Revista Colombiana de Estadística
Diciembre 2013, volumen 36, no. 2, pp. 303 a 317

On an Improved Bayesian Item Count Technique Using Different Priors

Técnica de conteo de items bayesiana mejorada usando diferentes distribuciones a priori

Zawar Hussain[1,3,a], Ejaz Ali Shah[2,b], Javid Shabbir[1,c], Muhammad Riaz[1,4,d]

[1]Department of Statistics, Faculty of Natural Sciences, Quaid-i-Azam University, Islamabad, Pakistan

[2]Department of Statistics, Faculty of Sciences, University of Hazara, Mansehra, Pakistan

[3]Department of Statistics, Faculty of Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

[4]Department of Mathematics and Statistics, Faculty of Sciences, King Fahad University of Petroleum and Minerals, Dhahran, Saudi Arabia

Abstract

Item Count Technique (ICT) serves the purpose of estimating the proportion of the people with stigmatizing attributes using the indirect questioning method. An improved ICT has been recently proposed in the literature (not requiring two subsamples and hence free from finding optimum subsample sizes unlike the usual ICT) in a classical framework that performs better than the usual ICT and the Warner method of Randomized Response (RR) technique. This study extends the scope of this recently proposed ICT in a Bayesian framework using different priors in order to derive posterior distributions, posterior means and posterior variances. The posterior means and variances are compared in order to study which prior is more helpful in updating the item count technique. Moreover, we have compared the Proposed Bayesian estimation with Maximum Likelihood (ML) estimation. We have observed that simple and elicited Beta priors are superior choices (in terms of minimum variance), depending on the sample size, number of items and the sum of responses. Also, the Bayesian estimation provides relatively more precise estimators than the ML Estimation.

***Key words*:** Bayesian Estimation, Indirect Questioning, Item Count Technique, Population Proportion, Prior Information, Privacy Protection, Randomized Response Technique, Sensitive Attributes.

[a]Professor. E-mail: zhlangah@yahoo.com
[b]Professor. E-mail: alishahejaz@yahoo.com
[c]Professor. E-mail: javidshabbir@gmail.com
[d]Professor. E-mail: riaz76qau@yahoo.com

Resumen

La técnica de conteo de ítems (ICT, por sus siglas en inglés) es útil para estimar la proporción de personas que poseen atributos que pueden tener algún grado de estigmatización mediante el uso de un método de preguntas indirectas. Una ICT mejorada ha sido propuesta recientemente en la literatura bajo la inferencia clásica (la cual no requiere dos submuestras y libre de la necesidad de encontrartamaños de muestra óptimos para cada una de ellas como sucede en la ICT usual). Esta ICT mejorada se desempeña mejor que la ICT usual y que el método de Respuesta Aleatorizada (RR, por sus siglas en inglés) de Warner. Este artículo extiende su estudio bajo una visión Bayesiana usando diferentes a priori con el fin de derivar distribuciones, medias y varianzas a posteriori.Las medias y varianzas a posteriori son comparadas con el fin de estudiar cuál a priori es más útil en mejorar la técnica de conteo de ítems. Se observa que a priori simples y Beta elicitadas son las mejores escogencias (en términos dela varianza mínima) dependiendo del tamaño de muestra, el número de ítems y la suma de la respuesta. También, la estimación bayesiana proporciona estimadores relativamente más precisas que la estimación ML.

***Palabras clave*:** atributos sensitivos, estimación Bayesiana, información a priori, preguntas indirectas, proporción poblacional, protección de la privacidad, técnica de conteo de ítems, técnica de respuesta aleatorizada.

1. Introduction

Survey techniques are now being utilized in almost every branch of physical and social sciences. These branches include medical, sociology, economics, agriculture, information technology, business, marketing, quality inspection, psychology, human behavior and many others. In surveys relating to these fields, especially, sociology, psychology, economics, people do not report their true status when the study question is sensitive in nature. Collection of trustworthy (truthful) data mainly depends upon the sensitivity of the study question, survey method, privacy (confidentiality) and cooperation of the respondents. The cooperation from the respondents will be low if the study question is sensitive and direct questioning method is applied. Consequently, the inferences made through direct questioning run the risk of response bias, non response (refusal) bias or both. An ingenious method pioneered by Warner (1965) was suggested in anticipation of reducing these biases and to provide more confidentiality to respondents.

The technique proposed by Warner (1965) is known as Randomized Response Technique (RRT). A comprehensive review of developments on Randomized Response (RR) techniques is given by Tracy & Mangat (1996) and Chaudhri & Mukerjee (1998). Some of the recent developments, among others, include Gupta, Gupta & Singh (2002), Ryu, Kim, Heo & Park (2005-2006), Bar-Lev, Bobovitch & Boukai (2004), Arnab & Dorffner (2006), Huang (2010), Hussain & Shabbir (2010), Barabesi & Marcheselli (2010) and Chaudhuri (2011). A number of applications of RRT can be found in the literature, for instance, Liu & chow. L. P. (1976), Reinmuth & Guerts (1975), Guerts (1980), Larkins, Hume & Garcha (1997), etc.

Although these studies were seen to be fruitful in the sense of estimation of the parameters, there are some applied difficulties associated with RRT as reported by Guerts (1980) and Larkins et al. (1997). Guerts (1980) found that RRT could have some limitations such the requirement of increased sample sizes in order to have confidence intervals as good as obtained through the direct questioning technique. More time is needed to administer and explain the procedure to the survey respondents. He further argued that, compilation of the results in the form of tables is somewhat protracted.

Larkins et al. (1997) were of the view that RRT was not suitable in the estimation of population proportion of tax payers/non-payers. Dalton & Metzger (1992) found that RRT might not be efficient in a mailed or telephonic survey. Similarly, Hubbard, Casper & Lesser (1989) argued that the major problem for RRT is to choose a randomization device to apply as a best one in specified circumstances and the very decisive feature of an RRT is about the respondent's acceptance of the technique. More recently, Chaudhuri & Christofides (2007) criticized RRT arguing that it is burdened with the respondent's ability to understand and handling of the device and also it asks respondents to report the information which may be useless or tricky. An intelligent interviewee may fear that his/her response can be traced back to his/her true status if he/she does not understand the mathematical logic behind the randomization device. Owing to these difficulties and limitations associated with $RRTs$, alternative techniques have been suggested. Some of these include the Item Count Technique by Droitcour, Casper, Hubbard, Parsley, Visscher & Ezzati (1991), the Three Card Method by Droitcour, Larson & Scheuren (2001) and the Nominative Technique by Miller (1985). These alternatives were suggested to avoid evasive answers on sensitive questions particularly concerning private issues, communally unexpected behaviors or illegitimate acts. Chaudhuri & Christofides (2007) also supplemented such an idea.

If some prior information is available about the mean of the study variable it may be used together with sample information. One of the methods using the prior knowledge is the Bayesian method of estimation where prior knowledge is used in the form of prior distribution. It has been established through many studies that when prior information is more informative the Bayesian estimation provides the more precise estimators.

In this paper, we plan to do a a Bayesian analysis of a recent item count technique by Hussain, Shah & Shabbir (2012) and provide the Bayesian estimators assuming that prior information is available through the past studies, past experience or simply through intelligent guess. Specifically, we will consider some prior distributions and compare the Bayesian estimator in case of each prior distribution used in this study. These comparisons will be in anticipation of finding the more suitable prior. The paper is organized as: Section 2 discusses the recent technique by Hussain et al. (2012); Section 3 provides Bayesian estimation using different priors; Section 4 presents a comparative analysis, concluding remarks are furnished in Section 5.

2. A Recent Item Count Technique

Hussain et al. (2012) proposed an improved item count technique based on single sample of size n in a classical framework showing an improvement over the usual ICT and the novel method of Randomized Response (RR) technique of Warner (1965). The said technique does not require two subsamples and consequently finding optimum subsample sizes is not needed. This study extends the scope of their study in a Bayesian framework and investigates the choice of a suitable prior to update the item count technique.

In the improved ICT of Hussain et al. (2012), each respondent is provided a list of g items and asked to report the number of items applicable to him/her, where each item is a combination of an unrelated item say F_j and a sensitive characteristic say S. The i^{th} respondent is asked to count 1, if he /she possess at least one of the characteristics F_j and S, and count 0 otherwise and finally report the total count. So, for a single respondent his/her response may be 0 to g. The response 1 for a single question or item means the respondent belongs either to non sensitive characteristic, sensitive characteristic or to both. Now the probability of 1 for j^{th} item is given by:

$$P(1) = \theta_j = \theta_{F_j} + \pi - \pi\theta_{F_j} \tag{1}$$

where θ_{F_j} denotes the proportion of j^{th} innocuous characteristic andπ denotes population proportion of individuals possessing a sensitive characteristic. Let Y_i be the response of i^{th} respondent, then it can be written as: $Y_i{=}\sum_{j=1}^{g}\alpha_j$, where α_j is a Bernoulli random variable taking values 1 and 0 with probabilities θ_j and $(1-\theta_j)$ respectively. The unbiased moment (and ML) estimator for proportion of people bearing sensitive behavior is given as:

$$\hat{\pi}_M = \left(\bar{y} - \sum_{j=1}^{g}\theta_{F_j}\right)\left(g - \sum_{j=1}^{g}\theta_{F_j}\right)^{-1} \tag{2}$$

with variance given by:

$$Var(\hat{\pi}_M) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)}{n(g-\sum_{j=1}^{g}\theta_{F_j})^2}\{\sum_{j=1}^{g}\theta_{F_j}(1-\sum_{j=1}^{g}\theta_{F_j}) + 2\sum_{j<k}^{g}\theta_{F_j}\theta_{F_k}\} \tag{3}$$

In order to have Y_i as a binomial random variable we take $\theta_j = \theta$ (or equivalently $\theta_{F_j} = \theta_F$) for all $j = 1, 2, \ldots, g$ such that $\theta_{F_j} = \frac{1}{g}$. In this case variance of ML estimator turns out to be

$$Var(\hat{\pi}_M) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)}{ng(g-1)} \tag{4}$$

Now we develop Bayesian estimation of population proportion through the above mentioned item count technique of Hussain et al. (2012) by assuming that $\theta_j = \theta$ for all $j = 1, 2, \ldots, g$. We use different prior distributions for deriving

posterior distributions in order to find which posterior distribution gives high posterior probability for higher estimates of π. Prior distributions used here are Beta distribution with known hyper parameters, Non-informative Uniform distribution, Non-informative Haldane distribution, Mixture of Beta distributions and a Beta distribution with elicited hyperparameters. The posterior distribution using density kernel is defined as:

$$P(\pi|y) \propto L(y,\pi)P(\pi) \tag{5}$$

where $L\left(y,\pi\right)$ is the likelihood function and $P(\pi)$ is the prior distribution. Since α_j is the Bernoulli random variable with parameter $\theta_j = \theta$ the response variable Y_i is a binomial random variable with parameter g and θ. Thus the likelihood function becomes:

$$L\left(y,\pi\right) = \prod_{i=1}^{n} \left\{ \binom{g}{y_i} \theta^{y_i} \left(1-\theta\right)^{g-y_i} \right\} \tag{6}$$

where $\theta = \theta_F + \pi\left(1-\theta_F\right)$ Substituting $\theta = \theta_F + \pi\left(1-\theta_F\right)$ in above equation and taking $d = \frac{\theta_F}{(1-\theta_F)}$, we get

$$L\left(y,\pi\right) = \left(1-\theta_F\right)^{ng} \left\{ \prod_{i=1}^{n} \binom{g}{y_i} \right\} \left(d+\pi\right)^{n\bar{y}} \left(1-\pi\right)^{ng-n\bar{y}} \tag{7}$$

3. Bayesian Estimation using Different Priors

In this section, we derive the Bayesian estimators of π assuming different prior distributions mentioned above in Section 2.

3.1. Beta Prior

Suppose the prior distribution of $\pi$is given by:

$$P(\pi) = \frac{1}{B\left(a,b\right)} \pi^{a-1} \left(1-\pi\right)^{b-1}, \qquad 0 < \pi < 1 \tag{8}$$

where $B\left(a,b\right) = \int_0^1 \pi^{a-1}\left(1-\pi\right)^{b-1} d\pi$ is a complete Beta function.
Thus, using (7) and (8) in (5) the posterior distribution of π is derived as:

$$P\left(\pi|\, y\right) \propto \left(1-\theta_F\right)^{ng} \left\{ \prod_{i=1}^{n} \binom{g}{y_i} \right\} \left(d+\pi\right)^{n\bar{y}} \left(1-\pi\right)^{ng-n\bar{y}} \left\{ \pi^{a-1}\left(1-\pi\right)^{b-1} \right\}$$

$$P\left(\pi|\, y\right) \propto \left(1-\theta_F\right)^{ng} \left\{ \prod_{i=1}^{n} \binom{g}{y_i} \right\} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} \pi^{a+i-1} \left(1-\pi\right)^{b+ng-n\bar{y}-1}$$

Now we find the normalizing constant say k. As we know that for posterior distribution we must have

$$k\left(1-\theta_F\right)^{ng}\left\{\prod_{i=1}^{n}\binom{g}{y_i}\right\}\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}\int_0^1\pi^{a+i-1}\left(1-\pi\right)^{b+ng-n\bar{y}-1}d\pi=1$$

This gives

$$k=\left[\left(1-\theta_F\right)^{ng}\left\{\prod_{i=1}^{n}\binom{g}{y_i}\right\}\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i,b+ng-n\bar{y}\right)\right]^{-1}$$

Thus, the posterior distribution of $\pi$is given by:

$$P\left(\pi|y\right)=\frac{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}\pi^{a+i-1}\left(1-\pi\right)^{b+ng-n\bar{y}-1}}{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i,b+ng-n\bar{y}\right)} \tag{9}$$

Now the Bayesian estimator (posterior mean) is given by:

$$E\left(\pi|y\right)=\frac{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}\int_0^1\pi^{a+i+1-1}\left(1-\pi\right)^{b+ng-n\bar{y}-1}d\pi}{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i,b+ng-n\bar{y}\right)}$$

$$E\left(\pi|\,y\right)=\frac{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i+1,b+ng-n\bar{y}\right)}{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i,b+ng-n\bar{y}\right)} \tag{10}$$

While, the posterior variance is given as:

$$Var\left(\pi|\,y\right)=\frac{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i+2,b+ng-n\bar{y}\right)}{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i,b+ng-n\bar{y}\right)}$$

$$-\left(\frac{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i+1,b+ng-n\bar{y}\right)}{\sum_{i=0}^{n\bar{y}}\binom{n\bar{y}}{i}d^{n\bar{y}-i}B\left(a+i,b+ng-n\bar{y}\right)}\right)^2 \tag{11}$$

3.2. Non-informative Uniform Prior

The non-informative uniform prior distribution is given as:

$$P\left(\pi\right)\propto 1. \tag{12}$$

Using (12) and (7) in (5), the posterior distribution is derived as:

$$P(\pi|y) = \frac{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} \pi^{i+1-1} (1-\pi)^{ng-n\bar{y}+1-1}}{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i+1, ng-n\bar{y}+1)}. \tag{13}$$

Under the non-informative prior, the posterior mean and variance are given by:

$$E(\pi|y) = \frac{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i+2, ng-n\bar{y}+1)}{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i+1, ng-n\bar{y}+1)} \tag{14}$$

$$Var(\pi|y) = \frac{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i+3, ng-n\bar{y}+1)}{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i+1, ng-n\bar{y}+1)}$$

$$- \left(\frac{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i+2, ng-n\bar{y}+1)}{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} Beta(i+1, ng-n\bar{y}+1)} \right)^2 \tag{15}$$

3.3. Non-informative Haldane Prior

Another non-informative prior used here is the Haldane prior (Zellner 1996) which has the probability distribution defines as:

$$P(\pi) \propto \frac{1}{p(1-p)} \tag{16}$$

It is also defined as $B(0,0)$. Thus the posterior distribution is give as:

$$P(\pi|y) = \frac{\sum_{i=1}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} \pi^{i-1} (1-\pi)^{ng-n\bar{y}-1}}{\sum_{i=1}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i, ng-n\bar{y})} \tag{17}$$

Posterior mean and variance are, now, given as:

$$E(\pi|y) = \frac{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i+1, ng-n\bar{y})}{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(i, ng-n\bar{y})} \tag{18}$$

$$Var\left(\pi|\, y\right) = \frac{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} Beta\left(i+2, ng-n\bar{y}\right)}{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} Beta\left(i, ng-n\bar{y}\right)}$$

$$- \left(\frac{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(i+1, ng-n\bar{y}\right)}{\sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(i, ng-n\bar{y}\right)} \right)^2 \tag{19}$$

3.4. Mixture of Beta Priors

We assume that prior information come as a mixture of different Beta distributions. The mixture of Beta distributions with H components is given as:

$$P(\pi) = \sum_{h=1}^{H} \frac{W_h}{B(a_{h,} b_h)} \pi^{a_h - 1} (1-\pi)^{b_h - 1} \tag{20}$$

where W_h are the weights such that $\sum_{h=1}^{H} W_h = 1$, anda_h b_h are the hyper-parameters of h^{th}component Beta distribution.

The posterior distribution, now, is given by:

$$P\left(\pi|\, y\right) = \frac{\sum_{h=1}^{H} \frac{W_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} \pi^{a_h+i-1} \left(1-\pi\right)^{b_h+ng-n\bar{y}-1}}{\sum_{h=1}^{H} \frac{W_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B(a_h+i, b_h+ng-n\bar{y})} \tag{21}$$

Posterior mean and variance, under the assumption of a mixture of Beta distributions, are given as:

$$\left(\pi|\, y\right) = \frac{\sum_{h=1}^{H} \frac{W_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(a_h+i+1, b_h+ng-n\bar{y}\right)}{\sum_{h=1}^{H} \frac{W_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(a_h+i, b_h+ng-n\bar{y}\right)} \tag{22}$$

$$Var\left(\pi|\, y\right) = \frac{\sum_{h=1}^{H} \frac{W_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(a_h+i+2, b_h+ng-n\bar{y}\right)}{\sum_{h=1}^{H} \frac{W_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(a_h+i, b_h+ng-n\bar{y}\right)}$$

$$- \left(\frac{\sum_{h=1}^{H} \frac{W_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(a_h+i+1, b_h+ng-n\bar{y}\right)}{\sum_{h=1}^{H} \frac{w_h}{B(a_h, b_h)} \sum_{i=0}^{n\bar{y}} \binom{n\bar{y}}{i} d^{n\bar{y}-i} B\left(a_h+i, b_h+ng-n\bar{y}\right)} \right)^2 \tag{23}$$

3.5. Beta Prior with Elicited Hyperparameters

There are many methods for eliciting parameters of prior distributions. The method we have used for eliciting the hyperparameters is the method of prior predictive distribution (Aslam 2003, cf.). We first derived the prior predictive distribution and then by using "SAS" we elicited the hyperparameters. Then we have derived Posterior mean and Posterior variance.

The prior predictive distribution is given as:

$$P(y) = \frac{\binom{g}{y}(1-\theta_F)^g \sum_{i=0}^{y} \binom{y}{i} B(a+i, b+g-y)}{B(a,b)} \tag{24}$$

We solved this equation further for different values of g and y and then by using "SAS" we elicited the hyperparameters a and b. For every g we have different values of a and b. Although according to our calculations, for different values of g and y, we got same value for a, but b changed accordingly. The derived expressions for posterior distribution, posterior mean, and posterior variance are same as we have derived for posterior distribution using Beta prior with known hyperparameters, but the numerical values obtained for hyperparameters are now different.

4. Comparative Analysis

In this section, we provide a comparative analysis of posterior means and posterior variances obtained through different prior distributions assumed in this study. We should mention that under the squared error loss function posterior mean is taken as Bayesian estimator while posterior variance is taken as the measure of precision. Also, under Uniform and Haldane prior distributions, posterior distributions are not defined for $ng = n\bar{y}$. If $ng < n\bar{y}$, posterior distributions under all the priors considered here are not defined. That is why, some entries in the Tables 3 and 4 are not given. For different values of sum of responses, $n\bar{y}$, number of items g and sample size n, we have computed posterior means and variances under different prior distributions and results are displayed in Tables 1-12 given below.

We compare ML estimator and proposed Bayesian estimators in terms of variablity. To compare proposed Bayesian estimators with ML estimator, we selected $g = 7$ and $\theta_F = \frac{1}{g} \simeq 0.143$ and computed variance of ML estimator for $n = 20, 30, 40$ and 50. The variances of ML estimator for the different values of π are presented in Table 13.

From Tables 1-12 it is observed that when $n\bar{y}$, n and g are small, posterior means are larger under mixture and elicited Beta prior distributions compare to posterior means under other prior distributions considered here. For a fixed g, posterior distribution using elicited Beta prior produces larger means than the others with the increase in $n\bar{y}$. As n increases posterior means under all priors

Table 1: Posterior means for $n\bar{y} = 30$, $\theta_F = 0.33$ and $g = 3$.

	Prior distribution				
n	Beta	Uniform	Hadlane	Mixture	Elicited Beta
20	0.2793	0.2666	0.2187	0.3083	0.2990
30	0.1403	0.0659	0.0391	0.1453	0.1023
40	0.0849	0.0264	0.0177	0.0818	0.0480
50	0.0588	0.0154	0.0114	0.0541	0.0293

Table 2: Posterior means for $n\bar{y} = 50$, $\theta_F = 0.33$ and $g = 3$.

	Prior distribution				
n	Beta	Uniform	Hadlane	Mixture	Elicited Beta
20	0.6188	0.7553	0.7484	0.6525	0.7608
30	0.3299	0.3439	0.3238	0.3546	0.3608
40	0.1865	0.1395	0.0970	0.1974	0.1683
50	0.1136	0.0507	0.0283	0.1140	0.0786

Table 3: Posterior means for $n\bar{y} = 60$, $\theta_F = 0.33$ and $g = 3$.

	Prior distribution				
n	Beta	Uniform	Hadlane	Mixture	Elicited Beta
20	0.8074	-	-	0.8613	0.9987
30	0.4549	0.5079	0.4967	0.4800	0.5179
40	0.2687	0.2599	0.2383	0.2867	0.2771
50	0.1636	0.1144	0.0755	0.1705	0.1411

Table 4: Posterior means for $n\bar{y} = 90$, $\theta_F = 0.33$ and $g = 3$.

	Prior distribution				
n	Beta	Uniform	Hadlane	Mixture	Elicited Beta
20	-	-	-	-	-
30	0.8609	-	-	0.9058	0.9991
40	0.5691	0.6300	0.6243	0.5880	0.6349
50	0.3864	0.4069	0.3978	0.4035	0.4152

Table 5: Posterior means for $n\bar{y} = 30$, $\theta_F = 0.143$ and $g = 7$.

	Prior distribution				
n	Beta	Uniform	Hadlane	Mixture	Elicited Beta
20	0.1288	0.0905	0.0657	0.1306	0.1079
30	0.0646	0.0250	0.0141	0.0612	0.0399
40	0.0387	0.0108	0.0069	0.0349	0.0198
50	0.0266	0.0064	0.0046	0.0232	0.0123

Table 6: Posterior means for $n\bar{y} = 50$, $\theta_F = 0.143$ and $g = 7$.

	Prior distribution				
n	Beta	Uniform	Hadlane	Mixture	Elicited Beta
20	0.2599	0.2551	0.2460	0.2701	0.2639
30	0.1381	0.1152	0.1026	0.1401	0.1254
40	0.0797	0.0108	0.0293	0.0778	0.0602
50	0.0495	0.0189	0.0101	0.0462	0.0301

TABLE 7: Posterior means for $n\bar{y} = 30$, $\theta_F = 0.33$ and $g = 3$.

	Prior distribution				
n	Beta	Uniform	Hadlane	Mixture	Elicited Beta
20	0.5519	0.5862	0.5824	0.5637	0.5899
30	0.3319	0.3364	0.3316	0.3401	0.3411
40	0.2173	0.2110	0.2058	0.2219	0.2160
50	0.1478	0.1357	0.1298	0.1495	0.1411

TABLE 8: Posterior means for $n\bar{y} = 30$, $\theta_F = 0.33$ and $g = 3$.

	Prior distribution				
n	Beta	Uniform	Haldane	Mixture	Elicited Beta
20	0.0052	0.0086	0.0106	0.0057	0.0078
30	0.0023	0.0023	0.0014	0.0027	0.0028
40	0.0011	0.0005	0.0003	0.0012	0.0009
50	0.0006	0.0002	0.0001	0.0006	0.0004

TABLE 9: Posterior means for $n\bar{y} = 50$, $\theta_F = 0.33$ and $g = 3$.

	Prior distribution				
n	Beta	Uniform	Haldane	Mixture	Elicited Beta
20	0.0052	0.0052	0.0052	0.0050	0.0047
30	0.0043	0.0059	0.0065	0.0044	0.0056
40	0.0027	0.0039	0.0042	0.0030	0.0036
50	0.0015	0.0014	0.0008	0.0017	0.0016

TABLE 10: Posterior variances for $n\bar{y} = 30$, $\theta_F = 0.143$ and $g = 7$.

	Prior distribution				
n	Beta	Uniform	Haldane	Mixture	Elicited Beta
20	0.0013	0.0015	0.0017	0.0014	0.0015
30	0.0005	0.0004	0.0002	0.0005	0.0005
40	0.0002	0.00009	0.00005	0.0002	0.0001
50	0.0001	0.00004	0.00002	0.0001	0.00007

TABLE 11: Posterior variances for $n\bar{y} = 50$, $\theta_F = 0.143$ and $g = 7$.

	Prior distribution				
n	Beta	Uniform	Haldane	Mixture	Elicited Beta
20	0.0019	0.0021	0.0022	0.0020	0.0022
30	0.0010	0.0012	0.00129	0.0010	0.0011
40	0.0005	0.0006	0.0005	0.00057	0.0006
50	0.0003	0.0002	0.0001	0.0003	0.0003

TABLE 12: Posterior variances for $n\bar{y} = 90$, $\theta_F = 0.143$ and $g = 7$.

	Prior distribution				
n	Beta	Uniform	Haldane	Mixture	Elicited Beta
20	0.0020	0.0022	0.0022	0.0020	0.0022
30	0.0014	0.0015	0.0016	0.0014	0.0015
40	0.0010	0.0010	0.0010	0.0010	0.0010
50	0.0006	0.0007	0.0007	0.0007	0.0007

TABLE 13: Variances of ML estimator for different values of π, n, $\theta_F = \frac{1}{g}$ and $g = 7$.

	π								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0.005	0.008	0.011	0.012	0.013	0.012	0.010	0.008	0.004
30	0.003	0.005	0.007	0.008	0.008	0.008	0.007	0.005	0.003
40	0.002	0.004	0.005	0.006	0.006	0.006	0.005	0.004	0.002
50	0.002	0.003	0.004	0.005	0.005	0.005	0.004	0.003	0.001

decrease rapidly and posterior means using mixture and simple Beta prior distributions turn out to be larger for a relatively smaller $n\bar{y}$. The reason being their dependence upon the data and hyperparameters (see Tables 1-3 and 5-7). From Tables 1-7, it is also observed that as $n\bar{y}$ increases, posterior means under all the priors considered here become larger. The reason being they mainly depend on the magnitude of $n\bar{y}$. For a given g and larger n, if we observe the maximum $n\bar{y}$, posterior distribution using elicited Beta prior yields larger means than those provided by the other prior. We also observed that as g increases, posterior means under all prior distributions decrease. Comparatively, posterior mean using a mixture of Beta priors and Beta distributions with assumed hyperparameters have larger means than the others. However, posterior mean increases under Uniform, Heldane and mixture priors, as $n\bar{y}$ increases. for larger n, they are still smaller than posterior means using mixture and simple Beta priors. It is also observed that for increased $n\bar{y}$, posterior mean using elicited Beta prior is larger but for using large value of n it is smaller than posterior mean using simple Beta and mixture priors (see Tables 5-7).

Tables 8 and 9 show that for smaller $n\bar{y}$ and g posterior variances using Beta prior with assumed hyperparameters and mixture prior are relatively smaller than the posterior variances under other priors. For fixed values of $n\bar{y}$ and g, as n increases the posterior variance with Haldane and Uniform priors remaining smaller than that obtained under other priors.The posterior variance under Haldane and Uniform priors depend only on the $n\bar{y}$. As $n\bar{y}$, increases for given n, posterior variance under elicited Beta prior remains smaller than the posterior variance obtained under other priors. As it is largely affected by $n\bar{y}$, for larger n and for fixed $n\bar{y}$ and g, posterior variance under mixture and simple Beta priors remains smaller than the posterior variances obtained under other priors.

It is also observed that for larger g, posterior distributions using Beta prior with assumed hyperparameters and mixture prior have the smaller variances as compared to the others. But, again, for larger n, posterior distributions using Haldane, Uniform and elicited Beta prior have smaller variances than other two. But as $n\bar{y}$ is increased posterior distributions under elicited Beta, Uniform and mixture prior have smaller variances than the other two (see Tables 10-12). From expression (4), it is obvious that variance of ML estimator does not depend upon $n\bar{y}$. Thus comparison of ML estimator and proposed Bayesian estimators can be made using Tables 10-13. From Tables 10-13, it is observed that when $g = 0.7$, $n\bar{y} = 30, 60, 90$ and $\theta_F = 0.143$, posterior variances under each prior are smaller than variance of ML estimator over the whole range of π. It shows a better performance of the proposed Bayesian estimation.

5. A Real Application of Proposed Methodology

A survey was designed to collect the data from the students at Quaid-i-Azam University Islamabad. Visiting websites containing adult contents was taken as the sensitive characteristic of interest. Finding unrelated characteristics with equal known proportions among the students was observed to be difficult. Alternatively, we took three boxes containing red and white cards with equal proportion ($\theta_F = 0.33$) of red cards in each box (that is, we took $g = 3$). A simple random sample of 50 students was selected from the university. Each student was asked to randomly draw a card form each box and count 1 if he/she have ever visited a website containing adult material or if the card selected from the j^{th} $(j = 1, 2, 3)$ box is a red card. Each respondent, then, was asked to report his/her total count (which may be any value from $0 - 3$). The actual data (Y_i, $i = 1, 2, \ldots, 50$) gathered from the sample students are given in table 13 below. Thus, we have $n\overline{y} = 90$. To obtain the Bayesian estimates of proportion of students who have ever visited a website containing adult material we considered five different prior distributions: (a) simple Beta prior with hyper-parameters $a = 5, b = 10$, (b) noninformative uniform prior, (c) Haldane prior, (d) a mixture prior of 4 Beta distributions with hyperparameters; (i) $a = 1, b = 2$, (ii) $a = 2, b = 4$, (iii) $a = 3, b = 6$, (iv) $a = 4, b = 8.$, (e) Beta prior with hyperparameters $(a = 2, b = 0.0540)$ elicited from the data. Findings of the survey are summarized in Table 14.

Table 14: Actual data obtained from 50 students using $\theta_F = 0.33$ and $g = 3$

Student	1	2	3	4	5	6	7	8	9	10
Response	2	2	2	3	2	1	2	2	3	3
Student	11	12	13	14	15	16	17	18	19	20
Response	3	1	0	2	2	2	2	2	1	2
Student	21	22	23	24	25	26	27	28	29	30
Response	2	2	3	1	0	0	3	2	1	1
Student	31	32	33	34	35	36	37	38	39	40
Response	2	2	3	2	1	3	1	1	2	2
Student	41	42	43	44	45	46	47	48	49	50
Response	2	0	2	3	1	1	3	1	2	2

Table 15: Summary of the survey results

Estimates	Simple Beta	Uniform	Haldane	Mixture priors	Beta prior
Proportion	0.386	0.406	0.397	0.4035	0.4152
Variance	0.0030	0.0035	0.0036	0.0030	0.0034
95% C.I	0.278-0.492	0.293-0.523	0.284-0.523	0.284-0.507	0.292-0.522

From table 15, it is observed that the simple Beta prior with assumed known hyperparametrs and mixture prior of Beta distributions yielded relatively more precise estimators with narrower 95% confidence intervals.

6. Concluding Remarks

This study investigates a recent item count technique in a Bayesian framework using different priors in order to study which prior is more helpful in updating the item count technique. We have compared the posterior means and variances in order to check which posterior performs better than other under different conditions. In case of large values of g and n, in general, we have observed that if large sum of responses, $n\bar{y}$, are observed, posterior distribution with elicited Beta prior comes up as the most suitable choice. However the sum of response, $n\bar{y}$, is not large then posterior distribution with simple beta prior a more suitable choice. Compared to ML estimator, in terms of precision, the proposed Bayesian estimators under each prior distribution (considered in this study) perform relatively better.

[Recibido: octubre de 2012 — Aceptado: agosto de 2013]

References

Arnab, R. & Dorffner, G. (2006), 'Randomized response technique for complex survey design', *Statistical Papers* (48), 131–141.

Aslam, M. (2003), 'An application of prior predictive distribution to elicit the prior density', *Journal of Statistical Theory and Application* (2), 70–83.

Bar-Lev, S. K., Bobovitch, E. & Boukai, B. (2004), 'A note on randomized response models for quantitative data', *Metrika* (60), 255–260.

Barabesi, L. & Marcheselli, M. (2010), 'Bayesian estimation of proportion and sensitivity level in randomized response procedures', *Metrika* (72), 75–88.

Chaudhri, A. & Mukerjee, R. (1998), *Randomized Response: Theory and Methods*, Marcel-Decker, New York.

Chaudhuri, A. (2011), *Randomized Response and Indirect Questioning Techniques in Surveys*, Chapman & Hall, Florida, United States.

Chaudhuri, A. & Christofides, T. C. (2007), 'Item count technique in estimating proportion of people with sensitive feature', *Journal of Statistical Planning and Inference* (137), 589–593.

Dalton, D. R. & Metzger, M. B. (1992), 'Integrity testing for personal selection: An unsparing perspective', *Journal of Business Ethics* (12), 147–156.

Droitcour, J. A., Casper, R. A., Hubbard, M. L., Parsley, T., Visscher, W. & Ezzati, T. M. (1991), The item count technique as a method of indirect questioning: a review of its development and a case study application, *in* P. P. Biemer, R. M. Groves, L. Lyberg, N. Mathiowetz & S. Sudeman, eds, 'Measurement Errors in Surveys', Willey, New york.

Droitcour, J. A., Larson, E. M. & Scheuren, F. J. (2001), The three card method: estimating sensitive survey items with permanent anonymity of response, *in* 'Proceedings of the Social Statistics Section', American Statistical Association, Alexandria, Virginia.

Guerts, M. D. (1980), 'Using a randomized response design to eliminate nonresponse and response biases in business research', *Journal of the Academy of Marketing Science* (8), 83–91.

Gupta, S., Gupta, B. & Singh, S. (2002), 'Estimation of sensitivity level of personal interview survey questions', *Journal of Statistical Planning and Inference* **100**, 239–247.

Huang, K. C. (2010), 'Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling', *Metrika* (71), 341–352.

Hubbard, M. L., Casper, R. A. & Lesser, J. T. (1989), Respondent's reactions to item count list and randomized response, *in* 'Proceeding of the Survey Research Section of the American Statistical Association', Washington, D. C., pp. 544–448.

Hussain, Z. & Shabbir, J. (2010), 'Three stage quantitative randomized response model', *Journal of Probability and Statistical Sciences* (8), 223–235.

Hussain, Z., Shah, E. A. & Shabbir, J. (2012), 'An alternative item count technique in sensitive surveys', *Revista Colombiana de Estadistica* (35), 39–54.

Larkins, E. R., Hume, E. C. & Garcha, B. (1997), 'The validity of randomized response method in tax ethics research', *Journal of the Applied Business Research* **13**(3), 25–32.

Liu, P. T. & chow. L. P. (1976), 'A new discrete quantitative randomized response model', *Journal of the American Statistical Association* (71), 72–73.

Reinmuth, J. E. & Guerts, M. D. (1975), 'The collection of sensitive information using a two stage randomized response model', *Journal of Marketing Research* (12), 402–407.

Ryu, J. B., Kim, J. M., Heo, T. Y. & Park, C. G. (2005-2006), 'On stratified randomized response sampling', *Model Assisted Statistics and Applications* (1), 31–36.

Tracy, D. & Mangat, N. (1996), 'Some development in randomized response sampling during the last decade-A follow up of review by Chaudhuri and Mukerjee', *Journal of Applied Statistical Science* (4), 533–544.

Warner, S. L. (1965), 'Randomized response: A survey for eliminating evasive answer bias', *Journal of the American Statistical Association* (60), 63–69.

Zellner, A. (1996), *An Introduction to Bayesian Inference in Econometrics*, Chichester, John Wiley, New York.