DigitalCommons @WayneState

Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1

Article 17

5-1-2003

Bayesian Analysis Of Poverty Rates: The Case Of Vietnamese Provinces

Dominique Haughton
Bentley College, dhaughton@bentley.edu

Nguyen Phong
General Statistics Office, nphong@gso.gov.vn

Recommended Citation

Haughton, Dominique and Phong, Nguyen (2003) "Bayesian Analysis Of Poverty Rates: The Case Of Vietnamese Provinces," *Journal of Modern Applied Statistical Methods*: Vol. 2 : Iss. 1 , Article 17.
DOI: 10.22237/jmasm/1051748220
Available at: http://digitalcommons.wayne.edu/jmasm/vol2/iss1/17

Journal of Modern Applied Statistical Methods
May, 2003, Vol. 2, No. 1, 189-194

1538 – 9472/03/$30.00

Bayesian Analysis Of Poverty Rates: The Case Of Vietnamese Provinces

Dominique Haughton
Bentley College, USA

Nguyen Phong
General Statistics Office, Vietnam

This paper presents a Bayesian analysis of poverty rates in urban Ho Chi Minh City and rural Nghe An province in Vietnam. Using mixtures of beta distributions as priors for the poverty rates, we find that, when the prior is reasonably informative, our approach yields more accurate estimated poverty rates than a frequentist approach. On the other hand, we find that, in the presence of poor/non-poor misclassification, average probabilities of posterior credible intervals for poverty rates can fall well short of .95 even with sample sizes such as 2000 or 3000 when the width of the interval is for example four percentage points. In general, we suggest reporting prior and posterior means and standard deviations along with traditional frequentist measures. Our results rely on techniques due to Nandram and Sedransk (1993) and Rahme, Joseph and Gyorkos (2000), and make use of the software WINBUGS.

Key words: Vietnamese poverty, Bayesian analysis, WINBUGS

Introduction

The problem of estimating the binomial parameter has attracted a lot of attention among statisticians and others in the business of estimating proportions. It is widely known that, informally speaking, large sample sizes are needed to get acceptable accuracies when estimating proportions.

Sample size estimations are often based on classical computations of confidence intervals, sometimes adjusted to take into account special survey designs. Recent work of Brown (2001) has focused attention on the shortcomings of such confidence intervals, notably on the fact that “95% confidence intervals” have less than 95% coverage in a number of cases.

In the context of the estimation of poverty rates, we are led to the estimation of a binomial parameter, since the poverty rate is in general defined as the proportion of households whose annual expenditure per capita falls below a given poverty line. In most of this paper we will assume that this poverty line is non-random, and that the classification poor/non-poor is known accurately. We will discuss the implications of an inaccurately known poverty line in the latter part of the paper.

The estimation of poverty rates for Vietnamese provinces lends itself very well to a Bayesian analysis: informative prior information is frequently available; moreover sample sizes tend to be fairly small, since surveys are expensive and prone to non-sampling errors. Sampling statisticians and others involved in the design and analysis of such surveys (in Vietnam or elsewhere) have to date not performed a Bayesian analysis of poverty rates (see, for example, Glewwe & Yansaneh, 2001, for an exposition of a typical

Dominique Haughton is Professor of Mathematical Sciences at Bentley College. Her research interests include model selection, statistics applied to marketing and the analysis of living standards in Vietnam. With Jonathan Haughton and Nguyen Phong, she is a co-editor of the book “Living standards during an economic boom: the case of Vietnam”. Email her at DHAUGHTON@bentley.edu. Nguyen Phong is the director of the Social and Environmental Department in the General Statistics Office in Hanoi, Vietnam. He leads the implementation of large nationwide surveys on living standards. With Dominique Haughton and Jonathan Haughton, he is a co-editor of the book “Living standards during an economic boom: the case of Vietnam”. Email him at nphong@gso.gov.vn.

analysis in this context).

We will show in this article that a gain in accuracy is obtained when a reasonably informative prior is used, and when the poverty line is assumed known. We will illustrate this result with a wealthier urban sample (urban Ho Chi Minh City), and a poorer rural sample (rural Nghe An). However, to qualify these results, one should keep in mind that when poor/non-poor misclassification occurs, as it almost certainly does, the average coverage of four-percentage-point-wide probability intervals does not reach .95, even asymptotically in large sample sizes, while it is likely to do so for an eight-percentage-point-wide probability interval.

Methodology

Bayesian Estimation Of Poverty Rates When The Poverty Line Is Known

In urban Ho Chi Minh City, our sample from the Vietnam Living Standards Survey of 1998 has 433 households, 2 of which are poor. Frequentist weighted (according to sampling weights) computations yield a poverty rate of .00462, with a standard deviation of .00334 (yielding a coefficient of variation of about .7). In order to perform the Bayesian analysis, we use a mixture of beta distributions as a prior for the unknown poverty rate as suggested in Nandram and Sedransk. This is justified by the work of Dalal and Hall (1983), who showed that any prior can be approximated by such a mixture. We then apply the closed form formulas of Nandram and Sedransk for the posterior mean and posterior standard deviation of the poverty rate for a two-stage cluster sample design.

In our case, we assume that a commune is randomly selected, then a household randomly selected from the commune; in reality there is an additional step in the sampling design – a village is randomly selected from the commune – and then a household is randomly selected from the village. We expect to address the issue of three-level clustering in future work; no closed form formula is available in this case for the posterior mean and standard deviation of the poverty rate. The present analysis is a close approximation of reality, though; we don't expect the addition of the third level to make a large difference. We then simulate the posterior distribution using WINBUGS, with the code published in Congdon (2001; example 5.18 p. 196). In addition to the data on poor/non-poor households from surveyed communes, the analysis makes use of the number of households in each commune of urban Ho Chi Minh City and rural Nghe An respectively; the model specifies an individual poverty rate for each commune and then combines these poverty rates into an overall poverty rate for the province.

Results

In Table 1 and Figure 1, we present the results from four different priors for urban Ho Chi Minh City. In Table 2 and Figure 2, we present the results from two different priors for rural Nghe An. The posterior means and standard deviations are those of the overall poverty rate for the whole area (urban Ho Chi Minh City and rural Nghe An, respectively). The mixture of beta distributions used for the prior for a vector θ of N poverty rates for N communes is given by Nandram and Sedransk (1993) as:

$$\pi(\theta \mid \tau) = \sum_{r=1}^{R} \omega_r B(a_r, \tau - a_r)^{-N} \prod_{k=1}^{N} \theta_k^{a_r - 1} (1 - \theta_k)^{\tau - a_r - 1} \quad ,$$

where θ_k is the poverty rate for the kth province, and B denotes the Beta function. The values of ω_r, a_r and τ must be chosen when eliciting the prior. Note that the means of the beta distributions in the mixture are a_r/τ, and that the value of τ controls the standard deviation of the beta distributions; the higher τ is, the smaller the standard deviation.

The two first priors for urban Ho Chi Minh City are based loosely on poverty rates and their standard deviations for Vietnamese provinces estimated in Baulch et al. (2002), using data from the Census of 1999 and regression equations based on VLSS data. These estimates were used to define 4 bins centered at the values indicated in the column "Mean" in Table 1 for each of 4 components, and prior probabilities of .07, .43, .43 and .07 for each of the 4 bins. Note that the value of 4 for R was chosen somewhat arbitrarily for convenience and flexibility. Priors 1 and 2 differ by the value of τ, and thus by the standard deviations.

The components are less separated in prior 2, as seen on Figure 1. The results from both priors are close, a posterior poverty rate of about .01, with a standard deviation of about .005, an improvement (coefficient of variation of about .5) over the frequentist estimation. Figure 1 shows the two posterior densities from priors 1 and 2 to be close, and to give most of the posterior probability to two components, conceivably corresponding to more and less affluent communes. Prior 3 corresponds to a prior elicited from the expert opinion that "we are 95% certain that the poverty rate for urban Ho Chi Minh City is between .01 and .03". As for priors 1 and 2, 4 bins were created for prior 3, centered at values given in Table 1 and with widths consistent with the expert prior belief. The summary statistics for the posterior poverty rate are quite similar to those for priors 1 and 2. Prior 4 is a very diffuse prior, and in this case, the posterior poverty rate is not accurate (standard deviation of .008) as can be expected.

In this case, we have both closed form expressions for the posterior means and standard deviations, as well as the option of using WINBUGS to generate a sample from the posterior. The results from both analyses should be, and are, close. We note here that we have found that if the beta components are too well separated or if one of the components is too close to 0, it can happen that the MCMC chain in WINBUGS gets "stuck" in a component, and gives an incorrect posterior mean. This in fact is not surprising to the authors of WINBUGS (N. Best, personal communication), and could be remedied by checking the WINBUGS results against the closed form formulas for a two-level cluster sample design for a given prior, and then moving on to more complex survey designs if desired.

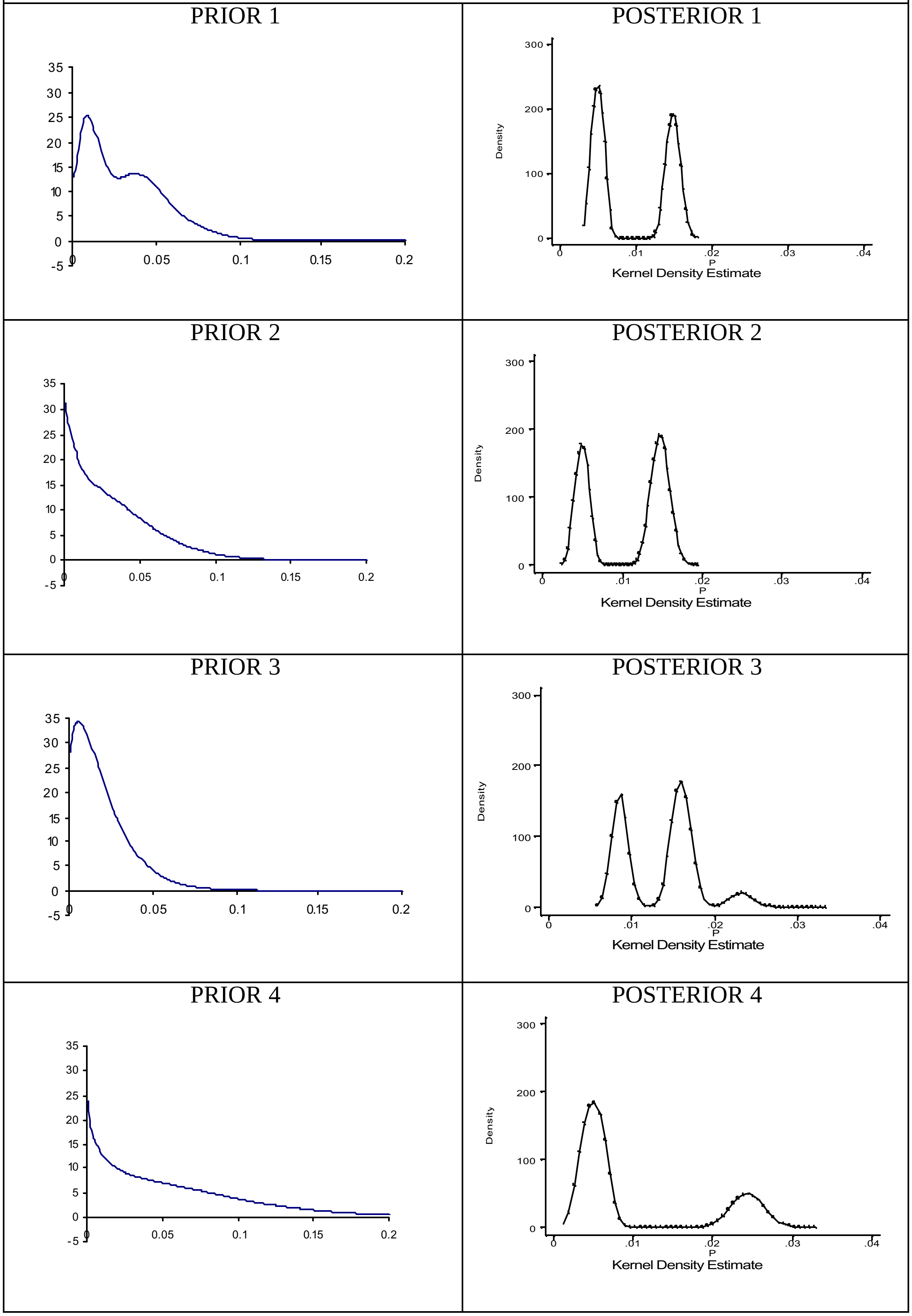
Table 1: Prior And Posterior Means And Standard Deviations; Ho Chi Minh City Urban

		Prior 1, τ = 200		Prior 2 τ, = 80	
ω_i		***Mean***	***St. Dev.***	***Mean***	***St. Dev.***
.07	Comp. 1	.005	.005	.005	.008
.43	Comp. 2	.015	.009	.015	.014
.43	Comp. 3	.045	.015	.045	.023
.07	Comp. 4	.075	.019	.075	.029
	Overall	.031	.023	.031	.027
		Post. Mean	Post. St. Dev	Post. Mean	Post. St. Dev.
	Closed form	.009872	.004982	.010765	.004911
	Winbugs	.009664	.004964	.010611	.004910

Table 1 (continued)

		Prior 3, τ = 80		Prior 4, τ = 40	
ω_i		***Mean***	***St. Dev.***	***Mean***	***St. Dev.***
.07	Comp. 1	.009	.010	.005	.011
.43	Comp. 2	.016	.014	.025	.024
.43	Comp. 3	.024	.017	.080	.042
.07	Comp. 4	.031	.019	.140	.054
	Overall	.020	.017	.055	.051
		Post. Mean	Post. St. Dev	Post. Mean	Post. St. Dev.
	Closed form	.013684	.004561	.008841	.007801
	Winbugs	.013530	.004508	.010130	.008632

FIGURE 1: PRIOR DENSITIES AND POSTERIOR KERNEL DENSITIES;
HO CHI MINH CITY URBAN



PRIOR 1

POSTERIOR 1

PRIOR 2

POSTERIOR 2

PRIOR 3

POSTERIOR 3

PRIOR 4

POSTERIOR 4

Table 2: Prior And Posterior Means And Standard Deviations; Nghe An Rural

		Prior 1, $\tau = 40$		Prior 2, $\tau = 30$	
ω_i		***Mean***	***St. Dev.***	***Mean***	***St. Dev.***
.07	Comp. 1	.225	.065	.050	.039
.43	Comp. 2	.375	.076	.125	.059
.43	Comp. 3	.525	.078	.275	.080
.07	Comp. 4	.675	.073	.425	.089
	Overall	.450	.133	.205	.122
		Post. Mean	Post. St. Dev	Post. Mean	Post. St. Dev.
	Closed form	.499810	.055138	.424697	.008203
	Winbugs	.503400	.051560	.424500	.009934

For rural Nghe An, we have 225 sampled households, among which 110 are poor. Weighted frequentist estimations give an estimated poverty rate of .489, with a standard deviation of .104. Prior 1 is again based loosely on the estimations in Baulch et al. (2002); it yields a posterior mean for the poverty rate of about .5, with a posterior standard deviation of .05, an improvement in accuracy over the frequentist analysis.

Prior 2 is based on an estimated poverty rate of about .2 from MOLISA (Ministry of Labour, Invalids and Social Affairs), used to create 4 bins of about the same width as in prior 1. The prior poverty rate of .2 is probably too low, and it is interesting to see how the Bayesian analysis uses the data to correct this prior information: the MCMC chain concentrates almost exclusively on one higher component to yield a posterior mean of .42 with a standard deviation of about .01 for the poverty rate.

Bayesian Estimation Of Sample Sizes In The Presence Of Misclassification

We now consider the case where it is in fact not known exactly which households are poor and which are not. Poverty lines are difficult to establish, in large part because of the difficulty in getting accurate data on the prices of basic goods. So the problem of identifying poor households is similar to the problem of diagnosing a disease on the basis of an imperfect test.

We use here work of Rahme et al. (2000) where Bayesian sample size determinations are performed for the binomial parameter subject to misclassification, and applied to a situation in the medical area. In this context, the test for poverty has a sensitivity (probability of a poor household being classified as poor) and a specificity (probability of a non-poor household being classified as non-poor), both with a beta prior distribution following Rahme et al. (2000), and the prevalence of poverty (the poverty rate) is also given a beta prior distribution.

We illustrate this approach in the case of rural Nghe An. We define a prior distribution of a beta with parameters $\alpha = 70.32$ and $\beta = 77.1$ for the poverty rate, on the basis of the estimates for the poverty rate and its standard deviation in Baulch et al. (2002), and elicit beta distributions as priors for the sensitivity and specificity of the poor/non-poor classification from the opinion that the mean sensitivity (and specificity) is about .95 and that we are 95% certain that the sensitivity (or specificity) is between .9 and 1. This opinion yields the values for the beta parameters given in Table 3.

The table gives average coverages of probability intervals for two different interval widths and three different sample sizes, calculated from an S-plus program made available by Rahme (2000) et al. It is clear that the coverage will not attain .95 for a width of 4 percentage points, even with very large sample sizes. Such a coverage might be feasible with an interval of width .08, with large sample sizes. However, we note that the techniques in Rahme et al. (2002) assume i.i.d. samples, so the situation is likely to be somewhat worse in a situation where a more complex survey design was used. We also note that less informative priors on the poverty rates and/or the sensitivity and the specificity of the poor/non-poor classification would be likely to yield even smaller average coverage probabilities.

Table 3: Average Coverage Of Probability Intervals For Poverty Rates For Nghe An Rural Assuming I.I.D. Samples

<table>
<tr><th colspan="3">$\alpha_{sens} = \alpha_{spe\ c} = 71.25; \beta_{sens} = \beta_{spe\ c} = 3.75; \alpha = 70.32; \beta = 77.1$</th></tr>
<tr><th>Width of interval</th><th>Sample size</th><th>Prob. coverage</th></tr>
<tr><td>.04</td><td>1000</td><td>.6439</td></tr>
<tr><td>.04</td><td>2000</td><td>.6924</td></tr>
<tr><td>.04</td><td>3000</td><td>.6995</td></tr>
<tr><td>.08</td><td>1000</td><td>.9261</td></tr>
<tr><td>.08</td><td>2000</td><td>.9471</td></tr>
<tr><td>.08</td><td>3000</td><td>.9587</td></tr>
</table>

Conclusion

We have shown in this paper the benefits of a Bayesian approach to the estimation of poverty rates. Poverty rates are often calculated – and reported – as sample proportions. In some cases, a measure of accuracy such as standard deviation is reported as well.

In our analyses, the use of sensible prior information has provided a significant improvement in the accuracy of the poverty rates, as measured by their posterior standard deviation, provided that the poverty line is known exactly. Results tend to be robust with respect to the choice of a sensible prior.

Our Bayesian analysis has also shown that whenever there is uncertainty in the poor/non-poor classification, the accuracy of poverty rates as measured by the width of posterior credible intervals is significantly negatively affected. For example, coverage probabilities of about 95% may require interval widths of about 8 percentage points, implying poverty rates known only up to four percentage points.

In general we suggest that posterior means and standard deviations be reported along with more traditional measures, and that a discussion of the accuracy of poverty lines accompany poverty rates reports.

References

Baulch, B. & Minot, N. (2002). The spatial distribution of poverty in Vietnam and the potential for targeting. *World Bank working paper* 2829, Washington, DC.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101-133.

Congdon, P (2001). *Bayesian statistical modelling.* NY:Wiley.

Dalal, S. R., & Hall, W. J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society, B45*(2), 287-286.

General Statistical Office (1999). *Vietnam living standards survey,* Hanoi, Vietnam.

Glewwe, P., & Yansaneh, I. (2001). *Mission report: Recommendations for multi-purpose household surveys from 2002 to 2010*. World Bank, Washington, DC.

Nandram, B., & Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society, B55*(2), 399-408.

Rahme, E., Joseph L., & Gyorkos, T. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics, 49*, 119-128.