Pak. J. Statist.
2023 Vol. 39(3), 377-385

INTRODUCING A WEIGHTED MEASURE OF PRIVACY AND EFFICIENCY FOR COMPARISON OF QUANTITATIVE RANDOMIZED RESPONSE MODELS

Muhammad Azeem
Department of Statistics, University of Malakand,
Khyber Pakhtunkhwa, Pakistan
Email: azeemstats@uom.edu.pk

ABSTRACT

In sample surveys on sensitive variables, the randomized response technique is commonly employed by researchers to ensure the respondents' privacy protection and hence minimizing the non-response rate. The privacy protection level and efficiency are the two important indicators of the quality of any randomized response technique. The existing unified measure of model efficiency and privacy level for comparison of randomized response models assigns equal weightage to respondents' privacy protection and model's efficiency. In practice, some situations may prefer privacy over efficiency whereas in other situations, efficiency may be more important than privacy. This paper introduces a unified weighted metric of privacy protection level and efficiency for comparison of quantitative randomized response techniques. The practical usefulness of the proposed measure is that it enables the researchers to assign different weights to privacy and efficiency as per requirement of the situation. Two existing randomized response models are compared using the proposed measure and the results are discussed.

KEYWORDS

Model efficiency, randomized response models, privacy protection, sensitive surveys, weighted measure.

1. INTRODUCTION

Researchers in almost every survey encounter refusal and/or false responses at data collection phase. This problem occurs in almost every survey especially in situations in which the variable of interest is of sensitive type. The sensitive characteristics include, but not limited to, income, salary, drugs usage, abortion, expenditure, and cheating in an examination. In an attempt to eliminate the problem of non-response or refusals to questions related to sensitive issues, Warner (1965) laid the foundation of a method commonly known as randomized response. Originally, the Warner's (1965) method was designed to be used in sample surveys where the variable under study is of binary type. A quantitative version of the Warner's (1965) technique was studied by Warner (1971). The Warner's (1971) technique used the addition of a random noise to the true response, thus protecting the privacy of the respondents. Gupta et al. (2002) suggested the use of an optional quantitative technique where the individuals have the freedom to either report their true response or report a random response. Gjestvang and Singh (2009) presented an efficient quantitative randomized response strategy for sensitive variables which was further improved by the model of Narjis and Shabbir (2021). Unlike the Warner's (1971) technique which used only additive scrambling, Diana and Perri (2011) developed a randomization technique which utilized additive and multiplicative noise. The technique suggested by Al-Sobhi et al. (2016) uses additive and subtractive scrambling variables.

Yan et al. (2008) suggested a metric for quantification of the level of respondents' privacy in randomized response models. Gupta et al. (2018) developed a unified metric of the privacy level and efficiency for randomized response models. The study of Khalil et al. (2021) is based on the analysis of the impact of measurement error on estimators of parameters. Gupta et al. (2022) presented a new efficient variant of the scrambled randomized response models where the respondents can opt for either an additive scrambling only or use both additive and multiplicative randomization simultaneously. The Gupta et al. (2022) model was found superior to the model of Diana and Perri (2011). Different aspects of quantitative randomization techniques have been studied by survey statisticians in past few decades. For details related to studies on randomized response techniques, one may refer to Kalucha et al. (2016), Young et al. (2019), Murtaza et al. (2021), Zhang et al. (2021), and Azeem (2023), etc.

Recently, Azeem and Salam (2023) developed an improved randomized response strategy which achieved improvement in efficiency over the previous randomized response techniques. Another recent study of Azeem et al. (2023) presented a quantitative randomized response model which improved the Narjis and Shabbir (2021) model.

This paper presents a unified weighted measure for quantifying the privacy protection level and efficiency for comparison of randomized response techniques. The proposed measure gives the researchers the choice to assign weights to privacy protection and efficiency while comparing randomized response models. The choice of assigning weights to privacy and efficiency makes the proposed measure practically more suitable than the Yan et al. (2008) and the Gupta et al. (2018) measures.

2. EXISTING MEASURES OF EFFICIENCY AND PRIVACY COMPARISON

In this section, the available measures of privacy and efficiency for comparison of quantitative randomized response models are presented.

2.1 The Relative Efficiency Measure

The most commonly used method of the performance comparison is the relative efficiency. This method has been used by almost all researchers for comparison of randomized response models since 1965. Let Model P be the randomized response model under consideration, and let Model O be any other randomized response model with which the efficiency comparison is desired. Let MSE_P and MSE_O be the mean squared error of the mean estimator under Model P and Model O, respectively. Then the relative efficiency can be expressed as:

$$E = \frac{MSE_O}{MSE_P} . \tag{1}$$

Since the mean square error is always a positive quantity, so $E \geq 0$. Clearly, $E > 1$ indicates that Model P is more efficient than Model O. The larger the value of *E*, the more efficient the Model P is, compared to Model O. If the mean estimator is unbiased, one can replace MSE_P and MSE_O by Var_P and Var_O, respectively, in equation (1). In comparing the performance of randomized response models, one may clearly notice that the above formula for relative efficiency completely ignores the level of privacy provided by the two models.

2.2 The Yan et al. (2008) Measure of Privacy

Yan et al. (2008) proposed the following metric for quantifying the privacy level of a given quantitative model under consideration.

$$\nabla = E\left[Z - Y\right]^2. \tag{2}$$

It is clear that for any randomized response model, $\nabla \geq 0$. Larger values of ∇ indicate a higher level of protection of privacy offered by a model under consideration. The Yan et al. (2008) metric gives full weightage to privacy protection but ignores the other important aspect of model quality – its efficiency.

2.3 The Gupta et al. (2018) Unified Measure

Until 2018, only separate measures for quantification of privacy level and efficiency were available. In an attempt to develop a single quantitative metric of the privacy level and efficiency, Gupta et al. (2018) proposed a combined measure, given as follows:

$$\delta = \frac{MSE}{\nabla}. \tag{3}$$

Examining equation (3), one may clearly observe that the value of δ ranges from 0 to ∞. Moreover, a smaller value of δ shows either a smaller mean square error, or a higher privacy level, or both. This measure quantifies the overall performance of a randomized response model, expressed as a single δ value. Moreover, this measure doesn't assign relative weights to privacy or efficiency, which limits its application.

3. THE PROPOSED WEIGHTED MEASURE

Let ∇_P and ∇_O be the Yan et al. (2008) measure of privacy protection for Model P and Model O, respectively. Let

$$P = \frac{\nabla_P}{\nabla_O}. \tag{4}$$

Since ∇_P is in the numerator, so a higher value of *P* indicates higher privacy protection for Model P than Model O. If the two models under consideration offer equal privacy protection, then *P* = 1. Using equation (1) and (4), the following weighted measure of privacy protection and efficiency is proposed.

$$\phi = \frac{w_1 E + w_2 P}{w_1 + w_2}, \tag{5}$$

where w_1 and w_2 denote the weights of efficiency and privacy, respectively.

4. PROPERTIES OF THE PROPOSED MEASURE

The proposed measure ϕ has the following properties.

i. If Model P and Model O are equally efficient and offer equal privacy protection, then $E = 1$, and $P = 1$, and hence

$$\phi = \frac{w_1 + w_2}{w_1 + w_2} = 1.$$

Thus $\phi = 1$ indicates *equal quality* of both models.

ii. The value of the proposed measure, ϕ, varies from 0 to ∞, with $\phi > 1$ indicates that Model P is better than Model O. Likewise, if $0 < \phi < 1$, this will indicate that Model P is *worse* than Model O in terms of *overall quality*.

iii. For $w_1 = 1, w_2 = 0$, the suggested measure ϕ reduces to the simple measure of relative efficiency.

iv. For $w_1 = 0, w_2 = 1$, the suggested measure ϕ reduces to the measure of privacy protection defined in equation (4).

v. In order to have a symmetric measure, it can be more convenient to use $\log\phi$ in place of ϕ so that equal quality translates to $\log\phi = 0$. The measure $\log\phi$ ranges from $-\infty$ to $+\infty$ and is symmetric around 0. A positive value of $\log\phi$ indicates that Model P is better than Model O, whereas a negative value indicates the opposite situation. Any departure of $\log\phi$ from zero in any direction indicates the improvement in the quality of the model in that direction. The mathematical expression of $\log\phi$ is given as:

$$\log\phi = \log\left(\frac{w_1 E + w_2 P}{w_1 + w_2}\right). \tag{6}$$

5. COMPARISON OF MODELS USING THE PROPOSED MEASURE

In this section, the comparative performance of the Gupta et al. (2022) and the Diana and Perri (2011) models is analyzed using the proposed measure. Suppose our population under consideration consists of N units and let a simple random sample consisting of n units is selected from the population with replacement. Let Y be the sensitive variable of interest and let S be an additive-type scrambling variable. We also assume that $E(Y_i) = \mu_Y$

, $E(S)=0$, $V(Y_i)=\sigma_Y^2$, $V(S)=\sigma_S^2$. Further, let T be another scrambling variable which is of multiplicative-type, such that $E(T)=1$, and $V(T)=\sigma_T^2$, where σ_Y^2,σ_T^2, and σ_S^2 denote population variance of variables Y, T, and S, respectively, and μ_Y denotes the mean of the entire population for variable Y. To ensure protection of the respondents' privacy, we further assume that all variables under consideration work independently of each other.

The observed responses using the Diana and Perri (2011) scrambling model are given by:

$$Z = TY + S \,. \tag{7}$$

An unbiased mean estimator based on the Diana and Perri (2011) technique can be expressed as:

$$\hat{\mu}_{DP} = \frac{1}{n}\sum_{i=1}^{n} Z_i \,. \tag{8}$$

The sampling variance of $\hat{\mu}_{DP}$ can be obtained as:

$$Var\left(\hat{\mu}_{DP}\right) = \frac{1}{n}\left[\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_Y^2+\sigma_S^2\right]. \tag{9}$$

The Yan et al. (2008) metric of privacy using the Diana and Perri (2011) model is given by:

$$\nabla_{DP} = E\left[TY+S-Y\right]^2 = \sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_S^2 \,. \tag{10}$$

The observed responses provided by the Gupta et al. (2022) optional scrambling model are:

$$Z = \begin{cases} Y & \text{with probability } 1-W \\ Y+S & \text{with probability } WA \\ TY+S & \text{with probability } W(1-A). \end{cases} \tag{11}$$

An unbiased mean estimator using the Gupta et al. (2022) technique can be expressed as:

$$\hat{\mu}_{G} = \frac{1}{n}\sum_{i=1}^{n} Z_i \,. \tag{12}$$

The variance of $\hat{\mu}_G$ can be written as:

$$Var\left(\hat{\mu}_{G}\right) = \frac{1}{n}\left[W\left(1-A\right)\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_Y^2+W\sigma_S^2\right]. \tag{13}$$

The Yan et al. (2008) metric of privacy using the Gupta et al. (2022) model can be derived as:

$$\nabla_G = (1-A)\left[\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)\right]+\sigma_S^2 . \tag{14}$$

The efficiency of the Gupta et al. (2022) model with respect to the Diana and Perri (2011) model can be expressed as:

$$E = \frac{\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_Y^2+\sigma_S^2}{W(1-A)\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_Y^2+W\sigma_S^2} . \tag{15}$$

The ratio of the privacy measures is given as:

$$P = \frac{(1-A)\left[\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)\right]+\sigma_S^2}{\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_S^2} . \tag{16}$$

Using equation (15) and (16) in (5), the mathematical expression for ϕ is given as:

$$\phi = \frac{w_1\left[\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_Y^2+\sigma_S^2\right]}{(w_1+w_2)\left[W(1-A)\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_Y^2+W\sigma_S^2\right]} + \frac{w_2\left[(1-A)\left\{\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)\right\}+\sigma_S^2\right]}{(w_1+w_2)\left[\sigma_T^2\left(\sigma_Y^2+\mu_Y^2\right)+\sigma_S^2\right]} . \tag{17}$$

6. RESULTS AND DISCUSSION

The values of $\log\phi$ for various choices of parameters and constants have been presented in Table 1. It is clear that the values of $\log\phi$ do not change as W changes. The table also shows that for $w_1 = 0.2, w_2 = 0.8$, the values of $\log\phi$ are negative in most cases, meaning that the Diana and Perri (2011) model is better than the Gupta et al. (2022) model. It is also clear that if $w_1 = 0.5, w_2 = 0.5$, the Gupta et al. (2022) model is slightly better than the Diana and Perri (2011) model. However, for $w_1 = 0.8, w_2 = 0.2$, the Gupta et al. (2022) model outperforms the Diana and Perri (2011) model. Table 1 also shows that as the relative weight of model efficiency, w_1, increases, the value of $\log\phi$ also increases. Thus, the findings suggest that if the privacy protection of the respondents is a priority over efficiency, then the Diana and Perri (2011) model is preferable. However, if efficiency of the model is more important than privacy, then the Gupta et al. (2022) model is desirable.

It is concluded that the proposed measure gives a detailed comparison of the two randomized response models for various choices of weights. The findings of the current study guides the researchers about which model is better in a particular situation. Hence it is recommended for researchers to use the proposed measure when comparing randomized response models by using the relative weights of privacy and efficiency.

Table 1
Values of $\log\phi$ using the Gupta et al. (2022) Model with respect to the Diana and Perri (2011) Model for $\mu_Y = 5$, $\sigma_Y^2 = 2$

W	A	w_1	w_2	$\sigma_S^2 = 3$			$\sigma_S^2 = 8$		
				$\sigma_T^2 = 1$	$\sigma_T^2 = 4$	$\sigma_T^2 = 8$	$\sigma_T^2 = 1$	$\sigma_T^2 = 4$	$\sigma_T^2 = 8$
0.8	0.8	0.2	0.8	-0.028	0.033	0.048	-0.082	-0.005	0.024
		0.5	0.5	0.285	0.374	0.394	0.175	0.320	0.362
		0.8	0.2	0.464	0.563	0.584	0.336	0.503	0.550
	0.5	0.2	0.8	-0.095	-0.097	-0.097	-0.088	-0.096	-0.097
		0.5	0.5	0.073	0.090	0.093	0.050	0.080	0.088
		0.8	0.2	0.194	0.220	0.225	0.154	0.205	0.217
	0.2	0.2	0.8	-0.046	-0.049	-0.050	-0.040	-0.047	-0.049
		0.5	0.5	0.008	0.010	0.10	0.006	0.009	0.010
		0.8	0.2	0.057	0.062	0.063	0.047	0.059	0.062
0.5	0.8	0.2	0.8	-0.028	0.033	0.048	-0.082	-0.005	0.024
		0.5	0.5	0.285	0.374	0.394	0.175	0.320	0.362
		0.8	0.2	0.464	0.563	0.584	0.336	0.503	0.550
	0.5	0.2	0.8	-0.095	-0.097	-0.097	-0.088	-0.096	-0.097
		0.5	0.5	0.073	0.090	0.093	0.050	0.080	0.088
		0.8	0.2	0.194	0.220	0.225	0.154	0.205	0.217
	0.2	0.2	0.8	-0.046	-0.049	-0.050	-0.040	-0.047	-0.049
		0.5	0.5	0.008	0.010	0.10	0.006	0.009	0.010
		0.8	0.2	0.057	0.062	0.063	0.047	0.059	0.062
0.2	0.8	0.2	0.8	-0.028	0.033	0.048	-0.082	-0.005	0.024
		0.5	0.5	0.285	0.374	0.394	0.175	0.320	0.362
		0.8	0.2	0.464	0.563	0.584	0.336	0.503	0.550
	0.5	0.2	0.8	-0.095	-0.097	-0.097	-0.088	-0.096	-0.097
		0.5	0.5	0.073	0.090	0.093	0.050	0.080	0.088
		0.8	0.2	0.194	0.220	0.225	0.154	0.205	0.217
	0.2	0.2	0.8	-0.046	-0.049	-0.050	-0.040	-0.047	-0.049
		0.5	0.5	0.008	0.010	0.10	0.006	0.009	0.010
		0.8	0.2	0.057	0.062	0.063	0.047	0.059	0.062

Data Availability Statement

All relevant data is available within the manuscript.

Funding Source

The author received no funding for this study.

Conflict of Interest

The author has no conflict of interest to declare.

REFERENCES

1. Al-Sobhi, M.M., Hussain, Z., Al-Zahrani, B., Singh, H.P. and Tarray, T.A. (2016). Improved randomized response approaches for additive scrambling models. *Mathematical Population Studies*, 23(4), 205-221. https://doi.org/10.1080/08898480.2015.1087773
2. Azeem, M. (2023). Using the exponential function of scrambling variable in quantitative randomized response models. *Mathematical Methods in the Applied Sciences*. https://doi.org/10.1002/mma.9295
3. Azeem, M., Salam, A. (2023). Introducing an efficient alternative technique to optional quantitative randomized response models. *Methodology*, 19, 24-42. https://doi.org/10.5964/meth.9921
4. Azeem, M., Hussain, S., Ijaz, M. and Salahuddin, N. (2023). An improved quantitative randomized response technique for data collection in sensitive surveys. *Quality and Quantity*, https://doi.org/10.1007/s11135-023-01652-5
5. Diana, G. and Perri, P.F. (2011). A class of estimators of quantitative sensitive data. *Statistical Papers*, 52(3), 633-650. https://doi.org/10.1007/s00362-009-0273-1
6. Gjestvang, C.R. and Singh, S. (2009). An improved randomized response model: Estimation of mean. *Journal of Applied Statistics*, 36(12), 1361-1367. https://doi.org/10.1080/02664760802684151
7. Gupta, S., Gupta, B. and Singh, S. (2002). Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference*, 100(2), 239-247. https://doi.org/10.1016/S0378-3758(01)00137-9.
8. Gupta, S., Mehta, S., Shabbir, J. and Khalil, S. (2018). A unified measure of respondent privacy and model efficiency in quantitative rrt models. *Journal of Statistical Theory and Practice*, 12(3), 506-511. https://doi.org/10.1080/15598608.2017.1415175
9. Gupta, S., Zhang, J., Khalil, S. and Sapra, P. (2022). Mitigating lack of trust in quantitative randomized response technique models. *Communications in Statistics – Simulation and Computation*, 1-9. https://doi.org/10.1080/03610918.2022.2082477
10. Kalucha, G., Gupta, S. and Shabbir, J. (2016). A two-step approach to ratio and regression estimation of finite population mean using optional randomized response models. *Hacettepe Journal of Mathematics and Statistics*, 45, 1819-1830.
11. Khalil, S., Zhang, Q. and Gupta, S. (2021). Mean estimation of sensitive variables under measurement errors using optional rrt models. *Communications in Statistics – Simulation and Computation*, 50(5), 1417-1426. https://doi.org/10.1080/03610918.2019.1584298

12. Murtaza, M., Singh, S. and Hussain, Z. (2021). Use of correlated scrambling variables in quantitative randomized response technique. *Biometrical Journal*, 63(1), 134-147.
13. Narjis, G. and Shabbir, J. (2021). An efficient new scrambled response model for estimating sensitive population mean in successive sampling. *Communications in Statistics – Simulation and Computation*, 1-18. https://doi.org/10.1080/03610918.2021.1986528
14. Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63-69. https://doi.org/10.1080/01621459.1965.10480775
15. Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66(336), 884-888. https://doi.org/10.1080/01621459.1971.10482364
16. Yan, Z., Wang, J. and Lai, J. (2008). An efficiency and protection degree-based comparison among the quantitative randomized response strategies. *Communications in Statistics – Theory and Methods*, 38(3), 400-408. https://doi.org/10.1080/03610920802220785
17. Young, A., Gupta, S. and Parks, R. (2019). A binary unrelated-question rrt model accounting for untruthful responding. *Involve, A Journal of Mathematics*, 12(7), 1163-1173. https://doi.org/10.2140/involve.2019.12.1163
18. Zhang, Q., Khalil, S. and Gupta, S. (2021). Mean estimation in the simultaneous presence of measurement errors and non-response using optional RRT models under stratified sampling. *Journal of Statistical Computation and Simulation*, 91(17), 3492-3504. https://doi.org/10.1080/00949655.2021.1941018