Repeated Trials in Randomized Response Sampling

ABSTRACT

In this article, a new estimator for estimating the proportion of sensitive attribute is introduced based on Warner (1965). The results are derived in case respondents are selected by simple random sampling with replacement (SRSWR) and without replacement (SRSWOR). The proposed estimator is unbiased in both the cases and is more efficient than the estimator proposed by and Singh and Joarder (1997) estimators.

Keywords: *Randomized response, Sensitive variable, Estimation of proportion, Relative efficiency*

AMS Subject Classification: 62D05

1. INTRODUCTION

The data relating to issues such as induced abortions, drug abuse, family income etc. that could lead to stigmatization on personality are tedious to obtain because respondents very often report untrue values or even refuse to respond. Warner (1965) suggested an ingenious method of collecting information on sensitive characters. According to the method, each interviewee in the sample is furnished with an identical randomization device. One such device could be a spinner or a deck of cards with each card having one of the following two statements: (i) "I belong to group A"; (ii) "I do not belong to group A." The statements occur with relative frequencies p and $(1-p)$, respectively, in the deck of cards. Assuming truthful reporting, the probability of getting "yes" response is

$$\lambda = \pi p + (1-\pi)(1-p)$$

where π = the true probability of A in the population,

For estimating the population proportion π possessing the sensitive character A, a simple random sample with replacement of size n is drawn from the population of size N. Each respondent in the sample is asked to select a card at random from the well-shuffled deck. The respondent answers "yes" if the outcome of the randomization device tallies with his/her actual status otherwise he/she answers "no". The maximum likelihood estimator of π given by Warner (1965) is given by

$$\hat{\pi}_w = \frac{(n_1/n)-(1-p)}{2p-1} \tag{1}$$

where n_1 is the number of "yes" responses from a sample of n individuals.

The above estimator is unbiased with variance

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

In case of simple random sample without replacement, Kim and Fleuk (1978) modified the Warner's model and shown that the estimator $\hat{\pi}_w$ is still unbiased with the following variance:

$$V(\hat{\pi}_w) = \left(\frac{N-n}{N-1}\right)\frac{\pi(1-\pi)}{n} + \frac{(1-p)p}{n(2p-1)^2} \tag{3}$$

Singh and Joarder (1997) suggested an unknown repeated trials in randomized response sampling in which if a person belongs to the sensitive group A, he/she is requested to repeat the trial if in the first

trial he/she does not get the statement according to his/her status. Assuming truthful reporting, the probability of getting "yes" response is

$$\theta_1 = \pi\left[p + p(1-p)\right] + (1-\pi)(1-p)$$

and they suggested the following estimator:

$$\hat{\pi}_s = \frac{n_1/n - (1-p)}{2p-1+p(1-p)} \tag{4}$$

where n_1 is the number of "yes" responses from a sample of n individuals selected under simple random sampling with replacement.

The estimator $\hat{\pi}_s$ is unbiased with variance:

$$V(\hat{\pi}_s) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n\left(2p-1+p(1-p)\right)^2} - \frac{\pi p(1-p)}{n\left(2p-1+p(1-p)\right)} \tag{5}$$

If the n individuals are selected under simple random sampling without replacement and n_1 is the number of "yes" responses, then the estimator proposed by Singh and Joarder (1997) still remains unbiased with variance:.

$$V(\hat{\pi}_s)_{WOR} = \left(\frac{N-n}{N-1}\right)\frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n\left(2p-1+p(1-p)\right)^2} - \frac{\pi p(1-p)}{n\left(2p-1+p(1-p)\right)} \tag{6}$$

2. PROPOSED STRATEGY

In this proposed strategy, we follow Kim and Fleuk (1978) and utilize the following notation:

$$X_i = \begin{cases} 1, & \text{if i}^{\text{th}} \text{ respondent belongs to A} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_i = \begin{cases} 1, & \text{if i}^{\text{th}} \text{ respondent selects sensitive question} \\ 0, & \text{otherwise} \end{cases}$$

$$Z_i = \begin{cases} 1, & \text{if i}^{\text{th}} \text{ respondent's answer is yes} \\ 0, & \text{otherwise} \end{cases}$$

Then, $Z_i = X_i Y_i + (1-X_i)(1-Y_i)$

We also define $X = \sum_{i=1}^{n} X_i$, $Y = \sum_{i=1}^{n} Y_i$ and $Z = \sum_{i=1}^{n} Z_i$, where X is the random variable representing the total number of respondents in the sample having sensitive characteristic, Y is the random variable representing total number of times sensitive question is selected by the n respondents, and Z is the random variable representing the total number of "yeses" reported by the n respondents.

$$V(Z) = E\left(\sum_{i=1}^{n} Z_i^2\right) + E\left(\sum_{i\neq j}^{n} Z_i Z_j\right) - \left[E\left(\sum_{i=1}^{n} Z_i\right)\right]^2 \tag{7}$$

In the proposed method, the respondent is requested to repeat the trial twice in the Warner's randomization device if in the first and second trials the respondent does not get the statement according to his status. The rest of the procedure remains the same. The repetition of the trial is known to the interviewee but remains unknown to the interviewer. Assuming completely truthful reporting by the respondents, the probability of getting "yes" response is

$$\theta = \pi\left[\, p + (1-p)\, p + (1-p)^2\, p \,\right] + (1-\pi)(1-p) \tag{8}$$

The maximum likelihood estimator of π exists for $p \neq 0.5$ and is given by

$$\hat{\pi}_G = \frac{(Z/n) - (1-p)}{2p - 1 + (1-p)\, p + (1-p)^2\, p} \tag{9}$$

Respondents are selected with SRSWR

Here X follows Binomial distribution with parameters n and π, Y follows Binomial distribution with parameters n and p, and Z follows Binomial distribution with parameters n and θ

Theorem 1: The estimator $\hat{\pi}_G$ is unbiased for the population proportion π i.e. $E(\hat{\pi}_G) = \pi$.

Proof: Taking expectation on both sides of equation (9), we get

$$E(\hat{\pi}_G) = \frac{\left(E(Z)/n\right) - (1-p)}{2p - 1 + (1-p)\, p + (1-p)^2\, p}$$

Since Z being the Binomial random variable with parameter (n, θ)

So, $E(Z) = n\theta$, $V(Z) = n\theta(1-\theta)$

Now substitute the value of θ from equation (8)

$$E(\hat{\pi}_G) = \frac{\theta - (1-p)}{2p - 1 + (1-p)\, p + (1-p)^2\, p}$$

$$= \frac{\pi\left[\, p + (1-p)\, p + (1-p)^2\, p \,\right] + (1-\pi)(1-p) - (1-p)}{2p - 1 + (1-p)\, p + (1-p)^2\, p}$$

$$= \pi$$

Hence the theorem.

Theorem 2: The variance of the estimator $\hat{\pi}_G$ is given by

$$V(\hat{\pi}_G)_{WR} = \frac{\pi(1-\pi)}{n} + \frac{(1-p)\, p}{n\left[\, 2p - 1 + (1-p)\, p + (1-p)^2\, p \,\right]^2} - \frac{\left[\, (1-p)\, p + (1-p)^2\, p \,\right]\pi}{n\left[\, 2p - 1 + (1-p)\, p + (1-p)^2\, p \,\right]}$$

Proof: $$V(\hat{\pi}_G)_{WR} = \frac{V(Z)}{n^2\left[\, 2p - 1 + (1-p)\, p + (1-p)^2\, p \,\right]^2}$$

$$V(Z) = E\left(\sum_{i=1}^{n} Z_i^2\right) + E\left(\sum_{i\neq j}^{n} Z_i Z_j\right) - \left[E\left(\sum_{i=1}^{n} Z_i\right)\right]^2$$

$$E\left(\sum_{i=1}^{n} Z_i^2\right) = E\left(\sum_{i=1}^{n} Z_i\right) = n\theta\,, \qquad E\left(\sum_{i\neq j}^{n} Z_i Z_j\right) = n(n-1)\theta^2$$

$$V(Z) = n\theta(1-\theta)$$

$$V(\hat{\pi}_G)_{WR} = \frac{n\theta(1-\theta)}{n^2\left[2p-1+(1-p)p+(1-p)^2 p\right]^2} \tag{10}$$

Substituting the value of θ from equation (8) to equation (10) and after some simplification, we get

$$V(\hat{\pi}_G)_{WR} = \frac{\pi(1-\pi)}{n} + \frac{(1-p)p}{n\left[2p-1+(1-p)p+(1-p)^2 p\right]^2} - \frac{\left[(1-p)p+(1-p)^2 p\right]\pi}{n\left[2p-1+(1-p)p+(1-p)^2 p\right]} \tag{11}$$

Hence the theorem.

Theorem 3: The estimate of the variance is given by

$$\hat{V}(\hat{\pi}_G) = \frac{\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)}{(n-1)\left[2p-1+(1-p)p+(1-p)^2 p\right]^2}$$

The proof being straightforward is omitted.

Respondents are selected with SRSWOR

Theorem 4: The estimator $\hat{\pi}_G$ is unbiased for the population proportion π i.e. $E(\hat{\pi}_G) = \pi$.

Proof: Taking expectation on both sides of equation (9), we get

$$E(\hat{\pi}_G) = \frac{\left(E(Z)/n\right) - (1-p)}{2p-1+(1-p)p+(1-p)^2 p}$$

Since $E(Z) = n\theta$

Now substitute the value of θ from equation (8)

$$E(\hat{\pi}_G) = \frac{\theta-(1-p)}{2p-1+(1-p)p+(1-p)^2 p}$$

$$= \frac{\pi\left[p+(1-p)p+(1-p)^2 p\right]+(1-\pi)(1-p)-(1-p)}{2p-1+(1-p)p+(1-p)^2 p}$$

$$= \pi$$

Hence the theorem.

Theorem 5: The variance of the estimator $\hat{\pi}_G$ is given by

$$V(\hat{\pi}_G)_{WOR} = \frac{\pi(1-\pi)}{n}\frac{(N-n)}{N-1} + \frac{(1-p)p}{n\left[2p-1+(1-p)p+(1-p)^2 p\right]^2} - \frac{\left[(1-p)p+(1-p)^2 p\right]\pi}{n\left[2p-1+(1-p)p+(1-p)^2 p\right]}$$

Proof: $$V(\hat{\pi}_G)_{WOR} = \frac{V(Z)}{n^2\left[2p-1+(1-p)p+(1-p)^2 p\right]^2} \quad (12)$$

$$V(Z) = E\left(\sum_{i=1}^{n} Z_i^2\right) + E\left(\sum_{i\neq j}^{n} Z_i Z_j\right) - \left[E\left(\sum_{i=1}^{n} Z_i\right)\right]^2$$

$$E\left(\sum_{i=1}^{n} Z_i^2\right) = E\left(\sum_{i=1}^{n} Z_i\right) = n\theta$$

$$E\left(\sum_{i\neq j}^{n} Z_i Z_j\right) = n(n-1)\left[\frac{\pi(N\pi-1)}{N-1}(2p-1)^2 + 2\pi(2p-1)(1-p) + (1-p)^2\right]$$

$$V(Z) = n\theta + n(n-1)\left[\frac{\pi(N\pi-1)}{N-1}(2p-1)^2 + 2\pi(2p-1)(1-p) + (1-p)^2\right] - n^2\theta^2$$

Substituting the value of $V(Z)$ in equation (11) and after some simplification, we get

$$V(\hat{\pi}_G)_{WOR} = \frac{\pi(1-\pi)}{n}\frac{(N-n)}{N-1} + \frac{(1-p)p}{n\left[2p-1+(1-p)p+(1-p)^2 p\right]^2} - \frac{\left[(1-p)p+(1-p)^2 p\right]\pi}{n\left[2p-1+(1-p)p+(1-p)^2 p\right]} \quad (13)$$

Hence the theorem.

3. EFFICIENCY COMPARISON

The percentage relative efficiency (PRE) of the proposed strategy with respect to the Singh and Joarder (1997) in case of simple random sampling with replacement is given by

$$E = \frac{V(\hat{\pi}_s)}{V(\hat{\pi}_G)_{WR}} X100$$

Table 1. Percentage relative efficiency of the proposed estimator over Singh and Joarder (1997) in SRSWR

	p			
	0.6	0.7	0.8	0.9
π	PRE	PRE	PRE	PRE
0.1	147.81	120.83	107.83	101.66

π	p			
	0.6	0.7	0.8	0.9
	PRE	PRE	PRE	PRE
0.2	148.34	120.83	107.72	101.63
0.3	150.01	121.58	108.03	101.72
0.4	152.98	123.14	108.71	101.91
0.5	157.65	125.72	109.87	102.21
0.6	164.80	129.88	111.75	102.69
0.7	176.05	136.87	114.98	103.51
0.8	195.05	150.09	121.46	105.16
0.9	232.22	182.71	139.91	110.13

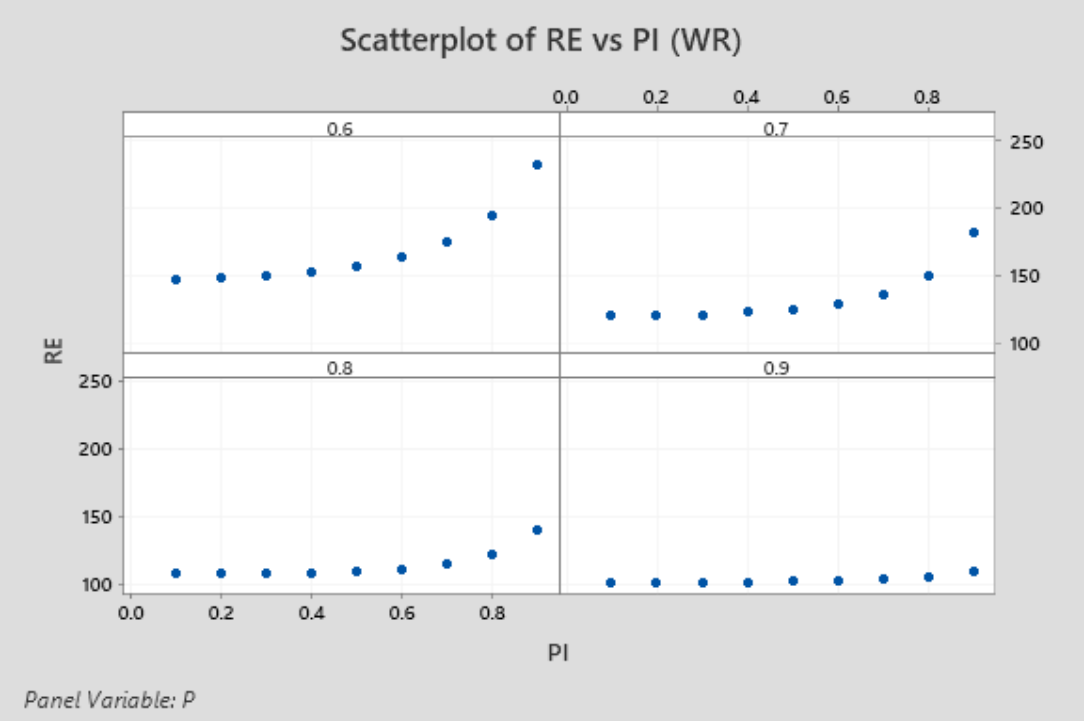

Fig. 1: Scatterplot of relative efficiency of the proposed estimator for different values of π in SRSWR

The percentage relative efficiency (PRE) of the proposed strategy with respect to the Singh and Joarder (1997) in case of simple random sampling without replacement is given by

$$E = \frac{V(\hat{\pi}_s)_{WOR}}{V(\hat{\pi}_G)_{WOR}} X100$$

Table 2. Percentage relative efficiency of the proposed estimator over Singh and Joarder (1997) in SRSWOR

π	P			
	0.6 PRE	0.7 PRE	0.8 PRE	0.9 PRE
0.1	153.38	125.26	110.88	103.15
0.2	159.24	129.54	113.70	104.55
0.3	166.23	134.84	117.28	106.34
0.4	174.72	141.56	121.94	108.71
0.5	185.26	150.38	128.30	112.01
0.6	198.67	162.45	137.47	116.92
0.7	216.32	179.96	151.84	124.98
0.8	240.60	207.68	177.61	140.68
0.9	276.10	258.23	237.26	184.60

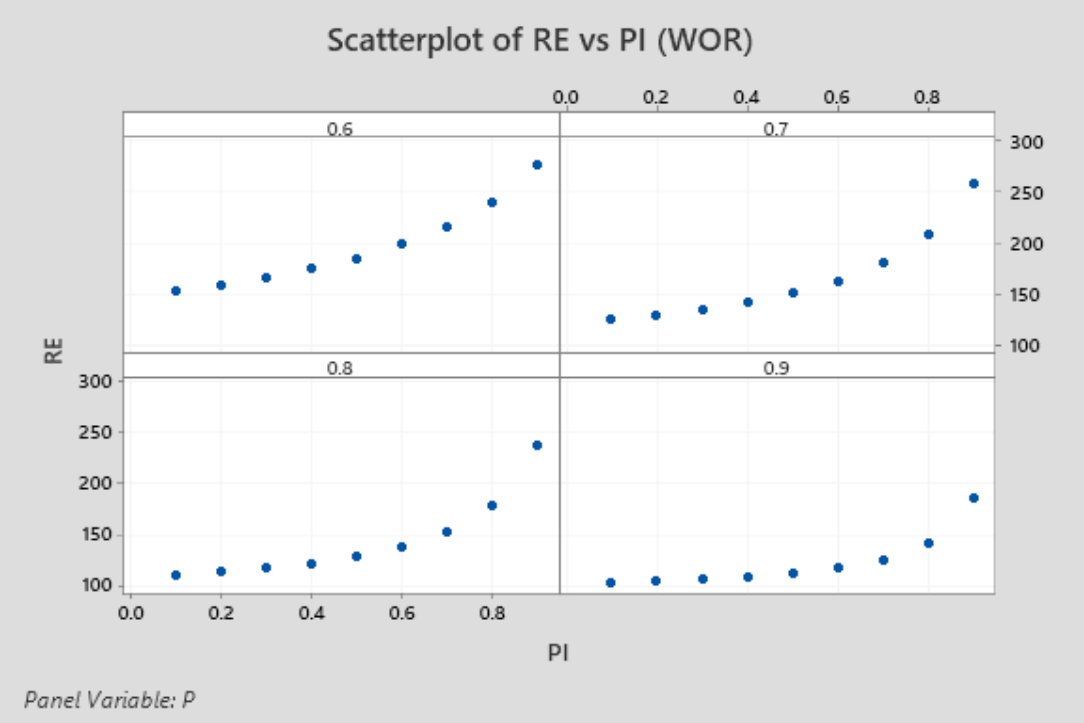

Fig. 2: Scatterplot of relative efficiency of the proposed estimator for different values of π in SRSWOR

4. CONCLUSION

The results in Table 1 and Table 2 shows that the proposed strategy can bring substantial gain in efficiency up to 232.22% in case of simple random sampling with replacement and up to 276.10% in case of simple random sampling without replacement over the estimator proposed by Singh and Joarder (1997).

Remarks: The propose strategy can be easily generalized by asking the respondent to use the randomization device until he/she gets the statement according to his/her status. But in that case the respondent may get irritated and may refuse to respond.

REFERENCES

1. Kim J. and Fleuk J.A. (1978). Modifications of the randomized response technique for sampling without replacement. *Amer. Stat. Assoc*. 346-350.
2. Singh, Sarjinder and Joarder, A.H. (1997). Unknown repeated trials in randomized response sampling. *J. of Indian Soc. of Agri. Stat*., 50(1), 103-105
3. Warener, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. of Amer. Stat. Assoc*. 60, 63-69.