The Statistician (1995)
44, *No.* 2, *pp.* 167–171

Some comments on Bayesian sample size determination

By LAWRENCE JOSEPH†,

Montreal General Hospital and McGill University, Montreal, Canada

DAVID B. WOLFSON

McGill University, Montreal, Canada

and ROXANE DU BERGER

Montreal General Hospital, Canada

[Received May 1994. Revised September 1994]

SUMMARY

Several criteria for Bayesian sample size determination have recently been proposed. Criteria based on highest posterior density (HPD) intervals from the exact posterior distribution in general lead to smaller sample sizes than those based on non-HPD intervals and/or normal approximations to the exact density. The economies are variable, however, and depend both on the prior inputs and the desired posterior accuracy and coverage probability. In our reply we review several properties of sample size methods and discuss the importance of these properties in the context of a binomial experiment. A general algorithm for Bayesian sample size determination that is useful for more complex sampling situations based on Monte Carlo simulations is briefly described.

1. Introduction

We are grateful to Dr Adcock and Professor Pham-Gia for their commentaries, and to the Editor of *The Statistician* for providing us with the opportunity to respond. Our position is this: HPD regions *always* have the smallest volume for a given credibility level $1 - \alpha$. Sample size calculations for binomial experiments should therefore be based on HPD criteria. The complexity of the computations required to meet such criteria is no longer an issue as the software now exists for this problem and calculations can be carried out very quickly. Other methods should be viewed as a compromise until the HPD software becomes widely available. To obtain a copy of the Fortran software, send the message 'send bhpd1 from general' to statlib@lib.stat.cmu.edu. We shall elaborate on this and other points.

2. Bayesian sample size determination for experiments with binomial outcomes

For convenience in Table 1 we summarize the references on Bayesian binomial sample size determination along with the criteria on which these determinations are based:

(a) $\mathscr{P}_1$—the criterion is satisfied on average over the preposterior distribution of the data;
(b) $\mathscr{P}_2$—the criterion is conservative, in that it is satisfied even if the worst possible outcome occurs.

†*Address for correspondence*: Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada.
E-mail: joseph@binky.epi.mcgill.ca

0039–0526/95/44167

TABLE 1
Properties of the various sample size criteria

Criterion	*Reference*	*Property*						
		$\mathscr{P}_1$	$\mathscr{P}_2$	$\mathscr{P}_3$	$\mathscr{P}_4$	$\mathscr{P}_5$	$\mathscr{P}_6$	$\mathscr{P}_7$
PGT(i)	Pham-Gia and Turkkan (1992)		√			√		
PGT(ii)	Pham-Gia and Turkkan (1992)	√		√		√		
CA87	Adcock (1987)		√				√	
CA95(i)	Adcock (1995)	√		√		√		
CA95(ii)	Adcock (1995)	√			√	√		
ALC	Joseph *et al.* (1995)	√		√		√		√
ACC	Joseph *et al.* (1995)	√			√	√		√
WOC	Joseph *et al.* (1995)		√			√		√

Criteria that average over the preposterior distribution of the data $\mathscr{P}_1$ can be further subdivided into

(i) $\mathscr{P}_3$—averages of variable lengths of fixed coverage intervals or
(ii) $\mathscr{P}_4$—averages of variable coverages of fixed length.

Other properties of interest are

(a) $\mathscr{P}_5$—the criterion is based on the exact posterior distribution,
(b) $\mathscr{P}_6$—the criterion is based on an approximation to the posterior distribution and
(c) $\mathscr{P}_7$—HPD intervals are used in calculating interval lengths and coverages.

Decision theoretic and sample information criteria are omitted from the discussion. The labelling scheme of Adcock (1992) has been retained in Table 1 with the addition of CA95(i) and CA95(ii), which indicate the criteria defined by equations (15) and (17) of Adcock (1995) respectively.

Clearly it is theoretically advantageous to use exact rather than approximate densities in any statistical calculation, and HPD intervals are highly desirable owing to their minimum length property. We agree with Adcock (1995) that there is a need for balance between accuracy and complexity in any given problem. However, the use of exact HPD intervals does not entail any conceptual difficulties in the present problem, whereas the computational complexities have been overcome in Joseph *et al.* (1995). As the beta(a, b) density is asymmetric whenever $a \neq b$, surely the use of methods based on the normal approximation could lead, unnecessarily, to a loss of efficiency. In particular, for *a priori* very rare or very common outcomes, the use of the exact posterior density along with HPD regions has been shown to result in considerable sample size reductions (Joseph *et al.*, 1995). When the two prior beta parameters are approximately equal, Adcock (1995) has shown in his Table 1 that reductions in sample size are essentially due to the use of an exact posterior distribution, after which it makes little difference whether HPD or symmetric intervals are used. This can be explained by the symmetry of the preposterior distributions when the prior parameters are assumed equal; the xs sampled from this distribution will produce a family of posterior distributions, which will be highly asymmetric for x close to 0 or n. However, most of the xs will occur away from these extremes, where the resulting posterior distributions will be approximately symmetric. Symmetric and HPD intervals will roughly coincide for such posterior distributions. Also, when the prior parameters are unequal but such that very small sample sizes suffice, the choice of criterion is not important. However, in cases of high *a priori* asymmetry with larger sample size requirements, such as when $a = 1, b = 100, \alpha = 0.05$ and $l = 2d = 0.02$, say, significant further reductions are obtained by using HPD rather than symmetric intervals, even when the symmetric intervals are derived from the exact posterior distribution.

To summarize, it makes little difference whether the criterion used has properties $\mathscr{P}_5$, $\mathscr{P}_6$ or $\mathscr{P}_7$ in the case of approximately symmetric prior distributions. For asymmetric prior distributions, methods based on the exact posterior density are more efficient than those based on normal approximations; the efficiency increases directly with the degree of asymmetry. Further reductions from HPD intervals result for moderate or large sample sizes in the latter case.

The choice between an average coverage or an average length criterion appears to be somewhat arbitrary, although sample sizes derived from these two criteria can be substantially different, even for symmetric priors (see Table 1 of Joseph *et al.* (1995)). Average length methods may be somewhat more conventional, since fixed coverage (usually 95%) intervals are most often reported, regardless of their length.

We have some additional comments regarding sample size determination for binomial experiments.

(a) Pham-Gia (1995) raises the question of how to choose between an averaging or worst outcome type of criterion. In part, this will depend on the degree of risk that one is willing to take, and how serious the losses might be in the event of an unfortunate outcome. Graphical summaries that present the coverages or interval lengths over the outcome space (see Figs 1 and 2 of Joseph *et al.* (1995)) can be useful in decision-making.

(b) We agree with Pham-Gia (1995) that the methods based on variances and/or normal approximations have the advantage of analytical tractability, allowing an algebraic investigation of their properties. It is also pointed out that many practitioners will use the mean ±2 standard deviations rule of thumb in calculating 95% posterior credible sets. However, widespread practice does not necessarily imply good practice. It is a theoretical fact that this rule of thumb will rarely if ever provide true 95% coverage probabilities and will in many cases provide substantially different coverage. Should statisticians not be advising against the use of such rules of thumb for known asymmetric distributions, especially when exact quantiles are easily obtained from tables or many statistical packages? Numerical approaches are increasingly common in applied statistics and are especially important to Bayesian analysis. Researchers wishing to maximize the efficiency of their experiments should be advised to use HPD intervals or at least intervals based on exact beta quantiles.

(c) It is true that a program based on HPD intervals is more complex than a program based on variances. However, once the programs have been written, it becomes no more difficult to run one than the other.

(d) Another advantage of numerical techniques can be their generalizability to other situations. In the next section an efficient Monte Carlo algorithm is described that can be used to find sample sizes for complex experiments.

3. Bayesian sample size determination for more complex situations

Binomial sampling is simple in that it is one dimensional, and exact closed form expressions are available for the preposterior and posterior densities. Adcock (1995) raises the issue of the evaluation of Bayesian sample size criteria for the multinomial distribution, and in general for any case where the above densities are intractable. In addition to standard Monte Carlo simulations, recent advances in Bayesian numerical algorithms such as the Gibbs sampler (Gelfand and Smith, 1990) or sampling–importance resampling (Rubin, 1987), combined with the increasing speed of modern computers, have shown that it is possible to deal with samples from posterior densities in cases where the direct use of these densities is not feasible. Suppose, for example, that we wished to evaluate an average coverage criterion, which has the general form

$$\int_{\mathscr{X}} \left\{ \int_{\mathscr{R}(x,n)} f(\theta \mid x, n) \, \mathrm{d}\theta \right\} f(x) \, \mathrm{d}x \geqslant 1 - \alpha, \qquad (1)$$

where θ is the parameter or parameter vector of interest, $\mathscr{X}$ is the outcome space and $\mathscr{R}(x, n)$ is a prespecified type of region (e.g. ellipsoid) with an *a priori* fixed volume. Although either of these integrals may be intractable, the following algorithm may be feasible.

(a) Draw a random sample $(x_1, x_2, \ldots, x_n)$ from $f(x)$, the preposterior distribution. This can be performed by using any of the above-mentioned techniques. The left-hand side of inequality (1) can then be approximated by Monte Carlo integration, i.e.

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \int_{\mathscr{R}(x_i, n)} f(\theta | x_i, n) \, \mathrm{d}\theta \right\}. \tag{2}$$

(b) If the remaining integral in expression (2) is also intractable, it can similarly be approximated by a numerical or Monte Carlo technique. For example, a sample from $f(\theta | x_i, n)$ can be drawn, and the integral approximated by the number of points falling inside $\mathscr{R}(x_i, n)$. Tanner (1991) presented methods of calculating the boundary and probability content of HPD regions for multivariate distributions based on random samples.

(c) As in previous methods, a bisectional or other search method can be used to find the minimum sample size that satisfies inequality (1).

Note that, in each iteration of the search, the integral in expression (2) is evaluated n times, so that random errors from each individual calculation may cancel out considerably when summed. Of course, the accuracy will depend on the Monte Carlo sample sizes drawn at each step in the procedure. Whether this algorithm is worthwhile, compared with, for example, a multivariate normal approximation depends on the degree of similarity between $f(\theta | x, n)$ and the multivariate normal distribution, as well as the accuracy required and computer programming and running times. The running times will also vary depending on the ease of drawing the appropriate samples required.

We have tested this algorithm in single and two independent binomial sampling experiments. In the former case it is almost as accurate as our exact method for all three of our HPD criteria. In the latter case we have shown it to be feasible in the more complex (although still one-dimensional) situation of estimating sample sizes for $\theta = \pi_1 - \pi_2$, the difference between two independent binomial parameters (Joseph *et al.*, 1995). Further work is required to determine whether these or similar techniques are worth pursuing for determining sample sizes for multinomial and other more complex experiments. Other non-Monte-Carlo techniques may also be explored, such as those presented in Smith *et al.* (1985) or Tierney *et al.* (1989).

4. Discussion

Many Bayesian criteria for sample size determination have been proposed, and still others can be defined. For example, one could investigate a mixed Bayesian–likelihood approach, where the prior information is used to calculate the preposterior marginal distribution but only the likelihood is used for final inferences. Thus in equation (1) $f(x)$ would include prior information, but $f(\theta | x, n)$ would not. This would accommodate investigators who need to plan according to the best available prior information, but who would prefer to use only the information in the data when reporting the results of a study.

Another criterion of interest might be a modification of the worst outcome criterion. Rather than the worst $x \in \mathscr{X}$, we could instead choose the worst $x \in \mathscr{S} \subset \mathscr{X}$, where $\mathscr{S}$ is a subset of $\mathscr{X}$ that covers say 95% of the most likely values of x according to $f(x)$. Thus we avoid choosing unnecessarily high sample size values when the worst outcome is very unlikely, such as in example 2 of Joseph *et al.* (1995).

The choice of which of the many criteria to use in the planning of any particular experiment appears situation specific, and no general rule can be given. We believe that the current evolution in statistics towards computer-intensive methods is pertinent to Bayesian sample size calculations. Although much work remains to be done for more complex and especially for multivariate situations, the benefits of these methods have been clearly demonstrated.

References

Adcock, C. J. (1987) A Bayesian approach to calculating sample sizes for multinomial sampling. *Statistician*, **36**, 155–159.

——— (1992) Bayesian approaches to the determination of sample sizes for binomial and multinomial sampling—some comments on the paper by Pham-Gia and Turkkan. *Statistician*, **41**, 399–404.

——— (1995) The Bayesian approach to determination of sample sizes—some comments on the paper by Joseph, Wolfson and du Berger. *Statistician*, **44**, 155–161.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Joseph, L., du Berger, R. and Bélisle, A. (1995) Bayesian and mixed Bayesian/likelihood criteria for sample size determination. To be published.

Joseph, L., Wolfson, D. B. and du Berger, R. (1995) Sample size calculations for binomial proportions via highest posterior density intervals. *Statistician*, **44**, 143–154.

Pham-Gia, T. (1995) Sample size determination in Bayesian statistics—a commentary. *Statistician*, **44**, 163–166.

Pham-Gia, T. and Turkkan, N. (1992) Sample size determination in Bayesian analysis. *Statistician*, **41**, 389–397.

Rubin, D. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C. and Dransfield, M. (1985) The implementation of the Bayesian paradigm. *Communs Statist. Theory Meth.*, **14**, 1079–1102.

Tanner, M. A. (1991) *Tools for Statistical Inference.* New York: Springer.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989) Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Am. Statist. Ass.*, **84**, 710–716.