The Statistician (1997)
46, *No.* 2, *pp.* 145–149

Interval-based *versus* decision theoretic criteria for the choice of sample size

By LAWRENCE JOSEPH†

Montreal General Hospital and McGill University, Montreal, Canada

and DAVID B. WOLFSON

McGill University, Montreal, Canada

[Received January 1996. Revised February 1996]

SUMMARY

Several different criteria for Bayesian sample size determination have recently been proposed. Bayesian approaches are natural, since at the planning stage of an experiment one is forced to consider prior notions about unknown parameter values that may affect the choice of a final sample size. For this, all the methods consider a prior distribution over the unknown parameters. Differences between the methods have been driven by the type of inferences that will be made, e.g. hypothesis testing or interval estimation, the latter based on posterior means and variances or highest posterior density regions. A more fundamental question, however, is whether to introduce formally a loss or utility function to aid in choosing the sample size. In this paper, we discuss the advantages and disadvantages of taking a fully decision theoretic approach *versus* one of the simpler approaches, which only implicitly consider utilities in balancing increased precision against the increased costs associated with larger sample sizes. Throughout, we emphasize the practical aspects of sample size estimation, raising issues that would face the consumer of statistics in selecting a sample size in a given experiment.

Keywords: Bayesian design; Decision theory; Highest posterior density; Sample size determination

1. Introduction

The issue of sample size selection for a given experiment is frequently encountered by applied statisticians. The general problem is to select a sample size to draw inferences about or to make a decision regarding an unknown parameter θ. Standard (non-Bayesian) solutions to this problem, many of which are summarized in Desu and Raghavarao (1990), are deficient in that the resulting sample size formula will typically require a point estimate $\hat{\theta}$ of the unknown θ. This is problematic, since θ will not usually be known with high precision at the planning stage of the experiment, and the sample size formula can be very sensitive to the choice of $\hat{\theta}$. Bayesian approaches replace the need to specify a point estimate by using a prior distribution over the range of values for θ, allowing for a more satisfactory utilization of the available information. The prior distribution leads to a predictive (marginal) distribution for the future data x that incorporates the uncertainty of both the unknown θ and the sampling variation of x averaged over θ. Sample size criteria can then be defined by taking expectations of various quantities over this predictive distribution.

This Bayesian approach has been employed by Spiegelhalter and Freedman (1986) and Spiegelhalter *et al.* (1994) in deriving the predictive power of a hypothesis test, and by Adcock

†*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.ri.mcgill.ca

© 1997 Royal Statistical Society 0039–0526/97/46145

(1987, 1988, 1992, 1995) and Pham-Gia and Turkkan (1992) in the context of interval estimation based on normal approximations to the exact posterior densities or intervals based on posterior means and variances. Joseph *et al.* (1995a) proposed exact sample sizes leading to highest posterior density (HPD) regions, whereas Joseph *et al.* (1995b) summarized the differences between the interval-based approaches. In particular, for a given α, an average coverage criterion (ACC) can be defined that ensures that the mean coverage of posterior credible intervals of length l, weighted by the predictive distribution, is at least $1 - \alpha$. Analogously, by considering fixed coverages, an average length criterion (ALC) can be defined. A conservative criterion, labelled the worst outcome criterion (WOC), is defined by considering the set $\mathscr{S}$ that consists of, for example, 80%, 95% or even 100% of the most likely data values according to the predictive distribution, and simultaneously ensuring a sufficiently small l and α for all data vectors in $\mathscr{S}$. A 100% WOC is analogous to the choice of $p = 0.5$ in a binomial frequentist sample size calculation.

Such interval-based approaches are simple to apply, since all that need to be specified are l and α, and of course the prior information. As recommended by Joseph *et al.* (1995a), the sample sizes corresponding to the ACC, ALC and WOC can then be computed, with a final choice based on the information that the calculations provide, while also balancing increasing 'costs' against increased precision of the estimates as the sample size increases. These costs would not form part of the statistical calculation.

Although much thought may go into the final choice, the decision would typically be made only informally, which makes it possible (or even likely) that a different sample size would have been 'optimal' if a formal utility function had been introduced. Lindley (1997) summarizes a decision theoretic approach to sample size determination, recommending selection of the sample size that maximizes a utility function that depends not only on the random quantities θ and x but also on the decision. As in Lindley (1997), we shall term this method 'maximization of expected utility' (MEU).

Theoretically, it is very difficult to argue against the incorporation of a utility into any decision problem; coherent decisions can only be guaranteed by introducing utilities. For a good discussion of such issues see Lindley (1985). However, the difficulty in specifying, communicating and understanding utilities in practice means that simpler criteria may often be preferred. Trade-offs between what may be axiomatically 'correct' and what is practical are often required. A judgment must be made about whether a compromise solution is 'real world' optimal, or whether we can afford the luxury (in terms of the time, costs and other resources that would be spent to make the decision) of a utility-based solution. We shall argue that, largely owing to the nature of many sample size problems in practice, the former is often the preferred option.

2. Comparison of interval-based and decision theoretic Bayesian approaches

We now compare the above two approaches to sample size determination by listing their main contrasts, and we discuss the implications of each.

2.1. *Maximization of expected utility formally incorporates a utility function, whereas interval-based methods do not*

We agree with Lindley (1997) that the incorporation of a utility function is a strong point in favour of MEU over any other method that is not based on utilities. However, we must be able to propose a useful and meaningful utility function to operationalize the MEU criterion. It is this step that may cause several problems in practice.

(a) In many cases, reasonable candidates for utility functions may not be apparent. Most experiments involve much more than simply accepting or rejecting a batch of items and the relatively simple costs associated with such actions; the true situation can be very complex indeed. For example, we may wish to minimize harm to prospective participants in setting up a clinical trial to evaluate a new drug. How many subjects should we choose to balance the risks to each individual participant, while still ensuring that the experiment has a good chance of providing sufficient information to aid in the treatment of future patients? Further compounding the issue is that there may be multiple risks and benefits of the drug, some of which may be unforeseen. How would we choose a 'cost' in such a setting? More generally, how do we place a cost on the effects of a drug on a life even if we know what the effects are? Although Lindley and Singpurwalla (1991) have explicitly addressed decision theoretic sample sizes for such situations, the suggested forms of the utilities do not generally approach the complexities of the true situation. Thus, it is reasonable to wonder whether we are further ahead maximizing a utility that may fail to capture important aspects of the problem adequately. Interval-based methods side-step rather than address this problem, of course, but they do allow a sample size to be selected with all considerations in mind, without the perhaps hopeless task of choosing a particular utility to be maximized.

(b) A related point is that, even if reasonable utility functions are available, there is no guarantee that everyone will agree on which should be used. This can be very important, since often many interested parties may be involved in any given experiment, all of whom will have many different uses for the experimental results, and who may be bearing different parts of the costs (both monetary and otherwise). For example, a pharmaceutical company may be interested in evaluating a new drug. From a corporate point of view, they may wish to maximize profits, i.e. they wish to carry out an experiment at minimum cost that will help them to decide how to market the drug (even this is a gross oversimplification). Participants, future users and government regulatory agencies will surely have different priorities, and coming to a consensus on maximizing a particular utility would be very difficult. It may be much easier to agree on the degree of accuracy required for the experiment to provide useful information to all involved.

(c) In addition, utilities usually must reduce several different aspects to a single unitless dimension, whose meaning may be difficult to interpret. For example, although much effort has recently gone into utilities of various health states, the issue is very far from settled. If we are not satisfied that we have a good utility measure, how confident should we be in using them to plan studies?

2.2. *Maximization of expected utility can be used for planning both inferential and operational experiments, whereas interval-based methods directly apply only to inferential sample size calculations*

The dual planning generality is clearly an advantage of MEU, provided that the problem lends itself to a clear utility function. In practice, this may be difficult to specify, since, as Lindley (1997) notes, it may 'depend on the practicalities of the situation', and these may be complicated, as in the examples above. However, if knowing θ to a specified degree of accuracy ensures that a good decision will probably be made, then the potentially difficult step of specifying the utility can be avoided. Although there is potential for loss of optimality in any given experiment, there are great gains in ease of implementation for standard situations. For example, solutions have been derived and easy-to-use software has been made available for the ACC, ALC and WOC for single binomial (Joseph *et al.*, 1995a), difference between two binomial (Joseph *et al.*, 1997) and normal and difference between two normal (Joseph and Bélisle, 1997) sampling situations. To obtain copies of this software, send the electronic mail messages 'send bhpd1 from general', 'send samplesize-prop from S' and 'send samplesize-mean from S' to `statlib@lib.stat.cmu.edu`, for single binomial, two binomial and normal sampling respectively.

2.3. *Average coverage criterion, average length criterion and maximization of expected utility are Bayesian in that they average over random quantities of interest, whereas worst outcome criterion does not have this property*

Although Lindley (1997) uses the Bayesian argument as a reason for dismissing the WOC, others have found it very useful in practice (DasGupta, 1995). Along with the ACC or ALC, the 100% WOC may be useful in setting an upper bound for the sample size. Other precentage WOC sizes can aid in clearly exposing the trade-offs between the accuracy of the inferences and the sample size. Remember that a sample size from the ACC or ALC that is selected for a given l and α will attain or exceed these limits only approximately 50% of the time, so that one may wish to attain the target values more often than 'on average'.

2.4. *Average coverage criterion, average length criterion and worst outcome criterion prescribe the decision to be taken for each possible x in advance, since they will all report a highest posterior density region with l and/or α fixed in advance as the final inferential decision: maximization of expected utility, in contrast, maximizes over the decision as well, for example, allowing an optimal balance between coverage and length to be part of the sample size decision; thus, choosing the 'best estimator' is an unnecessary question*

This is an illustration of the additional flexibility of allowing the decision d to depend on the data, rather than specifying the decision in advance. However, in practice, it is difficult to think of a case where this would be a serious concern, for several reasons. If the posterior distributions are reasonably smooth, there will not normally be large reductions in lengths of intervals for small changes in coverage. Any such relationships may be found by calculating sizes from interval-based methods for a range of α- and l-values. Similarly, α and l may be adjusted depending on the prior information about θ; for example, we may wish to decrease l for θ near 0 or 1. It is also difficult to think of situations where reporting non-HPD regions would be preferred to HPD regions, although if this were the case the interval-based criteria could also be adjusted accordingly.

2.5. *Maximization of expected utility is necessarily constructed to be coherent in the sense of Savage (1954), whereas average coverage criterion, average length criterion and worst outcome criterion may possibly be incoherent in this sense*

As Lindley (1997) admits, at the moment coherence remains only an intriguing possibility, and it is not clear whether it would have any practical importance in most situations.

2.6. *Maximization of expected utility introduces a specific cost of sampling, whereas other approaches do not*

Again, although a specific cost of sampling is desirable in theory, difficulties in specifying costs may often make it less so, especially since costs must be represented on the same scale as the utility measure. Neither of the two specific examples of such costs provided by Lindley seem to be very persuasive. In one instance, a cost of $c = 0.000211$ is deduced while comparing the ALC with MEU for binomial sampling, meaning, for example, that one is willing to invest in 47 observations to reduce the length of the HPD interval by 1%. Although this may be a useful point to consider, since the value of c changes with l, it is not obvious how it could have been deduced as the value of c on which to have calculated an MEU sample size in the first place. A utility based on Shannon information is used to deduce that $n = c/2$ in the case of normal sampling. Thus, if we are willing to pay £40 for a unit increase in information where each individual costs £1, then we should take 20 observations. However, this appears to assume that the utility of an increase in information is constant, regardless of the information already gathered, which surely is not correct in most situations, so that again the choice of c is not apparent.

3. Concluding remarks

There is no doubt that we should make coherent decisions whenever possible. Realistic utilities may often be difficult to specify for many practical courses of action. Therefore, inferential formulations will often be used to make sample size decisions. We can either attempt to form a reasonable inferential utility function to maximize, or more informally consider the information provided by the various interval-based criteria together with all other relevant information at hand. As Lindley (1997) points out, decision theoretic criteria were first proposed more than 35 years ago (Raiffa and Schlaifer, 1961). That virtually all sample size calculations performed today are not based on these criteria is potent evidence that there is strong resistance to applying them in practice, which we suggest is largely due to the problems in deriving utility functions and specifying their parameters in particular applications.

Of course, we are in full agreement with much of what Lindley (1997) suggests. Although we have focused on the possible practical problems that may arise in implementing decision theoretic sample size criteria, the utility functions contained in Lindley (1997) may be perfectly adequate for some situations. It is also possible that further research into the idea of utility would make its consideration seem as natural to applied researchers as criteria based on interval lengths and posterior probability coverages, which one can argue are 'natural' only because they are familiar. Lindley (1997) has pointed out that the ACC, ALC and WOC do not represent the final word in sample size computation. There is no doubt that further work on the implementation of coherent utility-based methods needs to be done.

References

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

———(1988) A Bayesian approach to calculating sample sizes. *Statistician*, **37**, 433–439.

———(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.

DasGupta, A. (1995) Review of optimal Bayes designs. *Technical Report 95-4*. Department of Statistics, Purdue University, West Lafayette.

Desu, M. M. and Raghavarao, D. (1990) *Sample Size Methodology*. Boston: Academic Press.

Joseph, L. and Bélisle, P. (1997) Bayesian sample size determination for normal means and differences between normal means. *Statistician* **46**, 209–226.

Joseph, L., du Berger, R. and Bélisle, P. (1997) Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statist. Med.*, to be published.

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

———(1995b) Some comments on Bayesian sample size determination. *Statistician*, **44**, 167–171.

Lindley, D. V. (1985) *Making Decisions*, 2nd edn. New York: Wiley.

———(1997) The choice of sample size. *Statistician*, **46**, 129–138.

Lindley, D. V. and Singpurwalla, N. D. (1991) On the evidence needed to reach agreed action between adversaries with application to acceptance sampling. *J. Am. Statist. Ass.*, **86**, 933–937.

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

Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory*. Boston: Harvard University Graduate School of Business Administration.

Savage, L. J. (1954) *The Foundations of Statistics*. New York: Wiley.

Spiegelhalter, D. J. and Freedman, L. S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statist. Med.*, **5**, 1–13.

Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. (1994) Bayesian approaches to randomized trials (with discussion). *J. R. Statist. Soc.* A, **157**, 357–416.