Journal of Policy Research, 8(4), 427-438.
https://jprpk.com
https://doi.org/10.5281/zenodo.7726031

Effect of Misclassification on Test of Independence Using Different Randomized Response Techniques

Asma Halim[1], Irshad Ahmad Arshad[2], Summaira Haroon[3], Waqas Shair[4]

Abstract

In Survey Sampling, while we are dealing with sensitive issues, it is a challenge to get the accurate and unbiased answers from the respondents. To fulfill this need researcher exert on discovering such methodology which can help to get accurate responses from respondents in case of sensitive issues. One of the most famous techniques amongst these techniques is randomized response technique. In current research we discuss two dimensional tables and derive the misclassification probabilities for various randomized response techniques. The transition matrices of conditional misclassification probabilities are used to get perturbed or misclassified data and then Chi-square test of independence between two attributes is carried out. The results of chi-square test of independence calculated for perturbed data show that variables are found to be dependent which were independent originally, depicting that misclassification can change the status of dependence in the data. This paper has great contribution towards checking independence status of data while dealing with sensitive issues, where data can be misclassified. The derived matrices of conditional misclassification probabilities can be used acquire estimates of log-linear model for numerous randomized response techniques.

Keywords: Misclassification, Randomized Response, Perturbation, Independence, Chi-square test of association, Sensitive issues

1. Survey Sampling

Survey sampling is a technique of collecting information about any certain characteristic of population on the basis of a subset or a part of population. Survey sampling is very useful in the cases when our population is enough large but we do not have much time, money, human power and resources to scrutinize every individual of our population, to draw inference about certain characteristic. Our interest in sampling is to obtain reliable, authentic and accurate results and it becomes nearly impossible to do so in case of sensitive or socially stigmatized variables. So while we are dealing with sensitive variables, it is obvious that there arises a bias. This bias is specifically referred as Social Desirability Bias (SDB), when variable of interest makes respondents socially stigmatized. SDB is stated as a bias which arises when respondent conceals or hides the true response of a sensitive or highly controversial issue due to the fear that if he/she discloses the right information so that makes him/her socially undesirable or stigmatized. Sensitive issues can be like harassment at workplace, practicing a fraud, use of illegal drugs or alcoholic beverages, having extra marital affairs, earning or gaining illegal income, evading income tax and having savings in forms of prize bonds. In such cases we search for some other data collection technique.

2. Randomized Response Technique (RRT)

As we discussed above that in case of sensitive variable, our usual survey methods fail to estimate the parameters of concerned distribution. Because, we know the possession of sensitive attribute results in biased and inaccurate results and our usual survey methods which are applied in case of un-harmful / innocuous questions are not appropriate to apply in case of sensitive issues and if applied then there is a great chance of un-true results. So, it was desired to find some other methods that raise the factual responses from respondents and also lessen the response error. To fulfill this need researcher worked on finding some methodology, which will be successfully dealing with sensitive issues. One of the most famous techniques among these techniques is randomized response techniques (RRT). Warner (1965) is pioneer, who introduced RRT to deal with sensitive variables. He introduced this technique to find the proportion of sensitive/ socially stigmatized variables. This technique is very useful to lessen SDB or evasive answer bias up to great extent. It can also be useful to enhance the collaboration between interviewer and interviewee to work together at one end and to keep up privacy and confidentiality level of interviewee at other end. The respondent answers to the questions, which appear to him, using RRT of cards or spinner as specified by the researchers. Respondents have to answer in the form of a 'yes' or 'no'. As respondent answers to the questions independently un-seen from the interviewer so privacy of respondent remains intact and he answers truthfully and as a result bias reduces. Warner (1965) randomized response (RR) methodology is extended by many researchers like Greenberg et al., (1969), Moors (1971), Mangat and Singh (1990), Mangat (1994), Mahmood et al., (1998) and Christofides (2003), Huang (2004) Kim and Warde (2004, 2005), Kim and Elam (2007), Singh and Tarray (2012, 2014), Narjis and Shabbir (2022) and Hsieh et al., (2022). Singh et al., (2021) and Singh and Singh (2022) presented their papers for estimating population proportion in case of sensitive character using negative binomial and poisson as a randomization device.

3. Misclassification

3.1. Two dimensional or contingency table

When every member of a population can be divided into two categories, we say that the categories are mutually exclusive and exhaustive as a whole. A member chosen at random will fall into one of the two groups, with probability

[1]Corresponding Author, Ph. D Scholar, Allama Iqbal Open University, Islamabad, Pakistan; Assistant Professor, Department of Business Administration, Iqra University, Islamabad, Pakistan
[2]Professor, Department of Statistics, Allama Iqbal Open University, Islamabad, Pakistan
[3] Senior Lecturer, Department of Business Studies, Bahria University, Islamabad, Pakistan
[4]Lecturer, Minhaj University Lahore, Pakistan

pi. A structure is imposed when cells are described in terms of groups of two variables. A rectangular array with rows belonging to one category and columns belonging to the other category represents the natural organization of two variables. A detailed analysis on contingency table has been done by many researchers like Johnson and Wichern (2006) and Fujisawa and Tahata (2022). The position of a specific cell talks about the characteristic of an individual falling into them. The probability of falling an observation into ith row and jth column is denoted by p_{ij} and marginal probabilities by $p_{i.}$and $p_{.j}$. Table 1 and 2 depict observed values and marginal probabilities respectively.

Table 1. Structure of 2x2 table showing observed frequencies and marginal totals.

A	***B***	***1***	***2***	$\sum$
1		n_{11}	n_{12}	$n_{1.}$
2		n_{21}	n_{22}	$n_{2.}$
$\sum$		$n_{.1}$	$n_{.2}$	N

Table 2.Theoretical Probabilities of a Contingency Table

A	***B***	**1**	**2**	$\sum$
1		p_{11}	p_{12}	$p_{1.}$
2		p_{21}	p_{22}	$p_{2.}$
$\sum$		$p_{.1}$	$p_{.2}$	*1*

Our interest in RR is to find the misclassification probabilities. The marginal probabilities $p_{i.}$and $p_{.j}$ are the unconditional probabilities of belonging to category i of attribute A and category j of attribute B respectively. Van den Hout et al., (2010) states that RR is a misclassification design, which is used in survey sampling, when sensitive questions are asked. The randomized response data may be described as misclassified data (Shair & Majeed, 2020). "The purpose of classification is to arrange observations into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes".(Johnson and Wichern, 2006). Fujisawa and Tahata (2022) have worked on finding quasi association models for square contingency tables. Ko and Kim (2016) proposed a study to recognize misclassification objects in discriminant model. Ngailo and Ngaruye (2022), Ngailo and Chuma (2022) and Egleston et al., (2011) has done work on approximations and calculations of misclassification probabilities under different scenarios.

3.2. Misclassification

A good classification procedure also results in few misclassifications or we may say that probabilities of misclassification should be small if a good classification procedure is applied. The chances of misclassification are always there especially, in case of randomized response techniques. Randomized response variables can be considered as misclassified categorical variables where conditional misclassifications probabilities are known, so we may say that randomized response is a misclassification design that is used in sample surveys of sensitive issues. Misclassification in survey is included by the interviewee but interviewer specifies conditional misclassification probabilities. Misclassification occurs, if observed category i is while true category is j for $i \neq j$. The primary thought behind randomized response is unsettling influence made by misclassification design, which is utilized to safeguard the privacy of respondents. A general randomized response design Van den Hout et al., (2010) is given as

$$\underline{\boldsymbol{\theta}} = \underline{\boldsymbol{P}}\,\underline{\boldsymbol{\pi}} \tag{1}$$

Where $\underline{\boldsymbol{\theta}} = (\theta_1, \theta_2, \ldots, \theta_k)$ is a vector of probabilities of observed responses, $\underline{\boldsymbol{\pi}} = (\pi_1, \pi_2, \ldots, \pi_k)$ is the vector of probabilities of the true response with categories 1, 2, 3,…, k and $\underline{\boldsymbol{P}}$ is the kxk transition matrix containing conditional misclassification probabilities. The RR transition matrix contains misclassification probabilities p_{ij}.

3.3. Conditional misclassification probabilities

Let the random variable *Y* denote true status and Y^* denote the observed status where *Y* and Y^* have the same set of categories {1, 2, 3, …, k}. Let $\underline{\boldsymbol{P}}$ denote the *K*x*K* nonsingular transition matrix

$$p_{kl} = P(Y^* = k \mid Y = l) \tag{2}$$

for all $k \in \{1,\ 2,\ 3, \ldots,\ k\}$. In multivariate design the transition matrix $\underline{\boldsymbol{P}}_k$ can be obtained by taking Kronecker product of uni-variate transition matrices. Let $\pi^*_k = P(Y^* = k)$ and $\pi_k = P(Y = k)$ for all $k \in \{1,\ 2,\ 3, \ldots,\ k\}$. So the general randomized response model in matrix notation by Van den Hout et al., (2002) is as under

$$\underline{\boldsymbol{\pi}}^* = \underline{\boldsymbol{P}}\,\underline{\boldsymbol{\pi}} \tag{3}$$

Where $\underline{\pi}$ and $\underline{\pi*}$ are vectors of true and observed response probabilities and $\underline{\boldsymbol{P}}_k$ is the matrix of misclassification probabilities. It is very advantageous to use matrix notation in a multivariate RR design. If the misclassification of Y_1

is described by $\underline{\boldsymbol{P}}_{Y_1}$ and the misclassification of Y_2 is described by $\underline{\boldsymbol{P}}_{Y_2}$ the misclassification of the cartesian product $\underline{\boldsymbol{Y}} = (Y_1, Y_2)$ is described by

$$\underline{\boldsymbol{P}} = \underline{\boldsymbol{P}}_{Y_1} \otimes \underline{\boldsymbol{P}}_{Y_2} \tag{4}$$

Where $\otimes$ denotes the Kronecker product. One thing is to be noted in RR transition matrix that columns or rows of the matrix add up to 1. In case of RRT for a binary variable, if we consider *Y* and *Y** as true and observed answer (1 = yes, 2 = no) respectively, then misclassification probabilities according to above defined conditional misclassification probabilities become as

$$p_{11} = P(1|1) = P(Y^{*} = 1 \mid Y = 1) \tag{5}$$

$$p_{12} = P(1|2) = P(Y^{*} = 1 \mid Y = 2) \tag{6}$$

$$p_{21} = P(2|1) = P(Y^{*} = 2 \mid Y = 1) \tag{7}$$

$$p_{22} = P(2|2) = P(Y^{*} = 2 \mid Y = 2). \tag{8}$$

The transition matrix of conditional misclassification probabilities is given by:

$$\underline{\boldsymbol{P}} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}. \tag{9}$$

Next we work on acquiring the transition matrix of conditional misclassification probabilities for different RRT's.

4. Misclassification Probabilities for few RRT's

4.1. Misclassification Probabilities for Warner (1965) RRT

Warner (1965) is the pioneer, who introduced RRT to deal with sensitive attribute. The randomization device used by Warner (1965) is spinner having two statements of belonging to Group A or B, where A is the group f sensitive attribute. In a sample of size n, each respondent has to answer the sensitive question using specified randomized response device. In randomized response a privacy protection atmosphere is created in such a way that respondent only answers in a yes or no depending on his status of possessing sensitive attribution. Probability of yes and no response by Warner (1965) are as under:

$$P(Y = 1) = p\pi + (1-p)(1-\pi), \quad p \neq 1/2 \tag{10}$$

We know that conditional probability is defined by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{11}$$

So by using equation (11) we find conditional misclassification probabilities for Warner (1965*)* RRT as under:

$$p_{11} = \frac{p\pi}{\pi}$$

$$p_{11} = p$$

$$p_{12} = \frac{(1-p)(1-\pi)}{(1-\pi)}$$

$$p_{12} = 1-p$$

$$p_{21} = \frac{\pi(1-p)}{\pi}$$

$$p_{21} = 1-p$$

$$p_{22} = \frac{p(1-\pi)}{(1-\pi)}$$

$$p_{22} = p.$$

Finally the transition matrix of conditional misclassification probabilities for Warner (1965) RRT is derived as under:

$$\underline{\boldsymbol{P}}_{w} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}. \tag{12}$$

4.2. Misclassification Probabilities for Mangat and Singh (1990) RRT

In the proposed RRT by Mangat and Singh (1990) a new randomizing device is introduced by utilizing two randomizing devices named R_1 and R_2. They discussed a model for two cases when respondents are making truthful

reporting and less than truthful reporting. They used sample of size n without replacement and individuals are used to use R_1 having two statements of belonging to sensitive group A or Go to R_2, with probabilities T and 1-T respectively. The R_2 also includes two statements of belonging to or not belonging to sensitive Group A, with probabilities p and 1-p. Here R_2 is same like Warner (1965) RRT. The probability of a yes response is

$$P(yes) = T\pi + (1-T)\{p\pi + (1-p)(1-\pi)\}. \quad (13)$$

Now we find conditional misclassification probabilities for Mangat and Singh (1990) RRT as under

$$p_{11} = \frac{T\pi}{\pi} + \frac{(1-T)p\pi}{\pi}$$

$$p_{11} = T + (1-T)p$$

$$p_{12} = \frac{(1-p)(1-\pi)(1-T)}{(1-\pi)}$$

$$p_{12} = (1-p)(1-T)$$

$$p_{21} = \frac{\pi(1-p)(1-T)}{\pi}$$

$$p_{21} = (1-p)(1-T)$$

$$p_{22} = \frac{T(1-\pi)}{(1-\pi)} + \frac{p(1-\pi)(1-T)}{(1-\pi)}$$

$$p_{22} = T + (1-T)p.$$

Finally the transition matrix of conditional misclassification probabilities for Mangat and Singh (1990) RRT is

$$\underline{\boldsymbol{P}}_{ms} = \begin{pmatrix} T + (1-T)p & (1-p)(1-T) \\ (1-p)(1-T) & T + (1-T)p \end{pmatrix}. \quad (14)$$

.

4.3. Misclassification Probabilities for Mangat (1994) RRT

Mangat (1994) has pointed that two stage randomized response technique used by Mangat and Singh (1990) can be a bit confusing for the respondents while reporting. So to solve this issue, Mangat (1994) proposed a simpler technique. In proposed technique respondents in a sample size *n* are advised to use the same Warner's (1965) randomized device having belonging and non-belonging to group A with 'p' and '1- p' probabilities respectively and they have to answer 'yes' or 'no' according to outcomes of randomized device and actual status which they posses. The probability of a 'yes' response for Mangat (1994) RRT is given as under:

$$p(yes) = \pi_m + (1-\pi_m)(1-p). \quad (15)$$

We obtain conditional misclassification probabilities for Mangat (1994) RRT as under:

$$p_{11} = \frac{\pi(1)}{\pi}$$

$$p_{11} = 1$$

$$p_{12} = \frac{(1-p)(1-\pi)}{1-\pi}$$

$$p_{11} = (1-p)$$

$$p_{21} = \frac{0}{\pi}$$

$$p_{21} = 0$$

$$p_{22} = \frac{p(1-\pi)}{(1-\pi)}$$

$$p_{22} = p.$$

Finally the transition matrix of conditional misclassification probabilities for Mangat (1994) RRT is

$$\underline{\boldsymbol{P}}_m = \begin{pmatrix} 1 & 1-p \\ 0 & p \end{pmatrix}. \quad (16)$$

4.4. Misclassification Probabilities for Corstange (2004) RRT

Corstange (2004) proposed a new RRT and used hidden logit estimation procedure. In the RRT proposed by Corstange (2004) the respondent is instructed to toss a coin if head appears, he/she is requested to report 'yes' unreservedly, but if tail appears they have to answer a yes/no question. If φ is the probability of unconditional 'yes' and 'p' is the true proportion of respondents saying 'yes', then the probability of a 'yes' response is given as:

$$p(yes) = \varphi + (1-\varphi)\, p \tag{17}$$

For the derivation purpose of conditional misclassification probabilities, we will interchange φ with π, so the probability of a 'yes' response is given as:

$$p(yes) = \pi + (1-\pi)\, p \tag{18}$$

Using conditional probability formula, we find conditional misclassification probabilities for Corstange (2004) RRT as under:

$$P(1|1) = \frac{\pi(1)}{\pi}$$

$$P(1|1) = 1$$

$$P(1|2) = \frac{(1-\pi)\,p}{1-\pi}$$

$$P(1|2) = p$$

$$P(2|1) = \frac{0}{\pi}$$

$$P(2|1) = 0$$

$$P(2|2) = \frac{(1-\pi)(1-p)}{(1-\pi)}$$

$$P(2|2) = 1-p$$

The transition matrix of conditional misclassification probabilities for Corstange (2004) RRT is given by:

$$\underline{\boldsymbol{P}}_c = \begin{pmatrix} 1 & p \\ 0 & 1-p \end{pmatrix}. \tag{19}$$

4.5. Misclassification Probabilities for Huang (2004) RRT

Huang (2004) introduced a straightforward survey method that could be used to estimate the sensitivity of survey questions. His suggested method can also be used to calculate the probability that a respondent will honestly state that they have a sensitive characteristic even in the case of a direct response survey. Taking a simple random sample of size n with replacement, in response to a direct question, the respondent must state whether or not they belong to the sensitive group A. If the respondent selects "no," he or she is given a randomization device with two statements of belonging or not belonging to a sensitive group with p and 1-p probabilities respectively (Shair & Anwar, 2023). No matter if a direct response process is used, the respondent has no motivation to tell a lie because they are a member of an innocent group. In this instance, it is presumptive that respondents in the sensitive group will react in full candor using the RRT, but with probability T and the customary direct response procedure.

Using the proposed technique, the probability of a 'yes' response in the direct response procedure is

$$p(yes) = \pi T \tag{20}$$

The probability of a 'yes' response in the RR procedure is given as:

$$p(yes) = p\pi(1-T) + (1-p)(1-\pi) \tag{21}$$

Next we find conditional misclassification probabilities for Huang (2005) RRT as under:

$$P(1|1) = \frac{p\pi(1-T)}{\pi}$$

$$P(1|1) = p - pT$$

$$P(1|2) = \frac{(1-p)(1-\pi)}{(1-\pi)}$$

$$P(1|2) = (1-p)$$

$$P(2|1) = \frac{\pi\{(1-p) + pT\}}{\pi}$$

$$P(2|1) = 1 - p + pT$$

$$P(2|2) = \frac{p(1-\pi)}{(1-\pi)}$$

$$P(2|2) = p$$

Finally the transition matrix of conditional misclassification probabilities for Huang (2004) RRT is given by

$$\underline{\boldsymbol{P}}_h = \begin{pmatrix} p - pT & 1-p \\ 1-p+pT & p \end{pmatrix} \quad (22)$$

Next we use these derived transition matrices of conditional misclassification probabilities to get perturbed data and then Chi-square test of independence between two variables, is carried out, where variables are subject to misclassification due to RR.

5. Chi-Square Test of Independence

The current section examines testing independence between the two variables, when one or both categorical variables may be misclassified as a result of a RR design. We know the cross-tabulation of the variables A and B from above section. Let $p_{ij} = P(A=i, B=j)$ for each $i \in \{1, 2, .., J\}$. and $j \in \{1, 2, .., J\}$. The data is assumed to be distributed multinomially. The null hypothesis of independence is $H_0 : p_{ij} = p_{i+}\, p_{+j}$, where the plus sign denotes summation over the related index, e.g., $p_{i+} = p_{i1} + p_{i2} + ... + p_{ij}$

5.1. Materials

We take into account the 2x2 table with two cross-classified variables from research into breaking regulations of social benefit (Van Gils et al. 2001), when estimating the chi-square test of independence. The variables Y_1 and Y_2 stand for gender and fraud, respectively. The question is, whether the respondents made money from odd jobs without alerting the office that gives their social benefit. Here Y_1 denotes sex (men=1, women= 2) and Y_2 denotes the latent status as to whether or not the respondent committed fraud (yes= 1, no= 2). Data is considered to be calculated from direct question. Counts are given by

$$\underline{\boldsymbol{y}}^* = (y^*_{11}, y^*_{12}, y^*_{21}, y^*_{22})^t$$

$$\underline{\boldsymbol{y}}^* = (218,\ 500, 152,\ 438)^t \quad (23)$$

Table 3. Classification by Gender (Y_1) and Fraud (Y_2)

Y_1	Y_2		
	Yes	No	Totals
Male	218	500	718
Female	152	438	590
Totals	370	938	1308

Consider Table 3, without misclassification, to apply Chi-Square test, the expected frequencies in the (i,j) cell under H_0 are estimated by $m_{ij} = \frac{n_{i+} n_{+j}}{n}$, where n_{ij} denotes the observed frequencies in the (i,j) cell of the cross-tabulation of A and B, and n is the sample size. The usual chi-square test of independence can be used on the observed table when one or two variables are incorrectly classified and the misclassification is non-differential and independent. Therefore, by using the chi-square test on the observed cross-classification of A* and B*, it is possible to draw conclusions about the independence between A and B when the misclassification is caused by RR. When the Chi-Square test is run on the data in Table 3, it yields $\chi^2 = 3.377$ with 1 degree of freedom (DF) and p-value as 0.066. When we choose a significance level of $\alpha = 0.05$, the data do not give a reason to reject the null hypothesis hence we conclude that variables are independent.

5.2. Chi-square test of independence on misclassified data for Warner (1965) RRT

Now taking the misclassification probabilities of different RRT's as calculated in section 4, we calculate Chi-Square and then check effect of Chi-Square test of independence on perturbed data due to RR design. We get the misclassified (perturbed) data in our calculation, using relationship given as under by Van den Hout et al., (2010):

$$\underline{\boldsymbol{y}} = \underline{\mathrm{P}}^{-1}\ \underline{\boldsymbol{y}}^* \quad (24)$$

We use the transition matrix of conditional misclassification probabilities for Warner (1965) RRT, for different values of p from 0.1-0 .9 to get perturbed or misclassified data using equation (24). Then Chi-Square test of independence is applied on misclassified data. The results are shown in Table 4.

Table 4. Chi-square test for Warner (1965) RRT

P	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
χ^2	5.3061	9.5458	22.2376	109.95	*	109.95	22.2376	9.5458	5.3061
P-Value	0.0213	0.002	0.000	0.000	*	0.000	0.000	0.002	0.0213

This table shows chi-square and p-values of misclassified data taking p = 0.1 - 0.9 with 1 DF. We can see that this table is symmetric around 0.5 and notice that as p increases and tends to 1, the chi square value decreases and move towards original value of chi square. The data gives us a reason to reject the null hypothesis, when we choose a significance level of 0.05 as we can see that p-values are less than 0.05, indicating that the variables are dependent. Chi- square cannot be calculated for p = 0.5 as the transition matrix of conditional misclassification probabilities for Warner (1965) RRT becomes singular, so perturbed data cannot be calculated. Considering p = 1, the matrix of conditional misclassification probabilities yield in identity matrix so we get same value of chi-square as original.

5.3. Chi-Square Test of Independence on misclassified data for Mangat and Singh (1990) RRT

We use the transition matrix of conditional misclassification probabilities for Mangat and Singh (1990) RRT, for different combinations of p and T from 0.1-0.9 to get perturbed or misclassified data using relation (24). Then Chi-Square test of independence is applied on misclassified data.

We can notice that Table 5 is a symmetric table. This table shows chi-square and p-values of misclassified data taking combinational values of p and T from 0.1-0.9 with 1 DF. Values in "*" show that misclassified data using transition matrix of conditional misclassification probabilities for Mangat (1990) RRT yields in negative counts and chi-square cannot be calculated. We can notice that as p and T increase and tend to 1, the chi square values decrease and move towards original value of chi square. When we choose a significance level of $\alpha = 0.05$ then we reject the null hypothesis of independence, as p-values is less than α, for all combinational values of p and T less than or equals to 0.8. Hence we conclude that variables are found to be dependent. Important values to be noted in this table are for taking combinations of T and p as 0.8, 0.9 which yields chi square equals to 3.6678 with 1 DF and p-values as 0.0555. Lastly taking T and p as 0.9 yields which yields chi square equals to 3.5142 with 1 DF and p-values as with 1 DF and p-values as 0.0607. When we choose a significance level of $\alpha = 0.05$ then we do not have enough strong reason to reject the null hypothesis, so variables are found to be independent. Hence we may say that for values of p and T which are greater than or equals to 0.8, variables are independent or we may say that data is accurately classified. Considering p and T = 1, the matrix of conditional misclassification probabilities yields in identity matrix so we get same value of chi-square as original

5.4. Chi-square test of independence on misclassified data for Mangat (1994) RRT

We use the transition matrix of conditional misclassification probabilities for Mangat (1994) RRT for different values of p from 0.1 - 0.9 to get perturbed or misclassified data using relation (24). Finally chi-square test of independence is applied on misclassified data.

Table 6 shows chi-square and p-values of misclassified data taking p = 0.5 - 0.9 with 1 DF. As p increases and tends to 1, the chi square value decreases and move towards original value of chi square. The table gives us a reason to reject the null hypothesis of independence, when we choose a significance level of 0.05 as we can see that p-values are less than 0.05, indicating that the variables are dependent. Chi-square cannot be calculated for p = 0.1 – 0.4 as the transition matrix of conditional misclassification probabilities for Mangat and Singh (1994) RRT yields in negative counts. Considering p = 1, the matrix of conditional misclassification probabilities yield in identity matrix so we get same value of chi-square as original.

5.5. Chi-Square Test of Independence on misclassified data for Corstange (2004) RRT

We use the transition matrix of conditional misclassification probabilities for Corstange (2004) RRT, for different values of p from 0.1 - 0.9 to get perturbed or misclassified data using relation (24). After getting misclassified data, chi-square test of independence is applied in table 7.

Table 7 depicts chi-square and p-values of misclassified data taking p = 0.1 - 0.5 with 1 DF. As p decreases and tends to 0, the chi square value also decreases and move towards original value of chi square. The data gives us a reason to reject the null hypothesis, when we choose a significance level of 0.05 as we can see that p-values are less than 0.05, indicating that the variables are dependent. Chi-square cannot be calculated for p = 0.6 – 0.9 as the transition matrix of conditional misclassification probabilities for Corstange (2004) RRT yields in negative counts. Considering p = 1, the matrix of conditional misclassification probabilities yields in symmetric matrix so we cannot get perturbed or misclassified data. Considering p = 0, the matrix of conditional misclassification probabilities yield in identity matrix so we get same value of chi-square as original. Graphical comparison of χ^2 values is shown in figure 1 against different values of p, where we can clearly see that Corstange (2004) and Mangat (1994) has totally opposite behavior.

Table 5. Chi-Square Test for Mangat and Singh (1990) RRT

T		P								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	χ^2	8.9244	18.1773	57.6500	*	*	48.6034	16.5578	8.3622	5.0467
	P-Value	0.0028	0.0000	0.0000	*	*	0.0000	0.0000	0.0038	0.0247
0.2	χ^2	18.1773	48.6034	*	*	109.947	27.8700	12.8250	7.3871	4.8060
	P-Value	0.0000	0.0000	*	*	0.0000	0.0000	0.0003	0.0066	0.0284
0.3	χ^2	57.6500	*	*	*	41.5930	18.1774	10.2351	6.5741	4.5821
	P-Value	0.0000	*	*	*	0.0000	0.0000	0.0014	0.0104	0.0323
0.4	χ^2	*	*	*	48.6034	22.2376	12.8250	8.3622	5.8890	4.3851
	P-Value	*	*	*	0.0000	0.0000	0.0003	0.0038	0.0152	0.0363
0.5	χ^2	*	109.9497	41.5930	22.2376	13.9132	9.5458	6.9627	5.3061	4.1791
	P-Value	*	0.0000	0.0000	0.0000	0.0002	0.0020	0.0083	0.0213	0.0409
0.6	χ^2	48.6034	27.8700	18.1774	12.8250	9.5458	7.3871	5.8890	4.8060	3.9973
	P-Value	0.0000	0.0000	0.0000	0.0003	0.0020	0.0066	0.0152	0.0284	0.04557
0.7	χ^2	16.5578	12.8250	10.2351	8.3622	6.9627	5.8890	5.0467	4.3851	3.8272
	P-Value	0.0000	0.0003	0.0014	0.0038	0.0083	0.0152	0.0247	0.0363	0.0504
0.8	χ^2	8.3622	7.3871	6.5741	5.8890	5.3061	4.8060	4.3851	3.9973	3.6678
	P-Value	0.0038	0.0066	0.0104	0.0152	0.0213	0.0284	0.0363	0.0456	0.0555
0.9	χ^2	5.0467	4.8060	4.5821	4.3851	4.1791	3.9973	3.8272	3.6677	3.5181
	P-Value	0.0247	0.0284	0.0323	0.0363	0.0409	0.0456	0.0504	0.0555	0.0607

Table 6. Chi-square test for Mangat (1994) RRT

P	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
χ^2	*	*	*	*	37.8902	12.4485	7.4477	5.3133	4.1297
P-Value	*	*	*	*	0.0000	0.0004	0.0064	0.0212	0.0421

Table 7. Chi-square test for Corstange (2004) RRT

P	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
χ^2	4.1297	5.3133	7.4477	12.4485	37.8902	*	*	*	*
P-Value	0.0421	0.0212	0.0064	0.0004	0.0000	*	*	*	*

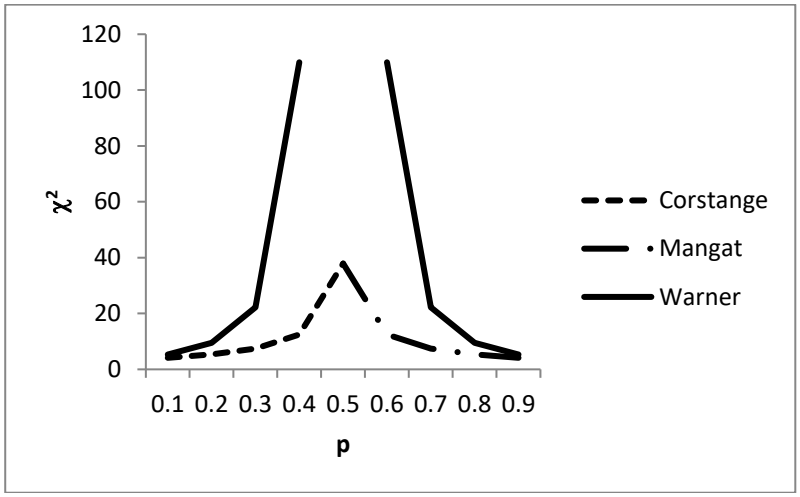

Figure 1: Comparison of χ^2 against different values of *p*

5.6. Chi-Square Test of Independence on misclassified data for Huang (2004) RRT

We use the transition matrix of conditional misclassification probabilities for Huang (2004) RRT for different combinations of p and T from 0.1-0.9 to get perturbed or misclassified data using relation (24). Then Chi-Square test of independence is applied on misclassified data. The results are shown in Table 8. We can notice that Table 8 is not a symmetric table. This table shows chi-square and p-values of misclassified data taking combinational values of p and T from 0.1-0.9 with 1 DF. Values in "*" show that misclassified data using transition matrix of conditional misclassification probabilities for Huang (2004) RRT yields in negative counts and chi-square cannot be calculated. We can notice that when p is higher and T is lower p so the chi square values decrease and move towards original value of chi square and vice versa. When we choose a significance level as 0.05 then we reject the null hypothesis, as p-values is less than α, for all combinational values of p and T. Hence we conclude that variables are found to be dependent. Taking T and p above 0.5 gives negative counts of misclassified data. Hence chi-square cannot be calculated. Important values to be noted in this table are for taking all values of T and p = 0.1, where highest value of T yields chi square equals to 4.42 with 1 DF and p-value as 0.0355, which is near to the original value of chi square. But the data gives us a reason to reject the null hypothesis, when we choose a significance level of 0.05 as we can see that p-values are less than 0.05, indicating that the variables are dependent. Taking into consideration p = 1 and T = 0, the matrix of conditional misclassification probabilities yields in identity matrix so we get same value of chi-square as original.

Table 8. Chi-Square Test for Huang (2004) RRT

T		P								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	χ^2	5.1904	9.0283	19.8459	86.6713	*	*	41.4386	14.0098	7.1352
	P-Value	0.0227	0.0027	0.0000	0.0000	*	*	0.0000	0.0002	0.0076
0.2	χ^2	5.0797	8.5640	17.9186	71.5275	*	*	*	26.3166	10.8891
	P-Value	0.0242	0.0034	0.0000	0.0000	*	*	*	0.0000	0.0010
0.3	χ^2	4.9737	8.1452	16.3326	60.8887	*	*	*	*	22.9775
	P-Value	0.0257	0.0043	0.0001	0.0000	*	*	*	*	0.0000
0.4	χ^2	4.8720	7.7654	15.0045	53.0048	*	12.8250	8.3622	*	*
	P-Value	0.0273	0.0053	0.0001	0.0000	*	0.0003	0.0038	*	*
0.5	χ^2	4.7743	7.4194	13.8761	46.9286	*	*	*	*	*
	P-Value	0.0289	0.0065	0.0002	0.0000	*	*	*	*	*
0.6	χ^2	4.6805	7.1030	12.9056	42.1021	*	*	*	*	*
	P-Value	0.0305	0.0077	0.0003	0.0000	*	*	*	*	*
0.7	χ^2	4.5903	6.8124	12.0620	38.1759	*	*	*	*	*
	P-Value	0.0322	0.0091	0.0005	0.0000	*	*	*	*	*
0.8	χ^2	4.5035	6.5447	11.3219	34.9195	*	*	*	*	*
	P-Value	0.0338	0.0105	0.0008	0.0000	*	*	*	*	*
0.9	χ^2	4.4200	6.2972	10.6673	32.1749	*	*	*	*	*
	P-Value	0.0355	0.0121	0.0011	0.0000	*	*	*	*	*

6. Conclusions and Recommendations

We may observe that Mangat (1994) and Corstange (2004) RRT have opposite behavior in terms of their chi-square values but Warner (1965) shows a symmetric behavior around 0.5. Finally we may say that chi-square can be calculated for perturbed data taking misclassification probabilities as 0.1- 0.9. From 0.8 or higher shows that data is accurately classified, so chi-square can be calculated easily. Finally we may say that chi-square can be calculated for perturbed data taking misclassification probability as p = 0.8 or higher in all RRT's except Corstange (2004). In case of Mangat and Singh (1990) chi-square can be calculated for all values of p and for T = 0.8 but for p = 0.8 gives chi-square value near to the value for the original data which is irrespective of misclassification. So as a comparative statement we may say that Mangat and Singh (1990)is fund to be efficient in results. We may conclude that when the data is misclassified, the results of different estimates are altered. Chi-square test of independence calculated for perturbed data shows that variables are dependent but for original data, variables were independent. We can conclude that misclassification can change the status of dependence in the data. We recommend that misclassification probabilities of further RR models can be derived using the same methodology given in the current work. Furthermore the derived matrices of conditional misclassification probabilities can be used acquire estimates of log-linear model for numerous randomized response techniques.

References

Christofides, T.C., (2003).A generalized randomized response technique, *Metrika* 57, 195-200.

Corstange, D., (2004). Sensitive questions, truthful responses? Randomized response and hidden logit as a procedure to estimate it, Annual Meeting of the American Political Science Association 2 - 5 September 2004.

Egleston, B. L., Miller, S. M., Meropol, N. J., (2011) The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects, National Library of Medicine 30(30), 3560-3572.

Fujisawa, K., Tahata K., Quasi (2022). Association Models for Square Contingency Tables with Ordinal Categories, Symmetry 14(4), 805.

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.

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) 439-461.

Huang. K.C., (2004). A survey technique for estimating the proportion and sensitivity in a dichotomous finite population, Statistica Neerlandica 58(1) 75-82.

Johnson, R. A., Wichern, D. W., (2006). Applied Multivariate Statistical Analysis, Pearson Education, Inc. Dorling Kindersley, India. 3, 581-598

Kim, J.M., Elam, M.E., (2007). A stratified unrelated question randomized response model, Statistical Papers 48(2), 215–33.

Kim, J.M., Warde, W.D., (2004). A Mixed Randomized Response Model. Journal of Statistical Planning Inference, 133(1) 211-221.

Kim, J.M., Warde, W.D., (2005. Some new results on the multinomial randomized response model, Communications in Statistics—Theory and Methods 34(4) 847-856.

Ko, M. M., Kim H., (2016). The Study of Misclassification Probability in Discriminant Model of Pattern Identification for Stroke, Evidence-Based Complementary and Alternative Medicine Article ID 1912897

Mahmood, M., Singh, S., Horn, S., (1998). On the confidentiality guaranteed under randomized response sampling: a comparison with several new techniques, *Biometrical Journal* 40(2) 237-242.

Mangat, N.S., (1994). An improved randomized response strategy, Journal of Royal Statistical Society, B 56(1) 93-95.

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

Moors, J.J.A., (1971). Optimization of the unrelated question randomized response model. Journal of the American Statistical Association 66(335), 627–629.

Narjis, G., Shabbir, J., (2022). An improved two-stage randomized response model for estimating the proportion of sensitive attribute. Sociological Methods & Research 52(1) 335-355.

Ngailo, E. K., Chuma, F., ((2022). Approximation of misclassification probabilities in linear discriminant analysis based on repeated measurements, Communications in Statistics - Theory and Methods.

Ngailo, E. K., Ngaruye, I., (2022). Asymptotic results for expected probability of misclassifications in linear discriminant analysis with repeated measurements, Communications in Statistics - Theory and Methods.

Shair, W., & Anwar, M. (2023). Effect of internal and external remittances on expenditure inequality in Pakistan. *Cogent Economics & Finance*, *11*(1), 2178121.

Shair, W., Majeed, M.T., (2020). Labor Market Outcomes of Non-migrant Members in Response to Remittances: Evidence from Provincial capital of Punjab and Khyber Pakhtunkhawa (KPK). *Review of Socio-Economic Perspectives*. 5, 1–22.

Singh, C., Singh, G. N., Kim, J.M., (2021). A randomized response model for sensitive attribute with privacy measure using Poisson distribution, *Ain Shams Engineering Journal* 12(4), 4051-4061.

Singh, G., Singh, C., (2022). Proficient randomized response model based on blank card strategy to estimate the sensitive parameter under negative binomial distribution, *Ain Shams Engineering Journal* 13(5).

Singh, H.P., Tarray, T.A., (2014). A stratified Mangat and Singh's optional randomized response model using proportional and optimal allocation, Statistica 74(1) 65–83.

Singh, H.P., Tarray, T.A., (2012). A Stratified Unknown repeated trials in randomized response sampling, Communications for Statistical Applications and Methods 19(6), 751–759.

Van den Hout, A., Gilchrist, R., Van der Heijden, P.G.M., (2010). The Randomized Response Model as a Composite Link Model. Statistical Modeling 10, 57-67.

Van den Hout, A., Peter G. M. van der Heijden., (2002). Randomized Response, Statistical Disclosure Control and Misclassification: A Review. *International Statistical Review / Revue Internationale de Statistique* 70(2) 269–288.

Van Gils, G., Van der Heijden, P.G.M., Rosebeek, A., (2001). Onderzoek naar regelovertreding. Resultaten ABW, WAO en WW. Amsterdam: NIPO. (In Dutch)

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