Statistics and Applications {ISSN 2452-7395 (online)}
Volume 15 Nos. 1&2, 2017 (New Series), pp 193-199

Two-Stage Randomized Response Group-Testing Model

Neeraj Tiwari and Prachi Mehta
Department of Statistics, Kumaun University, S.S.J.Campus, Almora, India

Received: January 14, 2017; Reviewed: February 04, 2017 ; Accepted: February 16, 2017

Abstract

Randomized Response Technique (RRT) is a survey technique for improving the reliability of responses to sensitive questions. RRT has been first developed by Warner (1965). Group-testing is a sampling scheme in which measurements are taken simultaneously in groups. Kim and Heo (2013) incorporated the group testing method into the Warner's randomized response (RR) model and unrelated question RR model. In this paper, a new estimator is proposed by applying the theory of group-testing in two-stage randomized response (RR) model given by Mangat and Singh (1990). It has been empirically established that the proposed estimator is more efficient than the estimator developed by Mangat and Singh (1990). By using group testing in RR model, the cost of the survey is also reduced and the respondent's privacy is increased.

Keywords: Randomized response technique, Group testing, sensitive information, cost of survey, privacy of the respondent.

1. Introduction

Survey is carried out to extract specific data from particular group of the people which will subsequently serve various purposes in field of politics, public health, professional organization, advertising etc. Direct questioning method may not give required results, specially in case of sensitive questions that involved social stigma and taboo.

In the year 1965, S. L. Warner proposed a research method that allowed respondent to respond the questions of sensitive nature while maintaining the confidentiality of the respondent. This technique was known as Randomized Response Technique (RRT). Under this technique the response of the respondent towards any sensitive question was randomized so that respondent's privacy was protected. RRT uses probability theory to protect the respondent's privacy. Warner (1965) used a randomization device, by which each respondent chooses one of the two questions:

1) "Do you belong to A?" and
2) "Do you not belong to A?",

with probabilities p and (1-p) respectively, where A denotes the sensitive group and p is not equal to 0.5. After getting the 'yes' or 'no' answers, researcher estimates the value of required population proportion using the estimator suggested by Warner (1965).

Corresponding author: Neeraj Tiwari
E-mail: neerajtiwari.amo@gmail.com

In the model suggested by Warner (1965), both the questions have same sensitivity level because the second question was just a negation of the first question, thus both questions were interrelated. Greenberg et al. (1969) introduced an unrelated question technique, where one question possessed the sensitivity whereas the other question was different and unrelated to the sensitive question. For example one question may be- “Do you take drugs?” and second one may be- “Is red your favourite colour”? Greenberg et al. (1971) proposed and developed the unrelated question randomized response design for estimating the mean and the variance of the distribution of a quantitative variable. In this model the first question may be – “How many abortions have you had during your lifetime?” and second one may be- “How many magazines do you subscribe to?”.

Mangat and Singh (1990) developed a more efficient randomized response technique known as two-stage randomized response technique, in which instead of taking one randomization device they used two different randomization devices to get more efficient estimator and respondent’s cooperation.

An Optional Randomized Response (ORR) model was introduced by Gupta (2001). In this model, the respondent gives a scrambled answer if he/she feels that the survey question is sensitive and a true response if it is non-sensitive. So, there is a choice or an option for the respondents to give their answers. Using the idea of ORR model given by Gupta (2001), an improved ORR model was suggested by Gupta et al. (2002). Gupta et al. (2002) showed that their estimator was more efficient than the usual estimators based on randomized response technique.

Ryu et al. (2006) proposed a new quantitative randomized response model based on two-stage randomized response model developed by Mangat and Singh (1990) and showed that their model is more efficient than the randomized response (RR) model suggested by Gupta et al. (2002).

Tiwari and Mehta (2016) proposed an improved methodology for RRT, in which the sensitivity level was considered to be known and the RR technique was applied only for those respondents who considered the particular question to be sensitive. Their method was simpler and more efficient compared to the existing methods. Tiwari and Mehta (2017) used the idea of known sensitivity level proposed by Tiwari and Mehta (2016) to One-Stage Optional RRT, Two-Stage Optional RRT and Three-Stage Optional RRT based on quantitative data and established that the proposed estimators were more efficient than the estimators based on ORRT models. While Tiwari and Mehta (2016) used the idea of known sensitivity level to qualitative data, it was applied for quantitative data by Tiwari and Mehta (2017).

Group-testing is a sampling scheme in which simultaneous measurements are obtained, rather than obtaining measurements on individuals. Under this scheme several units are pooled and a single test applied to the entire group. Group testing improves the estimator’s efficiency as well as reduces the cost for the data collection.

Hughes-Oliver and Swallow (1994) developed a method for adaptively estimating a proportion using group testing procedure and analyzed with emphasis placed on a two-stage procedure. Hughes-Oliver and Rosenberger (2000) extended the work of Hughes-Oliver and Swallow (1994) and developed an adaptive two-stage design for group testing of multiple rare traits. They derived the optimum group sizes and applied their design to a problem of estimating the prevalence of HIV, Chlamydia and syphilis in Ethiopian women. Kim and Heo

(2013) used the concept of group testing in the models suggested by Warner (1965) and Greenberg et al. (1969) and showed that the randomized response group-testing (GT) models proposed by them were more efficient than the traditional randomized response models.

In this manuscript, an attempt has been made to develop a more efficient and cost effective technique, known as two-stage randomized response group-testing model. In this technique, we have incorporated group-testing method in two-stage RR model proposed by Mangat and Singh (1990). It has been empirically established that the proposed estimator appears to be more efficient than the estimator suggested by Mangat and Singh (1990).

In Section 2, group-testing and two-stage RR model have been discussed. Two-stage randomized response group-testing (RR-GT) model has been proposed in Section 3. The proposed model has been empirically compared with the existing model in Section 4. The findings of the paper have been concluded in Section 5.

2. Group-Testing Method and Two-Stage Randomized Response Model

In this section, a brief description about group-testing method and two-stage randomized response (RR) model suggested by Mangat and Singh (1990) is given.

2.1 Group-testing method

The group-testing (GT) refers to a sampling procedure in which simultaneous measurements are obtained rather than obtaining measurements on individuals in any population. Group-testing starts with the premise that experimental units (or individuals) can be combined into groups of size g >1. For simplicity, we have taken common group size or same size for all groups as in the case of the model suggested by Kim and Heo (2013).

In group-testing model, a simple random sample of m groups is selected, each group having a common group size of *g* (*g* >1). Then one unit is selected from each group to test whether it is sensitive or not. If the selected unit is sensitive in any group, we discard the entire group and if selected unit is not sensitive we take the entire group as it is. This process is repeated for each group. Assuming there are no reporting errors, the number of sensitive groups observed, say *T*, has a binomial distribution with parameters m and $1-(1-\pi)^g$, where π denotes the true proportion of persons possessing the sensitive characteristics. Under the group-testing model, the maximum likelihood estimator (MLE) of π is given by

$$\hat{\pi}_{MLE} = 1-(1-T/m)^{1/g}. \tag{1}$$

2.2 Two-Stage Randomized Response (RR) model [Mangat and Singh (1990)]

In the two-stage model suggested by Mangat and Singh (1990), one more randomization device was added into the method given by warner (1965), in order to have more truthful answers. This process included two stages. In the first stage, first randomization device has two options:

(1) "Do you possess the sensitive character A?" and

(2) "Go to the second randomization device",

with probabilities 'Q' and '(1-Q)' respectively. The second randomization device is the same as the Warner's randomization device discussed earlier. Thus in the second stage, respondent follows Warner's method and chooses one of the two questions:

(1) “Do you possess the sensitive characteristic?” and
(2) “Do you not possess the sensitive characteristic?”
with probabilities ‘P’ and ‘(1-P)’ respectively and P is not equal to 0.5. Then the probability of ‘yes’ responses (θ_1) is given by

$$\theta_1 = Q\pi + (1-Q)[P\pi + (1-P)(1-\pi)], \tag{2}$$

where π is the population proportion with the sensitive characteristic. Solving equation (2) for estimating π, we get

$$\hat{\pi}_M = \frac{\hat{\theta}_1 - (1-P)(1-Q)}{(2P-1)+2Q(1-P)}, \tag{3}$$

where $\hat{\pi}_M$ is the estimator proposed by Mangat and Singh (1990) and $\hat{\theta}_1$ is the proportion of “yes” responses in the survey. Here P and Q should be chosen in such a way that the denominator should not be equal to 0.

The variance of the estimator $\overset{\wedge}{\pi_M}$ is given by-

$$V(\hat{\pi}_M) = \frac{1}{[(2P-1)+2Q(1-P)]^2} V(\hat{\theta}_1)$$

$$V(\hat{\pi}_M) = \frac{1}{[(2P-1)+2Q(1-P)]^2} \frac{\theta_1(1-\theta_1)}{n}, \tag{4}$$

where $P \neq 0.5$.

3. The Proposed Two-Stage Randomized Response Group-Testing (RR-GT) Model

In the proposed model, we have applied the group-testing (GT) method to two-stage randomized response model suggested by Mangat and Singh (1990). The proposed model gives the greater efficiency as compared to the estimator suggested by Mangat and Singh (1990) as well as reduces the cost of the survey.

In this method, first we divide the whole population (of size N) in to ‘M’ homogeneous groups (of common group size ‘g’), based on some prior or auxiliary information which is already available with us. For example – a researcher wants to select a sample of students to estimate the proportion of students who have used cocaine in the campus, the researcher gets the prior information like GPAs of the students and divide the whole population into groups according their GPAs. Then we select a simple random sample of ‘m’ homogeneous groups out of ‘M’ homogeneous groups and randomly select a representative from each group. The group representative follows the two-stage randomized response technique with the help of two randomization devices, as discussed earlier in Section 2.2. By following two-stage method, the question which comes to the group representative is put before everyone in the group by the group representative and instructed to respond ‘yes’ or ‘no’. The entire information is passed to the interviewer by the group representative. The entire procedure is completed by the group representatives, unobserved by the interviewer and thus the individual’s privacy is maintained. Assuming all responses are true and there are no reporting errors, then the observed number of groups reporting ‘yes’, say T_1, has a binomial distribution with parameters ‘m’ and ‘π_g’ where $\pi_g = 1-(1-\theta_1)^g$ and θ_1 is the probability of having ‘yes’ responses. The value of θ_1 is-$\theta_1 = Q\pi + (1-Q)[P\pi + (1-P)(1-\pi)]$. Using the method of moments, we can derive

$$\hat{\pi}_g = \frac{T_1}{m} \tag{5}$$

Solving equation (5), we get,

$$\hat{\pi}_{M_N} = \frac{(P+Q-PQ)-(1-T_1/m)^{1/g}}{(2Q+2P-2PQ-1)} \tag{6}$$

where $\overset{\Lambda}{\pi}_{M_N}$ denotes the proposed estimator based on group-testing and P and Q are selected in such a way that the denominator is not equal to zero. Expanding equation (6) by the binomial expansion we have

$$\hat{\pi}_{M_N} \approx \frac{(P+Q-PQ)}{(2Q+2P-2PQ-1)} - \left[\frac{1-g^{-1}\left(\frac{T_1}{m}\right)+\frac{g^{-1}(g^{-1}-1)}{2}\left(\frac{T_1}{m}\right)^2-\frac{g^{-1}(g^{-1}-1)(g^{-1}-2)}{6}\left(\frac{T_1}{m}\right)^3+......}{(2Q+2P-2PQ-1)}\right] \tag{7}$$

Neglecting the terms having power of (T_1/m) more than 2 in equation (7), the variance of the estimator $\hat{\pi}_{M_N}$ is given by

$$V(\hat{\pi}_{M_N}) = \frac{1}{(2Q+2P-2PQ-1)^2} V\left[g^{-1}\frac{T_1}{m}-\frac{g^{-1}(g^{-1}-1)}{2}\left(\frac{T_1}{m}\right)^2\right]$$

or

$$V(\hat{\pi}_{M_N}) = \frac{1}{(2Q+2P-2PQ-1)^2}\left[\left(\frac{g^{-1}}{m}\right)^2 V(T_1)+\left\{\frac{g^{-1}(g^{-1}-1)}{2m^2}\right\}^2 V(T_1^2)-2\left\{\frac{g^{-2}(g^{-1}-1)}{2m^3}\right\}\mathrm{cov}(T_1,T_1^2)\right] \tag{8}$$

where $V(T_1) = E(T_1^2)-[E(T_1)]^2$

$$V(T_1^2) = E(T_1^4)-[E(T_1^2)]^2$$

$$Cov(T_1,T_1^2) = E(T_1^3)-E(T_1)E(T_1^2)$$

Expected values of T_1, T_1^2, T_1^3 *and* T_1^4 can be obtained using moment generating function and are given as,

$$E(T_1) = m\pi_g$$

$$E(T_1^2) = m\pi_g(1-\pi_g+m\pi_g)$$

$$E(T_1^3) = m\pi_g(1-3\pi_g+3m\pi_g+2\pi_g^2-3m\pi_g^2+m^2\pi_g^2)$$

$$E(T_1^4) = m\pi_g(1-7\pi_g+7m\pi_g+12\pi_g^2-18m\pi_g^2+6m^2\pi_g^2-6\pi_g^3+11m\pi_g^3-6m^2\pi_g^3+m^3\pi_g^3)$$

4. Efficiency comparison

In this section, we have compared the proposed estimator $(\hat{\pi}_{M_N})$ with the estimator $(\hat{\pi}_M)$ given by Mangat and Singh (1990) using an artificial data. In this study 'n=m.g', where 'm' is the number of groups selected in the sample out of total 'M' groups in the population and 'g' is the group size (g>1), which is assumed to be same for all m groups. Note that 'n' is the total number of individuals in the sample and 'N' is the total number of individuals in the population (N=M.g). Here 'm' and 'g' are used in the proposed estimator and 'n' is used in the traditional estimator.

The relative efficiency (RE) of proposed estimator $(\hat{\pi}_{M_N})$ as compared to the estimator $(\hat{\pi}_M\ )$ given by Mangat and Singh (1990) is given by,

$$RE(\hat{\pi}_{M_N}, \hat{\pi}_M) = \frac{V(\hat{\pi}_M)}{V(\hat{\pi}_{M_N})} \tag{9}$$

To demonstrate the utility of the proposed estimator, we have considered an artificial data and obtained the relative efficiency of the proposed estimator compared to the estimator suggested by Mangat and Singh (1990). The relative efficiency of the proposed estimator $(\hat{\pi}_{M_N})$ in comparison to the estimator$(\hat{\pi}_M)$proposed by Mangat and Singh (1990) for different sample sizes are given in Table 1.

Table 1: The Relative Efficiency of the Proposed Estimator $(\hat{\pi}_{M_N})$ in Comparison to the Estimator $(\hat{\pi}_M)$ Proposed by Mangat and Singh (1990) for Different Values of Groups (*m*)

g	P	Q	π	m=20	m=25	m=30
				RE	RE	RE
2	0.7	0.4	0.1	1.43742	1.43800	1.43839
2	0.7	0.4	0.2	1.59802	1.59812	1.59834
2	0.7	0.4	0.3	1.79231	1.79219	1.79213
2	0.7	0.4	0.4	2.03467	2.03393	2.03345
2	0.7	0.4	0.5	2.34552	2.34494	2.34289
2	0.7	0.4	0.6	2.75754	2.75480	2.75300
2	0.7	0.4	0.7	3.32596	3.32216	3.31879
2	0.7	0.4	0.8	4.15222	4.14553	4.14117
2	0.7	0.4	0.9	5.44498	5.43439	5.42755

From Table 1, it is evident that for all sample of groups i.e. for $m = 20$ ($n = 40$), $m = 25$ ($n = 50$), $m = 30$ ($n = 60$) with $g = 2$, $P = 0.7$, and $Q = 0.4$ the relative efficiency (RE) is greater than 1 and increases as the value of π increases from 0.1 to 0.9. Thus, we can say that the proposed estimator $(\hat{\pi}_{M_N})$ appears to be more efficient than the traditional estimator $(\hat{\pi}_M)$ proposed by Mangat and Singh (1990) and the efficiency of the proposed estimator increases as π increases from 0.1 to 0.9. There is a slight gain in relative efficiency (RE) as the number of groups (m) increases from 20 to 30 when $\pi<0.3$. In addition to gain in efficiency, the group testing model also reduces the cost of survey and provides more protection against privacy of the respondent.

5. Conclusion

In this article, we have implemented the concept of group-testing in two-stage randomized response (RR) model developed by Mangat and Singh (1990) and suggested an improved model, known as two-stage randomized response group-testing (RR-GT) model. Empirically, it has been established that the proposed two-stage RR-GT model is more efficient than the simple two-stage RR model given by Mangat and Singh (1990). Another advantage of the group testing model is in reducing the cost of the survey and increasing the protection against privacy of the respondent, as the data collection is done on the basis of groups.

Acknowledgements

The authors are grateful to one of the referees for his/her constructive comments, which have led to considerable improvement in presentation of this manuscript.

References

Greenberg, B.G., Abul-Ela, A.L.A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model-theoretical framework. *Journal of the American Statistical Association*, **64 (326)**, 520-539.

Greenberg, B.G., Kuebler Jr., 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.

Gupta, S.N. (2001). Qualifying the sensitivity level of binary response personal interview survey questions. *Journal of Combinatorics, Information and System Sciences*, **26(1-4)**, 101-109.

Gupta, S.N., Gupta, B.C., Singh, S. (2002). Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inferences*, **100**, 239-247.

Hughes-Oliver, J.M. and Rosenberger, W.F. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika*, **87(2)**, 315-327.

Hughes-Oliver, J.M. and Swallow, W.H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association*, **89(427)**, 982-993.

Kim, J.M. and Elam, M.E. (2005). A two-stage stratified Warner's randomized response model using optimal allocation. *Metrika*, **61**, 1-7.

Kim, J.M. and Heo, T.Y. (2013). Randomized response group testing model. *Journal of Statistical Theory and Practice*, **7(1),** 33-48.

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

Ryu, J.B., Kim, J.M., Heo, T.Y., Heo, T.Y. and Park, C.G. (2006). On stratified randomized response sampling. *Model Assisted Statistics and Applications*, **1**, 31-36.

Tiwari, N. and Mehta, P. (2016). An improved two stage optional RRT model. *Journal of Indian Society of Agricultural Statistics*, **70(3)**, 197-203.

Tiwari, N. and Mehta, P. (2017). Additive optional randomized response model with known sensitivity level. *Communications in Statistics-Theory and Methods*. Under Review.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60 (309)**, s63-69.