Original Research Article

Mangat Randomized Response Group Testing Theory

Abstract Pls write this in one paragraph

~~Aims/ objectives:~~ This paper proposed a new Randomized Response Group Testing (RRGT) model that estimates proportion of people characterizing a sensitive variable ~~(S)~~ under study.
~~Methodology:~~ Simple random sampling with replacement, binomial probability distribution and maximum likelihood were used as randomization procedure, sampling model and estimation technique, respectively.
~~Results:~~ The distributional properties (expectation and variance) of the proposed estimator, efficiency comparison of the model with some existing models, and numerical illustration of all the competing models were also explored.
~~Conclusion:~~ The study found that the developed model outperformed competing existing orthodox RRMs and earlier RRGT model in terms of efficiency, privacy protection and it is economically advantageous.

Keywords: Randomized response; Group testing; ~~Proportion~~; ~~sensitive variable~~; Binomial probability distribution; maximum likelihood estimation

2010 Mathematics Subject Classification: 62D05

1 Introduction

Noise is intrinsic and inevitable in survey exercises. When noise/disturbance arises due to sensitive nature of the survey question(s), randomised response theory invented by (8) gives a setting to construct techniques where these problems become well posed. To obtain estimator of specific quality (say, efficiency) and computational advantage (ease of mathematical workloads), measurement noise is often assumed to follow Gaussian distribution in a wide range of applications (1). The conventional

method of data collection usually employed in ordinary statistical surveys, may suffer a surprising failure when applied on noisy data without choosing appropriate values for parameters to minimize the noise because obtaining truthful responses is challenging in all types of surveys, particularly, when sensitive subject matters are being investigated. Sensitive variables such as raping, sexual assault, pilfering etc are thought to be threatening to respondents (8), (2). In RR design, a yes-answer does not imply admission of guilty. Consequently, RRM reduce non-response and response bias, increase cooperation and ensures absolute protection of privacy.

write in full

Warner technique requires every person in a population be partitioned into two mutually exclusive and exhaustive groups (say, sensitive group S and non-sensitive group S'). The objective is to estimate π (the proportion of respondent belonging to the sensitive group). The unbiased maximum likelihood estimator of π (proportion of sensitive group in the population) is:

$$\hat{\pi}_w = \frac{\frac{n_o}{n} - (1-p)}{2p-1} = \frac{\hat{\lambda} - (1-p)}{2p-1}, \qquad \text{with } p \neq \tfrac{1}{2};\ \hat{\lambda} = \tfrac{n_o}{n} \tag{1.1}$$

and variance

$$Var(\hat{\pi}_w) = \frac{\pi_s(1-\pi_s)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{1.2}$$

Subsequently, several attempts, in theory and application, with similar objective, aimed at improving the efficiency and effectiveness of Warner's RRM, have been observed in literature. Another popular work is (5) improved two-step procedure, an optimization of [(4)], instructed respondent to answer yes if he or she is a $\in U_s$, otherwise use [(8)] design. In this case, the proportion of yes-answers is obviously contaminated with magnitude

$$\lambda = \pi + (1-\pi)(1-p). \tag{1.3}$$

Hence, impartial estimator of the population proportion is given by

$$\hat{\pi}_m = \frac{\hat{\lambda} - (1-p)}{p} = \frac{\frac{n_o}{n} - 1 + p}{p} \tag{1.4}$$

where $\hat{\lambda} = \frac{n_o}{n}$ is the remarked proportion of yes-responses with

$$Var(\hat{\pi}_m) = \frac{\hat{\lambda}(1-\hat{\lambda})}{np^2} = \frac{\pi(1-\pi)}{n} + \frac{(1-p)(1-\pi)}{np}. \tag{1.5}$$

Group Testing (GT) is a sampling scheme where concurrent readings are recorded for a cluster of units instead of obtaining measurements on individuals (7); (6). The reading is often taken to be binary, with a positive test signifying the appearance of at least a positive units in the cluster. This (GT) sampling strategy provides significant benefits such as reduction in the classification cost of all elements of a universe in consonance to whether or not they carry a certain characteristics when the incidence rate is infinitesimal or rare. That is, introducing group testing in RR survey makes the cost of survey becomes cheaper and the respondents' privacy protected. GT begins with the presupposition that survey units, whose replies are identically and independently Bernoulli (π) random variables, can be merged into clusters of size $k > 1$ (k can be equal or unequal). GT requires selecting a srs of m groups, each of size k and from each observed cluster, select one individual to examine whether it is negative or not. If the observed element from a given cluster is negative, we discard the whole class and select another cluster for testing. Supposing all responses are truthful, the size of positive categories recorded, say L, has a binomial probability density function with parameters $m, 1-(1-\pi)^k$. Under the GT model, the MLE of π is:

$$\hat{\pi}_{gt} = 1 - \left(1 - \frac{L}{m}\right)^{\frac{1}{k}} \tag{1.6}$$

where
$\pi =$ population proportion of units having the ignominious trait,
$m =$ number of batch being appraised, and
$n =$ number of units belonging to m clusters each of size k.

write in full

(3) incorporated GT into RRT using (8) RR model. The derived model possessed the protection/confidentiality property of the RR model, as well as the economic advantage (cost reduction) of the GT model. (3) estimator under Warner design is

$$\hat{\pi}_{wg} = \frac{p - \left(1 - \frac{L_1}{m}\right)^{\frac{1}{k}}}{2p - 1} \tag{1.7}$$

with variance

$$Var(\hat{\pi})_{wg} = \frac{1}{(2p-1)^2}\left[\left(\frac{k^{-1}}{m}\right)^2 Var(L_1) + \left\{\frac{D_1}{2m^2}\right\}^2 Var(L_1^2) - 2\left\{\frac{D_2}{2m^3}\right\}^2 Cov(L_1, L_1^2)\right] \tag{1.8}$$

where,
$D_1 = k^{-1}\left(k^{-1} - 1\right)$,
$D_2 = k^{-2}\left(k^{-1} - 1\right)$,
$Var(L_1) = E(L_1^2) - E^2(L_1)$,
$Var(L_1^2) = E(L_1^4) - E^2(L_1^2)$ and
$Cov(L_1, L_1^2) = E(L_1^3) - E(L_1)E(L_1^2)$.

Other sophisticated works on RRGT-models are: Two-stage randomised response group-testing model (7), M-stage hierarchical group testing model for estimating the occurrence of rare Multiple traits in a finite population (6). This paper aimed at developing a robust and more efficient randomized response group testing model that performs better than existing conventional RRMs and frontier RRGT models.

2 Methodology

This study incorporates group testing (GT) method to the randomized response (RR) design suggested by (5). First, the whole population (of size N) is divided into M homogeneous subgroups (equal group size $k(k \geq 2)$), based on some prior or auxiliary information which assumed available to the researcher. Suppose an investigator is interested in estimating the proportion of students who used Tyramadol in a College, student academic performance based on cumulated grade point average (CGPA) can be used for stratification factor such that students in the same group belong to same homogeneous category of CGPA. Then, a simple random sample of m-groups are selected from M homogeneous groups from which a response is recorded according to Mangat randomized response technique. By following Mangat's method, the sensitive question under study is put before everyone in the group with randomization device unobserved by the interviewer and thus the individual's privacy is maintained. Assuming there are no reporting errors, the number of sensitive trait groups observed, say L , has a binomial distribution with parameters m and π_k where

$$\pi_k = 1 - [1 - \{\pi_{mg} + (1-p)(1-\pi_{mg})\}]^k \tag{2.1}$$

By using the method of moments for the Mangat RR-GT model

$$\hat{\pi}_k = \frac{L}{m} \tag{2.2}$$

Putting (2.1) in (2.2) gives the following results

$$1-[1-\{\pi_{mg}+(1-p)(1-\pi_{mg})\}]^k=\frac{L}{m}$$
$$1-\pi_{mg}-(1-p)(1-\pi_{mg})=\left(1-\frac{L}{m}\right)^{\frac{1}{k}}$$
$$p-p\pi_{mg}=\left(1-\frac{L}{m}\right)^{\frac{1}{k}}$$

which finally gives

$$\hat{\pi}_{mg}=\frac{p-\left(1-\frac{L}{m}\right)^{\frac{1}{k}}}{p};\quad p\neq 0 \tag{2.3}$$

The estimator π_{mg} can be expanded or approximated by the binomial expansion as

$$\hat{\pi}_{mg}=1-\left[\frac{1-k^{-1}\left(\frac{L}{m}\right)+\frac{k^{-1}(k^{-1}-1)}{2}\left(\frac{L}{m}\right)^2-\frac{k^{-1}(k^{-1}-1)(k^{-1}-2)}{6}\left(\frac{L}{m}\right)^3+\frac{k^{-1}(k^{-1}-1)(k^{-1}-2)(k^{-1}-3)}{24}\left(\frac{L}{m}\right)^4}{p}\right] \tag{2.4}$$

Neglecting the terms having the power of $\frac{L}{m}$ more than 2 in (2.4), the variance of estimator π_{mg} can be approximated as

$$Var(\hat{\pi}_{mg})\approx\frac{1}{p^2}Var\left[-k^{-1}\left(\frac{L}{m}\right)+\frac{k^{-1}(k^{-1}-1)}{2}\left(\frac{L}{m}\right)^2\right] \tag{2.5}$$

This is equivalent to

$$Var(\hat{\pi})_{mg}=\frac{1}{p^2}\left[\left(\frac{k^{-1}}{m}\right)^2Var(L_1)+\left\{\frac{D_1}{2m^2}\right\}^2Var(L_1^2)-2\left\{\frac{D_2}{2m^3}\right\}^2Cov(L_1,L_1^2)\right] \tag{2.6}$$

where,
$D_1=k^{-1}\left(k^{-1}-1\right)$,
$D_2=k^{-2}\left(k^{-1}-1\right)$,
$Var(L_1)=E(L_1^2)-E^2(L_1)=m\pi_g(1-\pi_g+m\pi_g)-[m\pi_g]^2$,
$Var(L_1^2)=E(L_1^4)-E^2(L_1^2)=m\pi_g(1-7\pi_g+7m\pi_g+12\pi_g^2-18m\pi_g^2+6m^2\pi_g^2-6\pi_g^3+11m\pi_g^3-6m^2\pi_g^3+m^3\pi_g^3)-[m\pi_g(1-\pi_g+m\pi_g)]^2$ and
$Cov(L_1,L_1^2)=E(L_1^3)-E(L_1)E(L_1^2)=m\pi_g(1-3\pi_g+3m\pi_g+2\pi_g^2-3m\pi_g^2+m^2\pi_g^2)-[m\pi_g][m\pi_g(1-\pi_g+m\pi_g)]$.

2.1 Efficiency Comparison

The proposed Mangat RRGT model is compared with Direct Questioning Technique (DQT), (5) RRM and Warner's RRGT model developed by (3) using Mean Square Error (MSE) criterion and validated with an artificial data. Note that Warnar, Mangat, Kim and Heo, and the proposed estimators are unbiased, then the criteria will be limited to variance comparison. Elementary statistical theory asserts that the variance of the estimate of proportion π is

$$Var(\hat{\pi}_{dqt})=\frac{\hat{\pi}_s(1-\hat{\pi}_s)}{n} \tag{2.7}$$

Equation (2.7) is a special case of the (8) model for value of $p=1$. The relative efficiency (RE) of the proposed estimator ($\hat{\pi}_{mg}$) with respect to direct question approach is defined using (2.7) and (2.6) as

$$RE(\pi_{dqt},\pi_{mg})=\frac{Var(\hat{\pi}_{dqt})}{Var(\hat{\pi})_{mg}}\quad \text{where } p,\pi_s\in[0,1]. \tag{2.8}$$

Furthermore, the developed estimator is compared with conventional Mangat using (1.5) and (2.6) and defined as

$$RE(\pi_m, \pi_{mg}) = \frac{Var(\hat{\pi}_m)}{Var(\hat{\pi})_{mg}} \quad \forall\, p, \pi_s \in [0, 1]. \tag{2.9}$$

Lastly, Warner-RRGT model developed by [(3)] is compared with the proposed Mangat-RRGT technique using (1.8) and (2.6) as

$$RE(\pi_{wg}, \pi_{mg}) = \frac{Var(\hat{\pi}_{wg})}{Var(\hat{\pi})_{mg}} \quad \forall\, p, \pi_s \in [0, 1]. \tag{2.10}$$

The proposed Mangat-RRGT model is more efficient than DQT, Mangat orthodox RRM and Warner-RRGT models if and only if (2.8), (2.9) and (2.10) are each > 1. The tables and figures showing the results are presented in the next section.

3 Results and Discussion

This section presents analysis of data and the discussions of results. The study generated the relative efficiency for all values of $\pi = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$; $m = 20, 25$ and 30 and group size $k = 2$ (assumed equal for all clusters). The numerical results showing effect of varying design parameter on the relative efficiency of the proposed RRGT versus DQT, original (5) and (3) models are presented both in tabular form and graphically. Table 1 and Figure 1 depict the efficiency of the model over DQT. Table 2 demonstrates the relative efficiency of the model versus conventional Mangat-RRM and the developed model efficiency against Warner-RRGT model propounded by (3). Figure 2 graphically illustrated Table 2, and the salient features of the results were discussed.

Table 1: Performance of the proposed Mangat-RRGT versus the Direct Questioning Technique (DQT) when $p = 1$

π	$m = 20,\ n = 40$	$m = 25,\ n = 50$	$m = 30,\ n = 60$
0.1	1.7791467	1.7860396	1.7906595
0.2	1.6293196	1.6339816	1.6371041
0.3	1.4978223	1.5006858	1.5026029
0.4	1.3817582	1.3831680	1.3841123
0.5	1.2787852	1.2790177	1.2791756
0.6	1.1869921	1.1862706	1.1857922
0.7	1.1048064	1.1033119	1.1023189
0.8	1.0309245	1.0288045	1.0273958
0.9	0.9642577	0.9616328	0.9598893

Pls amend the font size

small p

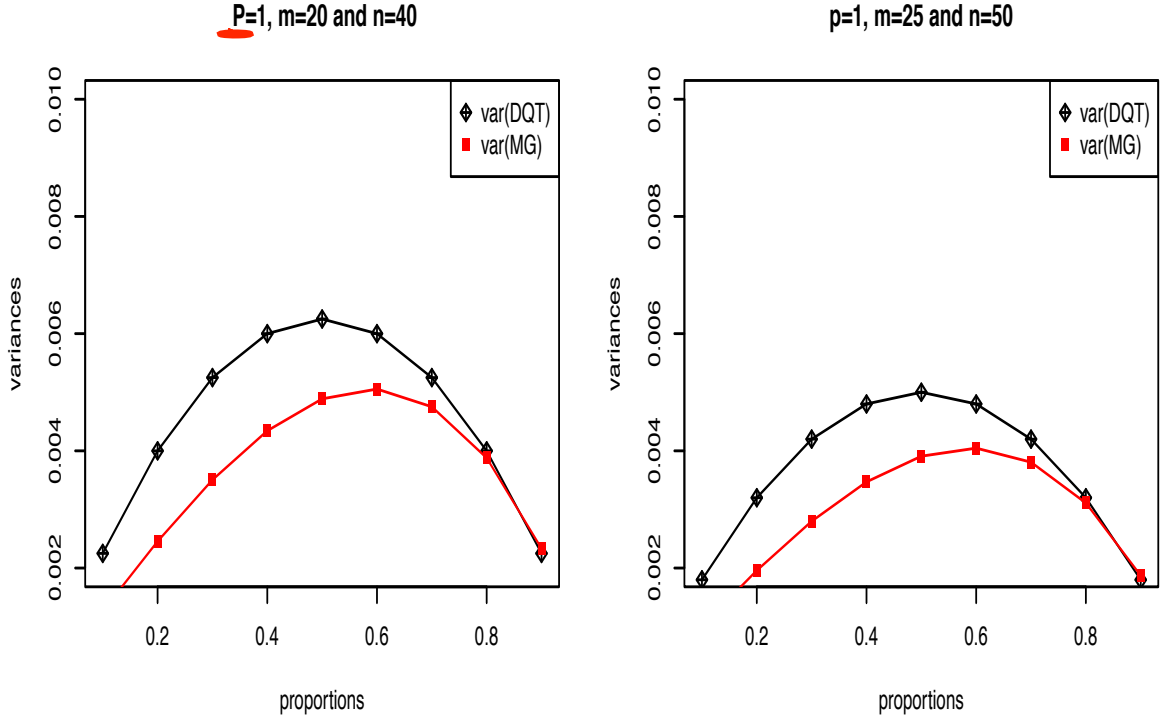

Figure 1: Efficiency of the proposed Mangat-RRGT versus the Direct Questioning Technique (DQT)

why no graph for m = 30 & n = 60 ?

empty page

Table 2: Performance of the proposed Mangat-RRGT versus the conventional Mangat and Warner-RRGT RRMs

		Mangat-RRGT vs Mangat			Mangat-RRGT vs Warner-RRGT		
p	π	$m = 20, n = 40$	$m = 25, n = 50$	$m = 30, n = 60$	$m = 20, n = 40$	$m = 25, n = 50$	$m = 30, n = 60$
0.7	0.1	4.6079900	4.6258426	4.6378080	2.735579	2.745253	2.751745
	0.2	2.5091522	2.5163316	2.5211402	2.526093	2.532067	2.536068
	0.3	1.7824085	1.7858161	1.7880975	2.351497	2.354736	2.356901
	0.4	1.4024846	1.4039155	1.4048740	2.205587	2.206834	2.207664
	0.5	1.1636945	1.1639061	1.1640498	2.083422	2.083248	2.083129
	0.6	0.9970733	0.9964673	0.9960655	1.981031	1.979875	1.979103
	0.7	0.8727971	0.8716164	0.8708320	1.895194	1.893398	1.892202
	0.8	0.7757707	0.7741754	0.7731153	1.823287	1.821114	1.819669
	0.9	0.6974797	0.6955810	0.6943199	1.763153	1.760805	1.759247
0.8	0.1	3.9852887	4.0007287	4.0110772	1.588001	1.593617	1.597385
	0.2	2.3462202	2.3529335	2.3574298	1.466394	1.469862	1.472184
	0.3	1.7574448	1.7608047	1.7630541	1.365041	1.366922	1.368178
	0.4	1.4370285	1.4384947	1.4394768	1.280341	1.281065	1.281547
	0.5	1.2276337	1.2278570	1.2280086	1.209424	1.209323	1.209254
	0.6	1.0762061	1.0755520	1.0751183	1.149986	1.149315	1.148867
	0.7	0.9596033	0.9583052	0.9574427	1.100158	1.099116	1.098421
	0.8	0.8659766	0.8641958	0.8630125	1.058416	1.057155	1.056316
	0.9	0.7885485	0.7864019	0.7849761	1.023508	1.022145	1.021241
0.9	0.1	3.0423409	3.0541277	3.0620277	1.130520	1.134518	1.137201
	0.2	2.0529427	2.0588168	2.0627511	1.043947	1.046415	1.048069
	0.3	1.6625828	1.6657613	1.6678893	0.971792	0.973131	0.974025
	0.4	1.4301197	1.4315788	1.4325562	0.911493	0.912008	0.912351
	0.5	1.2659973	1.2662275	1.2663839	0.861006	0.860934	0.860885
	0.6	1.1395124	1.1388198	1.1383605	0.818691	0.818214	0.817895
	0.7	1.0369397	1.0355370	1.0346050	0.783218	0.782476	0.781982
	0.8	0.9510279	0.9490721	0.9477726	0.753501	0.752603	0.752006
	0.9	0.8774745	0.8750858	0.8734993	0.728650	0.727680	0.727036

pls amend the font size

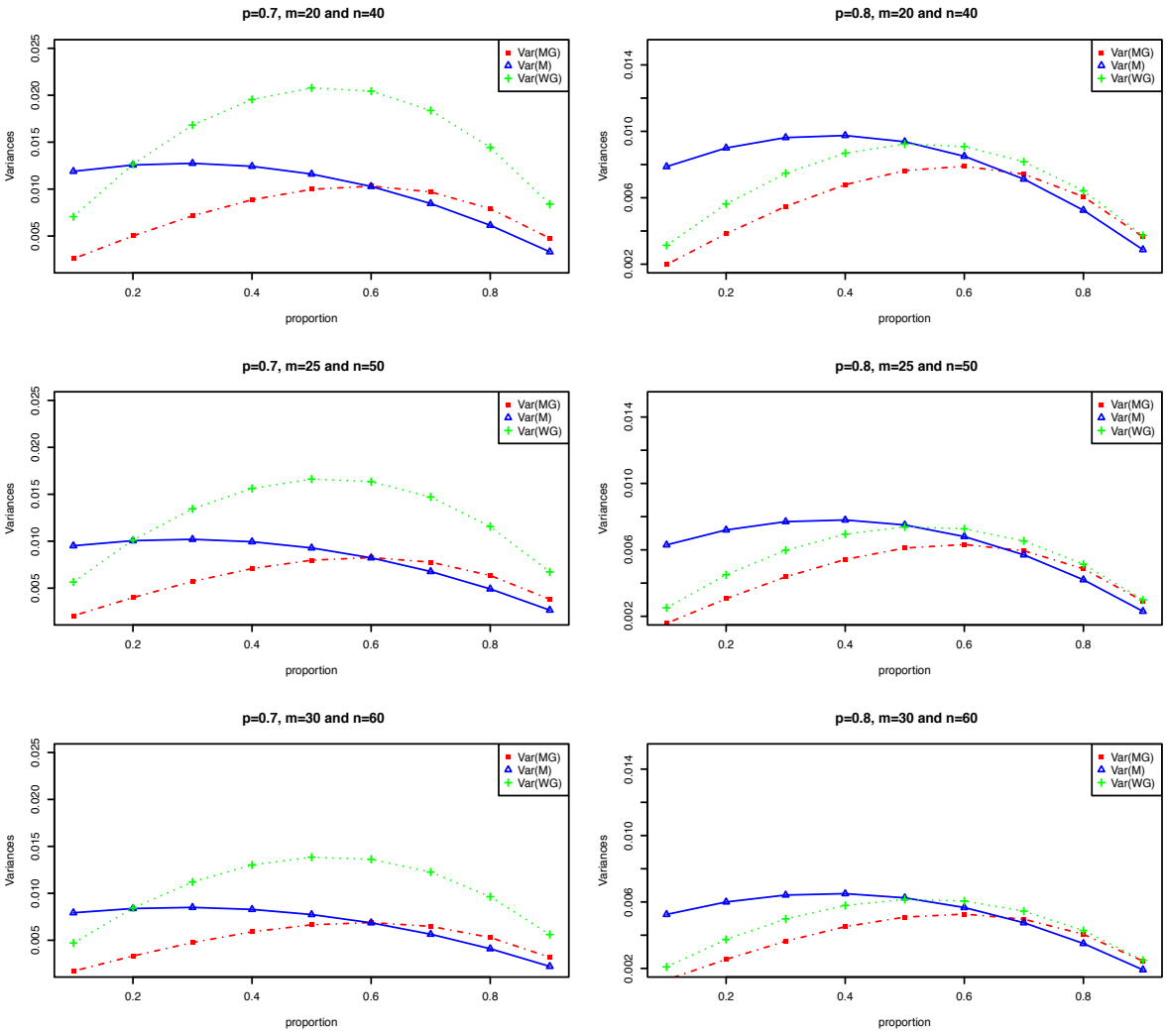

Figure 2: Efficiency of the proposed Mangat-RRGT with Conventional Mangat and Warner-RRGT models

why no graph for p = 0.9

delete this → ~~4 Discussion~~

this paragraph should be after Figure 1

From Table 1, for all sample of groups i.e., when $m = 20, n = 40$, $m = 25, n = 50$, and $m = 30, n = 60$, with $p = 1$ for DQT, $k = 2$ and $n = mk$, the relative efficiency $RE(\pi, \pi_{mg})$ is greater than 1 and decreases as the value of π increases from 0.1 to 0.9. Thus, the proposed group testing method performs better than DQT $\forall\, p \in [0.1, 0.8]$ which holds in most practical situations. While a high level of randomisation parameter (p) may reflect the extent at which respondents characterise the sensitive traits under survey, there is little practical application of $p \approx 1$ in a real life study of RR survey, since increase in randomization parameter decelerates the truthful response. Hence, careful attention must

be paid to the choice of p so as not to compromise quality information and design effectiveness for model efficiency. As m increases from 20 to 30, there is a slight gain in efficiency.

this paragraph after Figure 2

Table 2 depicts that for all sample of groups, that is, when $m = 20, n = 40$, $m = 25, n = 50$, $m = 30, n = 60$, $g = 2$ and $n = mg$, with $P = 0.7$, 0.8, 0.9 the relative efficiency $RE(\pi_m, \pi_{mg})$ is greater than 1 and decreases as the value of π increases from 0.1 to 0.9. Hence, the proposed estimator π_{mg} is more efficient than the (5) traditional estimator π_m. As m is increasing from 20 to 30, there is a great gain in efficiency. The developed estimator is less efficient when $p-0.9$ which is in line with the rule of RRM because privacy protection will be lost as $p \to 1$. This is valid in most practical situations. Also from Table 2, the relative efficiency of Warner Group Testing with the proposed model for $P = 0.7$, 0.8, 0.9, $m = 20, n = 40$, $m = 25, n = 50$, $m = 30, n = 60$, $g = 2$, $n = mg$ as the value of π increases from 0.1 to 0.9. It is crystal clear that the relative efficiency $RE(\pi_{wg}, \pi_{mg})$ is greater than 1 and decreases as the value of π increases from 0.1 to 0.9. Thus, the study inferred that the proposed estimator π_{mg} performs better in terms of efficiency than the earlier existing Warner group testing estimator π_{wg} proposed by (3)Kim and Heo (2013). Similarly, inefficiency of the proposed estimator when $\pi = 0.9$ is an extra benefit to the respondents is in agreement with general understanding that the use of $p = 1$ or ≈ 1 is practically impossible in any effective RR-survey. Evidenced from Figure 2 which gives visual description of possible results of the relative efficiency of the proposed Mangat-RRGT versus conventional Mangat and Warner-RRGT model, it becomes apparent that the developed model is better than the two competing models $\forall\, \pi \in [0.1, 0.8]$.

small letter

4

5 Conclusions

This study adopted the concept of group-testing to RR-model developed by (5) and proposed an improved estimator known as Mangat randomized response group-testing (Mangat-RRGT) model. Empirically, it is established that the proposed model is more efficient than the Direct Questioning Technique (DQT), conventional RR model given by (5) and RRGT model established by (3). Extra merits of the developed Mangat-RRGT model is in cost reduction of the survey (as the data collection is done on group basis), increment in the privacy protection of the respondents and model efficiency over the earlier existing ones.

References

[1] Adeniran, A. T., Faweya, O., Ogunlade, T.O. and Balogun, K. O. (2020a). Derivation of Gaussian Probability Distribution: A New Approach, *Scientific Research Publishing: Applied Mathematics*, 11, 436-446,DOI: https://doi.org/10.4236/am.2020.116031

[2] Adeniran, A. T., Sodipo, A. A. and Udomboso, C. G. (2020b). A Modified Forced Randomized Response Model, *Journal of the Nigerian Society of Physical Sciences (JNSPS)*, **2**(1), 36-50, DOI: https://doi.org/10.46481/jnsps.2020.1

[3] Kim, J. M. and Heo, T. Y. (2013). Randomised response group testing model. *Journal of Statistical Theory and Practice*, **7**(1), 33-48, DOI: 10.1080/15598608. 2013.756324.

[4] Mangat, N. S. and Singh, R. (1990). An alternative randomised response procedure. *Biometrika*, **77**(2), 439-442, DOI: 10.1093/biomet/77.2.439.

[5] Mangat, N. S. (1994). An improved randomised response strategy. *Journal of Royal Statistical Society*, Series B, **56**(1), 93-95, DOI:10.0035-9246/94/56093.

[6] Sirengo, J. L., Kennedy, N. L. and Shem, A. (2020). An m-stage hierachical group testing model for estimating multiple traits in a population. *IOSR Journal of Mathematics (IOSR-JM)*, **16**(3), 29-34, DOI: 10.9790/5728-1603022934, e-ISSN: 2278-5728, p-ISSN: 2319-765X.

[7] Tiwari, N. and Mehta, P. (2017). Two-Stage Randomised Response Group-Testing Model, *Statistics and Applications*, **15**(1 and2), 193-199, ISSN 2452-7395.

[8] Warner, S. L. (1965). Randomised Response: A survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, **60**(309), 63-69, http://www.jstor.org/stable/2283137, DOI: 10.2307/2283137.