The Statistician (1997)
46, *No.* 2, *pp.* 129–138

The choice of sample size

By DENNIS V. LINDLEY†

Minehead, UK

[Received August 1995. Revised November 1995]

SUMMARY
This paper discusses the problem of how large a sample to take from a population in order to make an inference about, or to take a decision concerning, some feature of the population. It first describes a fully Bayesian treatment. It then compares this with other methods described in recent papers in *The Statistician*. The major contrast lies in the use of a utility function in the Bayesian approach, whereas other methods use constraints, such as fixing an error probability.

Keywords: Binomial sampling; Coherence; Coverage probabilities; Highest posterior density intervals; Information; Log-odds; Maximization of expected utility; Sample size determination

1. The Bayesian approach

This paper addresses the problem of how large a sample to take from a population in order to make an inference about, or to take a decision concerning, some feature of the population. The situation is modelled by supposing that there is a random quantity X having a density of known form dependent on an unknown parameter θ. For any non-negative integer n, it is possible to observe n realizations of X yielding $x_1, x_2, \ldots, x_n$, which, given θ, are independent and identically distributed as X. $n = 0$ means that no observations are made. The set is denoted $x = (x_1, x_2, \ldots, x_n)$. It is then required, after observing x, to make a decision d concerning θ. Both the choices, of n and d, are made by the decision maker. It will later be explained how inference may be incorporated into this framework. Generalizations of this model can easily be envisaged but will not be pursued here. The simple practical question, 'What size of sample should I take?', is often posed to a statistician, and it is a question that is embarrassingly difficult to answer. The Bayesian resolution is straightforward but there are difficulties in implementation.

Perhaps the first, fully Bayesian treatment of the problem was provided by Raiffa and Schlaifer (1961), in the general context of experimental design in section 1.2, and for a binomial, two-decision problem in section 5.6. Their method deals with the sequence (n, x, d, θ) in temporal order: first the sample size is chosen, the data are collected and then the terminal decision made. Finally the unknown θ is considered. Their method specifies a utility function $u(n, x, d, \theta)$ describing the merit of choosing n, obtaining the result x and taking the decision d, when the parameter has the value θ. It also supposes that θ has a density $p(\theta)$ before embarking on the sequence. Their resolution of the problem is to proceed in reverse time order, θ, d, x, n, taking expectations of utility over random quantities θ and x, and maximizing over deterministic quantities d and n. The density $p(\theta|d, x, n)$ can be found by first remarking that it equals $p(\theta|x, n)$, since your choice of d does not influence your opinion of θ, given x and n. An application of Bayes rule to θ and x for fixed n gives

$$p(\theta|x, n) = p(x|\theta, n)\, p(\theta|n)/p(x|n) = p(x|\theta, n)\, p(\theta)/p(x|n), \tag{1}$$

†*Address for correspondence*: 2 Periton Lane, Minehead, Somerset, TA24 8AQ, UK.

 0039–0526/97/46129

since similarly your choice of n does not influence your opinion of θ. Here

$$p(x|n) = \int p(x|\theta,\ n)\ p(\theta)\,\mathrm{d}\theta, \tag{2}$$

your density for x before taking the sample and without knowledge of θ. As usual

$$p(x|\theta,\ n) = \prod p(x_i|\theta)$$

in terms of the densities of each x_i, given θ.

With $p(\theta|d,\ x,\ n)$ available from equation (1), the expectation over θ of the utility of d, given $(x,\ n)$, can be calculated and then maximized over d to select the terminal decision. Going backwards in time, the expectation over x of this maximized value can be found by using $p(x|n)$. Finally, this expectation can be maximized over n to answer the original question. The optimum sample size is therefore the solution of

$$\max_n \left[\sum_x \max_d \left\{ \int u(n,\ x,\ d,\ \theta)\ p(\theta|d,\ x,\ n)\,\mathrm{d}\theta \right\} p(x|n) \right]. \tag{3}$$

Here x has been supposed discrete for clarity. In the continuous case the summation would be replaced by another integration.

Raiffa and Schlaifer (1961) now assumed, and we shall follow them, that the utility both does not depend on x and is additive and linear in n,

$$u(n,\ x,\ d,\ \theta) = u(d,\ \theta) - cn, \tag{4}$$

c being the cost in utiles of each observation. The lack of dependence on x can be violated if some values of x effectively lose utility and others do not. For example, if testing may lead to destruction of an item, any destroyed item is worthless, but a survivor may be capable of further use. Such cases will not be considered here. It is usual for each observation to cost the same, so justifying the linearity in n. There is often a set-up cost which introduces an additional term subtracted from equation (4) whenever $n \neq 0$. This refinement will not be introduced here since it only affects the comparison of $n = 0$ with other values of n. This has little effect on the issues that we discuss, though it can be of some practical importance.

With the special form of utility (4), expression (3) becomes

$$\max_n \left[\sum_x \max_d \left\{ \int u(d,\ \theta)\ p(\theta|x,\ n)\,\mathrm{d}\theta \right\} p(x|n) - cn \right]. \tag{5}$$

Use of equation (1) enables the probabilities to be written in terms of those prescribed initially, with the result that it is necessary to find

$$\max_n \left[\sum_x \max_d \left\{ \int u(d,\ \theta)\ p(x|\theta,\ n)\ p(\theta)\,\mathrm{d}\theta \right\} - cn \right]. \tag{6}$$

This method of determining sample size will be termed maximization of expected utility (MEU). In fact, MEU is used twice, first over θ and d, then over x and n. Notice the theoretically important point that MEU does *not* introduce a decision function $\delta(x)$ that, for each x, describes the terminal decision to be selected. In particular, it does not include the comparison of decision functions, or the judgment that some function is best. Replacing 'decision function' by 'estimator' may bring the point home. Of course, after the calculations in expression (6) have been made, you do have an estimator, since it tells you what d to select for each x, but that is the *only* estimator considered, and it is automatically the best. In this view, the question of what is the best estimator of a parameter is an unnecessary question.

A merit of MEU is that essentially *only* solutions produced this way can be guaranteed to

be coherent in the sense of Savage (1954). (There are a few delicate issues, such as whether $p(\theta)$ need integrate to 1. We pass these over.) In other words, any other procedure for the determination of sample size must, to be sensible, agree with MEU for some utility function and some density for the parameter. The use of the density $p(\theta)$ in this problem is almost forced on us because, at the time when the choice of n is made, there are only prior opinions of θ to guide us. Not until the data x are available can you afford to discard these opinions and to work with methods that depend only on the likelihood $p(x|\theta, n)$. Attention therefore concentrates on the utility function when comparing MEU with other procedures. By the assumptions, this reduces to discussion of $u(d, \theta)$, the utility of selecting d when the true value of the parameter is θ, and c.

2. Utility functions

The first case to consider is where d is some practical course of action, of which obvious examples are acceptance or rejection of material. There is little that can be said in general about this; the utility will depend on the practicalities of the situation. For example, if d is the decision to accept and θ is a measure of the quality of the product, higher values corresponding to better quality, then the utility of d will increase with θ. The situation is not so obvious when the problem is one of inference, rather than action.

It is first necessary to consider the desired form of the inference, that is, the form of your final statement d. Since a probability density was admitted at the beginning of the sequence as a description of your uncertainty concerning θ, it seems natural to use probability again at the end. Hence d might be identified with a distribution for θ. If so, we must consider $u\{p(\cdot), \theta\}$, the utility of stating a density $p(\cdot)$ when the parameter has the value θ.

Such a utility function is said to be *proper* if it results in the optimum choice of distribution being the posterior distribution. It is said to be *local* if it depends only on the value of the distribution at the true value of θ. Bernardo (1979) showed that, with mild restrictions on smoothness, the only proper, local, utility function is of the logarithmic form

$$u\{p(\cdot), \theta\} = A \log p(\theta) + B(\theta) \tag{7}$$

for some constant A and function $B(\cdot)$. In most applications the utility will not depend on θ but only on the relationship between the quoted $p(\cdot)$ and θ. If so, $B(\theta)$ is a constant that may be taken to be 0, reflecting the origin of utility. Similarly A may be put equal to 1, reflecting the scale. It is conceivable that information is more important at some values of θ than at others; if so, a suitable choice of $B(\cdot)$ could reflect this. However, here we suppose that $B = 0$.

Use of this utility leads, as in expression (5), to an expected utility, given (x, n), of

$$\int \log p(\theta|x, n)\, p(\theta|x, n)\, \mathrm{d}\theta, \tag{8}$$

the Shannon information about θ, given (x, n). This measure has been suggested by Lindley (1956) and used in the sequential, binomial context; Lindley (1957). In the present situation with a sample of fixed size n, its use would lead to the maximization of expected (over x) Shannon information, less allowance for the cost of sampling. Of the two concepts used to support Shannon information, propriety seems compelling because its violation would lead to a statement about the parameter that does not accord with one's beliefs. Locality is less compelling because it ignores the notion of how near one value of θ is to another. This may not matter since typical distributions are smooth, so that the value of $p(\theta)$, for θ near to θ', is near to $p(\theta')$.

If these assumptions are made, the overall utility function (4) is $\log p(\theta) - cn$. The consequences of using this when each observation has a normal distribution of mean θ and known variance σ^2, and $p(\theta)$ is also normal with variance τ^2, are easily found. Simple calculations show that the expected information (8) from a sample of size n is, apart from a constant, $\frac{1}{2}\log(1 + n\tau^2/\sigma^2)$. Exceptionally this does not depend on the result x of the sampling.

The choice of sample size is made by maximizing this expression, less cn, the cost of sampling, over n. The result is

$$n = 1/2c - \sigma^2/\tau^2. \tag{9}$$

If, as is often supposed, τ^2 is large in comparison with σ^2, reflecting the fact that one's prior uncertainty τ^2 about θ is larger than that of a single observation σ^2, then the optimum n is about $1/2c$. Further discussion of this result, and its extension to other members of the exponential family, will be found in Bernardo (1975).

Before considering other formulations of inference, two points need to be made. First, the statement of $p(\theta|x, n)$ is a complete form of inference, in the sense that it captures all that you know about θ, and any other form, like the intervals considered below, must contain less information. It is also information in the form needed for any subsequent decision problem concerning θ, since this will involve an expectation over θ. All this is within the Bayesian framework. The second point is that the utility function requires information and the cost of sampling to be measured in the *same* units. In equation (9), c is the cost of each observation, where the unit of measurement is a utile of information. The assessment of c is a practical difficulty which we return to later.

Instead of making an inference in the form of a probability distribution for θ, you may be content with stating an interval (a, b) within which θ must reasonably lie, i.e. $d = (a, b)$ and $u(d, \theta) = u(a, b, \theta)$. A possible form is $u(a, b) + \delta(\theta)$, where $\delta(\theta) = 1$ if θ lies in (a, b) and $\delta(\theta) = 0$ otherwise. This reflects the utility of θ lying within the stated interval, rather than outside it. Taking the expectation over θ in the integral part of expression (5) gives an expected utility of $u(a, b) + 1 - \alpha$, where $1 - \alpha$ is your probability, given (x, n), that θ lies in (a, b). The error probability is α, in conformity with common, notational practice.

A further restriction that might be reasonable in practice is to suppose that $u(a, b)$ depends only on $l = b - a$, the length of the stated interval, and not on where it is. If this dependence is linear, than the expected utility is $K - kl - \alpha$ for suitable constants K and k. This form will be considered in more detail later.

This type of utility is possibly too restrictive. The use of $\delta(\theta)$, equal to 1 or 0, like the familiar 0–1 loss function, pays no regard to how far outside, or how near the centre of, the interval the parameter lies. A more reasonable choice might be

$$C\sigma^{-1} \exp\{-\tfrac{1}{2}(\theta - \mu)^2/2\sigma^2\} + \delta(\theta), \tag{10}$$

where $\mu = \frac{1}{2}(a + b)$, the centre of the interval, and σ is a multiple of the length of the interval. C is a constant balancing being inside against considerations of where inside, or outside.

Another restriction on this type of utility function is that it uses the length of an interval of θ, rather than of some function of θ. For example, with binomial sampling discussed later, where $\theta = p(x_i = 1|\theta)$, an interval of length 0.05 may seem attractive when $\theta = 0.5$ but less so when $\theta = 0.05$ or less. When $\theta < 0.5$, proportional accuracy may be more apposite than fixed accuracy. This suggests consideration of $\log \theta$. To cover $\theta > 0.5$ as well, $\psi = \log\{\theta/(1 - \theta)\}$, the log-odds, may be a better parameter to consider. Calculation is then in terms of lengths of intervals in ψ. Notice that this issue does not arise when the posterior distribution, rather than the interval, is quoted as the inference, and when Shannon information is used, since the expected Shannon information is invariant under one-to-one transformations of the parameter.

Both the statement of a probability distribution and the use of an interval are related to what are usually thought of as estimation problems. It is also possible to use the utility approach to incorporate tests of hypotheses. Here, for example, one value of the parameter θ_0 refers to the null hypothesis and the inference d might be a probability that $\theta = \theta_0$. If this is to be the probability, given (x, n), then the utility function would have to be proper. This still leaves considerable choice. Another change from estimation would be to make the prior probability of θ_0 non-zero. The densities for θ would then need to be replaced by more general

forms. We do not explore the details here. Even in hypothesis testing, it would be better to use d as the whole posterior distribution and not as just one aspect of it, as with $p(\theta_0)$.

These considerations hopefully make it clear that within MEU the inference aspect of sample size determination is very flexible and that many ideas can be incorporated. The common feature is the specification of a single function, the utility, that includes all aspects except that of opinions about θ, and the subsequent maximizations over its expectations. In particular, the utility balances the cost of the sampling against the merits of the subsequent inference. Notice that the maximizations are all unrestricted; they are made without constraints. We now turn to other methods that have been suggested, which do not incorporate a cost of sampling, but do employ constraints.

3. Alternative methods

Four papers in a recent issue of *The Statistician*, Joseph *et al.* (1995a, b), Adcock (1995) and Pham-Gia (1995), addressed the problem of how large a binomial sample to take in order to acquire adequate knowledge of the binomial parameter θ. A major novelty was to suppose the final inferential statement about θ to be an interval which, according to the posterior density $p(\theta|x, n)$, had coverage $1 - \alpha$ and was of smallest possible length l. The restraint on the length means that the posterior density is greater inside the interval than outside. Such an interval is called a highest posterior density (HPD) interval. The viewpoint was Bayesian in so far as θ was treated as a random quantity with a probability distribution. It was not completely Bayesian in its failure to introduce a utility function.

We address the problem of comparing their methods with MEU. In particular, we consider a method based on an average length criterion (ALC). Similar remarks apply to the average coverage criterion and to the general methods described by Adcock (1995) in equation (14) of that paper. The other method, the worst outcome criterion, considered by Joseph *et al.* (1995a) is not discussed since, in its use of worst outcome scenarios, it offends Bayesian methods, which are confined to taking expectations of random quantities.

The terminal decision d in the ALC is an HPD interval described by its coverage probability $1 - \alpha$ and its length l. Once the posterior distribution of θ is available, l is determined by α and, in the ALC, α is fixed at a selected value α_0, l being the minimum length for coverge $1 - \alpha_0$. Until x is available, l is random and the method consists of choosing n so that $E(l|n) = l_0$, a preassigned value. (Owing to the discreteness of n, exact equality is not available but, for simplicity, this nicety is ignored.) By contrast, the coherent method would have a utility function as in equation (4), $u(a, b, \theta) - cn$, where the terminal decision d is the interval (a, b), which is not necessarily the HPD interval. This exposes two differences between the methods. First MEU introduces a cost of sampling, whereas the ALC does not, replacing it with constraints on l and α. A second distinction is that the ALC makes no explicit reference as to whether θ belongs to the interval or not, whereas MEU, through the utility, does. However, we have seen that a utility component of the form

$$u(a, b, \theta) = u(a, b) + \delta(\theta)$$

yields expected utility over θ of $u(a, b) + 1 - \alpha$, introducing the coverage concept present in the ALC. If, as is not unreasonable, $u(a, b)$ is a function of α and l, or even just of $l = b - a$, the quantity to be maximized at the next stage in expression (3) is a function

$$u^*(\alpha, l) = u(a, b) + 1 - \alpha.$$

Thus, there is a utility function $\delta(\theta)$ that at least relates the ALC and MEU, though it is not clear that it achieves a match between the two methods. Recall that such a match is desirable in order to ensure that the ALC is coherent.

A third distinction between MEU and the ALC emerges. The former maximizes $u^*(\alpha, l)$ over both α and l. The latter fixes α at α_0 and minimizes over l, resulting in the shortest

length and an HPD interval. (Recall that at this stage x is known.) Since both procedures take extremes over l, in order to compare them suppose that MEU is performed by first a maximization over l, followed by a maximization over α. The first will give $u^*\{\alpha, l(\alpha)\}$, where $l(\alpha)$ is the maximizing length. If, as will surely be true, u^* is a decreasing function of l for fixed α, this will give the interval of least length for that α, namely the HPD interval of the ALC, so drawing the methods together. However, there remains a distinction, because MEU maximizes $u^*\{\alpha, l(\alpha)\}$, whereas the ALC fixes α at the selected value α_0. The issue of whether this matters is delicate.

The continuous curves in Fig. 1 are curves of constant utility $u^*(\alpha, l)$ as functions of α and l. Since the objective is to decrease both α and l, these curves will have the property that any decrease in l will result in an increase in α. Also the utility will be larger the nearer the curve is to the origin. Now consider a particular posterior distribution. This will be capable of generating several pairs of values (α, l) but, by the argument of the previous paragraph, only pairs $(\alpha, l(\alpha))$, where $l(\alpha)$ is the length of the HPD interval of coverage $1 - \alpha$, are relevant, either in the ALC or MEU. Such values form a curve in the plane of (α, l). Two possible curves are shown as broken curves in Fig. 1. Think of them as possibly corresponding to two different values of x. Like the utility curves, these will be increasing in α as l decreases. Both curves pass through A, which has α co-ordinate equal to α_0, the value selected by the ALC, so that, in both cases, the ALC will select A. But MEU will in one case select the point B and in the other the point C, both having higher utility than A. This need not worry the adherent of the ALC for B is worse than A in that it has greater length, and C is worse than A because of its smaller coverage. However, there is another consideration that might worry the supporter of the ALC.

In Fig. 1, point A is preferred by the supporter of ALC to both B and C. Now take a 'mixture' of the two situations in which the curve AB is available with probability p and AC with probability $1 - p$. Such mixtures will give average errors and average lengths which lie on the straight-line segment BC. Choose p so that the mixture gives the point D with average error α_0. Then, according to a supporter of the ALC, B and C are both worse than A, but D, a mixture of them, is better because, for the selected α_0, it has average length less than A. This is an example of incoherence where a random choice between two inferior articles, B and C,

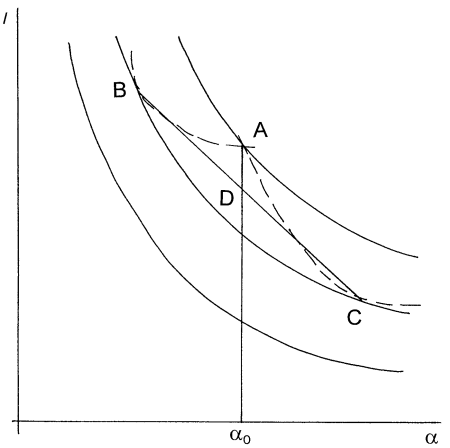

Fig. 1. $1 - \alpha$ is the coverage probability of an interval of length l; the continuous curves are curves of constant utility, the constant being larger the nearer the curve is to the origin; the broken curves are attainable values of (α, l) in two situations—on them, l will be the length of the HPD interval for that α; α_0 is the value of α selected by the ALC; D is a mixture of B and C with mixture probability chosen so that it has abscissa α_0

is preferred by the ALC method to A, which is preferred to both B and C. To most people who have thought about the matter, it seems obvious that no process of random mixing can turn inferior articles into a superior article. That it should not happen has been used as an axiom in decision analysis. The argument just used is due to Savage and Lindley and was employed by them to demonstrate incoherence of the minimax procedure; Lindley (1972). Of course, point D will only give average length, not exact length, less than A, but this presumably need not worry the supporter of the ALC since, at the next stage, the expectation of length is the criterion. If some other function of length were to be used at the next stage, then the whole argument could be rephrased in terms of that function and, in particular, used as the vertical axis in Fig. 1. Above we suggested the length of an interval in log-odds. Notice that, although the ALC fixes α, MEU allows it to vary with the result x of sampling.

So far the discussion has referred to a fixed posterior distribution (except in considering a mixture), resulting in a fixed value of the HPD length l for a given coverage. But l is a function of x, the random result of the binomial sampling. The ALC considers $E(l|n)$ and chooses n so that this has a selected value l_0. MEU, in contrast, maximizes $E[u^*\{\alpha, l(\alpha)|n\}] - cn$ over n, where α, a function of x and n, is the value selected at the previous stage. In fact, $(\alpha, l(\alpha))$ are the co-ordinates of a point like B or C in Fig. 1. This exposes another difference between the two methods: the ALC uses l; MEU uses a function of l. They might be drawn closer together by making the utility linear in l. Suppose then that the original utility, equation (4),

$$u(a, b, \theta) = K - kl + \delta(\theta),$$

where k is a positive constant balancing length against coverage, and K is another constant reflecting the origin of utility. Then $u^*(\alpha, l) = K - k\,l(\alpha) + 1 - \alpha$. (This might be more sensible if expressed in terms of log-odds, rather than θ, as suggested above.)

The distinction between the two methods may now be exposed by noting that the ALC solves the equation in n, $E(l|n, \alpha_0) = l_0$, where l_0 is the selected, average length, whereas MEU chooses n to maximize $E\{-k\,l(\alpha) + 1 - \alpha - cn|n\}$, where α and $l(\alpha)$ are the values selected at the previous stage. It is difficult, even with this restricted form of utility, to compare the methods, principally because MEU varies α whereas the ALC keeps it fixed at α_0. Suppose, however, that the decision problem is changed from a choice of interval (a, b) to the choice of an interval of coverage $1 - \alpha_0$. In other words, MEU is applied to the same restricted class of acts as the ALC. With the same utility, $K - l - cn$, omitting α because it is constant, and putting $k = 1$ without loss of generality, the contrast between the two methods can be easily appreciated. Write $E(l|n, \alpha_0) = \phi(n)$. Then MEU maximizes $K - \phi(n) - cn$, which reduces, on equating the derivative to 0, to solving the equation for n,

$$\phi'(n) = -c, \tag{11}$$

whereas the ALC has

$$\phi(n) = l_0. \tag{12}$$

From a mathematical perspective these two solutions are very different, one equating a function to a prescribed length, the other equating the function's derivative to a prescribed cost. Operationally they also differ: MEU uses a cost that is not mentioned in the ALC, which uses a choice of length that does not occur in MEU.

The distinctions between the two methods can be highlighted by considering the case where $E(l|n) = \phi(n) = tn^{-1/2}$, for a constant t. This is exactly true for normal sampling with known variance σ^2, where $t = \lambda_\alpha \sigma$. Thus, with $\alpha = 0.05$, $\lambda_\alpha = 3.92$. It is approximately true in the binomial case with $\sigma = E\{\theta^{1/2}(1 - \theta)^{1/2}\}$. MEU, from equation (11), gives the optimum n to be

$$n = (t/2c)^{2/3}, \tag{13}$$

whereas the ALC gives

$$n = (t/l_0)^2. \tag{14}$$

The contrast between the powers of t, and hence of σ in the normal case, is indicative of a real difference. Notice that with MEU and Shannon information the dependence on σ was small, equation (9), in marked contrast with equations (13) and (14).

Here is an argument that suggests they are not as different as the mathematics suggests. Suppose it happened that the two methods agreed for some c and l_0, in that the two equations (11) and (12) had the same value of n as solutions. Now suppose that the user of the ALC changes the length slightly from l_0. Differentiating equation (12) gives

$$\frac{\mathrm{d}n}{\mathrm{d}l} = \frac{1}{\phi'(n)}.$$

But, according to equation (11), $\phi'(n) = -c$, so $\mathrm{d}n/\mathrm{d}l = -c^{-1}$. This change is compatible with a utility function of the form suggested above, $K - l - cn$, for both methods would compensate a small change Δ in l with a small change $-\Delta/c$ in n. So although the ALC does not explicitly mention the cost of sampling, as does MEU, it does implicitly involve some pay-off between length and cost.

4. Approximate results

Some appreciation of the effectiveness of the various methods and, in particular, the choice of utility function can be obtained by considering the asymptotic results as $n \to \infty$, when the posterior distributions tend to normality. These can be viewed as providing approximations to the exact results. For example, Joseph *et al.* (1995a) give in their Table 1 results for the ALC in binomial sampling, when the prior distribution of θ is uniform in the unit interval. Our equation (14) shows that the optimum n is approximately $(\lambda_\alpha \sigma / l_0)^2$. The standard deviation σ is $E\{\theta^{1/2}(1-\theta)^{1/2}\}$ which, for the uniform distribution, is $\pi/8$. At $\alpha = 0.05$, $\lambda_\alpha = 3.92$. As an illustration, $l_0 = 0.01$ in Table 1 of Joseph *et al.* (1995a) gives an exact value of n of 23693, whereas the asymptotic result is $\{(3.92 \times \pi/8)/0.01\}^2 = 23697$, a trifling discrepancy. With the longer interval $l_0 = 0.1$, and consequent smaller samples, where the exact result is 234, the asymptotic value is 237, an error of only a little over 1%. Even at the large value $l_0 = 0.3$, the asymptotic value of 26 compares favourably with the exact sample size of 23.

It was suggested above that intervals in log-odds might be more suitable than those in θ. The posterior variance of $\log\{\theta/(1-\theta)\}$ is about $x^{-1} + (n-x)^{-1} = n/x(n-x)$, with expectation approximately $\{n\theta(1-\theta)\}^{-1}$. With a uniform prior, the expected standard deviation $E\{\theta^{-1/2}(1-\theta)^{-1/2}\} = \pi$, so that, again with $\alpha = 0.05$, the expected length is $3.92\pi n^{-1/2}$. To make a comparison with intervals for θ, it is necessary to compare lengths in θ with lengths in log-odds. Since the derivative of the log-odds is $\{\theta(1-\theta)\}^{-1}$, a small interval of length l in θ gives one of about length $l/\theta(1-\theta)$ in log-odds. At $\theta = \frac{1}{2}$, the latter is about four times the former: at $\theta = 0.1$, about 11 times. At $\theta = 0.15$, it is about 8, which is the same as the factor for the standard deviations ($\pi/8$ for θ, π for log-odds) so that the two factors cancel. Consequently the change to log-odds could have little effect.

These results for the ALC, derived from equation (14), are next compared with those for MEU, equation (13). Take the case of an interval in θ where $\alpha_0 = 0.05$ and $l_0 = 0.1$. We saw that the asymptotic value of n was 237. The comparison with MEU is made by asking what cost per observation would correspond to this choice of sample size using utility $K - l - cn$. From equation (13), we have $c = t/2n^{3/2}$. Inserting the numerical values, this yields a cost of $c = 0.000211$. Expressed differently, you are willing to invest in 47 $(= 0.01/0.000211)$ observations to reduce the length of the HPD interval by 0.01, or 10% of l_0. If l_0 is increased from 0.1 to 0.3, so that n reduces to 26, then c increases to 0.0058 and 5 $(= 0.03/0.0058)$

observations are all that you are willing to pay for to reduce the length of the interval by 0.03, or 10% of l_0.

Another possibility explored above was that of stating the posterior distribution, rather than just the incomplete summary of it provided by an HPD interval. We saw that it would then be reasonable to use Shannon information and to take the utility of providing a distribution $p(\cdot)$, from a sample of size n, when the true value is θ, as $\log p(\theta) - cn$. Calculations above showed that for the normal distribution this leads to the optimum sample size being about $1/2c$, equation (9), irrespective of σ. In contrast, the HPD methods considered above involved σ critically. The ALC had n proportional to σ^2, and the related MEU similarly used $\sigma^{2/3}$. But the result makes sense because c here measures the cost per unit of information, whereas in the other cases it measured the cost per unit reduction in length of the interval. In view of the strong arguments that exist for measuring the informal concept of information in terms of Shannon's function, rather than in terms of length or any other concept, the result is theoretically compelling. And practically it is extremely simple. If each observation costs £1, how much are you willing to pay for a unit increase in information? If £40, then $c = 1/40$ and it is worth your while to take 20 $(1/2c)$ observations.

5. General comments

On theoretical grounds, the use of MEU in the determination of sample size has everything in its favour. First, it is the only method that is guaranteed to be coherent: a user of MEU can never be exposed as making absurd comparisons (technically, comparisons that violate the axioms). Second, it is accompanied by a well-defined algorithm, expression (3), for its solution. This consists of a series of expectations and maximizations that can be performed on a computer. Third, it is very flexible, since a wide range of utility functions can be accommodated. Thus, it can balance a gain in knowledge with the cost of that knowledge. It can incorporate all aspects of how well the terminal decision relates to the true value of the unknown parameter, as was seen in the example of the HPD interval and the implicit use of the 0–1 loss function in the ALC.

On practical grounds, MEU is not so clearly advantageous. The demonstration in Fig. 1 of possible incoherence in the ALC and similar methods is suggestive but not conclusive. It has not been shown that the broken curves of the form given in Fig. 1 actually exist or, if they do, that the degree of incoherence exhibited is of serious consequence. For example, it might happen that these curves follow those of constant utility, when the mixture argument would fail. In general, this would not be possible since the broken curves result from the sampling plan, say binomial, which is separate from considerations of utility.

The main advantages of methods like the ALC lie in the simplicity and ease of comprehension of the inputs. Thus the ALC merely asks the user to specify a coverage probability and an expected length. Both these quantities are easily understood by a user. By contrast, MEU involves the concept of utility, which is alien to most users' way of thinking. It is not easy to balance the three quantities, coverage, length and cost of sampling, within a single function. It is even more difficult to compare cost with information, a concept with which few are familiar. One response to these important points is to call for more research into the idea of utility. We have a vast pool of experience in probability assessment but little in utility evaluation. Before MEU can be effectively employed, our understanding of utilities must be increased. The situation is somewhat easier if operational, as distinct from inferential, decisions are involved. Thus Raiffa and Schlaifer (1961), in their original treatment of the binomial case, considered a two-decision situation with the options of rejecting or accepting a batch. Lindley and Barnett (1965) described the sequential form of the same problem. The actions of acceptance and rejection are more easily thought of in terms of cost than are those that state an interval or a posterior distribution. Nevertheless, even in the pure inference

situation, the choice of sample size is clearly a decision, so the key ideas of decision analysis are relevant.

One difference between MEU and methods like the ALC lies in the extremal procedures used. For example, in the interval case there are three quantities over which the user has control, l, α and n. The remaining quantities, x and θ, are uncertain and both methods eliminate them by probabilistic arguments. But the control quantities are treated differently in the two approaches. MEU, having obtained l as a function of α, minimizes over both α and n through utility. The ALC in contrast fixes α and l, at α_0 and l_0, and finds n to attain these two values, or expectations of them. Fig. 1 attempts to show that this latter procedure may be unsound. What is also unclear is how the values α_0 and l_0 are to be selected. Admittedly statisticians have vast experience in using $\alpha_0 = 0.05$ but nevertheless it might be worth raising α_0 slightly if this would result in a substantially longer interval. It is this balance between coverage and length that MEU tackles and the ALC does not.

References

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

Bernardo, J. M. (1975) Tamaño optimo de una muestra: solucion Bayesiana. *Trab. Estadist.*, **26**, 83–92.

———(1979) Expected information as expected utility. *Ann. Statist.*, **7**, 686–690.

Joseph, L., Wolfson, D. B. 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. (1956) On a measure of the information provided by an experiment. *Ann. Math. Statist.*, **27**, 989–1006.

———(1957) Binomial sampling schemes and the concept of information. *Biometrika*, **44**, 179–186.

———(1972) *Bayesian Statistics, a Review.* Philadelphia: Society for Industrial and Applied Mathematics.

Lindley, D. V. and Barnett, B. N. (1965) Sequential sampling: two decision problems with linear losses for binomial and normal random variables. *Biometrika*, **52**, 507–532.

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

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.