Biometrika (1988) 75, 2, *pp.* 223-8
Printed in Great Britain

Inference for the binomial *N* parameter: A hierarchical Bayes approach

By ADRIAN E. RAFTERY

Department of Statistics, University of Washington, Seattle, Washington 98195, U.S.A.

Summary

A hierarchical Bayes approach to the problem of estimating N in the binomial distribution is presented. This provides a simple and flexible way of specifying prior information, and also allows a convenient representation of vague prior knowledge. It yields solutions to the problems of interval estimation, prediction and decision making, as well as that of point estimation. The Bayes estimator compares favourably with the best, previously proposed, point estimators in the literature.

Some key words: Bayes estimator; Decision making; Highest posterior density region; Population size estimation; Prediction; Vague prior.

1. Introduction

Suppose $x = (x_1, \ldots, x_n)$ is a set of success counts from a binomial distribution with unknown parameters N and θ. Most of the literature about statistical analysis of this model has focused on point estimation of N, while interval estimation, prediction and decision making have been little considered; see § 2.

I adopt a hierarchical Bayes approach. This provides a simple way of specifying prior information, and also allows a convenient representation of vague prior knowledge using limiting, improper, prior forms. It leads to solutions of the problems of interval estimation, prediction and decision making, as well as that of point estimation.

A difficulty with Bayesian analysis of this problem has been the absence of a sufficiently flexible and tractable family of prior distributions, mainly due to the fact that N is an integer. The present approach gets around this by first assuming that N has a Poisson distribution. The resulting hyperparameters are then continuous-valued, and one may use existing results about conjugate and vague priors in better understood settings.

I assume that N has a Poisson distribution with mean μ. Then $x_1, \ldots, x_n$ have, jointly, an exchangeable distribution such that, marginally, each x_i has a Poisson distribution with mean $\lambda = \mu\theta$. I specify the prior distribution in terms of (λ, θ) rather than (μ, θ). This is because, if the prior is based on past experience, it would seem easier to formulate prior information about λ, the unconditional expectation of the observations, than about μ, the mean of the unobserved quantity N. If this is so, the prior information about λ would be more precise than that about μ or θ, so that it may be more reasonable to assume λ and θ independent a priori than μ and θ. In this case, μ and θ would be negatively associated a priori.

The posterior distribution of N is for $N \geqslant x_{\max}$

$$p(N|x) \propto (N!)^{-1} \left\{ \prod_{i=1}^{n} \binom{N}{x_i} \right\} \int_0^1 \int_0^\infty \theta^{-N+S} (1-\theta)^{nN-S} \lambda^N \exp(-\lambda/\theta) p(\lambda, \theta)\, d\lambda\, d\theta, \tag{1·1}$$

where $S = x_1 + \ldots + x_n$ and $x_{\max} = \max(x_1, \ldots, x_n)$.

I now consider the case where vague prior knowledge about the model parameters is represented by limiting, improper, prior forms. I use the prior $p(\lambda, \theta) \propto \lambda^{-1}$, which is the product of the standard vague prior for λ (Jaynes, 1968) with a uniform prior for θ. This leads to the same solution as if a similar vague prior were used for (μ, θ), namely $p(\mu, \theta) \propto \mu^{-1}$. It is also equivalent to the prior $p(N, \theta) \propto N^{-1}$. The posterior is for $N \geqslant x_{\max}$

$$p(N \mid x) \propto \left\{ \frac{(nN-S)!}{(nN+1)!\,N} \right\} \left\{ \prod_{i=1}^{n} \binom{N}{x_i} \right\}. \tag{1·2}$$

When $n = 1$, expression (1·2) becomes

$$p(N \mid x) = x_1 / \{N(N+1)\} \quad (N \geqslant x_1).$$

Thus, when $n = 1$, the posterior median is $2x_1$. The same solution was obtained by Jeffreys (1961, § 4.8) to the related problem of estimating the number of bus lines in a town, having seen the number of a single bus. He argued that this was an intuitively reasonable solution, and it seems to be so in this case also.

If λ and θ are independent a priori, and λ has a gamma prior distribution, so that $p(\lambda, \theta) \propto \lambda^{\kappa_1 - 1} e^{-\kappa_2 \lambda} p(\theta)$, then λ can be integrated out analytically, and (1·1) becomes for $N \geqslant x_{\max}$

$$p(N \mid x) \propto (N!)^{-1} \Gamma(N + \kappa_1) \left\{ \prod_{i=1}^{n} \binom{N}{x_i} \right\} \int_0^1 \theta^{-N+S} (1-\theta)^{nN-S} (\theta^{-1} + \kappa_2)^{-(N+\kappa_1)} p(\theta)\, d\theta.$$

2. POINT ESTIMATION

Point estimation of N was first considered by Haldane (1942), who proposed the method of moments estimator, and Fisher (1942), who derived the maximum likelihood estimator, also used by Moran (1951). DeRiggi (1983) showed that the relevant likelihood function is unimodal. However, Olkin, Petkau & Zidek (1981) showed that both these estimators can be unstable in the sense that a small change in the data can cause a large change in the estimate of N. Under some circumstances a reasonable confidence set contains all large values of N, and it is no surprise that point estimators are unstable. What is wrong is asking for a point estimate except in a specific decision making context. In this spirit, later, I derive Bayes estimators which correspond to specific loss functions, and hence, implicitly, to specific decision problems.

Olkin et al. (1981) introduced modified estimators and showed that they are stable. On the basis of a simulation study, they recommended a stabilized method of moments estimator which they called MME:S, and which I denote here by $\hat{N}_{\text{MME:S}}$. Casella (1986) suggested a more refined way of deciding whether or not to use a stabilized estimator. Kappenman (1983) introduced the 'sample reuse' estimator; this performed similarly to $\hat{N}_{\text{MME:S}}$ in a simulation study, and is not further considered here. Dahiya (1980) used a closely related but different model to estimate the population sizes of different types of organism in a plankton sample by the maximum likelihood method; he did not investigate the stability of his estimators.

Draper & Guttman (1971) adopted a Bayesian approach, and gave a full solution for the case where N and θ are independent a priori, the prior distribution of N is uniform with a known upper bound, and that of θ is beta. Blumenthal & Dahiya (1981) suggested N^* as an estimator of N, where (N^*, θ^*) is the joint posterior mode of (N, θ) with the

Draper-Guttman prior. However, they did not say how the parameters of the beta prior for θ should be chosen. Carroll & Lombard (1985) recommended as an N estimator the posterior mode of N with the Draper-Guttman prior after integrating out θ, where the prior of θ has the form $p(\theta) \propto \theta(1-\theta)$ $(0 \leqslant \theta \leqslant 1)$. They called this estimator 'Mbeta(1, 1)'; here I denote it by $\hat{N}_{\text{MB}(1,1)}$. The Draper-Guttman prior has been criticized by Kahn (1987); see § 5. The simpler problem of estimating N where θ is known has been addressed by Feldman & Fox (1968), Hunter & Griffiths (1978) and Sadooghi-Alvandi (1986).

Bayes estimators of N may be obtained by combining (1·2) with appropriate loss functions; examples are the posterior mode of N, $\hat{N}_{\text{MOD}}$, and the posterior median of N, $\hat{N}_{\text{MED}}$. Previous authors, including Olkin et al. (1981), Carroll & Lombard (1985) and Casella (1986) have agreed that the relative mean squared error of an estimator $\hat{N}$, equal to $E\{(\hat{N}/N-1)^2\}$, is an appropriate loss function for this problem. The Bayes estimator corresponding to this loss function is

$$\hat{N}_{\text{MRE}} = \sum_{N=x_{\max}}^{\infty} N^{-1} p(N|x) \Big/ \sum_{N=x_{\max}}^{\infty} N^{-2} p(N|x).$$

The three Bayes estimators, $\hat{N}_{\text{MOD}}$, $\hat{N}_{\text{MED}}$ and $\hat{N}_{\text{MRE}}$, are reasonably stable, as can be verified by calculating them for the eight particularly difficult cases listed in Table 2 of Olkin et al. (1981). Also, $\hat{N}_{\text{MED}}$ is closer to the true value of N than the other estimators considered in four of these eight cases, while $\hat{N}_{\text{MOD}}$ is best in a further three cases. However, in the cases in which $\hat{N}_{\text{MOD}}$ is best, $\hat{N}_{\text{MED}}$ performs poorly; the converse is also true. The other three estimators fall between $\hat{N}_{\text{MOD}}$ and $\hat{N}_{\text{MED}}$ in all eight cases.

The results of a Monte Carlo study are shown in Table 1. I used the same design as Olkin et al. (1981) and Carroll & Lombard (1985). In each replication, N, θ and n were generated from uniform distributions on $\{1, \ldots, 100\}$, $[0, 1]$ and $\{3, \ldots, 22\}$ respectively. There were 2000 replications. Table 1 shows that $\hat{N}_{\text{MRE}}$ performed somewhat better than $\hat{N}_{\text{MME:S}}$ and $\hat{N}_{\text{MB}(1,1)}$ in both stable and unstable cases, with an overall efficiency gain of about 10% over $\hat{N}_{\text{MME:S}}$, and about 6% over $\hat{N}_{\text{MB}(1,1)}$. Here, following Olkin et al. (1981), a sample is defined to be stable if $\bar{x}/s^2 \geqslant 1 + 1/\surd 2$, and unstable otherwise, where $\bar{x} = \Sigma\, x_i/n$ and $s^2 = \Sigma\,(x_i - \bar{x})^2/n$.

Table 1. *Relative mean squared errors of the N estimators*

			Estimators	
Cases	No.	$\hat{N}_{\text{MME:S}}$	$\hat{N}_{\text{MB}(1,1)}$	$\hat{N}_{\text{MRE}}$
All cases	2000	0·171	0·165	0·156
Stable cases	1378	0·108	0·104	0·100
Unstable cases	622	0·312	0·300	0·281

3. Interval estimation

The posterior distribution of N given by (1·1) or (1·2) yields Bayesian estimation intervals for N, such as highest posterior density regions. Expression (1·1) is an exact Bayesian solution, while (1·2) provides an approximation to the exact solution when the prior information is vague. Highest posterior density regions based on (1·2) are always intervals. To my knowledge, no other interval estimator of N has been explicitly proposed.

In order to check the quality of the approximation provided by (1·2), note that, if the prior distribution also represents a distribution of values of the unknown parameters typical of those that occur in practice, then the average confidence coverage of the Bayesian interval is equal to its posterior probability (Rubin & Schenker, 1986). I therefore carried out a Monte Carlo study, designed in the same way as that reported in § 2. The empirical coverage of the 80% highest posterior density region was 0·82, that of the 90% interval was 0·91, and that of the 95% interval was 0·95.

The distribution from which N was simulated has a much shorter tail than the prior on which (1·2) is based, although it is fairly diffuse. Thus, these results are evidence that (1·2) does provide a reasonable approximation to an exact Bayesian solution when prior information is vague. It also supports the use of highest posterior density regions based on (1·2) as frequentist interval estimators. The interval estimators based on (1·2) are also reasonably stable, as can be verified by calculating them for the eight particularly difficult data sets of Olkin et al. (1981).

4. EXAMPLES

Carroll & Lombard (1985) analysed two examples, involving counts of impala herds and individual waterbuck. The observed numbers of impala herds were 15, 20, 21, 23 and 26. The observed numbers of waterbucks were 53, 57, 66, 67 and 72. The point and interval estimators are shown in Table 2. The stability of the Bayes estimators is again apparent; the stability of $\hat{N}_{\text{MRE}}$ for the waterbuck example is noteworthy given the highly unstable nature of this data set.

Table 2. *Point estimators and* 80% *highest posterior density regions for the impala and waterbuck examples*: *original and perturbed samples*

	Point estimators					Limits of 80% region	
Example	$\hat{N}_{\text{MME:S}}$	$\hat{N}_{\text{MB(1,1)}}$	$\hat{N}_{\text{MOD}}$	$\hat{N}_{\text{MED}}$	$\hat{N}_{\text{MRE}}$	Lower	Upper
Impala	54	42	37	67	49	26	166
	63	46	40	76	54	28	193
Waterbuck	199	140	122	223	131	80	598
	215	146	127	232	132	82	636

For each example, first entries are N estimates for original sample; second entries are N estimates for perturbed sample obtained by adding one to largest success count.

The posterior distributions obtained from (1·2) are shown in Fig. 1. The posterior distribution for the waterbuck example has a very long tail; this may be related to the extreme instability of this data set.

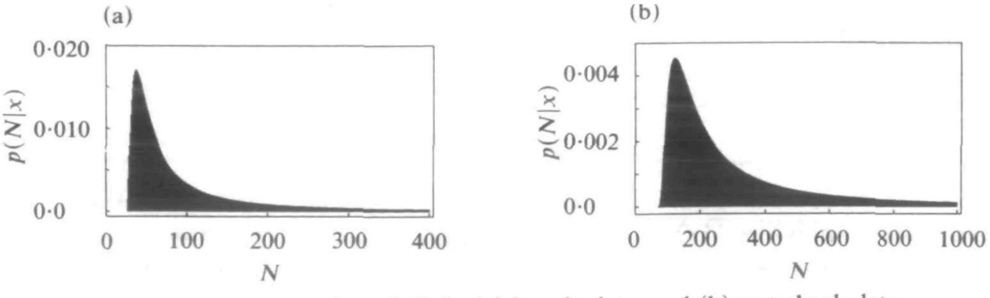

Fig. 1. Posterior distribution of N for (a) impala data, and (b) waterbuck data.

5. Discussion

The present approach can be used to solve the prediction problem. For example, the predictive distribution of a future observation, x_{n+1}, is simply

$$p(x_{n+1}|x) \propto \sum_{N=x_{\max}}^{\infty} \int_0^1 p(x_{n+1}, x | N, \theta) p(N, \theta)\, d\theta.$$

When the vague prior which leads to (1·2) is used, this becomes

$$p(x_{n+1}|x) \propto \sum_{N=x'_{\max}}^{\infty} \frac{S'!\{(n+1)N - S'\}!}{\{(n+1)N+1\}!\,N} \left\{ \prod_{i=1}^{n} \binom{N}{x_i} \right\},$$

where $S' = S + x_{n+1}$ and $x'_{\max} = \max(x_{\max}, x_{n+1})$.

No other solution to the prediction problem has, to my knowledge, been explicitly proposed in the literature. A standard non-Bayesian approach would be to use the predictive distribution conditional on point estimators of N and θ. As a general method, prediction conditional on the estimated values of the unknown parameters is widespread, and underlies, for example, the time series forecasting methods of Box & Jenkins (1976). For the present problem, however, it yields predictive distributions which are unsatisfactory because they attribute zero probability to possible outcomes.

The present approach also yields a full solution to the decision-making problem, by the usual method of minimizing posterior expected loss. It may often be easier to specify loss or utility in terms of future outcomes than of values of N, so that a predictive approach to loss specification may be helpful here.

Kahn (1987) has pointed out that in any Bayesian analysis of this problem, the asymptotic tail behaviour of the posterior distribution of N is determined by the prior. This is not, of course, the same as saying that inferences about N are determined by the prior. Indeed, in § 4, we have seen examples where different data lead to very different conclusions about N, in spite of the priors being the same, and the data sets being small, and of the same size. Kahn (1987) also pointed out that the posterior resulting from the prior used by Draper & Guttman (1971) depends crucially on the, assumed known, prior upper bound for N, contrary to a comment of Draper and Guttman (1971). The vague prior used here does not appear to suffer from such a drawback.

Acknowledgements

I am grateful to G. Casella, D. R. Cox, P. Guttorp, W. S. Jewell, D. V. Lindley, A. J. Petkau, C. E. Smith, P. J. Smith, P. Turet and the referee for very helpful comments and discussions. This research was supported by the Office of Naval Research.

References

Blumenthal, S. & Dahiya, R. C. (1981). Estimating the binomial parameter *n*. *J. Am. Statist. Assoc.* **76**, 903-9.

Box, G. E. P. & Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control*, 2nd ed. San Francisco: Holden-Day.

Carroll, R. J. & Lombard, F. (1985). A note on *N* estimators for the binomial distribution. *J. Am. Statist. Assoc.* **80**, 423-6,

Casella, G. (1986). Stabilizing binomial *n* estimators. *J. Am. Statist. Assoc.* **81**, 172-5.

Dahiya, R. C. (1980). Estimating the population sizes of different types of organisms in a plankton sample. *Biometrics* **36**, 437-46.

DERIGGI, D. F. (1983). Unimodality of likelihood functions for the binomial distribution. *J. Am. Statist. Assoc.* **78**, 181-3.

DRAPER, N. & GUTTMAN, I. (1971). Bayesian estimation of the binomial parameter. *Technometrics* **13**, 667-73.

FELDMAN, D. & FOX, M. (1968). Estimation of the parameter *n* in the binomial distribution. *J. Am. Statist. Assoc.* **63**, 150-8.

FISHER, R. A. (1942). The negative binomial distribution. *Ann. Eugenics* **11**, 182-7.

HALDANE, J. B. S. (1942). The fitting of binomial distributions. *Ann. Eugenics* **11**, 179-81.

HUNTER, A. J. & GRIFFITHS, H. J. (1978). Bayesian approach to estimation of insect population size. *Technometrics* **20**, 231-4.

JAYNES, E. T. (1968). Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.* SSC-**4**, 227-41.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Clarendon.

KAHN, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter *n*. *Am. Statistician* **41**, 38-9.

KAPPENMAN, R. F. (1983). Parameter estimation via sample reuse. *J. Statist. Comp. Simul.* **16**, 213-22.

MORAN, P. A. P. (1951). A mathematical theory of animal trapping. *Biometrika* **38**, 307-11.

OLKIN, I., PETKAU, J. & ZIDEK, J. V. (1981). A comparison of *n* estimators for the binomial distribution. *J. Am. Statist. Assoc.* **76**, 637-42.

RUBIN, D. B. & SCHENKER, N. (1986). Efficiently simulating the coverage properties of interval estimates. *Appl. Statist.* **35**, 159-67.

SADOOGHI-ALVANDI, S. M. (1986). Admissible estimation of the binomial parameter *n*. *Ann. Statist.* **14**, 1634-41.

[*Received June* 1987. *Revised September* 1987]