The Statistician (1995)
44, *No.* 2, *pp.* 143–154

Sample size calculations for binomial proportions via highest posterior density intervals

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 April 1993. Final revision February 1994]

SUMMARY
Three different Bayesian approaches to sample size calculations based on highest posterior density (HPD) intervals are discussed and illustrated in the context of a binomial experiment. The preposterior marginal distribution of the data is used to find the sample size needed to attain an expected HPD coverage probability for a given fixed interval length. Alternatively, one can find the sample size required to attain an expected HPD interval length for a fixed coverage. These two criteria can lead to different sample size requirements. In addition to averaging, a worst possible outcome scenario is also considered. The results presented here provide an exact solution to a problem recently addressed in the literature.

Keywords: Bayesian design; Binomial proportions; Sample size calculations

1. Introduction

Several criteria have been proposed recently for Bayesian sample size estimation. The particular application to the binomial parameter θ has been examined in detail in Pham-Gia and Turkkan (1992), whereas Adcock (1987) considered multinomial experiments, which of course include the binomial as a special case. Adcock (1992) compared the various approaches presented in the above two papers.

Although there is potential for using a decision theoretic approach to sample size estimation in any given problem (Berger, 1985), practical considerations dictate that simpler criteria concerning accuracy in the estimation of θ are often preferable. Therefore, many of the methods considered to date suggest sample sizes that satisfy criteria relating in some way to either the variance of the posterior distribution for θ or posterior coverage probabilities for intervals of prespecified length.

The purpose of the present paper is to re-examine the sample size question from the point of view of highest posterior density (HPD) regions. This approach is natural, since HPD regions result in the shortest intervals for any given coverage and provide a convenient summary of the posterior knowledge about θ. Of course, there is a relationship between the

† *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

 0039–0526/95/44143

posterior variance and HPD intervals. For example, if we let U be the posterior mean and W be the posterior variance, an approximate 95% HPD interval may be found by taking $U \pm 2\sqrt{W}$. This result should be adequate for large sample sizes, as it is based on a normal approximation to the posterior density for θ. However, for small or moderate sample sizes, or when θ may be concentrated near 0 or 1, the approximation is less accurate, and the resulting deviations may be enough to cause sample size estimates given by the approximation and the exact method to differ substantially. In fact, as observed by Pham-Gia and Turkkan (1992), the probability of an interval calculated by taking $U \pm 2\sqrt{W}$ can be as low as 0.75.

The three possible sample size criteria based on HPD regions are given in Section 2, and the particular application to binomial sampling is discussed in Section 3. Section 4 contains some examples as well as notes on the practical implementation of the methods, while the final section concludes with a discussion.

Throughout this paper, $f(\ )$ will be used to denote generically a probability density or probability function, and $f(\,|\,)$ will denote a conditional density or probability function. The random variables to which these distributions refer will be clear from their arguments and the context in which they appear.

2. Bayesian criteria for sample size

Let n denote the sample size, θ the parameter of interest, Θ the parameter space for θ and $f(\theta)$ the prior distribution of θ. The experiment will consist of observing n data points $x = (x_1, x_2, \ldots, x_n)$, where x is composed of n exchangeable components, from the data space $\mathcal{X}$. The preposterior marginal distribution of x is

$$f(x) = \int_{\Theta} f(x|\theta) f(\theta)\, \mathrm{d}\theta, \tag{1}$$

and the posterior distribution of θ given data x is

$$f(\theta|x, n) = \frac{f(x|\theta) f(\theta)}{\int_{\Theta} f(x|\theta) f(\theta)\, \mathrm{d}\theta}, \tag{2}$$

where $f(x|\theta)$ is the likelihood of the data.

The posterior coverage probability of the region R of volume $V = V(R)$ is given by

$$\int_R f(\theta|x, n)\, \mathrm{d}\theta.$$

The volume V is minimized for the given coverage if R is an HPD region. A necessary and sufficient condition for this is that $f(\theta_1|x, n) \geqslant f(\theta_2|x, n)$ for all θ_1 in R and all θ_2 not in R (see Box and Tiao (1973)).

In view of our application we shall consider θ to be a one-dimensional real-valued parameter. The theory extends easily to vector-valued parameters, such as the parameter $\mathbf{p}$ of the multinomial distribution, although in practice the calculations become much more difficult.

If θ is one dimensional and $f(\theta|x, n)$ is unimodal, (a, b) is an HPD interval if and only if $f(\theta_1|x, n) \geqslant f(\theta_2|x, n)$ for all θ_1 in (a, b) and all θ_2 not in (a, b). For monotone increasing posterior densities defined on the interval (u, v), the corresponding condition for (a, b) to be the HPD region is that $b = v$, and

$$\int_a^b f(\theta|x, n)\, \mathrm{d}\theta = 1 - \alpha.$$

A similar condition holds for monotone decreasing densities. These characterizations will be exploited in Section 4.

Under these conditions, an experimenter typically would specify that θ should fall in an HPD interval of length l with probability $1 - \alpha$. However, the fact that the posterior distribution depends on x whose uncertainty must be eliminated leads to the consideration of three criteria in the determination of sample size. These are as follows.

2.1. *Average coverage criterion*

For a given fixed HPD interval length l, find the minimum sample size n such that the expected coverage probability is at least $1 - \alpha$, i.e. the smallest n satisfying

$$\int_{\mathscr{X}} \left\{ \int_{a(x,n)}^{a(x,n)+l} f(\theta|x, n)\, d\theta \right\} f(x)\, dx \geqslant 1 - \alpha, \tag{3}$$

where $f(x)$ is given by equation (1), $f(\theta|x, n)$ is given by equation (2) and $a(x, n)$ is the lower limit of the HPD credible set of length l for the posterior density $f(\theta|x, n)$. Note that a will in general depend both on the data x and the sample size n. This criterion is similar to that proposed by Adcock (1987), equation (4), except that the average there was taken over coverages of tolerance regions $R(x)$, where $R(x)$ are not necessarily HPD regions, although they are asymptotically the HPD. This can lead to differences in sample size requirements under certain circumstances, since $R(x)$ can either be longer than the equivalent HPD interval of the same coverage or have smaller coverage for an interval of the same length. Further, Adcock's sample size equations for multinomial sampling are based on a normal approximation, whereas those presented here in the simpler binomial case are exact. Here the term exact is used in the sense that the only source of error is computer error, which can be made negligible. Since we can draw a very rough analogy between a multiple of the standard deviation and the coverage of HPD intervals, this criterion may be compared with the Bayes risk (as an average) of Pham-Gia and Turkkan (1992), equation (10), with $c = 1$.

2.2. *Average length criterion*

In a similar spirit, an alternative way to select a sample size would be to fix the coverage probability $1 - \alpha$ of the HPD credible set for θ. Then each possible outcome x will require a certain length $l'(x, n)$ to obtain the desired coverage probability.

The structure of this criterion then ensures that, for any x in $\mathscr{X}$,

$$\int_{a(x,n)}^{a(x,n)+l'(x,n)} f(\theta|x, n)\, d\theta = 1 - \alpha,$$

and the problem is then to select the sample size according to the smallest n such that the expectation (with respect to x) of these lengths is less than l, i.e. satisfies

$$\int_{\mathscr{X}} l'(x, n) f(x)\, dx \leqslant l, \tag{4}$$

where l is the prespecified average length that is desired. In contrast with the previous criterion where the length was fixed, here the average length of the HPD interval is fixed. This criterion can perhaps surprisingly lead to very different sample sizes from the average coverage criterion (ACC). The average length criterion (ALC) does not appear to have been discussed previously, although it may again be compared with the Bayes risk in that the posterior variance can be indirectly related to the length of the corresponding HPD interval.

2.3. *Worst outcome criterion*

Both of the above criteria may be criticized on the grounds that they ensure the desired coverages or lengths only on average. Thus they give no guarantee for any particular observed data x. The most conservative approach would be to ensure that the requirements for both length and coverage probability hold simultaneously over all possible data vectors x that may arise. Thus, rather than averaging, a minimum n is chosen such that

$$\inf_{x \in \mathcal{X}} \left\{ \int_{a(x,n)}^{a(x,n)+l} f(\theta|x, n)\, d\theta \right\} \geqslant 1 - \alpha, \tag{5}$$

where both l and α are fixed in advance. This specification is similar to the criterion of Pham-Gia and Turkkan (1992) that ensures that the maximum posterior variance over the sample space $\mathcal{X}$ does not exceed a prespecified value. As before, direct use of the HPD interval could lead to different sample sizes.

3. Bayesian sample sizes for binomial proportions

This section applies the three criteria discussed above to the estimation of θ, the probability of success from a binomial experiment. Here the space of possible outcomes, $\mathcal{X}$, is discrete, taking on values in the set $\{0, 1, \ldots, n\}$, where n is the sample size.

3.1. *Average coverage criterion*

In the context of a binomial parameter, the ACC can be written as the minimum n satisfying

$$\sum_{x=0}^{n} \Pr\{\theta \in (a(x, n), a(x, n) + l)\}\, p(x, n) \geqslant 1 - \alpha, \tag{6}$$

where

$$\Pr\{\theta \in (a(x, n), a(x, n) + l)\} \propto \int_{a(x,n)}^{a(x,n)+l} \theta^{x}(1 - \theta)^{(n - x)} f(\theta)\, \mathrm{d}\theta,$$

$a(x, n)$ is the lower limit of the HPD credible set given the sample size and observed number of successes, l is the user-specified length of the credible set, $f(\theta)$ is the prior distribution of θ and $p(x, n)$ is the preposterior probability function of the data. If $f(\theta)$ can be represented by a beta distribution with parameters (c, d), i.e.

$$f(\theta) = \frac{1}{B(c, d)} \theta^{(c - 1)}(1 - \theta)^{(d - 1)}, \qquad 0 < \theta < 1,$$

where $B(c, d)$ is the beta function with parameters (c, d), then

$$f(\theta|x, n, c, d) = \frac{1}{B(x + c, n - x + d)} \theta^{(x + c - 1)}(1 - \theta)^{(n - x + d - 1)}, \qquad 0 < \theta < 1, \tag{7}$$

is the posterior distribution for θ given the data x and

$$p(x, n) = \binom{n}{x} B(x + c, n - x + d) \Big/ B(c, d) \tag{8}$$

is the preposterior marginal distribution for x. ACC (6) then reduces to finding the minimum n that satisfies

$$\sum_{x=0}^{n} \left\{ \binom{n}{x} \Big/ B(c, d) \right\} \int_{a(x,n)}^{a(x,n)+l} \theta^{(x + c - 1)}(1 - \theta)^{(n - x + d - 1)}\, \mathrm{d}\theta \geqslant 1 - \alpha. \tag{9}$$

3.2. *Average length criterion*

Following inequality (4), the equation

$$\sum_{x=0}^{n} l'(x, n)\, p(x, n) \leqslant l \tag{10}$$

must be solved, where $p(x, n)$ is given by equation (8). The length $l'(x, n)$, corresponding to the HPD interval, is found for each given x and n by solving

$$\int_{a(x,n)}^{a(x,n)+l'(x,n)} f(\theta|x, n, c, d)\,\mathrm{d}\theta = 1 - \alpha,$$

where $f(\theta|x, n, c, d)$ is given by equation (7) and $a(x, n)$ and $a(x, n) + l'(x, n)$ are the lower and upper limits of the HPD interval of this distribution. Solution of the above equation does not by itself guarantee an HPD interval, so that for each x and n we must check a condition such as the condition previously referred to in Box and Tiao (1973). Details of a procedure for simultaneously searching for $a(x, n)$ and $l'(x, n)$ are deferred to Section 4.

3.3. *Worst outcome criterion*

The sum $n + c + d$ is often referred to as the 'effective sample size', i.e. the sum of the actual sample size and the sample size equivalent of the information contained in the prior distribution. It is intuitively reasonable but very difficult to prove that, for any fixed n, c, d and l, the value of the integral

$$\int_{a(x,n)}^{a(x,n)+l} f(\theta|x, n, c, d)\,\mathrm{d}\theta$$

is minimized for $x \in (0, 1, 2, \ldots, n)$ by taking

$$x^* = x^*(n, c, d) = \begin{cases} \dfrac{n+c+d+1}{2} - c \text{ or } \dfrac{n+c+d-1}{2} - c, & \text{if } n + c + d \text{ is odd and } n \geqslant |d - c|, \\ \dfrac{n+c+d}{2} - c, & \text{if } n + c + d \text{ is even and } n \geqslant |d - c|, \\ n & \text{if } 0 \leqslant n \leqslant |d - c|, \end{cases} \tag{11}$$

where $f(\theta|x, n, c, d)$ is given by equation (7). In other words, for any given sample size, the length of the HPD interval is maximized when the two beta parameters belonging to the posterior distribution are as close as the sample size will allow.

The following reasoning lends plausibility to conditions (11). Consider the family of beta distributions with parameters (u, v), where $u + v = k$ is a fixed constant. Since the variance of this family of distributions is given by $uv/k^2(k + 1)$, it is easy to show that this is maximized when $u = v$. If u, v and k are all positive integers with k odd, then it is not possible to have $u = v$ exactly, in which case choosing $|u - v| = 1$ will maximize the variance. In our problem, there is a further complication due to the constraints $k = n + c + d$, $u \geqslant c$ and $v \geqslant d$, leading to the statement that the variance will be maximized when $|u - v|$ is minimized subject to these constraints. Equations (11) follow directly from this, albeit with the leap of faith that maximum variance implies minimum HPD coverage for any given

interval length. This supposition is intuitively acceptable for unimodal distributions but appears difficult to prove.

This conjecture was checked via a computer simulation for all possible n, $1 \leqslant n \leqslant 1000$, and found to be correct throughout this range. Although $c = d = 1$ was used throughout the simulation, it also shows that the result holds for many non-uniform prior distributions as well. To illustrate, suppose that we are considering the case $c = 1$, $d = 1$ and $n = 4$. The simulation will investigate the five beta distributions beta(1, 5), beta(2, 4), beta(3, 3), beta(4, 2) and beta(5, 1), eventually choosing beta(3, 3) as the 'worst case'. If instead we wished to consider the case $c = 1$, $d = 2$ and $n = 3$, we would consider beta(1, 5), beta(2, 4), beta(3, 3) and beta(4, 2). Since this a subset of the first case, beta(3, 3) again will give rise to the worst case, and there is no need to run a separate simulation. The result should also hold asymptotically, since the lengths of HPD regions for the normal distribution depend only on the variance, and the result is true for variances of the beta distribution. This contrasts with the mathematical derivation of the location of the maximum posterior variance in Pham-Gia and Turkkan (1992). Assuming that x^* is correctly chosen, either through conditions (11) or by some other means, the worst outcome criterion (WOC) is satisfied by the minimum n with

$$\int_{a(x^*,n)}^{a(x^*,n)+l} f(\theta|x^*, n, c, d)\,\mathrm{d}\theta \geqslant 1 - \alpha. \tag{12}$$

4. Examples and practical implementation

4.1. *Example 1*

Consider the problem of the quality manager given by Pham-Gia and Turkkan (1992). Here $c = 13$ and $d = 57$ represent parameters of the beta prior distribution of the probability of a substandard item in a manufacturing process. The manager wishes to obtain the two-standard-deviation credible interval, which roughly corresponds to $1 - \alpha = 0.95$, which is less than $l = 0.20$ for any data x that may result. The value $l = 0.17$ was also considered. For $l = 0.20$, both methods agree that no sampling is required, as there is already sufficient prior information. However, for $l = 0.17$, Pham-Gia and Turkkan (1992) found $n = 68$, whereas $n_{\mathrm{WOC}} = 62$, a difference of almost 10%. This percentage can increase further for prior distributions which are more asymmetric. In part 2 of this example, the sample size required such that the average posterior variance is less than 0.0012 is $n = 55$, whereas $n_{\mathrm{ACC}} = 48$ and $n_{\mathrm{ALC}} = 47$, a difference of about 15%. In making these comparisons here and elsewhere, we have used the two-standard-deviation rule as suggested by Pham-Gia and Turkkan (1992), or $l = 2Z_{\alpha/2}\sqrt{W}$, where W is the posterior variance, and $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution.

4.2. *Example 2*

It is possible to produce wildly divergent estimates from the various criteria. Consider the situation of a known rare outcome, which commonly occurs, for example, in estimating the prevalence of a rare disease. Suitable prior input may be $c = 1$, $d = 200$, $1 - \alpha = 0.90$ and $l = 0.01$. The prior mean and standard deviation are thus both approximately 0.005. In this case, $n_{\mathrm{ACC}} = 164$, $n_{\mathrm{ALC}} = 96$ and $n_{\mathrm{WOC}} = 26\,854$. Here there is more than a hundredfold difference between the sample size estimates. This is due to the extreme *a priori* rarity of the worst outcome $x = 13526$, which will occur with probability 4.4×10^{-62} according to the prior information.

It is also important to note that in this example the sample size suggested by the average variance criterion is 332, which is more than double n_{ACC} and more than triple n_{ALC}.

4.3. *Example 3*

The ACC and ALC are averages, and it is often of interest to observe the individual lengths or coverages from which the average is calculated. Fig. 1 shows the coverages across the various possible values of x that may occur, for the case of $l = 0.1$, $1 - \alpha = 0.95$ and a uniform prior, $c = d = 1$. Under these conditions, the ACC yields a sample size of $n = 274$. The worst HPD coverage occurs at $x = 137$, where it is 0.9039. The probability of the event $\{x = 137\}$ is 0.00362. Further, increasing the coverage to the desired 0.95 at $x = 137$ and $n = 274$ requires an HPD length of 0.1177. A judgment can now be made on whether this is an acceptable trade-off for the savings on sample size, or if it is preferable to increase the size to the $n = 381$ required for $l \leqslant 0.1$ and coverage greater than 0.95 for all x, or somewhere in between.

4.4. *Example 4*

Fig. 2 shows the individual lengths for the same input, $l = 0.1$, $1 - \alpha = 0.95$ and a uniform prior, $c = d = 1$. This yields a sample size of $n = 234$ using the ALC. The widest length, $l = 0.1272$, occurs at $x = 117$, an event with a probability of 0.00426. Decreasing this length to $l = 0.1$ would entail a coverage of only 0.8761. The insight gained from considering the question of sample size from a variety of viewpoints should help in the selection of the optimal choice in any given application.

4.5. *Comparing various criteria*

Table 1 compares the required sample sizes for five different Bayesian criteria, assuming a uniform prior distribution. The criteria are enumerated below, in order of appearance in Table 1:

(a) n_{ACC}, given by inequality (6);
(b) n_{ALC}, given by inequality (10);
(c) n_{WOC}, given by inequality (12);

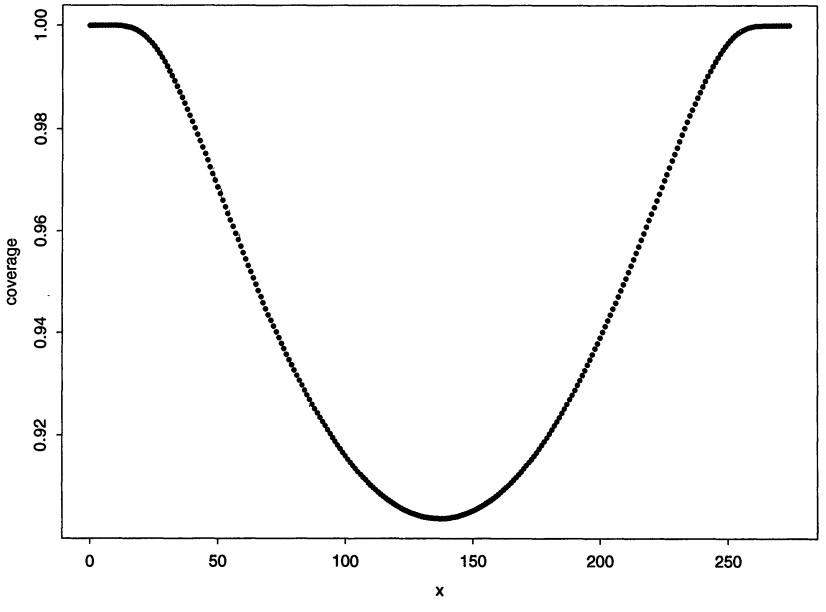

Fig. 1. Coverage *versus* x ($l = 0.1$; $c = d = 1$; average coverage, 0.95)

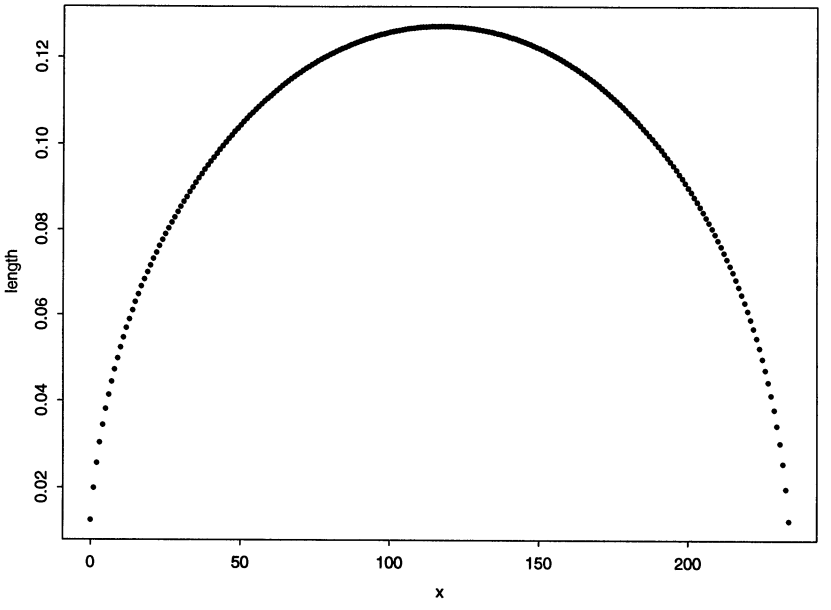

Fig. 2. Length *versus* x (coverage, 0.95; $c = d = 1$; average length, 0.1)

(d) n_{CONS}, from equation (15) of Adcock (1992), which gives a conservative region in a sense similar to the WOC;
(e) n_{VAR} from equation (10) of Pham-Gia and Turkkan (1992), the average posterior variance criterion.

The sample size given by n_{CONS} is equivalent to the maximum posterior variance criterion of Pham-Gia and Turkkan (1992), equation (3), if $\varepsilon = \chi^2_{1,\alpha}/d^2$, where $\chi^2_{1,\alpha}$ is the upper α percentage point of the χ^2-distribution with 1 degree of freedom and $l = 2d$.

From Table 1, two main observations can be made. Firstly, there is virtually no difference between the various conservative estimates when starting from a uniform prior. In fact, all are roughly equivalent to the standard conservative frequentist formula. Since, for example, n_{CONS} is based on the normal approximation to the binomial distribution and n_{WOC} is based on the beta distribution which is also asymptotically normal, this is hardly surprising. A similar observation was made concerning sample sizes derived from the normal distribution itself in Adcock (1988).

Secondly, the various Bayesian criteria can suggest very different sample sizes. It is intuitively obvious that both n_{ACC} and n_{ALC} are less than n_{WOC}, since if inequality (5) holds for any particular n then so will inequalities (3) and (4). Whether there are any general rules for comparing n_{ACC} with n_{ALC} is less evident, although some crude observations for a binomial outcome can be made from Table 1. Consider the ACC. Since coverage probabilities are bounded above by 1, we might expect that, as α moves towards 0, maintaining an average coverage of $1 - \alpha$ would become more difficult, not only because of the larger coverage probability required, but also because it becomes more difficult to balance the values which contribute less than average values to the mean coverage. A symmetrical argument can be made as the desired length l approaches 0 in the ALC. Thus we might expect to observe $n_{\text{ACC}} \geqslant n_{\text{ALC}}$ if α but not l is close to 0, and the reverse to hold if l but not α is close to 0. Indeed, Table 1 is consistent with these observations.

It is also worth observing from Table 1 that, for $1 - \alpha \geqslant 0.90$, $n_{\text{ALC}} \leqslant n_{\text{VAR}} \leqslant n_{\text{ACC}} \leqslant n_{\text{WOC}}$

TABLE 1
Comparison of the sample sizes given by the various criteria†

$1-\alpha$	l	n_{ACC}	n_{ALC}	n_{WOC}	n_{CONS}	n_{VAR}
0.50	0.01	2444	2804	4548	4547	3031
0.50	0.02	609	699	1136	1135	757
0.50	0.05	95	110	181	179	120
0.50	0.10	22	26	44	43	29
0.50	0.15	9	10	19	18	12
0.50	0.20	4	5	10	9	6
0.50	0.25	2	3	6	5	3
0.50	0.30	1	1	3	3	2
0.50	0.40	1	1	1	0	0
0.50	0.50	0	1	0	0	0
0.90	0.01	18533	16686	27053	27053	18035
0.90	0.02	4631	4169	6762	6761	4508
0.90	0.05	739	665	1080	1080	720
0.90	0.10	183	164	268	268	179
0.90	0.15	80	71	118	118	79
0.90	0.20	44	39	65	65	44
0.90	0.25	27	24	41	41	27
0.90	0.30	18	16	28	28	19
0.90	0.40	9	8	15	14	10
0.90	0.50	5	4	8	8	6
0.95	0.01	27691	23693	38412	38412	25608
0.95	0.02	6921	5921	9601	9601	6401
0.95	0.05	1105	945	1534	1534	1023
0.95	0.10	274	234	381	382	255
0.95	0.15	120	102	168	168	112
0.95	0.20	66	56	93	94	63
0.95	0.25	42	35	59	59	39
0.95	0.30	28	23	40	40	27
0.95	0.40	15	12	21	22	15
0.95	0.50	8	7	12	13	9
0.99	0.01	51552	40923	66345	66346	44231
0.99	0.02	12885	10228	16583	16585	11057
0.99	0.05	2058	1633	2650	2651	1768
0.99	0.10	512	405	659	661	441
0.99	0.15	225	178	291	292	195
0.99	0.20	125	98	162	163	109
0.99	0.25	79	62	102	104	69
0.99	0.30	53	42	69	71	48
0.99	0.40	28	22	37	39	26
0.99	0.50	17	13	22	24	16

†n_{ACC} is given by inequality (6), n_{ALC} is given by inequality (10), n_{WOC} is given by inequality (12), n_{CONS} is given by equation (15) of Adcock (1992) and n_{VAR} is given by equation (10) of Pham-Gia and Turkkan (1992). All sample sizes were computed starting from a uniform prior. Entries have been rounded upwards to the next highest integer.

$\approx n_{CONS}$. Finally, we note that $n_{VAR} \geqslant n_{ALC}$ across all cases considered in Table 1. This follows from the minimum length property of HPD regions.

4.6. *Practical implementation*

Finding the sample sizes in the above equations involves many intermediate steps, most of which have no closed form solution. However, computer algorithms may be devised that attain the sample sizes relatively quickly.

Each algorithm is composed of several subalgorithms. A description of the main subalgorithms is given below, followed by an outline of the steps required for each criterion.

4.6.1. *General strategy for finding sample size.* All algorithms employ a bisectional search strategy to arrive at the final sample size, which stops when the criterion is satisfied for n but not for $n - 1$. The lower bracketing limit was always assumed to be 0, whereas the upper limit was taken from the standard frequentist formula for binomial sample sizes. For each possible value of n, the relevant criterion was evaluated, and the next candidate was chosen depending on the result of the previous.

4.6.2. *Calculation of the incomplete beta function.* The frequent calculation of areas under the curve of a beta density between fixed upper and lower limits can be expressed in terms of the difference between two incomplete beta functions. Each of these integrals can be evaluated by using the continued fraction approximation of the incomplete beta function (Abramowitz and Stegun (1965), equation (26.5.8)). Sufficient terms in the continued fraction were retained to ensure accuracy to nine decimal places.

4.6.3. *Finding lower and upper limits of highest posterior density intervals.* Although the previous paragraph indicates how integrals with known limits can be evaluated, another frequent problem was to find the particular limits corresponding to HPD intervals. As indicated in Section 2, the method of solution depends on whether the beta density is unimodal or monotone increasing or decreasing. It also depends on whether a and l are both unknown, or whether l is given. In the latter case, lower (a), and upper ($a + l$), limits for unimodal densities can be found by solving the equation

$$f(a|x, n, c, d) - f(a + l|x, n, c, d) = 0,$$

where f is given by equation (7). Newton–Raphson iterations (Thisted, 1988) can be used to improve the speed of convergence. For monotone densities the problem is trivial, one end point being fixed at either 0 (if monotone decreasing) or 1 (if monotone increasing), and the second at a distance l from the first.

The case of a and l both unknown is two dimensional but can be approached through a combination of techniques already mentioned. A bisectional search strategy can be used to iterate towards the correct value of l, and a can be found for each l dictated by the search by the methods of the previous paragraph. Good starting points for the search are helpful in reducing the computing time. For example, first approximations for a and l can often be obtained from the symmetric credible set, for which a is simply the lower $\alpha/2$ percentile of the appropriate beta density, and $a + l$ is the upper $\alpha/2$ point. These quantities can be obtained from the approximation of beta quantiles given by equation (26.5.22) of Abramowitz and Stegun (1965).

4.6.4. *Algorithm for average coverage criterion.* For the ACC, c, d, α and l are fixed constants, and the coverage depends on x. For each n in the bisectional search, the left-hand side inequality (9) must be calculated and compared with the desired average coverage $1 - \alpha$. For each value of x, $x = 0, \ldots, n$ in the sum, the upper and lower limits, which depend only on a, as well as the resulting definite integral can be calculated as indicated earlier. The sum is compared with $1 - \alpha$, and the process continues until convergence.

4.6.5. *Algorithm for average length criterion.* For the ALC, the parameters c, d, α and l are again fixed constants and the length of the HPD interval depends on x. The minimum

n such that equation (10) is satisfied is sought. For each n indicated by the bisectional search, the vector $p(x, n)$, $x = 0, 1, \ldots, n$, is calculated by using equation (8). The length of the HPD interval, represented by the vector $l'(x, n)$, $x = 0, 1, \ldots, n$, is found by taking the difference between the upper and lower limits found as indicated above in the case that both a and l are unknown. The inner product of the resulting vectors can then be compared with $1 - \alpha$, and the search continues until convergence.

4.6.6. *Algorithm for worst outcome criterion.* For the WOC, c, d, l and α are fixed, with no averaging required. The value of x^* is determined by conditions (11), if we accept this conjecture. Thus, for each n, we need only to calculate the left-hand side of inequality (12), this integral being calculated as indicated earlier.

An easy-to-use Fortran program that implements all the HPD interval criteria discussed above is available for Sun SPARC workstations from the authors. In this program, all calculations are correct to at least nine decimal places. Finding sample sizes for all three criteria as well as calculating the minimum coverages, maximum lengths and preposterior probabilities of these values as discussed above typically took less than 1 min of computing time, and often only a few seconds.

5. Conclusion

Owing to their minimum length property and the feasible calculations for binomial outcomes, HPD regions are the preferred form of summary interval. The posterior variance can be used as a proxy, but caution must be exercised in its use. For conservative criteria, the answers provided by HPD regions and posterior variances are usually very similar. This is because these methods do not average over the predictive marginal distribution, and because the worst outcomes typically occur when the posterior parameters of the beta distribution are equal or nearly equal, so that the normal approximation to the beta distribution tends to be accurate. However, it is not difficult to construct examples where the average criteria can differ when HPD regions are replaced by posterior variances. In particular, for either rare or very common outcomes, there can be substantial savings in calculating sample sizes by using the HPD regions. Examples of this occur frequently, for instance in medicine when a rare disease is under study, or in quality control when defective items are uncommon.

It has been said (Berger, 1985) that design problems are naturally Bayesian, since before there are data there is no choice but to address planning issues by using prior information. The three exact methods discussed here are based on three different criteria, each leading to different sample sizes. All might contribute to the decision of the final size selected. Graphs such as those provided in Figs 1 and 2 are especially helpful in this regard.

Other criteria are also possible. A decision theoretic approach to finding the optimal sample size for both fixed and sequential sampling is discussed in Berger (1985), who also provided some examples. Pham-Gia and Turkkan (1992) discussed the problem in terms of the expected value of sample information. For estimating the mean of an arbitrary distribution, Goldstein (1981) suggested a Bayesian criterion based on the expected change of the point estimate for the mean over future sample values, and he provided an upper bound for the sample size.

The work presented here can be extended to other situations. One of the most commonly encountered sample size problems is to calculate the number of experimental units required for accurate estimation of the difference in response rates of the two independent groups when the outcome is dichotomous. Exact calculations may be very difficult owing to the complicated nature of the posterior distribution of the difference between two binomial random variables. Fortunately much progress has recently been made on the approximation of Bayesian posterior distributions, so that methods such as sampling–importance resampling (Rubin, 1987) can be used. We have already evaluated this algorithm as a substitute for exact

calculations in the context of the single binomial sample presented here. The results have encouraged us to investigate extensions to the two-sample problem.

Acknowledgements

The authors would like to thank the referees for their careful reading which led to a clearer and more complete paper, and Marie-Pierre Aoun for assistance in the preparation of the manuscript.

References

Abramowitz, M. and Stegun, I. A. (1965) *Handbook of Mathematical Functions*. New York: Dover Publications.
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.
Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.
Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. New York: Wiley.
Goldstein, M. (1981) A Bayesian criterion for sample size. *Ann. Statist.*, **9**, 670–672.
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.
Thisted, R. A. (1988) *Elements of Statistical Computing*. New York: Chapman and Hall.