

**BAYESIAN ESTIMATION OF POPULATION PROPORTION OF A  
SENSITIVE CHARACTERISTIC USING SIMPLE BETA PRIOR**

**Zawar Hussain and Javid Shabbir**

Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan  
Email: zhlangah@yahoo.com and jsqau@yahoo.com

**ABSTRACT**

In this study, we have developed the Bayes estimator of the population proportion of a sensitive characteristic when data are obtained through the randomized response technique (RRT) proposed by Hussain and Shabbir (2007). Using simple Beta prior information, superiority of the Bayes estimators is established for a wide range of the values of the population proportion. We observed that Bayes estimators are better than the Maximum Likelihood Estimator (*MLE*) and Kim et al. (2006) estimator. For small as well as moderate samples, it has been observed that Bayes estimators outperform the *MLE* and Kim et al. (2006) estimator in case of using RRT by Hussain and Shabbir (2007).

**KEY WORDS**

Bayesian estimation, randomized response technique, mixed prior information,

**1. INTRODUCTION**

Asking directly about a stigmatized attribute (induced abortion, drug usage, tax evasion, etc.) in a human population survey is a fiddly business. A surveyor may receive untruthful answers from the survey respondents when he/she uses direct questioning approach. Because of many reasons, information about incidence of stigmatized attributes, in the population, is necessary. For the first time, Warner (1965) proposed a nifty method of survey to collect information in relation to stigmatized attributes by providing privacy and anonymity to the respondents. To date, a large number of developments and variants of Warner's Randomized Response model (RRM) have been put forward by several researchers. Greenberg et al. (1969), Mangat and Singh (1990), Mangat (1994), Singh et al. (1998), Christofides (2003), Kim and Warde (2004) are some of the many to be cited. The interested readers may be asked to see Chaudhuri and Mukerjee (1988) and Tracy and Mangat (1996). In some situations, however, prior information about the unknown parameter is available and can be used along with the sample information for estimation of that unknown parameter called the Bayesian approach of estimation. Efforts made by researchers on Bayesian analysis of Randomized response models are not very enormous, nonetheless, attempts have been made on the Bayesian analysis of Randomized response techniques. Winkler and Franklin (1979), Pitz (1980), Spurrier and Padgett (1980), O'Hagan (1987), Oh (1994), Migon and Tachibana (1997), Unnikrishnan and Kunte (1999), Bar-Lev and Bobovich (2003), Barabesi and Marcheselli (2006), and Kim et al. (2006) are the major references on the Bayesian

analysis of the RRTs. The arrangement of the paper is as follows. In Section 2, we present the Warner (1965) and Hussain and Shabbir (2007) RRTs followed by Bayesian estimation of population proportion using Hussain and Shabbir (2007) RRT in Section 3. Section 4 contains the conclusion.

## 2. WARNER'S RANDOMIZED RESPONSE TECHNIQUE

Warner (1965) suggested the initiative of RRT. The basic thought is to develop a random rapport between the individual's response and the sensitive question. Warner's design consists of two kind of questions,  $A$  and  $A^c$ , to be answered on probability basis, where  $A$  is "do you possess stigmatized attribute", and  $A^c$  is "do you not possess stigmatized attribute". The two questions  $A$  and  $A^c$  are presented to respondents with preset probabilities  $P$  and  $1-P$ , respectively. The simple random sampling with replacement (SRSWR) is assumed. The  $i^{th}$  selected respondent is asked to select a question  $A$  or  $A^c$  and report *yes* if his/her actual status matches with selected question and *no* otherwise.

The probability of 'yes' for a particular respondent is then:

$$P(\text{yes}) = \theta = P\pi + (1-P)(1-\pi) \quad (2.1)$$

where  $P$  is the probability of selecting question  $A$ .

The maximum likelihood estimator of  $\pi$  is:

$$\hat{\pi} = \frac{\hat{\theta} - (1-P)}{2P-1}, \quad (2.2)$$

where  $\hat{\theta} = \frac{n'}{n}$  and  $n'$  is the number of *yes* responses in the sample of  $n$ . Moreover,

$$\text{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \quad (2.3)$$

### 2.1 Hussain and Shabbir Technique

Hussain and Shabbir (2007) proposed a RRT based on the random use of one of the two randomization devices  $R_1$  and  $R_2$ . In design, the two randomization devices  $R_1$  and  $R_2$  are same as that of Warner's (1965) device but with different probabilities of selecting the sensitive question. The idea behind this suggestion is to decrease the suspicion among the respondents by providing them choice to randomly choose the randomization device itself. As a result, respondents may divulge their true status. A simple random sample with replacement (SRSWR) sampling is assumed to select a sample of size  $n$ . Let  $\alpha$  and  $\beta$  be any two positive real numbers chosen such that

$q = \frac{\alpha}{\alpha+\beta}$ , ( $\alpha \neq \beta$ ) is the probability of using  $R_1$ , where  $R_1$  consists of the two

statements of Warner's device with preset probabilities  $P_1$  and  $1-P_1$  respectively and  $1-q = \frac{\beta}{\alpha+\beta}$  is the probability of using  $R_2$  consisting of the same two statements of Warner's device but with preset probabilities  $P_2$  and  $1-P_2$  respectively. For the  $i^{th}$  respondent, the probability of a *yes* response is given by

$$P(\text{yes}) = \phi = \frac{\alpha}{\alpha+\beta} \{P_1\pi + (1-P_1)(1-\pi)\} + \frac{\beta}{\alpha+\beta} \{P_2\pi + (1-P_2)(1-\pi)\}. \quad (2.1.1)$$

To provide the equal privacy protection in both the randomization devices  $R_1$  and  $R_2$ , it is suggested to set  $P_1 = 1 - P_2$ . With this setting (2.1.1) reduces to

$$\phi = \frac{\pi \{(2P_1 - 1)(\alpha - \beta)\} + P_1\beta + P_2\alpha}{(\alpha + \beta)} \quad (2.1.2)$$

hence

$$\pi = \frac{\phi(\alpha + \beta) - P_1\beta - P_2\alpha}{(2P_1 - 1)(\alpha - \beta)}, \quad P_1 \neq \frac{1}{2}, \quad \alpha \neq \beta. \quad (2.1.3)$$

This suggest defining an unbiased moment estimator of  $\pi$  as

$$\hat{\pi} = \frac{\hat{\phi}(\alpha + \beta) - P_1\beta - P_2\alpha}{(2P_1 - 1)(\alpha - \beta)}, \quad (2.1.4)$$

where  $\hat{\phi} = \frac{n'_1}{n}$  and  $n'_1$  is the number of respondents reporting a *yes* answer.

When  $P_1 = 1 - P_2$ , the variance of the proposed estimator is given by

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{(P_2\alpha + P_1\beta)(P_1\alpha + P_2\beta)}{n(2P_1 - 1)^2 (\alpha - \beta)^2 (\alpha + \beta)^2}. \quad (2.1.5)$$

### 3. BAYESIAN ESTIMATION OF $\pi$ USING HUSSAIN AND SHABBIR (2007) RRT

To go on for Bayesian estimation based on the data gathered by this technique, we assume a Beta prior distribution with parameters  $a$  and  $b$ , for the parameter to be estimated. That is prior distribution of  $\pi$  is given by

$$f(\pi) = \frac{1}{\beta(a,b)} \pi^{a-1} (1-\pi)^{b-1}, \quad 0 < \pi < 1, \text{ and } a, b > 0.$$

Let  $X = \sum_{i=1}^n x_i$  be the total number of the *yes* responses in a sample of size  $n$  drawn from the population with SRSWR sampling. Here  $x_i = 1$  with probability  $\phi$  and  $x_i = 0$  with probability  $1 - \phi$ , where  $\phi$  is defined as in (2.1.2). Then the conditional distribution of  $X$  given  $\pi$  is

$$\begin{aligned} f_{X/\pi}(x/\pi) &= \frac{n!}{x!(n-x)!} \phi^x (1-\phi)^{n-x} \\ &= \frac{n!}{x!(n-x)!} \left[ \frac{\pi \{(2P_1 - 1)(\alpha - \beta)\} + P_1\beta + P_2\alpha}{(\alpha + \beta)} \right]^x \left( 1 - \frac{\pi \{(2P_1 - 1)(\alpha - \beta)\} + P_1\beta + P_2\alpha}{(\alpha + \beta)} \right)^{n-x} \end{aligned}$$

which on simplification reduces to

$$\begin{aligned} &= \frac{n!}{x!(n-x)!} \left\{ \frac{(2P_1 - 1)(\alpha - \beta)}{\alpha + \beta} \right\}^x (\pi + f)^x (1 - \pi + h)^{n-x} \\ &= \frac{n!}{x!(n-x)!} \left\{ \frac{(2P_1 - 1)(\alpha - \beta)}{\alpha + \beta} \right\}^x \sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \pi^i (1-\pi)^j \end{aligned}$$

for  $x = 0, 1, 2, \dots, n_1$ , where  $f = \frac{P_1\beta + P_2\alpha}{(2P_1 - 1)(\alpha - \beta)}$  and  $h = \frac{3P_1(\beta - \alpha) + 3\alpha}{(2P_1 - 1)(\alpha - \beta)}$ . Thus the joint distribution of  $X$  and  $\pi$  is given by

$$\begin{aligned} f(X, \pi) &= \frac{1}{\beta(a, b)} \pi^{a-1} (1-\pi)^{b-1} \frac{n!}{x!(n-x)!} \left\{ \frac{(2P_1 - 1)(\alpha - \beta)}{\alpha + \beta} \right\}^x \\ &\quad \sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \pi^i (1-\pi)^j \end{aligned}$$

or

$$\begin{aligned} f(X, \pi) &= \frac{1}{\beta(a, b)} \frac{n!}{x!(n-x)!} \left\{ \frac{(2P_1 - 1)(\alpha - \beta)}{\alpha + \beta} \right\}^x \\ &\quad \sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \pi^{a+i-1} (1-\pi)^{b+j-1}. \end{aligned}$$

Now the marginal distribution of  $X$  can be obtained by integrating the joint distribution of  $X$  and  $\pi$  over  $\pi$ . Thus the marginal distribution of  $X$  is given by

$$f(X) = \frac{1}{\beta(a, b)} \frac{n!}{x!(n-x)!} \left\{ \frac{(2P_1 - 1)(\alpha - \beta)}{\alpha + \beta} \right\}^x$$

$$\sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \beta(a+i, b+j).$$

As we know that the posterior distribution of  $\pi$  given  $X$  is defined as

$$f_{\pi|X}(\pi|x) = \frac{f(\pi, X)}{f(X)}.$$

Thus the posterior distribution of  $\pi$  given  $X$  may be obtained as

$$f_{\rho/X}(\pi/x) = \frac{\sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \pi^{a+i-1} (1-\pi)^{b+j-1}}{\sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \beta(a+i, b+j)} I(0 < \pi < 1).$$

Under the squared error loss function the Bayes estimator of  $\pi$  is given by

$$\hat{\pi}_{Bayes} = \frac{\sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \beta(a+i+1, b+j)}{\sum_{i=0}^x \sum_{j=0}^{n-x} \frac{n-x!}{j!(n-x-j)!} \frac{x!}{i!(x-i)!} f^{x-i} h^{n-x-j} \beta(a+i, b+j)}. \quad (3.1)$$

Chaubey and Li (1995), and Kim et al. (2006) used classical approach to compare the Bayesian and classical estimators and did not use a loss function. In this paper the same approach of comparison is used. The mean squared errors (*MSEs*) of both the Bayesian and classical estimators are defined for a fixed value of  $\pi$  and are written as;

$$MSE(\hat{\pi}_{Bayes}) = E(\hat{\pi}_{Bayes} - \pi)^2 = \sum_{x=0}^n (\hat{\pi}_{Bayes} - \pi)^2 \phi^x (1-\phi)^{n-x}. \quad (3.2)$$

and

$$MSE(\hat{\pi}) = E(\hat{\pi} - \pi)^2 = \sum_{x=0}^n (\hat{\pi} - \pi)^2 \phi^x (1-\phi)^{n-x}. \quad (3.3)$$

In Figures 3.1-3.2, dotted line represents the *MSE* of the *MLE* and full line depicts the *MSE* of the Bayes estimator. We observe that Bayes estimator performs well compared to the usual *MLE* when the data are gathered through Hussain and Shabbir (2007) RRT. It is evident from the graphs that when the absolute difference between the design probabilities increases the relative efficiency of the Bayes estimator decreases for smaller as well as larger samples. This suggests setting the design probabilities in both the randomization devices almost equal.

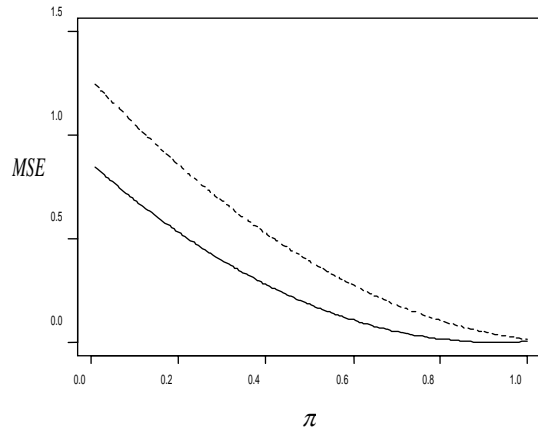


Fig. 3.1:  $MSE$  of  $\hat{\pi}_{Bayes}$ , and  $\hat{\pi}$  for  $n = 25$ ,  $P_1 = 0.9$ ,  $\alpha = 1$ ,  $\beta = 10$  and  $\pi$  ranges from 0 to 1.

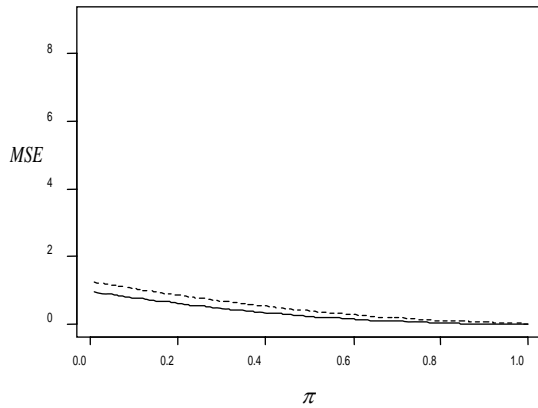


Fig. 3.2:  $MSE$  of  $\hat{\pi}_{Bayes}$ , and  $\hat{\pi}$  for  $n = 100$ ,  $P_1 = 0.9$ ,  $\alpha = 1$ ,  $\beta = 10$  and  $\pi$  ranges from 0 to 1.

Bar-Lev and Bobovich (2003) suggested a common conjugate prior structure for Warner (1965), Horvitz et al. (1967), Greenberg et al. (1969) and Mangat and Singh (1990) RRTs. They proposed to use truncated Beta prior distribution. The proposed Bayes estimator is not directly comparable with the estimators proposed by Bar-Lev and Bobovich (2003) estimators because they used the conjugate prior distributions. Using the Mangat (1994) RRT, Kim et al. (2006) proposed a Bayes estimator which can be compared with proposed Bayes estimator. Kim et al. (2006) estimator  $\hat{\pi}_{BK}$  is given by

$$\hat{\pi}_{BK} = \frac{\sum_{j=0}^x \frac{x!}{j!(x-j)!} d^{x-j} \beta(a+j+1, n+b-x)}{\sum_{j=0}^x \frac{x!}{j!(x-j)!} d^{x-j} \beta(a+j, n+b-x)}, \quad d = \frac{1-P_1}{P_1}. \quad (3.4)$$

The *MSE* the Kim et al. (2006) estimator  $\hat{\pi}_{BK}$  is given by

$$MSE(\hat{\pi}_{BK}) = E(\hat{\pi}_{BK} - \pi)^2 = \sum_{x=0}^n (\hat{\pi}_{BK} - \pi)^2 \phi^x (1-\phi)^{n-x} . \tag{3.5}$$

Behavior of *MSEs* of proposed Bayes estimator the Kim et al. (2006) estimator are given in the figures 3.3-3.4. In Figures 3.3-3.4, full line represents the *MSE* of proposed Bayes estimator and dotted line represents the *MSE* of the Kim et al. (2006) estimator. We observe that proposed Bayes estimator performs well compared to Kim et al. (2006) estimator when the data are gathered through our proposed RRT based on the randomized use of the Warner’s randomization device. We observed that when sample sizes increases the efficiency of the proposed Bayes estimator decreases. This guides to suggest that the proposed Bayes estimator should preferably be used in stratification. It is also observed that when the true proportion of the population is greater than 0.5, the proposed Bayes estimator is comparatively less efficient.

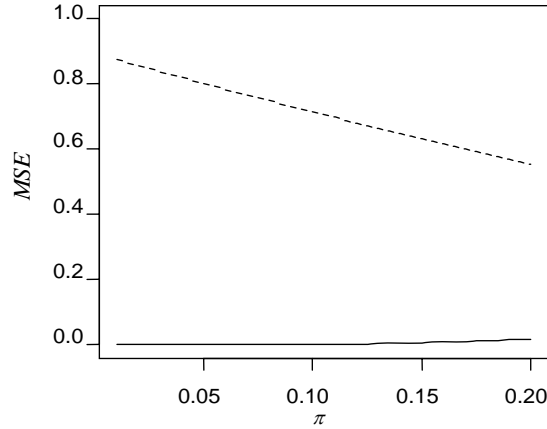


Fig. 3.3: *MSE* of  $\hat{\pi}_{Bayes}$  and to  $\hat{\pi}_{BK}$  for  $n = 50, a = 1, b = 2, \alpha = 1, \beta = 10, P_1 = P = 0.6$ .

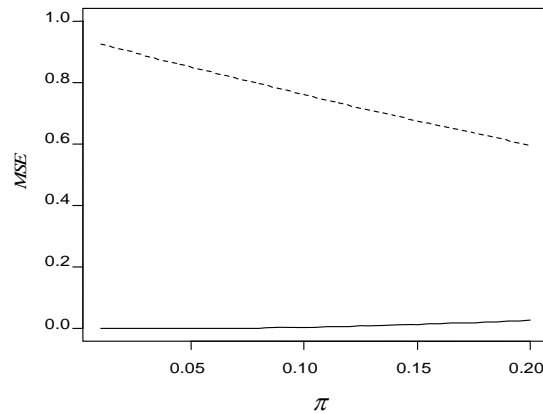


Fig. 3.4: *MSE* of  $\hat{\pi}_{Bayes}$  and  $\hat{\pi}_{BK}$  for  $n = 100, a = 1, b = 2, \alpha = 1, \beta = 10, P_1 = P = 0.6$ .

So, when it is anticipated that there would be a small proportion of individuals possessing characteristic of interest, we recommend using the proposed Bayes estimator.

#### 4. CONCLUSION

To sum up, in this study, we have presented the development of Bayesian estimation technique for Hussain and Shabbir (2007) RRT. Using the "R" software, we wrote program to study the performance of the Bayes estimators versus usual *MLE* and Kim et al. (2006) estimator. With these codes on hand, for many values of the design probabilities to study the behavior of *MLE* of the estimators: Bayes, *MLE* and Kim et al. (2006) estimator. It has been observed that in case of data obtained by the Hussain and Shabbir (2007) technique, the proposed Bayes estimators outperform the usual *MLE* and Kim et al. (2006) estimator for any value of the design probability  $P_1$  and for small samples. Bayes estimators continue to perform better, over the whole range of  $\pi$  when sample size is large. We tried each value of  $P_1$  for large samples and observed that Bayes estimator is more efficient than the *MLE*. For comparison with Kim et al. (2006) estimator, we observed that proposed estimator is not efficient when  $\pi$  is moderately large ( $> 0.45$ ). But we presented the graphs for some selected values of different parameters. It is noted that the relative efficiency of the Bayes estimators decreases when sample size increases. This suggests studying Bayesian estimation when stratification of the population is suitable. We can conclude from this study that when additional information about the unknown parameter is available, we should go for Bayes estimation in order to have the more precise estimators. Although the Bayes estimators and *MLE* become equally efficient when sample size increases indefinitely, but even for larger samples, we may use Bayes estimation due to their confinement in the range starting from 0 to 1. On contrary, the *MLE* can go beyond this interval when the number of yes responses is very low or very high. So we can recommend using Bayes estimators for randomized responses in sensitive surveys.

#### REFERENCES

1. Barabesi, L. and Marcheselli, M. (2006). A practical implementation and Bayesian estimation in Franklin's randomized response procedure. *Comm. Statist-Copmut. Simul.*, 35, 563-573.
2. Bar-Lev, S.K. Bobovich, E. and Boukai, B. (2003). A common conjugate prior structure for several randomized response models. *Test*, 12(1), 101-113
3. Chaudhuri, A. and Mukerjee, R. (1988). Randomized response: Theory and Methods. *Marcel- Decker*, New York.
4. Christofides, T.C. (2003). A generalized randomized response technique. *Metrika*, 57, 195-200.
5. Chaubey, Y. and Li, W. (1995). Comparison between maximum likelihood and Bayes methods of estimation for binomial probability with sample compositing. *J. Official Statist*, 11, 379-390.
6. Greenberg, B., Abul-Ela, A., Simmons, W. and Horvitz, D. (1969). The unrelated question randomized response: theoretical framework. *J. Amer. Statist. Assoc*, 64, 529-539.

7. Hussain, Z. and Shabbir, J. (2007). Randomized use of Warner's randomized response model. *InterStat*: April # 7. <http://interstat.statjournals.net/INDEX/Apr07.html>
8. Kim, J.M. and Warde, D.W. (2004). A stratified Warner's randomized response model. *J. Statist. Plann. Inference*, 120(1-2), 155-165.
9. Kim, J.M., Tebbs, J.M. and An, S.W. (2006). Extensions of Mangat's randomized response model. *J. Statist. Plann. Inference*, 36(4), 1554-1567.
10. Mangat, N.S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.
11. Mangat, N.S. (1994). An improved randomized response strategy. *J. Roy. Statist. Soc. Ser. B*, 56(1), 93-95.
12. Migon, H. and Tachibana, V. (1997). Bayesian approximations in randomized response models. *Comput. Statist. Data Anal.*, 24, 401-409.
13. O'Hagan, A. (1987). Bayes linear estimators for randomized response models. *J. Amer. Statist. Assoc.*, 82, 580-585.
14. Oh, M. (1994). Bayesian analysis of randomized response models: a Gibbs sampling approach. *J. Korean. Statist. Soc.*, 23, 463-482.
15. Pitz, G. (1980). Bayesian analysis of randomized response models. *J. Psychological Bull*, 87, 209-212.
16. Spurrier, J. and Padgett, W. (1980). The application of Bayesian techniques in randomized response. *Sociological Methodol*, 11, 533-544.
17. Singh, S., Horn, S. and Chowdhuri, S. (1998). Estimation of stigmatized characteristics of a hidden gang in finite population. *Austral. & New Zealand J. Statist*, 40(3), 291-297.
18. Tracy, D. and Mangat, N. (1996). Some development in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee. *J. Applied. Statist. Sci.*, 4, 533-544.
19. Unnikrishnan, N., Kunte, S. (1999). Bayesian analysis for randomized response models. *Sankhya*, Ser. B, 61, 422-432.
20. Winkler, R. and Franklin, L. (1979). Warner's randomized response model: A Bayesian approach. *J. Amer. Statist. Assoc.*, 74, 207-214.
21. Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, 60, 63-69.