 'yes' answer in the Kuk [1] model is given by:

$$\theta_{kuk} = \theta_1\pi + (1-\pi)\theta_2 \tag{1.3}$$

Further assume that a simple random with replacement sample of n respondents is selected from the population, and that n_1 is the number of observed "yes" answers. The number of people n_1 that answer "yes" is binomially distributed with parameters $\theta_{kuk} = \theta_1\pi + (1-\pi)\theta_2$ and n. For the Kuk [1] model, an unbiased estimator of the population proportion π is given by:

$$\hat{\pi}_{kuk} = \frac{n_1/n - \theta_2}{\theta_1 - \theta_2}, \theta_1 \neq \theta_2 \tag{1.4}$$

with variance, given by:

$$V(\hat{\pi}_{kuk}) = \frac{\theta_{kuk}(1-\theta_{kuk})}{n(\theta_1-\theta_2)^2} \tag{1.5}$$

Mangat [3] suggested a randomized response model in which respondents belonging to sensitive group A are instructed to report 'yes' without using any randomization device and if respondents do not belong to the sensitive group A then they are instructed using the Warner [5] device. The mode of response of the respondent remains unrevealed to the interviewer. Thus the probability of 'yes' answer in the Mangat [3] model is given by:

$$\theta_m = \pi + (1-\pi)(1-P) \tag{1.6}$$

Further assume a simple random with replacement sample of n respondents is selected from the population, and that n_m is the number of observed "yes" answers. The number of people n_m that answer "yes" is binomially distributed with parameters θ_m and n. For the Mangat [3] model, an unbiased estimator of the population proportion π is given by:

$$\hat{\pi}_m = \frac{n_m/n - (1-P)}{P} \tag{1.7}$$

with variance, given by:

$$V(\hat{\pi}_m) = \frac{\theta_m(1-\theta_m)}{nP^2} \tag{1.8}$$

In the Mangat [3] model, the respondents who report "no" are surely members of the non-sensitive group, and lose their privacy. Because part of respondents lose their privacy in the Mangat [3] model, it is not easy to develop a model which is more efficient than this one. Further note that if $\theta_1 = 1$ and $\theta_2 = (1-P)$, then the Kuk [1] model reduces to the Mangat [3] model.

In the next section, we suggest a new randomization device which could be adjusted to perform better than the [1,3,5] models.

2 Newly proposed randomization device

The newly proposed randomized response device consists of two urns: Urn-I contains N_1 balls, out of which r_1 balls bearing the statement, (i.) "I belong to the group$A$", and the remaining $(N_1 - r_1)$ balls are blank with no statement on them. Urn-II contains N_2 balls, out of which r_2 balls bearing the statement, (ii.) "I do not belong to the group$A$", and the remaining $(N_2 - r_2)$ balls are blank with no statement on them. Each respondent selected in the sample is instructed as follows: If he/she belongs to the group A, then he/she draws balls using without replacement sampling from the Urn-I until he/ she gets t_1 $(< r_1)$ balls bearing the statement (i.), and reports the total number of balls, say$X$, drawn by him/her. Thus X follows a negative hypergeometric distribution given by:

$$P(X=x|N_1, r_1, t_1) = \frac{\binom{x-1}{t_1-1}\binom{N_1-x}{r_1-t_1}}{\binom{N_1}{r_1}}, \quad x=t_1, (t_1+1), ..., (N_1-r_1+t_1) \tag{2.1}$$

If he/she does not belongs to the group A, then he/she draws balls using without replacement sampling from the Urn-II until he/she gets t_2 $(< r_2)$ balls bearing the statement (ii.), and reports the total number of balls, say$Y$, drawn by him/her. Thus Y also follows a negative hypergeometric distribution, but with different parameters, given by:

$$P(Y=y|N_2, r_2, t_2) = \frac{\binom{y-1}{t_2-1}\binom{N_2-y}{r_2-t_2}}{\binom{N_2}{r_2}}, \quad y=t_2, (t_2+1), ..., (N_2-r_2+t_2) \tag{2.2}$$

Obviously, the distribution of the observed responses Z_i is given by:

$$Z_i = \begin{cases} X & \text{if } i \in A \\ Y & \text{if } i \in A^c \end{cases} \tag{2.3}$$

Then, we have the following theorems:

Theorem 2.1 *An unbiased estimator of the population proportion π is given by:*

$$\hat{\pi}_{new} = \frac{\frac{(r_1+1)(r_2+1)}{n}\sum_{i=1}^{n} Z_i - t_2(r_1+1)(N_2+1)}{t_1(N_1+1)(r_2+1) - t_2(N_2+1)(r_1+1)} \tag{2.4}$$

Proof Obviously because the expected value of Z_i is given by

$$E(Z_i) = \frac{1}{(r_1+1)(r_2+1)}\left[\pi\ t_1(N_1+1)(r_2+1) + (1-\pi)t_2(N_2+1)(r_1+1)\right] \tag{2.5}$$

□

Theorem 2.2 *The variance of the new estimator $\hat{\pi}_{new}$ is given by:*

$$V(\hat{\pi}_{new}) = \frac{\pi(1-\pi)}{n} + \frac{\pi t_1(N_1+1)(N_1-r_1)(r_1+1-t_1)(r_2+1)^2(r_2+2) + (1-\pi)(N_2+1)(N_2-r_2)(r_2+1-t_2)(r_1+1)^2(r_1+2)}{n(r_1+2)(r_2+2)\left\{t_1(r_2+1)(N_1+1) - t_2(r_1+1)(N_2+1)\right\}^2} \tag{2.6}$$

Proof The expected value of Z_i^2 over the randomization device is given by:

$$E(Z_i^2) = \frac{\pi}{(r_1+1)^2(r_1+2)}\left\{t_1(N_1+1)(N_1-r_1)(r_1+1-t_1)+t_1^2(N_1+1)^2(r_1+2)\right\}$$
$$+\frac{(1-\pi)}{(r_2+1)^2(r_2+2)}\left\{t_2(N_2+1)(N_2-r_2)(r_2+1-t_2)+t_2^2(N_2+1)^2(r_2+2)\right\} \quad (2.7)$$

By the definition of variance, we have

$$V(\hat{\pi}_{new}) = V\left[\frac{\frac{(r_1+1)(r_2+1)}{n}\sum_{i=1}^{n} Z_i - t_2(r_1+1)(N_2+1)}{t_1(N_1+1)(r_2+1)-t_2(N_2+1)(r_1+1)}\right]$$
$$= \frac{\frac{(r_1+1)^2(r_2+1)^2}{n^2}\sum_{i=1}^{n} V(Z_i)}{\{t_1(N_1+1)(r_2+1)-t_2(N_2+1)(r_1+1)\}^2}$$
$$= \frac{\frac{(r_1+1)^2(r_2+1)^2}{n}\sigma_Z^2}{\{t_1(N_1+1)(r_2+1)-t_2(N_2+1)(r_1+1)\}^2} \quad (2.8)$$

where

$$\sigma_Z^2 = V(Z_i) = E(Z_i^2) - \{E(Z_i)\}^2 \quad (2.9)$$

On using (2.5) and (2.7) in (2.9), and later using it in (2.8), we get (2.6). Hence the theorem. □

3 Efficiency comparisons

In this section, we study the relative efficiency of the newly proposed randomization device over the [1], [3], and [5] estimators. Note that the balls in the two urns can be adjusted in many ways to get efficient results and co-operation with respondents. In this illustration, our aim is to beat the Mangat [3] model. We suggest choosing a randomization device in which Urn-I consists of $N_1 = 10$ identical balls (say, Ping pong balls), out of which $r_1 = 7$ bear the statement, “I belong to the sensitive group A”, and the rest of the three balls are blank. Urn-II consists of $N_2 = 12$ identical balls, out of which $r_2 = 6$ bear the statement, “I do not belong to statement A”, and the rest of the six balls are blank. Each respondent selected in the sample is requested to report the number of balls drawn by him/her once he/she collects $t_1 = t_2 = 5$ balls. Table 1 reports the relative efficiencies of the proposed estimator with respect to the Kuk’s model with parameters suggested by one of the reviewer i.e. $\theta_1 = 0.8, \theta_2 = 0.2$ (or $\theta_1 = 0.2, \theta_2 = 0.8$).

4 Discussion of results

Note that we are in the same boat as in the Mangat [3] model as far as the degree of protection of respondents is concerned, because if a respondent reports more than 10 balls, that is, either 11 balls or 12 balls, then this respondent surely belongs to the non-sensitive group. We noted that there is a gain over the Mangat [3] model for $P = 0.7$ and the value of π less than or equal to 0.35. In practice, it is usually the case that the proportion of those having the sensitive characteristic is small, so the proposed model seems to perform better than the Mangat [3] model in these realistic situations. No doubt it is always more efficient than the use of the Warner [5] and Kuk [1] models for the situations listed in Table 1. For the choice

Table 1 Relative efficiency with respect to the Kuk's model

	$\theta_1 = 0.8 \quad \theta_2 = 0.2$
π	RE(K)
0.05	129.40
0.10	129.43
0.15	129.78
0.20	130.41
0.25	131.31
0.30	132.47
0.35	133.90
0.40	135.65
0.45	137.74
0.50	140.25
0.55	143.29
0.60	146.98
0.65	151.53
0.70	157.23
0.75	164.57
0.80	174.30
0.85	187.80
0.90	207.69

$\theta_1 = 0.8, \theta_2 = 0.2$ in the Kuk's model, the relative efficiency changes from 129.40% to 207.69% as the value of π changes from 0.05 to 0.90. The relative efficiency values remain same for $\theta_1 = 0.2, \theta_2 = 0.8$ as for $\theta_1 = 0.8, \theta_2 = 0.2$. Thus, we conclude that the proposed new randomization device can be efficiently and cooperatively used in real practice to estimate the proportion of a sensitive characteristic.

Special case: We note that if the balls are replaced back into the urns before making the next draws, then X and Y follow Negative Binomial distribution and lead to a different estimator. With this one, no one losses privacy but the process of collecting data may go on for a very long time.

The following remarks address very constructive comments/suggestions given by one of the reviewers:

Remark 1 An alternative to having balls labelled with statements explicitly referring to the sensitive characteristic is to use ones with a more neutral distinction such as different colours. This has the advantage of not constantly bringing up the sensitive characteristic. It also allows the same device to be used for several questions—though the proportions are then forced to remain the same as well. A disadvantage is that the interviewer must be very clear about which colour ball the interviewee is to be looking for. It would be helpful to have it be the same for both urns as well. Our primary reason for having balls with explicit statements is to make comparisons with the other models more transparent.

Remark 2 Another alternative is to use a single urn, with neutral coloured balls, rather than two urns. An interviewee could be instructed to choose balls until he/she gets t_1 black ones (say) if he/she belongs to the sensitive group, or $t_2$white ones if he/she does not. This process must still be hidden from the interviewer. The advantage is that there is no special urn for the respondents belonging to the sensitive group. This might prevent respondents from the sensitive group to cheat by drawing balls from the urn for the non-sensitive group. One disadvantage is that the distributions used for the different groups cannot be chosen independently and it limits the possibilities of choices for the distributions.

Remark 3 No doubt in many RR designs the moment estimates can be outside the parameter space (being either negative or greater than one in the case of estimating a proportion) particularly if the sample size is not large enough. If the sample size is sufficiently large, then all the RR estimates have negligible chance of taking value outside the interval [0, 1]. (Please refer to Lee et al. [2]). Singh and Sedory [4] have also shown that maximum likelihood and chi-square estimates can also fail in RR sampling if the sample size is not sufficiently large and the proportion π of the sensitive characteristic is either close to zero or close to one.

Remark 4 It is conceivable that auxiliary information can be used with the device mentioned in this paper, but it is not apparently clear at this stage how it could be done.

5 Summary

The paper presents a new randomized response (RR) design with two urns. One urn is to be used by the respondents from the sensitive group, and the other by respondents from the non-sensitive group. Each urn is filled with two different kind of balls that are either blank or marked. In the proposed design, respondents are asked to draw as many balls as necessary from the urn belonging to their group until a specified number of marked balls is reached. The number of balls is then reported. By using different proportions of marked balls in each urn, we are able to estimate the probability of belonging to the sensitive group in the population. An efficiency study shows that the design compares well to the Warner, Kuk and Mangat designs. In the past decade many RR designs have been proposed that try to find a good balance between privacy protection and efficiency. Perceived privacy protection is essential to gain the respondents' compliance with the design. From this point of view, an interesting aspect of the model is that the nature of the response (a random number) is different from the nature of the question of interest (i.e., a yes/no question). It is not unreasonable to think that this would increase cooperation by an interviewee.

Acknowledgments The authors are thankful the Editor-in-Chief Pier Luigi Conti, Editorial Staff Marco Alfo and a learned referee for very constructive comments on the original version of this manuscript.

References

1. Kuk, A.Y.C.: Asking sensitive questions indirectly. Biomerika **77**(2), 436–438 (1990)
2. Lee C.-S., Sedory, S.A., Singh, S.: Simulated minimum sample sizes for various randomized response models. Commun. Stat. Simul. Comput. **42**, 771–789 (2013)
3. Mangat, N.S.: An improved randomized response strategy. J. R. Statist. Soc. B **56**, 93–95 (1994)
4. Singh, S., Sedory, S.A.: A true simulation study of three estimators at equal protection of respondents in randomized response sampling. Statist. Neersl. **66**(4), 442–451 (2012)
5. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**, 63–69 (1965)