ESTIMATION OF SENSITIVE PROPORTION USING MOMENT GENERATING FUNCTION

Zawar Hussain[1a], Zahid Pervez[2b], Said Farooq Shah[1c] and Javid Shabbir[1d],
[1]Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan
[2]Department of Statistics, Government Sadiq Egerton College, Bahawalpur, Pakistan
Corresponding author's email: [a]zhlangah@yahoo.com, Phone: 0092-51-90642185
Emails: [b]zahidpervez74@yahoo.com, [c]farooqshahqau@gmail.com,
[d]javidshabbir@gmail.com

***ABSTRACT**: Motivated by Sukhjinder et al. [1], an estimator of population proportion under a general class of randomized response models (RRM) is proposed. The proposed estimator is based on moment generating function and is compared with maximum likelihood estimator using the mean squared error as a performance criterion. It has been observed that proposed estimator performs better than the maximum likelihood estimator when the argument constant, say t, of the moment generating function tends to zero.*

Key words: Moment generating function, randomized response models, estimation of proportion, relative efficiency, simple random sampling.

1. INTRODUCTION

Having accurate answers to surveys of social issues is important, especially when questions are related to a respondent's privacy. Respondents frequently have incentives to not tell the truth when questions touch upon moral, legal, or other sensitive issues. If we ask sensitive questions, and ignore the possibility that the respondents will not tell the truth, this will cause estimation error and bias. Therefore, methods have been developed to minimize the likelihood of error and bias.

The randomized response technique (RRT) proposed by Warner [2] is perhaps the first attempt to obtain reliable information for estimating the proportion of a sensitive attribute in a population without revealing the respondent's actual status. Furthermore, different modifications of Warner's RRT [2] were developed by various authors, such as Greenberg et al. [3], Horvitz et al. [4], Moors [5], Raghavarao [6], Mangat and Singh [7], Kuk [8], Mangat [9], Mangat and Singh [10, 11], Bhargava and Singh [12], Singh et al. [13], Gjestvang and Singh [14], Zaizai [15] and Perri [16], Hussain and Shabbir [17, 18, 19, 20], Singh et al. [21], Huang [22], Kim and Warde [23], Chang et al. [24] and many others. For a comprehensive note on the randomized response models one may refer to Chaudhuri and Mukerjee [25], Tracy and Mangat [26] and Chaudhuri [27]. In RRT, a respondent is willing to answer and tell the truth to sensitive questions through some random devices (for instance, dice, playing cards, or coins). Thus, estimation of proportion of sensitive attribute is an important issue.

An interesting method of estimating the proportion has been suggested by Sukhjinder et al. [1] where they made use of moment generating function (*mgf*) of Binomial random variable. They also indicated possibility of using this technique in Warner [2] RRT.

Motivated by Sukhjinder et al. [1], we intend to apply *mgf* technique to estimate the proportion of a sensitive attribute. Specifically, the our objective in this paper is to propose an estimator of population proportion, π, of a sensitive attribute when the data are obtained through a generic class of the randomized response model with probability of a *yes* response of the form

$$P(yes) = \lambda = c\pi + d, \quad (1.1)$$

where c and d are the constants. These constants for the different models (discussed in this paper) are given in the Table 1 below. The moment (or maximum likelihood) estimator of π for this class of models is of the form

$$\hat{\pi}_{ML} = \frac{\hat{\lambda} - d}{c}, \quad (1.2)$$

where $\hat{\lambda} = \frac{n'}{n}$ (is the moment estimator or maximum likelihood estimator of λ) and n' is the number of *yes* responses in a sample of size n.

The variance of the moment estimator in (1.2) is given by

$$Var(\hat{\pi}_{ML}) = \frac{\lambda(1-\lambda)}{nc^2} = \frac{\pi(1-\pi)}{n} + \frac{\pi(c-c^2-2cd)}{nc^2} + \frac{d(1-d)}{nc^2}. \quad (1.3)$$

Table 1: The values of c and d for different randomized response models

Model	d	c
Warner (1965)	$(1-p)$	$(2p-1)$
Greenberg et al. (1969)	$(1-p)\pi_y$	p
Mangat and Singh (1990)	$(1-T)(1-p)$	$T+(1-T)(2p-1)$
Mangat et al. (1995)	p_2	$p_1 - p_2$
Bhargava and Singh (2000)	$p_2 + p_3$	$p_1 - p_2$
Singh et al. (2003)	$p_2\pi_y$	p_1
Zaizai (2006)	$(2p-1)$	$(1-p)$
Perri (2008) Tech. I	$p_2\pi_y + p_3$	p_1
Perri(2008) Tech. II	$\pi_y(1-p_1+\theta p_3)$	$p_1+\theta p_3$

In the next section, we intend to find another estimator of λ, say $\tilde{\lambda}$, which can replace $\hat{\lambda}$ in (1.2) and gives an estimator of π, say $\tilde{\pi}$, as an efficient estimator relative to $\hat{\pi}_{ML}$.

2. PROPOSED ESTIMATOR

Consider a randomization device with probability of a *yes* answer given in (1.1). The randomization device may be one of those given in the Table 1 (or it may be even different one. Assuming a simple random sampling with replacement, let a sample of size n is drawn from the population and each sample respondent is asked to report a *yes* or *no* according to outcome of the randomization device. Again, let Z_i be the binary variable taking value 1 if the response from the i^{th} respondent is *yes* and 0 otherwise,

then $Z_i \sim Bernoulli(\lambda)$. Suppose $X = \sum_{i=1}^{n} Z_i$, then $X \sim Binomial(n, \lambda)$. The moment generating function of X is given by

$$M_X(t) = E\left(e^{tx}\right) = \left(1-\lambda+\lambda e^t\right)^n, \quad (2.1)$$

where e is the exponential.

By the method of moments, an estimator of λ is given by

$$\tilde{\lambda} = \frac{e^{\frac{tx}{n}}-1}{e^t-1}, \quad t \neq 0. \quad (2.2)$$

Using (1.1), we define a new estimator of π as

$$\tilde{\pi} = \frac{\tilde{\lambda}-d}{c}. \quad (2.3)$$

Applying the L-Hospital rule after taking the limit $t \to 0$, it can be shown that the estimator $\tilde{\pi}$ is an unbiased estimator of π otherwise it is a biased estimator with bias given by

$$Bias(\tilde{\pi}) = E(\tilde{\pi}-\pi) = \frac{\left(1-\lambda+\lambda e^{\frac{t}{n}}\right)^n - 1 - \lambda\left(e^t-1\right)}{c\left(e^t-1\right)}. \quad (2.4)$$

Now as $t \to 0$, by L-Hospital rule, we get

$$Bias(\tilde{\pi}) = 0.$$

The variance of $\tilde{\pi}$ is given by

$$Var(\tilde{\pi}) = \frac{\left(1-\lambda+\lambda e^{\frac{2t}{n}}\right)^n - \left(1-\lambda+\lambda e^{\frac{t}{n}}\right)^{2n}}{\left(e^t-1\right)^2 c^2}. \quad (2.5)$$

Using (2.4) and (2.5) in the expression $MSE(\tilde{\pi}) = Var(\tilde{\pi}) + \left[Bias(\tilde{\pi})\right]^2$, we get

$$MSE(\tilde{\pi}) = \frac{\left(1-\lambda+\lambda e^{\frac{2t}{n}}\right)^n - \left(1-\lambda+\lambda e^{\frac{t}{n}}\right)^{2n} + \left\{\left(1-\lambda+\lambda e^{\frac{t}{n}}\right)^n - \left(1+\lambda\left(e^t-1\right)\right)\right\}^2}{c^2\left(e^t-1\right)^2} \quad (2.6)$$

Application of L-Hospital rule on (2.6) gives

$$\lim_{t\to 0} MSE(\tilde{\pi}) = \frac{\lambda(1-\lambda)}{nc^2} = \frac{\pi(1-\pi)}{n} + \frac{\pi\left(c-c^2-2cd\right)}{nc^2} + \frac{d(1-d)}{nc^2}, \quad (2.7)$$

which is exactly the variance of the maximum likelihood estimator $\hat{\pi}_{ML}$. From (1.3) and (2.7), it is obvious that when $t \to 0$, $\hat{\pi}_{ML}$ and $\tilde{\pi}$ are same with equal mean squared errors. The two estimators will have different mean squared errors when $t \neq 0$. So, it may be expected to have a value of t such that

$$MSE(\tilde{\pi}) \leq Var(\hat{\pi}_{ML}). \quad (2.8)$$

To find values (or range of values) of t, we used *R* language to write a code which produces range of values of t for which condition (2.8) is satisfied and we also have computed the Percent Relative Efficiency (*PRE*) of the estimator $\tilde{\pi}$ relative to $\hat{\pi}_{ML}$ with minimum and maximum *PRE*. To compute these results we have chosen the constants c and d according to Warner [2] model, that is, we take $c = (2p-1)$ and $d = (1-p)$.

R language was used to write the codes and results are given in Tables 2-5. From the Tables 2-5, following observations were made.

(i) When $\pi < 0.5$, the range of t remains negative for $p < 0.5$ and positive for $p > 0.5$, but it is symmetric around $p = 0.5$. When $\pi > 0.5$ the range of t remains positive for $p < 0.5$ and positive for $p > 0.5$, but again it symmetric around $p = 0.5$. In case $\pi = 0.5$, the proposed estimator works efficiently for t very close to 0 and the range of the values of t as well as the *PRE* do not depend upon on c and d but do depend upon the sample size. In this case, as the sample size increases, the *PRE* increases and range of t expands away from zero on both sides.

(ii) For a given value of π the range of t squeezes to 0 and *PRE* squeezes to 100, if $|p-0.5|$ decreases.

(iii) If $|\pi-0.5|$ decreases, range of t and *PRE* squeeze to 0 and 100 respectively.

(iv) For a given p, range of t is symmetric around π but with algebraic signs changed.

(v) When $|\pi-0.5|$ and $|p-0.5|$ are smaller (tend to zero), the two estimators become almost equally efficient when n is small.

(vi) When n is small, then to have better performance for proposed estimator both $|\pi - 0.5|$ and $|p - 0.5|$ should be larger.

(vii) For large n the usual likelihood estimator become more efficient than the proposed estimator if t is not close to zero. How large should n be in order to have better performance of maximum likelihood estimator depends on the parameters p and π e.g. if $p = 0.2$ and $\pi = 0.1$, then lower bound for n is greater than 25000, but if $\pi = p = 0.1$, then n should be much larger to have better performance of maximum likelihood estimator provided t is large in absolute value.

(viii) For any p and π the maximum efficiency was observed at t very close to 0.

Table 2: Range of t and range of PRE of $\tilde{\pi}$ relative $\hat{\pi}_{ML}$.

$n = 5$			
π	(p, d, c^2)	Range of t	Range of PRE
0.1	0.2, 0.8, 0.36	$-1.50 \to -0.01$	$100.38 \to 126.28$
	0.4, 0.6, 0.04	$-0.72 \to -0.01$	$100.08 \to 102.33$
	0.6, 0.4, 0.04	$0.01 \to 0.72$	$100.08 \to 102.33$
	0.8, 0.2, 0.36	$0.01 \to 1.50$	$100.38 \to 126.28$
0.3	0.2, 0.8, 0.36	$-1.12 \to -0.01$	$100.02 \to 105.41$
	0.4, 0.6, 0.04	$-0.35 \to -0.01$	$100.04 \to 100.57$
	0.6, 0.4, 0.04	$0.01 \to 0.35$	$100.04 \to 100.57$
	0.8, 0.2, 0.36	$0.01 \to 1.12$	$100.02 \to 105.41$
0.5	0.2, 0.8, 0.36	$-0.00048 \to 0.00062$	$100.00 \to 100.0054$
	0.4, 0.6, 0.04	$-0.00048 \to 0.00062$	$100.00 \to 100.0054$
	0.6, 0.4, 0.04	$-0.00048 \to 0.00062$	$100.00 \to 100.0054$
	0.8, 0.2, 0.36	$-0.00048 \to 0.00062$	$100.00 \to 100.0054$
0.7	0.2, 0.8, 0.36	$0.01 \to 1.12$	$100.02 \to 105.41$
	0.4, 0.6, 0.04	$0.01 \to 0.35$	$100.04 \to 100.57$
	0.6, 0.4, 0.04	$-0.35 \to -0.01$	$100.04 \to 100.57$
	0.8, 0.2, 0.36	$-1.12 \to -0.01$	$100.02 \to 105.41$
0.9	0.2, 0.8, 0.36	$0.01 \to 1.50$	$100.38 \to 126.28$
	0.4, 0.6, 0.04	$0.01 \to 0.72$	$100.08 \to 102.33$
	0.6, 0.4, 0.04	$-0.72 \to -0.01$	$100.08 \to 102.33$
	0.8, 0.2, 0.36	$-1.50 \to -0.01$	$100.38 \to 126.28$

Table 3: Range of t and range of PRE of $\tilde{\pi}$ relative $\hat{\pi}_{ML}$.

$n = 15$			
π	(p, d, c^2)	Range of t	Range of RE
0.1	0.2, 0.8, 0.36	$-0.74 \to -0.01$	$100.25 \to 108.57$
	0.4, 0.6, 0.04	$-0.20 \to -0.01$	$100.27 \to 100.75$
	0.6, 0.4, 0.04	$0.01 \to 0.20$	$100.27 \to 100.75$
	0.8, 0.2, 0.36	$0.01 \to 0.74$	$100.25 \to 108.57$
0.3	0.2, 0.8, 0.36	$-0.31 \to -0.01$	$100.04 \to 101.75$
	0.4, 0.6, 0.04	$-0.09 \to -0.01$	$100.06 \to 100.18$
	0.6, 0.4, 0.04	$0.01 \to 0.09$	$100.06 \to 100.18$
	0.8, 0.2, 0.36	$0.01 \to 0.31$	$100.04 \to 101.75$
0.5	0.2, 0.8, 0.36	$-0.00059 \to 0.00073$	$100.00 \to 100.0098$
	0.4, 0.6, 0.04	$-0.00059 \to 0.00073$	$100.00 \to 100.0098$
	0.6, 0.4, 0.04	$-0.00059 \to 0.00073$	$100.00 \to 100.0098$
	0.8, 0.2, 0.36	$-0.00059 \to 0.00073$	$100.00 \to 100.0098$
0.7	0.2, 0.8, 0.36	$0.01 \to 0.31$	$100.04 \to 101.75$
	0.4, 0.6, 0.04	$0.01 \to 0.09$	$100.06 \to 100.18$
	0.6, 0.4, 0.04	$-0.09 \to -0.01$	$100.06 \to 100.18$
	0.8, 0.2, 0.36	$-0.31 \to -0.01$	$100.04 \to 101.75$
0.9	0.2, 0.8, 0.36	$0.01 \to 0.74$	$100.25 \to 108.57$
	0.4, 0.6, 0.04	$0.01 \to 0.20$	$100.27 \to 100.75$
	0.6, 0.4, 0.04	$-0.20 \to -0.01$	$100.27 \to 100.75$
	0.8, 0.2, 0.36	$-0.74 \to -0.01$	$100.25 \to 108.57$

Table 4: Range of t and range of *PRE* of $\tilde{\pi}$ relative $\hat{\pi}_{ML}$.

$n=25$			
π	$\left(p,d,c^2\right)$	Range of t	Range of *RE*
0.1	0.2, 0.8, 0.36	−0.42→−0.01	100.37→105.02
	0.4, 0.6, 0.04	−0.11→−0.01	100.06→100.44
	0.6, 0.4, 0.04	0.01→0.11	100.06→100.44
	0.8, 0.2, 0.36	0.01→0.42	100.37→105.02
0.3	0.2, 0.8, 0.36	−0.17→−0.01	100.16→101.02
	0.4, 0.6, 0.04	−0.05→−0.01	100.04→100.10
	0.6, 0.4, 0.04	0.01→0.05	100.04→100.10
	0.8, 0.2, 0.36	0.01→0.17	100.16→101.02
0.5	0.2, 0.8, 0.36	−0.00060→0.00074	100.00→100.16
	0.4, 0.6, 0.04	−0.00060→0.00074	100.00→100.16
	0.6, 0.4, 0.04	−0.00060→0.00074	100.00→100.16
	0.8, 0.2, 0.36	−0.00060→0.00074	100.00→100.16
0.7	0.2, 0.8, 0.36	0.01→0.17	100.16→101.02
	0.4, 0.6, 0.04	0.01→0.05	100.04→100.10
	0.6, 0.4, 0.04	−0.05→−0.01	100.04→100.10
	0.8, 0.2, 0.36	−0.17→−0.01	100.16→101.02
0.9	0.2, 0.8, 0.36	0.01→0.42	100.37→105.02
	0.4, 0.6, 0.04	0.01→0.11	100.06→100.44
	0.6, 0.4, 0.04	−0.11→−0.01	100.06→100.44
	0.8, 0.2, 0.36	−0.42→−0.01	100.37→105.02

Table 5: Range of t and range of *PRE* of $\tilde{\pi}$ relative $\hat{\pi}_{ML}$.

$n=100$			
π	$\left(p,d,c^2\right)$	Range of t	Range of *RE*
0.1	0.2, 0.8, 0.36	−0.10→−0.01	100.07→101.21
	0.4, 0.6, 0.04	−0.02→−0.01	100.08→100.10
	0.6, 0.4, 0.04	0.01→0.02	100.08→100.10
	0.8, 0.2, 0.36	0.01→0.10	100.07→101.21
0.3	0.2, 0.8, 0.36	−0.04→−0.01	100.03→100.24
	0.4, 0.6, 0.04	−0.01→−0.01	100.02→100.02
	0.6, 0.4, 0.04	0.01→0.01	100.04→100.10
	0.8, 0.2, 0.36	0.01→0.04	100.03→100.24
0.5	0.2, 0.8, 0.36	−0.00102→0.00102	100.00→104.83
	0.4, 0.6, 0.04	−0.00102→0.00102	100.00→104.83
	0.6, 0.4, 0.04	−0.00102→0.00102	100.00→104.83
	0.8, 0.2, 0.36	−0.00102→0.00102	100.00→104.83
0.7	0.2, 0.8, 0.36	0.01→0.04	100.03→100.24
	0.4, 0.6, 0.04	0.01→0.01	100.04→100.10
	0.6, 0.4, 0.04	−0.01→−0.01	100.02→100.02
	0.8, 0.2, 0.36	−0.04→−0.01	100.03→100.24
0.9	0.2, 0.8, 0.36	0.01→0.10	100.07→101.21
	0.4, 0.6, 0.04	0.01→0.02	100.08→100.10
	0.6, 0.4, 0.04	−0.02→−0.01	100.08→100.10
	0.8, 0.2, 0.36	−0.10→−0.01	100.07→101.21

CONCLUSIONS

Using moment generating function, a new estimator of population proportion of the characteristic of interest has been studied. On the basis of above observations, it could be recommended to use proposed estimator for estimation of the population proportion through a RRM with probability of yes given in (1.1) when it is difficult/impossible to have a large sample and the parameter p should be set such that $|p-0.5|$ is large. Both estimators $\hat{\pi}_{ML}$ and $\tilde{\pi}$ are almost equally efficient if sample is large enough. It may be recommended that for large samples t should be fixed closer to zero. Moreover, the range of values of t does not depend on the constants c and d, so any randomized response model with probability of *yes* defined in (1.1) may be used in proposed estimation technique.

ACKNOWLEDGMENT:

The First author is highly thankful to Quaid-i-Azam University for funding this work through URF projects under project # FNS/14-1299.

REFERENCES

[1]. Sukhjinder, S. S., Rajesh, T., and Sarjinder, S., On the estimation of population proportion, *Applied Mathematical Sciences*, 3 (35), 1739-1744, (2009).

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

[3]. Greenberg, B. G., Kuebler, R. R., Jr., Abernathy, J. R. and Horvitz, D. G. The unrelated question randomized response model: theoretical framework, *Journal of the American Statistical Association*, 64, 520-539, (1969).

[4]. Horvitz, D. G, Shah, B. V. and Simmons, W. R., The unrelated question randomised response model, *Proceedings of the Social Statistics Section, American Statistical Association*, 65-72, (1967).

[5]. Moors, J. J. A., Optimization of the unrelated question randomized response model, *Journal of the American Statistical Association*, 66, 627-629, (1971).

[6]. Raghavarao, D., On an estimation problem in Warner's randomized response technique, *Biometrics*, 34, 87-90, (1978).

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

[8]. Kuk, A. Y. C., Asking sensitive questions indirectly, *Biometrika*, 77, 436-438, (1990).

[9]. Mangatn N. S., An improved randomized response strategy, *Journal of the Royal Statistical Society: Series B*, 56, 93-95, (1994).

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

[11]. Mangat, N. S. and Singh, R., A note on the inverse binomial randomized response procedure, *Journal of Indian Society of Agricultural Statistics*, 47(1), 21-25, (1995).

[12]. Bhargava, M. and Singh, R. A modified randomization device for Warner model, *Statistica*, 60, 315-321, (2000).

[13]. Singh, S., Horn, S., Singh, R. and Mangat, N. S., On the use of modified randomization device for estimating the prevalence of a sensitive attribute. *Statistics in Transition*, 6(4), 515-522, (2003).

[14]. Gjestvang, C. R. and Singh, S., A new randomized response model, *Journal of the Royal Statistical Society, Series B*, 68, 523-530, (2006).

[15]. Zaizai, Y., Ratio method of estimation of population proportion using randomized response technique, *Model Assisted Statistics and Applications*, 1, 125-130, (2006).

[16]. Perri, P.F., Modified randomized devices for Simmons' model, *Model Assisted Statistics and Applications*, 3, 233-239, (2008).

[17]. Hussain, Z. and Shabbir, J., An improvement of Christofides' randomized response technique, *Journal of Probability and Statistical Science*, 6(1), 85-90, (2008).

[18]. Hussain, Z. and Shabbir, J., Three stage quantitative randomized response models, *Journal of Probability and Statistical Science*, 8(2), 227-239, (2010).

[19]. Hussain, Z. and Shabbir, J., Some Randomized Response Models, *Journal of Probability and Statistical Sciences*, 10(2), 209-218, (2012).

[20]. Hussain, Z. and Shabbir, J., New Randomized response procedures, *Journal of Sciences-Islamic Republic of Iran*, 23(4), 347-356, (2012).

[21]. Singh, S, Singh R. and Mangat, N. S., Some alternative strategies to Moor's model in randomized response sampling, *Journal of Statistical Planning and Inference*, 83, 243-255, (2000).

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

[23]. Kim, J. M. and Warde, W. D., A stratfied Warner's randomized response model, *Journal of Statistical Planning Inference*, 120, 155-165, (2004).

[24]. Chang, H., Wang, C. and Huang, K. C., On estimating the proportion of a qualitative sensitive character using randomized response sampling, *Quality and Quantity*, 38, 675-580, (2004).

[25]. Chaudhuri, A. and Mukerjee, R., *Randomized response: Theory and Methods*, Marcel- Decker, New York, (1988).

[26]. Tracy, D. and Mangat, N., Some development in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee, *Journal of Applied Statistical Sciences*, 4 533-544, (1996).

[27]. Chaudhuri, A., *Randomized Response and Indirect Questioning Techniques in Surveys*, New York. Chapman and Hall, (2011).