DOI: https://doi.org/10.52131/pjhss.2023.1102.0408

Pakistan Journal of Humanities and Social Sciences

Volume 11, Number 02, 2023, Pages 971–978
Journal Homepage:
https://journals.internationalrasd.org/index.php/pjhss

Estimation of Sensitive Attributes Using Randomized Response Techniques in Sample Surveys

Muhammad Ahmed Shehzad[1], Ansa Bibi[2], Aamna Khan[3]

[1] Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan. Email: ahmad.shehzad@bzu.edu.pk
[2] Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan. Email: ansachohan898@gmail.com
[3] Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan. Email: aamnaa@bzu.edu.pk

ARTICLE INFO

Article History:
Received: April 10, 2023
Revised: May 19, 2023
Accepted: May 20, 2023
Available Online: May 20, 2023

Keywords:
Sensitive Questions
Randomized Response Technique
Horvitz Model
Cheater Detection

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The survey is a research technique that involves asking inquiries of a sample of components. Sample studies may draw attention to respondents' private or sensitive information. The tendency of respondents to respond on the basis of what is socially acceptable may harm the response rate of surveys. This may be because a large number of respondents decline to participate in the survey or give erroneous or conditional replies, which have a significant impact on the estimates' accuracy and dependability. To safeguard respondents' privacy and lessen the chance that they may avoid answering difficult questions, the Randomized Response (RR) approach was devised. Warner's work has spawned a large body of literature and been applied in numerous sectors, although these methods have challenges and drawbacks. The Horvitz Randomized Response method is employed in this article to address the sensitive properties. In order to obtain findings that are more accurate than those obtained by using previous RRT algorithms, we estimate two sensitive attributes: lying and bullying. To compute our findings, we employ the R codes.

Corresponding Author's Email: aamnaa@bzu.edu.pk

1. Introduction

A survey is a type of research that involves questioning from a group of people. In sample survey studies, the focus is frequently on complex or confidential characteristics of the interviewers. As a result, a common issue is to answer sensitive attributes, As a result, several respondent reject to contribute in the study or give incorrect or condition responses, significantly affecting the accuracy and reliability of the estimates, therefore this (RRT) technique are used to answer such type of complex questions. We desire to determine the quantity of cheating and bullying in students that is generally regarded as criminal, by applying the Randomized Response Technique (RRT). We might be able to gather accurate, trustworthy responses from these pupils using the RRT. Respondents won't be afraid to provide honest replies using this strategy because it will protect their identity. A possible method for preserving respondent anonymity is the Randomized Response Technique (RRT), which is used to questions on complex subjects to make it less likely that they would receive an incorrect or no response. Following Warner's finding, a growing body of research has been done on other methods for producing appropriate RR systems in direction to estimation a population proportion.

Studies that demand a double answer to a complex issue and pursue to evaluation the percentage of respondents that show a certain complex trait most frequently use standard RR techniques. Instead, some study has examined conditions where the answer to a complex query yields a quantifiable variable. Despite recent considerable advancements in RR methodology, the majority of research in this field still concentrates on straightforward random sampling, while actual studies still rely on intricate surveys. Recent R-packages for RR survey estimates have been created by certain authors on the basis of simple random sampling. The estimation of these procedures from intricate surveys in direction to approximation parameters for delicate appearances is not covered by any software currently on the market. This break has today been complete by the RRTCS package. This package provides estimators of means and sums using Randomized Response methods, as well as confidence interval estimation. The interest variable is the only factor that the majority of RRT studies explicitly employ to generate estimates for. To improve sample design and increase the accuracy of population parameter estimations, we include auxiliary variables for a huge class of estimators.

1.1. Limitations of the Study

The following list of key restrictions might serve as a summary:

- RRT surveys typically take longer and cost more money than other survey methods.
- RRT estimates are less effective than DQ estimates because they have more sampling variance. By removing response bias, this efficiency loss is offset by the acquisition of information that is more trustworthy. In order to produce estimates that are equivalent to those acquired using DQ, a substantially larger sample size may be required, which would raise the cost, which is seldom acceptable.
- The respondent population lacks understanding and trust. The RRT fundamentally encourages responders to provide information that may seem unimportant or even false, claim. The approach becomes suspect when the respondent doesn't understand the mathematical reasoning behind it, leading them to believe the interviewer can determine their precise position with relation to the complex attribute by examining their response. Additionally, respondents could not know how to use the RR gadget and/or have doubts about the claimed privacy protection.
- In Randomized response measures, a randomization mechanism is necessary to drive the solution. Physical barriers restrict Randomized response application to in person interviews with live subjects and may be more time and money consuming than DQ since each survey respondent must be informed about the approach. Other survey methods, there seem to be no room for methods like telephone interviews, self-administered postal surveys, or internet delivered interviews. Additionally, respondents may struggle to utilize a physical object because of poor motor skills or anxiety about using an item the interviewer provides.

2. Literature Review

Blair, Imai, and Zhou (2015) worked on the randomized response technique's design and analysis. In this method, we urged respondents to utilize a randomization tool that the interviewer would consider fair, like a coin flip. The approach protects responder privacy by masking individual responses with random noise. By looking at common designs that are readily available to researchers and a number of multivariate regression techniques being developed for in-depth analysis, we fill this methodological gap. Putting forth power analyses to aid in recovering research design; providing innovative, difficult designs that are based on less solid presumptions than traditional designs making all of the methods' source code available. We use an actual poll on terrorist groups in Nigeria to demonstrate some of these strategies. We provide examples of commonly employed designs, Showcase the use of the sensitive item as an outcome or predictor in multivariate regression analysis for each design, generate power calculations, and provide alternative designs that address particular departures from customary design protocols. Finally, in order to make these approaches more accessible, we provide open-source software. Overall, we believe that our study will benefit future methodological development as well as the efficient application of the randomized response strategy across disciplines.

Moreover, Hussain, Al-Sobhi, Al-Zahrani, Singh, and Tarray (2016) investigated ways to enhance additive scrambling models with randomized response. The mean and sensitivity level were accurately estimated by Gupta, Langridge, and Mir (2010) using partial and elective preservative scrambling methods. Hussain et al. (2016) upgraded the Gupta et al. (2010) scrambling model by combining additive and subtractive scrambling. Using information from Gupta et al. (2010), Hussain and Al-Zahrani created two methods for calculating the mean and sensitivity level. One approach uses two scrambling variables with two samples, while the other uses two scrambling variables with two replies from each responder.

The work of Lee, Peng, Tapsoba, and Hsieh (2017) regards to improved estimation methods for RRT using unrelated questions technique. In sensitive surveys, the randomized response technique (RRT) is a supportive device used for preventing biased replies while upholding the respondents' anonymity. In this Method, we offer a data collection strategy that takes into account both the RRT for irrelevant questions and the direct question design for surveys on sensitive topics. The direct inquiry approach is used to elicit responses to the innocent query RRT, which is tied to a non-sensitive topic. These responses provide more details that might help determine how common the sensitive behavior is. Should any information be lacking, we also offer two more methods for estimating the proportion of respondents who possess the sensitive feature.

The weighted conditional likelihood estimator, in particular, is what we construct. Greenberg's estimator is outperformed by the other two available estimators. Actual data from a survey study on unlicensed cable TV use in Taiwan are supplied to demonstrate our technique. When there are missing data, both approaches offer estimators that are superior to the Greenberg. But Chong, Wishart, and Xia (2019) worked on a scholarly study on the reasons for unlawful trash disposal called Asking Sensitive Questions Using the Randomized Response. In this survey study, we quantify demographic variables using a research approach that is often used in biostatistics and public health research, two fields where sensitive topics are frequently discussed. Respondents may not be completely honest in their responses if doing so could make other people uncomfortable talking about sensitive topic.

The work of Reiber, Pope, and Ulrich (2021) on Cheating Detection by means of the Unrelated Question Model Randomized response methods (RRTs) are survey procedures that may be used to estimate the frequency of sensitive subjects like doping in professional sports. Unrelated question model (UQM), a form of RRT, is often used due to its psychological appeal for study participants and outstanding statistical properties. The drawback of this paradigm is that it makes it impossible to identify survey cheaters who provide self-protective responses rather than responding as instructed. In this study, we give modified versions of the UQM that are designed to identify the frequency of dishonest answers. By offering explicit techniques for computing the parameters of these improved UQM versions, we demonstrate how to evaluate the empirical applicability of these changed UQM versions. The Appendices include all R-code.

Since it may be difficult to elicit genuine replies from survey respondents, Hsieh, Lee, and Li (2022) have been working on a two-stage multilevel randomized response approach combining proportional odds models and missing variables. To remove biases caused by underreporting or non-response, we provide a two-stage multilevel randomized response (MRR) approach that preserves personal anonymity while determining actual income levels. To deal with variables on particular people that are missing at random, we provide a proportional odds model for two-stage MRR data and use inverse probability weighting and several imputation techniques. To assess the effects of missing data and gauge the effectiveness of the offered approaches, a simulation study is done. Using the Taiwan Social Change project's regular monthly income data, the implementation of the suggested tactics is illustrated.

3. Methodologies

Both sensitive and non-sensitive questions are taken into account in this method. The likelihood of receiving a favorable answer is known for the non-sensitive inquiry. The interviewer is not aware of the question that the subject is responding to. It is supported by the following ideas.

1. Probability of choosing sensitive question.
2. The percentage of affirmative responses to the non-sensitive question.
3. The total number of student (who participated) who answer yes for the sensitive question or for the non-sensitive questions.

A section of the questionnaire's opening questions dealt with delicate conduct like lying and bullying. We applied a methodology developed by Horvitz et al. and expanded by Greenberg et al. to randomize the replies. This design is much more appropriate and less difficult when compared to other RRT variations. The programmed "Randomizers" with the "coin flipper" option was the randomizing tool utilized. The programmed is incredibly simple to use; all the user needs to do is click the "Randomize" button to see either the head or tail of the coin. In our study, the student responded to the sensitive question if "head" appeared and the non-sensitive question if "tail" occurred.

3.1. Horvitz Model

Greenberg, Abul-Ela, Simmons, and Horvitz (1969); Horvitz, Greenberg, and Abernathy (1976) refined Warner's method by combining a sensitive query (character y) with a non-sensitive (unrelated) inquiry (character x). The RR gadget displays the sampled person labeled in a box containing multiple similar cards, with a percentage p, (0p1)) bearing the mark A and the balance labeled B, an insignificant attribute whose population proportion is known. If I hold the card A, and the card drawn is marked A, or if I, hold the card B and the card drawn is marked B, the response requested represented by z_i takes the value y_i. If not, z_i is equal to 0.The transformed variable is

$r_i = \frac{\text{zi} - (1-\text{p})\alpha}{\rho}$ and the estimated variance is

$$V_R(r_i) = r_i(r_i - 1)$$

With the use of a randomized response approach created by Warner, participants in personal interview surveys can answer delicate or extremely private questions while maintaining their anonymity. Using a random mechanism, the respondent must choose one of two questions that are connected. Only "yes" or "no" responses are given in response to the selected question, with no indication of the actual question that was picked. The population is divided into the same two class's one mutually exclusive and the other complementary by the answers to any question. If the respondents answer honestly, the percentage of "yes" responses in the sample and the known probability of choosing either question is enough to give a fair estimate of the population's distribution among the two mutually exclusive classes.

This study reports on a modification of the Warner approach that Walt R. Simmons proposed in order to improve the respondents' participation and the accuracy of their answers. In order to avoid the Warner technique's mutually exclusive and complimentary features, it calls for respondents to choose one of two unrelated questions at random. Two samples are required, and for each sample, there must be a unique set of selection probabilities for the two questions. For two independent trials per respondent, the method for determining the parameters and variances for this alternate randomized response model is also provided.

$$\text{y}_\text{i} = \{ \text{ If the card marked the A and you really have to A}$$
$$y_i = \{\text{If the card marked the B and you really have to B}$$
$$Y_i = If\ the\ marked\ of\ the\ carddoes\ not\ match\ with\ your\ real\ characterstics$$

The Warner Model is designed to give a way to estimate the percentage of people who possess a sensitive attribute, like A, without needing each responder to disclose his classification to the interviewer (whether it be A or I). A random device is given to the respondent to select one of two statements of the type:

- "You have the attribute A"
- "You do have the attribute A"

The responder then responds "yes" or "no" depending on the statement he has chosen and whether he possesses attribute A or not, without telling the interviewer which statement he has chosen. Let,

π = true proportion with attribute A
p = probability that the first statement is selected (1-p)
x_i = If the i-th respondent says "yes" to the selected statement
x_i = 0, otherwise
n = sample size

Then, using one sample and one test with subjects who always tell the truth

Pr (x_i=1) = π_ρ + (1-п) (1-p)
Pr (x_i=0) = (1-п) p + (1-p)п

It follows that the maximum likelihood estimate of is

$$\hat{\pi} = \frac{p-1}{2p-1} + \frac{n_1}{n(2p-1)} \qquad p \neq \frac{1}{2}$$

Where $n_1 = \sum_{i=1}^{n} x_i$

If every responder provides a genuine response, and variation is shown by

Var ($\hat{\pi}$) = $\frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}$

3.2. An R Package for Randomized Response Techniques in Complex Surveys

The RRTCS software uses data from RR surveys with intricate sampling schemes to conduct point and interval estimate of linear parameters. The software supports a broad variety of sampling designs, including stratified sampling, cluster sampling, unequal probability sampling, simple random sampling with and without replacement (SRSWR and SRSWOR), and any combination of these. A second function, titled Resampling Variance, is also included in the package. It estimates the variance of the RR estimators using various resampling techniques Wolter and Wolter (2007) for stratified, clustered, and unequal probability sampling. This comprises the Jackknife approach Quenouille (1949), Escobar and Berger (2013) and Campbell Berger Skinner et al. (2014) methods Berger and Skinner (2005).

The package also contains 20 data sets that contain observations from various surveys that were done in both actual and simulated populations using various RRTs. We attempted to explore the two main issues that university students face, "cheating in Exams" and "want to Harm someone," using the Randomized Response Technique. We create a survey with this objective in mind. The questionnaire consists of two sets of questions, each with two questions. In our data we take a sample of 454 students through simple random sampling without replacement. We name the data as “MyData123” to estimate the proportion of student as listed in the data given below. We use the Horvitz model.

3.3. Design for a Horvitz Model Survey

These are the following approaches for questionnaire

- The question should not be sensitive.
- Sensitive question should not be asked directly to the respondent.

The survey is:

Head	Do you ever copy on exams?	YES	NO
Tail	Were you born in July?	YES	NO

Head	Have you ever bullied someone?	YES	NO
Tail	Does your ID end with 5?	YES	NO

This data collection includes findings from a university-wide survey with a randomized response approach to look into student cheating and bullying. By using cluster (by group) and stratified (by professor) sampling, the sample is selected. The Horvitz model (Greenberg et al., 1969; Horvitz et al., 1976) with parameter p = 0.5 is the randomized response strategy that is employed. Does the last number on your ID card end in an odd number? Are you a July baby? Does the last digit of your ID finish in 5 when = 1/12? With ɑ= 1/10. You need the response probabilities to non-sensitive questions in order to estimate results:

Table 1

#	Questions	Probabilities
1	Were you born in July?	1/12
2	Does your ID end with 5?	1/10

3.4. Horvitz Model Technique Description

The respondent was given a coin to toss and have to answer the question.

- Step 1: the respondent flips the coin and answer the question.
- Step 2: The side of the coin determines the question you must answer in each game:

-If the head is appearing then respondents have to answer the head question.
-If the tail is appearing then respondents have to answer the head question.

For every question the procedure must be repeated. Before depositing the survey form in an urn, the respondent just needed to tick off the replies they wanted to give.

3.5. Theoretical and Practical Implication

By using RRTCS Package in R software we perform the theoretical results of our data which are given below,

Library (RRTCS)
N=25000
n=454
data=read.csv ("D:\\MyData1.csv", header=T)
dim(data)
names(data)
P = 0.5
α = c (1/14, 1/12, 1/22, 1/10)
ρ = rep (n/N*n)
Confidence interval = 0.95

Result 1
Horvitz (data$ z, p, α [1], π, "μ", confidence interval, N)
Horvitz (z = data$ z, ρ = ρ, α = α [1], π = π, type = "mean", confidence interval = confidence interval, N = N)

ρ=0.5;
α=0.071
Estimated value: 0.9505979
Var: 0.002204468
CI (95%)
Lowest value: 0.8585741
Greater value: 1.042622

Result 2
Horvitz (data$ z, ρ, α [2], π, "μ", confidence interval, N)
Horvitz (z = data$ z, ρ = ρ, alpha = α [2], π= π, type = "μ", confidence interval = confidence interval, N = N)

ρ = 0.5;
α = 0.083
Estimated value: 0.9386931
Var: 0.002204044
CI (95%)
Lowest value: 0.8466782
Greater values: 1.030708

Result 3
Horvitz (data, p, α [3], π, "μ", confidence interval, N)
Horvitz (z = data, p = p, α = α [3], π = π, type = "μ", confidence interval = confidence interval, N = N)

ρ=0.5;
α=0.045
Estimated value: 0.9765719
Var: 0.002205431
CL (95%)
Lowest value: 0.884528
Greater value: 1.068616

Result 4

Horvitz (data$ z, ρ, ɑ [4], п, "μ", confidence interval, N)

Horvitz (z = data$ z, p = ρ, ɑ = ɑ [4], п = п, type = "μ", confidence interval = confidence interval, N = N)

ρ=0.5;
ɑ=0.1
Estimated value: 0.9220264
Var: 0.002203471
CL (95%)
Lowest value: 0.8300235
Greater value: 1.014029

4. Results Interpretation

Table 2 shows that the Horvitz estimation of Q1 is 0.938 and the Horvitz Variance is 0.0022 where ɑ is 0.083 and the lower bound is 84% and the upper bound is 30%. The result shows that 84% of the students cheated in Examination. Similarly, in Q2 the Horvitz estimation is 0.922 and the Horvitz variance is 0.2203 where ɑ is 0.1 and the upper bound is 83% and the lower bound is 14%. These results show that 14% of the student bullied.

Table 2

Questions	Horvitz Estimation	Horvitz Variance	Horvitz α	Horvitz Lower bound	Horvitz Upper bound
Q1T	0.9386931	0.002204044	0.083	0.8466782	1.030708
Q2T	0.9220264	0.002203471	0.1	0.8300235	1.014029

5. Conclusions

It is relatively common for wealthy countries to apply particular approaches and get specific outcomes, whereas developing countries have challenges because of their high rates of illiteracy, lack of modern technology, and unique demographic features.

In this study respondents were divided into two groups based on gender male and female with academic faculty. 454 students comprised the sample; among them 47.15 % were men and 52.8 % were women. By faculties, 44.6% of students studied in the faculty of social sciences and law, 29.7% in the faculty of health sciences, and 37.7% in the faculty of science and engineering, with 95% percent C.I, it was estimated that 8.3% to 14% of pupils had ever bullied another person, with C.I 95% the estimated percentage of students who had cheated on a test ranged from 84% to 30%. According to this exam cheating has become a serious issue at our university. The advent of mobile devices and laptops, which have given rise to high-tech cheating, may be the cause of this surge. On the other hand, our university's punitive policies for cheating are relatively lax.

5.1. Recommendations

We advise that only educated people who can speak and comprehend the interviewer's language should use RRT. There should be appropriate initiatives to teach the participants the RRT. Additionally, RRT should be used in private; this requirement is necessary for the RRT to produce proper results. As long as the responder is willing to participate outside of their home and no family members or friends should be present, the survey can be completed. The interview should last the allotted amount of time. According to studies, RRT makes respondents in underdeveloped nations distrustful; leading them to believe there was a trick involved. We discovered the highest degree of compliance in this research.

Additionally, it piqued the curiosity of several respondents. One harmless question served as a drawback in our study. Another unimportant question ought to be included to strengthen the study's validity. Even yet, this approach will be challenging for responders with low levels of education. By employing a team of interviewers, educating them to respect participant privacy throughout the interview, and ensuring that they understand the RRT, the validity of the procedure may be improved. It should be calculated how successful RRT findings are for various demographic categories, especially for respondents who are illiterate and literate. The sample size needs to be sufficient to verify the method's validity, and the requirements for lack of people, money, and time need to be met.

References

Berger, Y. G., & Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(1), 79-89. doi:https://doi.org/10.1111/j.1467-9868.2005.00489.x

Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association, 110*(511), 1304-1319. doi:https://doi.org/10.1080/01621459.2015.1050028

Chong, J., Wishart, D. S., & Xia, J. (2019). Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Current protocols in bioinformatics, 68*(1), e86. doi:https://doi.org/10.1002/cpbi.86

Escobar, E. L., & Berger, Y. G. (2013). A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica*, 595-613.

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

Gupta, P., Langridge, P., & Mir, R. (2010). Marker-assisted wheat breeding: present status and future possibilities. *Molecular Breeding, 26*, 145-161. doi:https://doi.org/10.1007/s11032-009-9359-7

Horvitz, D. G., Greenberg, B. G., & Abernathy, J. R. (1976). Randomized response: a data-gathering device for sensitive questions. *International Statistical Review/Revue Internationale de Statistique*, 181-196. doi:https://doi.org/10.2307/1403276

Hsieh, S.-H., Lee, S.-M., & Li, C.-S. (2022). A two-stage multilevel randomized response technique with proportional odds models and missing covariates. *Sociological Methods & Research, 51*(1), 439-467.

Hussain, Z., Al-Sobhi, M. M., Al-Zahrani, B., Singh, H. P., & Tarray, T. A. (2016). Improved randomized response in additive scrambling models. *Mathematical Population Studies, 23*(4), 205-221. doi:https://doi.org/10.1080/08898480.2015.1087773

Lee, S.-M., Peng, T.-C., Tapsoba, J. d. D., & Hsieh, S.-H. (2017). Improved estimation methods for unrelated question randomized response techniques. *Communications in Statistics-Theory and Methods, 46*(16), 8101-8112. doi:https://doi.org/10.1080/03610926.2016.1175626

Quenouille, M. H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics, 5*(2), 162-164. doi:https://doi.org/10.2307/3001917

Reiber, F., Pope, H., & Ulrich, R. (2021). 2.1 Unrelated Question Model-Cheating Extension. *Enhancing the Applicability of Randomized Response Techniques*.

Wolter, K. M., & Wolter, K. M. (2007). *Introduction to variance estimation* (Vol. 53): Springer.