Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 1, 2023 (New Series), pp 207–215

Randomized Response in Combination with Direct Response for Estimation of Incidence Parameters of Two Sensitive Qualitative Features

Opendra Salam
Department of Statistics
Manipur University, India

Received: 10 July 2021; Revised: 24 February 2022; Accepted: 09 July 2022

Abstract

We consider a situation wherein we are dealing with two sensitive qualitative features (SQlFs) say Q_1^* and Q_2^* with respective incidence proportions/parameters P_1^* and P_2^* [both unknown]. In a given sample of n respondents, some respondents will be comfortable with Q_1^*; some will be comfortable with Q_2^*; some respondents will be comfortable with both the features while some others will not be comfortable with either of the two. Here 'comfortable' refers to the situation wherein the respondent is agreeable to provide 'direct (yet, truthful) response' to the sensitive feature under consideration. Therefore, there are 4 obvious categories of respondents in a random sample of any reasonable size. The same categorization holds in the entire population of respondents in an analogous manner. Our objective is to provide unbiased estimates for the incidence parameters of the two categories, based on data [of responses from all the four types of respondents] accrued from a survey.

Key words: Qualitative features; Sensitive features; Direct response; Randomized response; Population proportion; Binary response; Designing the survey; Combination of estimates.

1. Introduction

In sample survey, the study variable may be sensitive in nature; as for example, it may be related to "addiction" to a drug, being a "habitual gambler", having a history of "abortion", "extramarital affairs" and the like. For such items of information, generally, the respondents may be reluctant to provide "Direct" and yet "Truthful" responses. It is also possible that a fraction of the respondents are quite 'comfortable' with such questions and are agreeable to respond truthfully without any social embarrassment. For two such sensitive features, naturally, there are 4 types of respondent-categories, as explained in the abstract.

As is well-known, Warner (1965) introduced "Randomized Response Techniques [RRTs]" to address such questions of eliciting information on sensitive qualitative features. In 2015, there was world-wide celebration of "Fifty Years of RRT" and Handbook of Statistics, Volume 34 was published. Research continues in this fascinating area of survey sampling - theory and applications - dealing with sensitive issues. We refer to the books / book chapters on RRT by Fox and Tracy (1986), Chaudhuri and Mukerjee (1988), Hedayat and Sinha (1991),

Corresponding Author: Opendra Salam
Email: sopendra@manipuruniv.ac.in

Chaudhuri (2011), Chaudhuri and Christofides (2013), and Mukherjee *et al.* (2018). For the work in this paper, we refer to (i) Nandy, Marcovitz and Sinha (2015), (ii) Sinha (2017), (iii) Nandy and Sinha (2020) and (iv) Salam and Sinha (2020). This last reference will be mainly used in the present study.

We contemplate a situation wherein we are dealing with two sensitive qualitative features [SQlFs] Q_1^* and Q_2^*, with corresponding incidence proportions P_1^* and P_2^* respectively in the population as a whole. However, a fraction of the sampled respondents are known to be 'comfortable' with 'Direct Response' to one or the other or both of the SQlFs Our aim is to unbiasedly estimate both these proportions P_1^* and P_2^*, based on the complete survey data.

Naturally, we have 4 different categories of responding units in the population of size N, as also in a randomly drawn sample of reasonably adequate size n. Without any loss of generality, we may start with the following table of classification of the population of N respondents.

Table 1: 2-Way classification of respondents

Type	Comfortable with Q_2^*	Uncomfortable with Q_2^*	Total size
Comfortable with Q_1^*	$N(C,C)$	$N(C,NC)$	$N(C,.)$
Uncomfortable with Q_1^*	$N(NC,C)$	$N(NC,NC)$	$N(NC,.)$
Total	$N(.,C)$	$N(.,NC)$	$N(.,.)$

In case of sample respondents, we use the obvious notations

$$n(C,C), n(C,NC), n(NC,C), n(NC,NC)$$

for the sample frequencies in the respective different categories. Under simple random sampling without replacement, it may be assumed that the sample counts in different categories are proportional to the respective population counts. Henceforth, we will assume simple random sampling of the respondents in each category.

It transpires that for respondents in the Category (C, C), we may freely address the two questions Q_1^* and Q_2^*, independent of one another, and extract truthful responses from each of the sampled respondents. Again, for the Category (C, NC) [respectively, (NC, C)], only the question Q_1^* [resp., Q_2^*] can be put forward directly and truthful responses may be extracted from the relevant respondents. The other question has to be handled by taking recourse to an RRT. For the Category (NC, NC), we must adopt some kind of RRT for simultaneous estimation of the underlying parameters. For this latest kind of subpopulation of respondent categories, we may take recourse to the procedures studied in recent years. Vide Salam and Sinha (2020), for example.

Remark 1: As a general principle, for unbiased estimation of a population proportion π [of "Yes" responses] based on the available responses on a binary [Yes/No] response feature, it is well-known that the sample proportion p [of 'Yes' responses] is an unbiased estimate for the corresponding population proportion π. Moreover, in order to combine information on the common proportion π arising out of different/disjoint independent samples, we collect head-counts of all the 'Yes' responses from different sources together and divide it by the total count of respondents. Recall the formula $\hat{p} = \frac{\sum_i x_i}{\sum_i n_i}$.

Remark 2: From the nature of the sampled respondents in three of the four categories, it transpires that RRT provides estimate(s) of the desired proportion(s) involving the sensitive feature(s). From there, we work out estimate(s) of the number of respondents in the relevant 'yes' category of the sensitive feature. Recall $\hat{p}_i = x_i/n_i$ from the i^{th} source so that $x_i = n_i\hat{p}_i$. Then we go by the technique of combination of evidences from different sources, as explained in Remark [1].

The rest of the paper is organized as follows. In Section 2, we briefly outline the general approach for tackling the problem stated above. Then, in Section 3, we consider an illustrative example to work out all the essential details. Finally, in Section 4, we provide some concluding remarks.

2. General approach for analysis of data

We refer to the general description of the 4 categories of respondents as laid down in Section 1.

Set $NP_i^* = N_i^*, i = 1, 2$ and, further, consider the natural and obvious decomposition of N_i^* as

$$N_i^* = N_i^*(C,C) + N_i^*(C,NC) + N_i^*(NC,C) + N_i^*(NC,NC), i = 1, 2.$$

We also express the above quantities - quite meaningfully - as

$$N_1^* = N_1^*(C,.) + N_1^*(NC,C) + N_1^*(NC,NC); N_2^* = N_2^*(.,C) + N_2^*(C,NC) + N_2^*(NC,NC).$$

Note that both the quantities $N_1^*(C,.)$ and $N_2^*(.,C)$ are amenable to unbiased estimation by direct questionnaire method. For $N_1^*(C,.) = N_1^*(C,C) + N_1^*(C,NC)$ population units, in view of simple random sampling, we know (i) the number of "Yes" respondents among $n_1^*(C,C)$ sampled respondents wrt the Category Q_1^* and also (ii) the number of "Yes" respondents among $n_1^*(C,NC)$ sampled respondents wrt the Category Q_1^*. Likewise, we have direct "Yes" responses for Q_2^* from $n_2^*(C,C)$ respondents randomly sampled from $N_2^*(C,C)$ respondents in the reference subpopulation and also, we have direct "Yes" responses for Q_2^* from $n_2^*(NC,C)$ respondents randomly sampled from $N_2^*(NC,C)$ respondents in the reference subpopulation.

These four separate count estimates of "Yes" categories are the ingredients for arriving at final estimates of P_1^* and P_2^*.

For unbiased estimation of $N_1^*(NC,C)$ or of $N_2^*(C,NC)$, we can follow the technique as in Section 2 [Subsections 2.1, 2.2 and 2.3] of Salam and Sinha (2020) - appropriately adjusted for our purpose- for unbiased estimation of the corresponding proportions *i.e.*, $P_1^*(NC,C) = \frac{N_1^*(NC,C)}{N_1(NC,C)}$ and $P_2^*(C,NC) = \frac{N_2^*(C,NC)}{N_2(C,NC)}$. Finally, for unbiased estimation of $N_1^*(NC,NC)$ and $N_2^*(NC,NC)$, we refer to Subsection 2.4 of Salam-Sinha (2020) paper which provides formulae for simultaneous unbiased estimation of the underlying population proportions.

3. Worked out example

We consider a large population of $N = 30,000$ respondents - broadly classified in the 4 categories as

$$N(C,C) = 2000, N(C,NC) = 3000, N(NC,C) = 2000, N(NC,NC) = 23000.$$

This information is a priori available to the investigating agency. In a random sample of $n = 3000$ respondents, the stratified sample sizes under proportional allocation are taken as $[200, 300, 200, 2300]$.

(i) Data Type : (C, C) Simple and direct implementation of the questionnaire yields: For Q_1^*, freq. count of "Yes" = 85; for Q_2^*, it is 56.

(ii) Data Type :(C, NC) (a) Direct implementation of Q_1^* yields : Freq. Count of "Yes" = 114.
(b) Implementation of technique adopted in Subsection 2.2 of Salam-Sinha (2020) paper yields: $\hat{P}_2^*(C,NC) = 0.35$.

(iii) Data Type :(NC, C) (a) Implementation of technique adopted in Subsection 2.2 of Salam-Sinha (2020) paper (Annexure 1) yields: $\hat{P}_1^*(NC,C) = 0.38$. (b) Direct implementation of Q_2^* yields: Freq. Count of "Yes" = 94.

(iv) Data Type :(NC, NC) (a) Implementation of technique adopted in Subsection 2.4 of Salam-Sinha (2020) paper yields: $\hat{P}_1^*(NC,NC) = 0.42$.
(b) Implementation of technique adopted in Subsection 2.4 of Salam-Sinha (2020) paper yields: $\hat{P}_2^*(NC,NC) = 0.33$.

Remark 3: The reader will find repeated use of a result from Salam-Sinha(20202) paper. One referee has suggested that it would be instructional to explain the methodology from that paper. Not to obscure the essential steps of reasoning, we will proceed through the following steps, using critical close arguments as in Salam-Sinha (2020) paper. For ready reference we reproduce the techniques and computations from the cited paper.

Now we are in a position to provide (unbiased) estimates for P_1^* and P_2^*.

We display all the four source information on each of the two sensitive features in the following table.

Table 2: 2-Way classification of observed / estimated number of "Yes" responses for the two features

Type	Comfortable with Q_2^*	Uncomfortable with Q_2^*	Total Figures for (Q_1^*)
Comfortable with Q_1^*	$(85/200, 56/200)$	$(114/300, 105/300)$	$(199/500)$
Uncomfortable with Q_1^*	$(76/200, 94/200)$	$(966/2300, 759/2300)$	$(1042/2500)$
Total Figures for (Q_2^*)	$(150/400)$	$(864/2600)$	$(1241/3000, 1014/3000)$

For the respondents who are comfortable with Q_1^* we may use the notation $N_1(C,C)$ and $N_1(C,NC)$ to denote the corresponding frequency counts with respect to Q_1^* . On the other hand for those who are not comfortable with Q_1^* we have to use RRT technique to estimate the proportions and hence the frequency counts in the two categories corresponding to (NC,C) and (NC,NC) have to be estimated indirectly. That is where Salam-Sinha (2020) technique has been used. Likewise we can use an analogous notation for cases involving Q_2^*.

We may thus conclude that

$$N_1^*(C,C) = 85; N_1^*(C,NC) = 114;$$

$$\hat{N}_1^*(NC,C) = 76; \hat{N}_1^*(NC,NC) = 966.$$

Therefore, $P_1^* = (85 + 114 + 76 + 966)/3000 = 1241/300 = 41.37$ *per cent*. Likewise, for estimation of P_2^*, we proceed similarly and derive an estimate as $P_2^* = 1014/3000 = 33.80$ *per cent*. Based on combined evidence of sample data covering both 'C', 'NC' for Q_1^* and Q_2^*, our estimation procedure produces estimates of proportions of Q_1^* and Q_2^* and we end up with $\hat{Q}_1^* = 0.41$ and $\hat{Q}_2^* = 0.34$ respectively.

4. Conclusion

For two Sensitive Qualitative Features along with a provision for Optional Randomization for either or both, we have considered a blend of the Randomised Response Technique and Direct Response Technique to estimate the two incidence parameters from a complex pattern of respondents' response profiles. Simple Random Sample of a reasonably large size is assumed to be available. For three or more Sensitive Qualitative Features with this kind of Optional Randomization for one or two or more of such features, it would be an interesting topic by itself to estimate all the incidence proportions. This seems to be a non-trivial generalization of our approach. Again, even for two such sensitive features with provision(s) for optional randomization, multi-category response profiles would be worth studying.

Acknowledgments

I am extremely grateful to Professor Bikas K Sinha, Retired Professor of ISI, Kolkata, for his continued interest in this work and for going through a preliminary version of this manuscript and offering suggestions to remove obscurities.

References

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

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys.* CRC Press, Chapman and Hall, Taylor and Francis Group, FL, USA.

Chaudhuri, A. and Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys.* Springer, Germany.

Coutts, E. and Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods and Research*, **40**, 169-193.

Fox, J. A. and Tracy, P. E. (1986). *A Method for Sensitive Surveys.* Sage University paper series. Quantitative Applications in the Social Sciences, 07-058. Sage Publication.

Hedayat, A. S. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling.* New York:Wiley.

Mukherjee, S. and Chattopadhyay, (2018). *Statistical Methods in Social Science Research.* Springer.

Nandy, K. and Sinha, B. K. (2020). Block total response technique for quantitative sensitive items in a finite population. *Statistics and Applications*, **18**, 85 − 95.

Nandy, K., Marcovitz, M. and Sinha, B. K. (2015). Eliciting information on sensitive features: Block total response technique and related inference. In *Handbook of Statistics*, **34**: Data gathering, analysis, and protection of privacy through randomized response techniques: qualitative and quantitative human traits, 317 - 328.

Opendra, S. and Sinha, B. K. Further thoughts on applications of block total response techniques in case of one or two sensitive binary features. Accepted; 2020 in press (publication scheduled for 2023 in *Thailand Statistician*)

Raghavarao, D. and Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society*, Series B, Methodological, 40-45.

Sinha, B. K. (2017). Some refinements of block total response technique in the context of RRT methodology. *Statistics and Applications*, **15**, 167-171.

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

APPENDIX

In 1965, Warner introduced a Randomized Response Technique/Methodology (RRT/RRM) and Raghavarao and Federer (1979) introduced a novel technique/methodology termed Block Total Response Technique/Methodology [BTRT/BTRM] to increase the degree of protection of confidentiality of the respondent's response. The technique is elaborated below.

Consider a collection of v Regular [Non-sensitive] Qualitative Features (RQlFs) $[Q_1, Q_2, \ldots, Q_v]$ and one Sensitive Qualitative Feature (SQlF) Q^*. Let b be the number of blocks (*i.e.*, sets of questions), each containing k (distinct) RQlFs and the SQlF Q^*. Each RQlF is replicated r times in the entire collection of b blocks and there is one block B_0 containing all RQlFs. Thus, there are $k + 1$ QlFs for each block and $(b + 1)$ blocks. The respondents in the sample are split into (b+1) sets of size $n^*, n^*, \ldots, n^*$, n_0 such that $n = (bn^* + n_0)$. That is, each of the b blocks received n^* respondents and block B_0 received n_0 respondents. A block of questions (which contains k NSBFs and the single SBF Q^*) is presented to each respondent. The respondent is to provide only the Block Total Response (BTR) in terms of the overall score (*i.e.*, only the total number of yes answers) without divulging any response to any specific Qs – be it NSBF or SBF. It is believed that this BTR Technique [BTRT] will adequately protect the privacy of the respondent, and hence, the correct response to the SBF will emerge. BTRT is an alternative method of RRT to increase respondent's anonymity and enable estimation of the parameter (*i.e.*, proportion of yes respondents in the population) involving sensitive binary feature. We know from BTRT that, in every block $[B_1 to B_b]$ we utilize only k of the v NSBFs, while the rest $(v - k)$ NSBFs are left unutilized. When k is small, respondents may feel uncomfortable responding truthfully since responding to Q^* is compulsory in each of the b blocks. Nandy and Sinha (2020) extended the above technique by bringing variations in the block compositions as:

1. List of k 'must respond' NSBFs are kept in Part A.

2. Remaining $(v-k)$ NSBFs and Q^* (SBF) are all kept in Part B.

A respondent is to choose one question from $(v-k+1)$ questions in part B and mix with the questions in Part A and supply BTR without divulging the identity of the question selected from Part B. Salam-Sinha (2020) introduced a purely random choice from both parts A and B as is explained below.

Suppose there are k_1 RQlFs in Part A and $k_2 = v-k_1+1$ RQlFs, including the sensitive question Q^* in Part B. Arrangement of k_1 RQlFs in Part A is the same as above. Respondent is to blend randomly selected s_1 RQlFs from k_1 RQlFs and s_2 from $k_2 = v-k_1+1$ RQlFs including the sensitive question Q^* and supply the BTR of (s_1+s_2) RQlFs possibly including the sensitive question without divulging any information about the selected questions. Let π_1 and π_2 denote the inclusion probabilities of i^{th} unit $[i = 1, 2, \ldots, k_1]$ RQlFs from Part A and inclusion probability of i^{th} RQlFs $[i = 1, 2, \ldots, (v-k_1+1)]$ from Part B [including Q^*] respectively. Therefore, every question in Parts A and B have an equal chance of inclusion *viz.*, $\pi_1 = s_1/k_1, \pi_2 = s_2/k_2$ respectively.

We have

$$\bar{x}_1 = \sum_{i=1}^{k_1} p_i(s_1/k_1) + P^*(s_2/k_2) + (\boldsymbol{\Delta} - \sum_{i=1}^{k_1} p_i)(s_2/k_2).$$

And, summing over all the b blocks, we derive the single estimating equation for P^* as

$$\sum_{i=1}^{b} \bar{x}_i = r\boldsymbol{\Delta}(s_1/k_1) + bP^*(s_2/k_2) + (b-r)\boldsymbol{\Delta}(s_2/k_2).$$

Therefore, the population proportion P^* of the sensitive qualitative question can be estimated by using the formula

$$\hat{P}^* = \frac{\sum_{i=1}^{b} \bar{x}_i - (s_1/k_1)2\boldsymbol{\Delta} - (b-r)\boldsymbol{\Delta} \times (s_2/k_2)}{b(s_2/k_2)}.$$

Note $\Delta = \sum_{i=1}^{10} p_i$ and $\hat{\Delta} = \bar{x}_0$ which stands for the sample mean of BTRs of the n_0 respondents in the block B_0.

Example

Suppose we have design parameters *i.e.*, $b = 5, v = 10, k = 4, r = 2$ and $n = 300$ respondents. we randomly split them into 6 sets, taking $n^* = 50$ and $n_0 = 50$. We adopt the same block compositions as are displayed above. We have $k_1 = k = 4$ and we select $s_1 = 3$ RQlFs from part A and we also select $s_2 = 3$ from part B. In order to implement the above scheme for Block 1, for example, we prepare a set of 11 identical cards of the same shape [square, say] and size. At the back of the cards, we write the numbers $1, 2, \ldots, 10$ and the symbol (*) one card for each. The procedure is : A respondent belonging to Block 1 is to draw three cards at random from the collection of first 4 cards [1 *to* 4]. Note that this is as good as selecting one card at random and discarding the same, and thereby, taking the rest at hand! Out of the remaining 7 cards, the respondent has to select any 3. Thus he/she will have a collection of 6 cards altogether from the two sets. Next, he/she will respond truthfully to all the 6 binary $[1-0]$ features selected and arrive at the Block Total Response and only the BTR score is supposed to be reported – without divulging any details. Note that the respondent may / may not have chosen the SBF (Q^*). Of course, he/she must respond truthfully and provide the BTR score – even if this has been selected. Likewise, we prepare 11 cards for use of the respondents in Block 2 and so on. Of course, each time we study the block composition before using the cards to form two designated sets. For the last block B_0, we do not need any cards. All NSBFs are compulsory. Having implemented the data-gathering tools, we end up with 'raw' scores of each of the 300 respondents – classified into 6 distinct groups. In each group, we calculate the group average of the scores and these are called 'summary statistics'. Assume that at the end we end up with the following results:

Table 1: Example : Summary statistics

Block	Total score	No. of respondents	Average score
1	128	50	2.56
2	136	50	2.72
3	146	50	2.92
4	125	50	2.50
5	115	50	2.30
6	220	50	4.40

Following Nandy and Sinha (2020), we may prepare the following table.

Table 2: Data analysis : Theory

Block	Sample size	Expected block average (EBA)	Average Score
1 (B_1)	n^*	EBA(1)	$\bar{x}_1$
2 (B_2)	n^*	EBA(2)	$\bar{x}_2$
3 (B_3)	n^*	EBA(3)	$\bar{x}_3$
4 (B_4)	n^*	EBA(4)	$\bar{x}_4$
5 (B_5)	n^*	EBA(5)	$\bar{x}_5$
6 (B_0)	n_0	$\mathbf{\Delta}$	$\bar{x}_0$

In the above,

$$EBA(1) = [(3/4)(p_1 + p_2 + p_3 + p_4) + (3/7)[P^* + (\mathbf{\Delta} - p_1 - p_2 - p_3 - p_4)],$$

$$EBA(2) = [(3/4)(p_5 + p_6 + p_7 + p_8) + (3/7)[P^* + (\mathbf{\Delta} - p_5 - p_6 - p_7 - p_8)],$$

$$EBA(3) = [(3/4)(p_9 + p_{10} + p_1 + p_2) + (3/7)[P^* + (\mathbf{\Delta} - p_1 - p_2 - p_9 - p_{10})],$$

$$EBA(4) = [(3/4)(p_3 + p_4 + p_5 + p_6) + (3/7)[P^* + (\mathbf{\Delta} - p_3 - p_4 - p_5 - p_6)],$$

$$EBA(5) = [(3/4)(p_7 + p_8 + p_9 + p_{10}) + (3/7)[P^* + (\mathbf{\Delta} - p_7 - p_8 - p_9 - p_{10})],$$

$$EBA(6) = \mathbf{\Delta} = \sum_{i=1}^{10} p_i.$$

Summing over all the first five block means, we obtain the estimating equation :

$$\sum_{i=1}^{5} \bar{x}_i = (3/4)2\mathbf{\Delta} + (3/7)5P^* + (3/7) \times 3\mathbf{\Delta}$$

$\bar{x}_0$ (Sample mean of BTRs of the n_0 respondents in the block B_0) = 4.40

Replacing $\mathbf{\Delta}$ by its estimate $\bar{x}_0$ and derive.

$$\hat{P}^* = [\sum_{i=1}^{5} \bar{x}_i - (39/14)\bar{x}_0] \times (7/15) = 0.35.$$

An estimate of the population proportion P^* of the sensitive qualitative question is 0.35.