An Objective Bayesian analysis of dichotomous sensitive data

Maria Maddalena Barbieri and Brunero Liseo

Abstract We consider a dichotomous population in which every person belongs either to a sensitive group A or to the non sensitive complement $\bar{A}$. The object of interest is to estimate the population proportion of individuals who are members of A. We refer to a randomized response model proposed by Huang (2004), where also another parameter is present, namely the probability that a respondent truthfully states that he/she belongs to A in a direct response survey. In the paper the posterior distribution of the parameters under the joint Jeffreys and Reference prior is derived. The properties of the noninformative priors are investigated through the frequentist coverage probabilities of posterior quantiles.

Key words: Jeffreys Prior, Randomized response models, Reference prior, Sensitive data.

1 Introduction

In survey sampling questions about sensitive issues are likely to lead to refusals or untruthful answers, making inferential results unreliable. To improve respondent cooperation and to encourage truthful answers, Warner [3] proposed a data collection procedure, the randomized response technique, that allows to obtain sensitive information preserving the confidentiality of the responses. Since Warner's [3] first paper several randomized response plans have been developed. We will focus on a procedure proposed by Huang [1] where each interviewee is fist asked directly if he/she is in the sensitive group A or in the non sensitive complement $\bar{A}$. Only the

Maria Maddalena Barbieri
Dipartimento di Economia, Università Roma Tre, e-mail: barbieri@uniroma3.it

Brunero Liseo
Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, e-mail: brunero.liseo@uniroma1.it

respondents bearing to belong to $\bar{A}$ undergo a randomization, implemented as proposed by Warner [3]. In [1] the model is considered from a classical perspective and the method of moments is used to estimate the parameters.

The model in [1] has been further developed in Barabesi and Marcheselli [2], who implemented a different randomization in the second stage and also proposed a Bayesian analysis of the model, using independent Beta priors on the parameters.

In section 2 we briefly review the model. In section 3 we introduce the prior distribution and derive the posterior for the parameters. We check the properties of the noninformative prior in terms of frequentist coverage of one tail credible intervals.

2 The model

Consider a dichotomous population in which every person belongs either to a sensitive group A or to the non-sensitive complement $\bar{A}$. The parameter of interest is θ, the population proportion of individuals in Group A. We assume that a simple random sample of size n is drawn from the population.

The randomized response procedure proposed by Huang [1] works in two stages. First the sample respondents are required to answer to the direct question whether they belongs to group A or not. Only the respondents answering that they do not belong to A procede to the second stage. Each of them is provided with a chance device to select one of the questions: "*Question 1: Do you belong to Group A?*" and "*Question 2: Do you belong to Group $\bar{A}$?*". Let p be the probability to select the first question, with p known. The respondents do not report the outcome of the random device. Thus, if a respondent says "Yes", the interviewer does not know whether the "Yes" refers to the first or the second question.

It is assumed that the respondents belonging to A answer the truth to the direct question in the first stage with probability ξ, but that they gives totally honest answers in the second stage, since they properly understand the random mechanism and feel their privacy preserved by it. On the other hand, the respondents not belonging to A have no reason to lie. Thus it is reasonable to expect that they will be completely truthful in their answers, no matter whether they have to answer to the direct question or if they undergo the randomized response stage.

We will denote by Y the Bernoulli random variable, observed on each individual in the sample, assuming value 1 if the answer to the direct question is "Yes" and 0 otherwise. For each respondent undergoing the randomized procedure, let Z be a Bernoulli random variable taking value 1 if the final answer is "Yes".

Since only respondents who belong to A and tell the truth answer "Yes" to the direct question, we have that $P(Y=1)=\theta\xi$, while $P(Z=1,Y=0)=(1-\xi)\theta p+(1-\theta)(1-p)$. In fact the randomized procedure gives a "Yes" if the respondent answers to *Question 1*, belongs to A and lied to the direct question or if he/she answers to *Question 2* and does not belong to A. Thus the likelihood function is:

$$L(\theta,\xi) = (\theta\xi)^{\sum_{i=1}^n y_i}[(1-\xi)\theta p+(1-\theta)(1-p)]^{\sum_{i=1}^n z_i(1-y_i)}\cdot \\ \cdot[(1-\xi)\theta(1-p)+(1-\theta)p]^{\sum_{i=1}^n (1-z_i)(1-y_i)}.$$

3 An Objective prior analysis

In order to derive the Jeffreys prior distribution for (θ, ξ) and to compute the posterior summaries of interest, we first reparametrize the model as follows.

Let $X_1 = \sum_{i=1}^n y_i$, $X_2 = \sum_{i=1}^n z_i(1-y_i)$, $X_3 = \sum_{i=1}^n (1-z_i)(1-y_i) = n - X_1 - X_2$, $\phi_1 = \theta\xi$ and $\phi_2 = (1-\xi)\theta p + (1-\theta)(1-p)$. Then (X_1, X_2, X_3) has a Trinomial distribution, namely a Multinomial distribution on three cells, with parameters $(\phi_1, \phi_2, 1-\phi_1-\phi_2, n)$ and parametric space defined by the following constraints: $\phi_1 \in [0,1]$ and $\phi_2 \in [(1-p)(1-\phi_1), p(1-\phi_1)]$, where, without loss in generality, we assume $p > \frac{1}{2}$. Note that when $p = 1$ the constraints on the parameter space vanish and $\phi_2 = \theta - \phi_1$, $1-\phi_1-\phi_2 = 1-\theta$. In particular $X_1 + X_2$ has a Binomial distribution with parameters n and θ and the sample procedure give the same information as a direct question sampling, when no respondent lies. However the closer p is to 1 the lower is the privacy allowance provided by the procedure and consequently the cooperation expected by the respondents. Note also that if $p = \frac{1}{2}$ there is an identifiability problem since $\phi_2 = 1-\phi_1-\phi_2 = \frac{1}{2}(1-\phi_1)$.

The usual computations give the Jeffreys prior distribution for (ϕ_1, ϕ_2):

$$\pi(\phi_1,\phi_2) = \frac{1}{2K_{1-p,p}\left(\frac{1}{2},\frac{1}{2}\right)} \frac{1}{[\phi_1\phi_2(1-\phi_1-\phi_2)]^{1/2}} \tag{1}$$

where $\phi_1 \in [0,1]$ and $\phi_2 \in [(1-p)(1-\phi_1), p(1-\phi_1)]$, $K_{a,b}(\alpha,\beta) = B_b(\alpha,\beta) - B_a(\alpha,\beta)$ and $B_z(\alpha,\beta)$ is the Incomplete Beta function. Equation (1) is also the Reference prior distribution for $\{\phi_1, \phi_2\}$.

The corresponding posterior distribution is

$$\pi(\phi_1,\phi_2|x_1,x_2,n) = \frac{\phi_1^{x_1-\frac{1}{2}}\phi_2^{x_2-\frac{1}{2}}(1-\phi_1-\phi_2)^{n-x_1-x_2-\frac{1}{2}}}{B(x_1+\frac{1}{2}, n-x_1+1)K_{1-p,p}\left(x_2+\frac{1}{2}, n-x_1-x_2+\frac{1}{2}\right)}$$

Since the relation between (θ,ξ) and (ϕ_1,ϕ_2) is one to one (apart on a set of null measure), the Jeffreys and Reference prior for (θ,ξ) can be easily obtained via a Jacobian argument:

$$\pi(\theta,\xi) = \frac{2p-1}{2K_{1-p,p}\left(\frac{1}{2},\frac{1}{2}\right)}\theta^{1/2}\xi^{-1/2}[(1-\xi)\theta p+(1-\theta)(1-p)]^{-1/2} \\ \cdot[(1-\xi)\theta(1-p)+(1-\theta)p]^{-1/2} \tag{2}$$

with $\theta \in [0,1]$ and $\xi \in [0,1]$, and the corresponding posterior distribution is

$$\pi(\theta,\xi|x_1,x_2,n) = \frac{2p-1}{B(x_1+\frac{1}{2},n-x_1+1)K_{1-p,p}\left(x_2+\frac{1}{2},n-x_1-x_2+\frac{1}{2}\right)} \cdot\theta^{x_1+\frac{1}{2}}\xi^{x_1-\frac{1}{2}}[(1-\xi)\theta p+(1-\theta)(1-p)]^{x_2-\frac{1}{2}} \cdot[(1-\xi)\theta(1-p)+(1-\theta)p]^{n-x_1-x_2-\frac{1}{2}}.$$

To evaluate the prior (2) from a frequentist perspective, we compute $P_\gamma(\theta,\xi) = Pr_{\theta,\xi}(\theta \leq \theta_\gamma)$, the frequentist probability that θ_γ is larger than the actual value, where θ_γ is the γth posterior quantile. Table 1 contains $P_{0.95}(\theta,\xi)$ for $n = 100$ and various values of θ, ξ and p. The calculations were done by simulation, generating $1,000$ samples for each (θ, ξ, p). The standard error for the entries in Table 1 is about 0.002. As expected the posterior quantiles perform poorly in frequentist terms near the boundary of the parameter space. However things should be considered in perspective: no satisfactory frequentist methods are available in this case. In the other situations the values of $P_{0.95}(\theta,\xi)$ are close to 0.95, indicating that $\theta_{0.95}$ exceeds θ the correct (from a frequentist point of view) proportion of time. Slightly better performances are obtained when p is larger.

Table 1 Frequentist coverage probabilities of 0.95 posterior quantiles

		$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 0.7$	$\theta = 0.9$
$\xi = 0$	$p = 0.7$	1	0.959	0.956	0.958	0.944
	$p = 0.9$	0.981	0.943	0.956	0.962	0.952
$\xi = 0.1$	$p = 0.7$	1	0.969	0.953	0.952	0.938
	$p = 0.9$	0.984	0.951	0.955	0.952	0.951
$\xi = 0.3$	$p = 0.7$	1	0.981	0.957	0.956	0.945
	$p = 0.9$	0.990	0.954	0.956	0.958	0.950
$\xi = 0.5$	$p = 0.7$	1	0.991	0.964	0.955	0.955
	$p = 0.9$	0.994	0.959	0.954	0.953	0.949
$\xi = 0.7$	$p = 0.7$	1	0.998	0.982	0.964	0.951
	$p = 0.9$	0.997	0.969	0.953	0.952	0.952
$\xi = 0.9$	$p = 0.7$	1	1	0.998	0.992	0.978
	$p = 0.9$	0.999	0.984	0.970	0.961	0.951
$\xi = 1$	$p = 0.7$	1	1	1	0.999	0.999
	$p = 0.9$	0.999	0.992	0.986	0.983	0.980

References

1. Huang, K.-C.: A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. Statistica Neerlandica **58**, 75-82 (2004)
2. Barabesi, L., Marcheselli, M.: Bayesian estimation of proportion and sensitivity level in randomized response procedures. Metrika **72**, 75-88 (2010)
3. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**, 63-69 (1965)