A NEW QUASI EMPIRICAL BAYES ESTIMATE IN RANDOMIZED RESPONSE SAMPLING

Augustus Jayaraj, MD
Cornell University
Ithaca, NY 14850

Oluseun Odumade
Enterprise Risk Services
Deloitte & Touche LLP
Minneapolis, MN

Sarjinder Singh
Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363

ABSTRACT

In this paper, a new notion of "quasi empirical" Bayes estimation is developed for estimating the proportion of a sensitive attribute in a population by making use of both a prior distribution of prevalence of the sensitive attribute in addition to the known prior distribution of an unrelated characteristic. The proposed quasi empirical Bayes estimate is compared with those of the unrelated question model due to Greenberg et al. (1969) by means of a simulation study. A quasi Cramer-Rao lower bound of variance is also suggested and compared to the variance of the Greenberg el al. (1969) estimator. Simulated situations are reported where the proposed lower bound of variance remains below the variance of the Greenberg et al (1969) estimator.

Key words: Randomized response sampling, estimation of sensitive characteristics, Bayes estimation, Cramer-Rao lower bound of variance, Simulation study.

1. INTRODUCTION

This section briefly reviews the literature available on randomized response techniques. Certain kinds of data are not available in research archives. For example, the true incidence of induced abortion in the United States is unknown because the vital statistic registration laws in most states do not require notification of fatal loss prior to the twentieth week of pregnancy. Induced abortions are generally performed prior to that time. Most of them are induced for non-medical reasons without legal sanction and, therefore, are not reported. Even if available, the data may be inaccessible to the investigator due to privacy and security reasons. Another instance is that the study population is geographically so dispersed, that use of available data may not be feasible. Further, when one has to make an inference about a large lot, it may not be practical to examine each and every individual from the lot. On account of these and a host of other reasons, the sample survey method for collecting information is generally preferred.

It is a fact that surveys of human populations receive good responses on non-sensitive questions than on personal, sensitive questions. The reason for this phenomenon include involving controversial assertions, stigmatizing and/or incriminating matters, for reasons of modesty, fear of being thought bigoted, or merely a reluctance to confide secretly to a stranger. Direct queries about sensitive questions often yield non-response or at most evasive and/or false response. Thus, estimation of relevant parameters at an intended level of accuracy becomes a problem.

Estimates involving sensitive topics are subject to two main sources of error: (I) Sampling error, and (II) Non-sampling error. The sampling error arises from the variation present because only a part of the population is studied instead of a complete enumeration of the population. Using appropriate estimation procedures can reduce sampling error. The second type of error, non-sampling error, can be of two types. One type is random error which reduces the reliability of measurements. Random error can be expected to cancel out over repeated measurements. The other type of non-sampling error is systematic error and is present because of two major sources. One source is refusal to respond, called “non-response bias”. It reduces sample size and thus variance of the estimate becomes greater. The other source of non-sampling error is response bias arising from purposeful falsification of answers. Unlike random error, the response and non-response biases do not cancel out over repeated measurements. Rather, the distortion between a respondent’s true and observed score persists. Philip (1971) has shown that response bias is likely to arise even in surveys relating to innocuous information, owing in part to a respondent’s perceptions and needs that can emerge during the data collection process. Thus, potential for bias to arise in surveys of sensitive information can be considered troublesome due to respondent’s concern over anonymity. It can lead to erroneous estimates of the sensitive phenomena under study.

Both non-response and response bias are serious concerns to survey statisticians. A review of survey research on the sensitive topic – ‘criminal behavior’ - indicates that few workers have acknowledged the problem of bias and even fewer have employed appropriate techniques to estimate the magnitude of bias and its effect on findings (Bridges, 1979). Borch and Cecil (1979) have argued that the essential problem concerns adherence to ethical standards without, at the same time, drastically constraining the research process. The privacy of respondents must be protected; it is the responsibility of the investigator. However, they further argue that the researcher may find it impossible to abide by both the promise of confidentiality made to a respondent and the demands of a court, legislative committee or executive agency for information about the respondents. The conclusion, therefore, necessitates some means to ensure confidentiality of the respondent and to intact the investigator from outside interference.

It was the theory of Warner (1965) that cooperation should be better when the information is requested from the informant in anonymity on probability basis rather than by using direct question. It makes a stochastic relationship between the question and the individual’s response and thus provides protection and confidentiality of the respondent. This increases both the willingness to respond and truthfulness of reporting, and hence reduces, if not eliminates, the evasive answer bias. He called the procedure “Randomized

Response (RR)" because the respondent selects a question on a probability basis from two or more questions without revealing to the interviewer which question has been chosen. Respondent's response is not fixed as in the case of direct or open surveys and can thus give various responses with different probabilities.

To illustrate hypothetically, every person appearing in a sample is asked to select question(s) unobserved by the interviewer from two or more questions by using a chance device (such as a box containing beads of two or more different colors, a deck consisting of two or more types of cards, a spinner having two or more types of outcomes, random number generator, a coin, etc.). The respondent is then required to answer the selected question in terms of 'yes' or 'no' without revealing to the interviewer which of the alternative questions is being answered. This way, the privacy of each respondent is protected and embarrassment to the respondent is thus removed. The information obtained from a sample of respondents is sufficient with the knowledge of the probability distribution generated by a random device to compute estimates of the required population proportion. The accuracy of the responses to RR queries depends, of course, on the extent to which the respondents perceive their individual privacy is protected. The Warner's model (say, W-model) considers the case where a proportion π of the population possesses a sensitive characteristic A (say, marijuana consumption) whereas the respondents $(1-\pi)$ of the population did not possess this characteristic. The randomizing device for eliciting information used by him/her consists of providing a spinner or some other suitable device with two mutually exclusive outcomes:

$$\text{(i)}\ Q_1\text{: I belong to group A,}\ \text{(ii)}\ Q_2\text{: I belong to group }\bar{A}$$

For every person appearing in a simple random sample (SRS) taken with replacement.

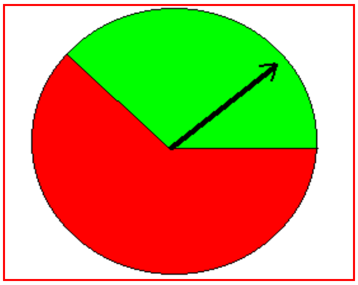

Fig. 2.1. *Spinner as a device*

The statements (i) and (ii) in the random device are represented with probability p and $(1-p)$, respectively. The respondent spins the spinner unobserved by the interviewer and answers 'yes' if he possesses the characteristic indicated by the pointer , and 'no' if otherwise. The interviewee does not report the group to which the spinner points and hence the interviewer does not known to which statement the reply refers to. The interviewer assumes that these 'yes' and 'no' answers are reported truthfully. Suppose n' persons in the sample answered 'yes' and $n-n'$ answered 'no'. Then an unbiased estimator $\hat{\theta}$ of the probability of 'yes' answer, θ, is given by n'/n. Also the true probability of 'yes' answer, θ, is given by:

$$\theta = p\pi + (1-p)(1-\pi) \tag{1.1}$$

If p is known, the maximum likelihood estimator (MLE) of π is given by:

$$\hat{\pi}_w = \frac{\hat{\theta}-(1-p)}{2p-1}, \qquad p \neq 0.5 \tag{1.2}$$

Since $\hat{\theta}$ follows a binomial distribution $B(n,\theta)$ with parameters n and θ, the estimator $\hat{\pi}_w$ is unbiased for π. The variance of the estimator $\hat{\pi}_w$ is given by:

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{1.3}$$

Clearly, the first part in (1.3) is the usual binomial variance associated with a direct question and truthful replies by all the respondents. The second part is the additional variance due to the random device.

The confidence of the respondents provided by randomized response might be further enhanced if one of the two questions is referred to a non-stigmatized attribute, say Y. Horvitz *et al.* (1967) developed an unrelated question randomized response model. While developing the theory for the unrelated question model, Greenberg *et al.* (1969) dealt with two situations when π_y, the proportion of people possessing the non-sensitive attribute Y is known or unknown. When π_y is known, each respondent selected in the sample of n individuals, by using simple random with replacement sampling, is provided with a random device, say R consisting of two possible outcomes: "(i) I am a member of group A", and "(ii) I am a member of group Y" , represented with probabilities P and $(1-P)$ respectively. The respondent randomly selects one of these two questions unobserved by the interviewer and reports "yes" or "no" according to the statement and actual status he possesses. If π is the true proportion of the sensitive group in the population, the probability of 'yes' answer is given by:

$$\theta_G = P\pi + (1-P)\pi_y \tag{1.4}$$

Then an unbiased estimator of π proposed by Greenberg et al. (1969) is given by:

$$\hat{\pi}_G = \frac{\hat{\theta}_G - (1-P)\pi_y}{P} \tag{1.5}$$

Where $\hat{\theta}_G = x/n$ is the observed proportion of 'yes' answers in the sample. The number of 'yes' answers, denoted by x , follows binomial distribution with parameters n and θ_G. Obviously, the variance of the estimator $\hat{\pi}_G$ is given by:

$$V(\hat{\pi}_G) = \frac{\theta_G(1-\theta_G)}{nP^2} \tag{1.6}$$

The second situation considered in Greenberg et al. (1969), when π_y is unknown, requires taking two independent samples and estimate π_y and π separately. Also the estimator when π_y is known remains more efficient than the situation when π_y is unknown. Thus the motivation is to improve the Greenberg et al. (1969) estimator when π_y is known and only one sample is required at the estimation stage.

Based on this motivation, we suggest a new quasi empirical Bayes estimation technique which requires only one sample to estimate both parameters π and π_y, but makes use of knowledge of prior information of these parameters.

There is a myriad of applications for this model; the basis of which starts with ethics. Table 1-A shows the various real-life scenarios:

	Table 1-A
	Application
Discipline	**Proportion of people who possess the Attribute**
Education	cheat on tests
Finance	participate in Insider Trading
Medicine	commit Medicaid fraud
Psychology	cheat on their significant other
Human Obesity	under report their caloric intake
Human Anorexia	over report their caloric intake
Pediatrics	experiment with drugs
Social Issues	patients who pursue a medical marijuana prescription inappropriately
Epidemiology	women who had an induced abortion
Psychiatry	have a substance abuse problem

2. PROPOSED QUASI EMPRIRICAL BAYES ESTIMATES

In Section 1, we assumed parameters π and π_y which we are interested in estimating are unknown constants and fixed in a population. In Bayesian estimation process, the parameters π and π_y are considered random variables with a prior distribution.

Let $h_1(\pi)$ and $h_2(\pi_y)$ be the prior distribution of π and π_y, respectively, given by:

$$h_1(\pi) = C_1 \pi^{\alpha_1} (1-\pi)^{\beta_1} \tag{2.1}$$

and

$$h_2(\pi_y) = C_2 \pi_y^{\alpha_2} (1-\pi_y)^{\beta_2} \tag{2.2}$$

where C_1 and C_2 are constants such that $\int_0^1 h_1(\pi) d\pi = 1$ and $\int_0^1 h_2(\pi_y) d\pi_y = 1$

.

Assume the sampling distribution of x given π, π_y and P is given by:

$$f(x|\pi,\pi_y,P)=\binom{n}{x}\theta_G^x(1-\theta_G)^{n-x}, \quad x=0,1,2,\ldots,n$$
(2.3)

The joint density of x, π and π_y given P is given by:

$$f(x,\pi,\pi_y|P)=f(x|\pi,\pi_y,P)h_1(\pi)h_2(\pi_y)$$

$$=C_1C_2\binom{n}{x}\theta_G^x(1-\theta_G)^{n-x}\pi^{\alpha_1}(1-\pi)^{\beta_1}\pi_y^{\alpha_2}(1-\pi_y)^{\beta_2} \qquad (2.4)$$

Then the sampling distribution of x given P is given by:

$$f(x|P)=\int_0^1\int_0^1 f(x,\pi,\pi_y|P)dy = C_1C_2\binom{n}{x}\int_0^1\int_0^1\theta_G^x(1-\theta_G)^{n-x}\pi^{\alpha_1}(1-\pi)^{\beta_1}d\pi d\pi_y \qquad (2.5)$$

Then the joint posterior distribution of π and π_y is given by:

$$f(\pi,\pi_y|x,P)=\frac{f(x,\pi,\pi_y|P)}{f(x|P)}=\frac{\theta_G^x(1-\theta_G)^{n-x}\pi^{\alpha_1}(1-\pi)^{\beta_1}\pi_y^{\alpha_2}(1-\pi_y)^{\beta_2}}{\int_0^1\int_0^1\theta_G^x(1-\theta_G)^{n-x}\pi^{\alpha_1}(1-\pi)^{\beta_1}\pi_y^{\alpha_2}(1-\pi_y)^{\beta_2}d\pi d\pi_y} \qquad (2.6)$$

Taking the log of both sides of (2.6) , we get

$$\begin{aligned}&\log[f(\pi,\pi_y|x,P)]\\&=x\log\theta_G+(n-x)\log(1-\theta_G)+\alpha_1\log\pi \qquad (2.7)\\&\quad+\beta_1\log(1-\pi)+\alpha_2\log\pi_y+\beta_2\log(1-\pi_y)-\log(\mathrm{K})\end{aligned}$$

where K is a constant.
On setting

$$\frac{\partial\log[f(\pi,\pi_y|x,P)]}{\partial\pi}=0$$
(2.8)

and

$$\frac{\partial\log[f(\pi,\pi_y|x,P)]}{\partial\pi_y}=0 \qquad (2.9)$$

The quasi empirical Bayes estimate of π and π_y is obtained by solving the non-linear equations:

$$P\left[\frac{x}{P\pi+(1-P)\pi_y}-\frac{(n-x)}{1-P\pi-(1-P)\pi_y}\right]+\frac{\alpha_1}{\pi}-\frac{\beta_1}{(1-\pi)}=0 \qquad (2.12)$$

and

$$\frac{x}{P\pi+(1-P)\pi_y}-\frac{(n-x)}{1-P\pi-(1-P)\pi_y}+\frac{\alpha_1}{\pi}-\frac{\beta_1}{(1-\pi)}+\frac{\alpha_2}{\pi_y}-\frac{\beta_2}{(1-\pi_y)}=0 \qquad (2.13)$$

In the next section, we used the joint posterior distribution of π and π_y to develop a quasi Cramer-Rao lower bound of variance.

3. QUASI CRAMER-RAO LOWER BOUND OF VARIANCE

We compute

$$-E\left[\frac{\partial^2 \log[f(\pi,\pi_y \mid x,P)]}{\partial\pi^2}\right]=\frac{nP^2}{\theta_G(1-\theta_G)}+\frac{\alpha_1}{\pi^2}+\frac{\beta_1}{(1-\pi)^2} \qquad (3.1)$$

$$-E\left[\frac{\partial^2 \log[f(\pi,\pi_y \mid x,P)]}{\partial\pi_y^2}\right]=\frac{n(1-P)^2}{\theta_G(1-\theta_G)}+\frac{\alpha_2}{\pi_y^2}+\frac{\beta_2}{(1-\pi_y)^2} \qquad (3.2)$$

and

$$-E\left[\frac{\partial^2 \log[f(\pi,\pi_y \mid x,P)]}{\partial\pi\,\partial\pi_y}\right]=\frac{nP(1-P)}{\theta_G(1-\theta_G)} \qquad (3.3)$$

Then a quasi Cramer-Rao lower bound of variance-covariance of π and π_y is given by:

$$V\begin{bmatrix}\hat{\pi}\\ \hat{\pi}_y\end{bmatrix}\geq\begin{bmatrix}\frac{nP^2}{\theta_G(1-\theta_G)}+\frac{\alpha_1}{\pi^2}+\frac{\beta_1}{(1-\pi)^2}, & \frac{P(1-P)}{\theta_G(1-\theta_G)}\\ \frac{P(1-P)}{\theta_G(1-\theta_G)}, & \frac{n(1-P)^2}{\theta_G(1-\theta_G)}+\frac{\alpha_2}{\pi_y^2}+\frac{\beta_2}{(1-\pi_y)^2}\end{bmatrix}^{-1} \qquad (3.4)$$

Since we are interested in the estimation of the parameter π, the quasi Cramer-Rao lower bound of variance is then given by:

$$V(\hat{\pi})_{LB}\geq\frac{\frac{n(1-P)^2}{\theta_G(1-\theta_G)}+\frac{\alpha_2}{\pi_y^2}+\frac{\beta_2}{(1-\pi_y)^2}}{\left[\frac{nP^2}{\theta_G(1-\theta_G)}+\frac{\alpha_1}{\pi^2}+\frac{\beta_1}{(1-\pi)^2}\right]\left[\frac{n(1-P)^2}{\theta_G(1-\theta_G)}+\frac{\alpha_2}{\pi_y^2}+\frac{\beta_2}{(1-\pi_y)^2}\right]-\left[\frac{nP(1-P)}{\theta_G(1-\theta_G)}\right]^2} \qquad (3.5)$$

In the next section, SAS was used to computationally illustrate the performance of the proposed quasi empirical Bayes estimates with respect to the Greenberg et al (1969) estimator.

4. SIMULATION STUDY

We performed two types of simulation studies, one based on comparing simulated mean squared errors and second based on comparing the variance and lower bound of variance for different choice of parameters.

In the first set-up of the simulation study, we proceed as follows. For given values of P, π and π_y, we first compute the value of θ_G. Then for a given value of θ_G and sample size n, we generate a binomial variate M giving us the observed number of 'yes' answers. We generate 500 samples each of size n. From each given sample, we compute the Greenberg et al (1969) estimate $\hat{\pi}_{G(j)}$, $j = 1,2,..,500$. Also we used **PROC MODEL** to solve the two non-linear equations to find quasi Bayes estimates of $\hat{\pi}_{B(j)}$, $j = 1,2,..,500$. Then we computed the percent relative efficiency of the proposed quasi Bayes estimates over the Greenberg et al (1969) estimator as:

$$\text{RE(Sim)} = \frac{\sum_{j=1}^{500}\left(\hat{\pi}_{G(j)} - \pi\right)^2}{\sum_{j=1}^{500}\left(\hat{\pi}_{B(j)} - \pi\right)^2} \times 100\% \qquad (4.1)$$

We also computed the true percent relative efficiency based on the variance of the Greenberg et al (1969) estimator and the quasi Cramer-Rao lower bound of variance as follows:

$$\text{RE(True)} = \frac{V(\hat{\pi}_G)}{V(\hat{\pi})_{LB}} \times 100\% \qquad (4.2)$$

The results obtained for different values of P, π, π_y, n, α_1, β_1, α_2 and β_2 are given in Table 4.1.

Table 4.1. Percent RE(Simulated) and Percent RE(True) values for the different choices of parameters where proposed estimator is more efficient.

P	PI	PY	A1	A2	B1	B2	N	RE	TH	RETRUE
0.7	0.1	0.15	2.8	1.6	2.5	1.7	50	120.363	0.115	180.024
0.7	0.1	0.15	2.8	1.6	2.5	1.7	75	132.635	0.115	130.952
0.7	0.1	0.15	2.8	1.6	2.5	1.7	100	141.171	0.115	104.176
0.7	0.1	0.15	2.8	1.6	2.5	1.7	200	158.801	0.115	58.7466
0.7	0.1	0.15	2.8	1.6	2.5	1.7	500	173.248	0.115	26.0075
0.7	0.2	0.15	2.8	1.6	2.5	1.7	50	187.272	0.185	116.593
0.7	0.2	0.15	2.8	1.6	2.5	1.7	75	203.876	0.185	92.4568
0.7	0.2	0.15	2.8	1.6	2.5	1.7	100	215.794	0.185	77.9128
0.7	0.2	0.15	2.8	1.6	2.5	1.7	200	241.169	0.185	49.4651
0.7	0.2	0.15	2.8	1.6	2.5	1.7	500	265.777	0.185	24.3008
0.7	0.3	0.25	2.8	1.6	2.5	1.7	50	207.214	0.285	86.5678
0.7	0.3	0.25	2.8	1.6	2.5	1.7	75	207.958	0.285	66.4339
0.7	0.3	0.25	2.8	1.6	2.5	1.7	100	209.116	0.285	54.3823
0.7	0.3	0.25	2.8	1.6	2.5	1.7	200	212.194	0.285	32.0028
0.7	0.3	0.25	2.8	1.6	2.5	1.7	500	215.926	0.285	14.4859
0.7	0.4	0.35	2.8	1.6	2.5	1.7	50	222.855	0.385	70.9631
0.7	0.4	0.35	2.8	1.6	2.5	1.7	75	212.708	0.385	53.2219
0.7	0.4	0.35	2.8	1.6	2.5	1.7	100	207.727	0.385	42.8214
0.7	0.4	0.35	2.8	1.6	2.5	1.7	200	200.318	0.385	24.2558
0.7	0.4	0.35	2.8	1.6	2.5	1.7	500	195.848	0.385	10.6104
0.7	0.5	0.45	2.8	1.6	2.5	1.7	50	247.12	0.485	64.4865
0.7	0.5	0.45	2.8	1.6	2.5	1.7	75	226.572	0.485	47.7568
0.7	0.5	0.45	2.8	1.6	2.5	1.7	100	216.574	0.485	38.0924
0.7	0.5	0.45	2.8	1.6	2.5	1.7	200	201.928	0.485	21.2015
0.7	0.5	0.45	2.8	1.6	2.5	1.7	500	193.337	0.485	9.14234
0.7	0.6	0.65	2.8	1.6	2.5	1.7	50	231.322	0.615	70.7888
0.7	0.6	0.65	2.8	1.6	2.5	1.7	75	204.895	0.615	53.3357
0.7	0.6	0.65	2.8	1.6	2.5	1.7	100	192.446	0.615	43.0363
0.7	0.6	0.65	2.8	1.6	2.5	1.7	200	174.812	0.615	24.5093
0.7	0.6	0.65	2.8	1.6	2.5	1.7	500	165.003	0.615	10.7658
0.7	0.7	0.67	2.8	1.6	2.5	1.7	50	337.583	0.691	76.8277
0.7	0.7	0.67	2.8	1.6	2.5	1.7	75	282.162	0.691	57.2138
0.7	0.7	0.67	2.8	1.6	2.5	1.7	100	257.756	0.691	45.8616
0.7	0.7	0.67	2.8	1.6	2.5	1.7	200	224.683	0.691	25.8278
0.7	0.7	0.67	2.8	1.6	2.5	1.7	500	206.884	0.691	11.2584
0.7	0.8	0.78	2.8	1.6	2.5	1.7	50	477.675	0.794	102.492
0.7	0.8	0.78	2.8	1.6	2.5	1.7	75	355.674	0.794	77.5405
0.7	0.8	0.78	2.8	1.6	2.5	1.7	100	308.156	0.794	63.0157

References

1. Boruch, Robert F., and Joe S. Cecil. Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press, 1979.

2. Bridges, G. S. Level of Effects of Response Error in Self-Reports of Crime and Delinquency. Unpublished Ph.D. Dissertation, University of Pennsylvania, 1979.

3. Greenberg, B., Abdul-Ela, A., Simmons, W., and Horvitz, D., The unrelated question randomized response model: theoretical framework. Journal of the American Statistical Association, June 1969, 520-539.

4. Hauser, Philip M. “The Issue of Privacy and Confidentiality.” Paper prepared for the Subcommittee on Privacy of the Decennial Census Review Committee. 1971. Unpublished.

5. Horwitz, D. G., Shah, B.V., & Simmons, W. R., The unrelated question randomized response model. Proceedings of Social Statistics Section, 1967 (pp. 65 – 72). American Statistical Association.

6. Warner, S., Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, March 1965, 63-69.