STATISTICS IN MEDICINE, VOL. 16, 769–781 (1997)

BAYESIAN AND MIXED BAYESIAN/LIKELIHOOD CRITERIA FOR SAMPLE SIZE DETERMINATION

LAWRENCE JOSEPH

Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada, and Department of Epidemiology and Biostatistics, 1020 Pine Avenue West, McGill University, Montreal, Quebec, H3A 1A2, Canada

AND

ROXANE DU BERGER AND PATRICK BÉLISLE

Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G 1A4, Canada

SUMMARY

Sample size estimation is a major component of the design of virtually every experiment in medicine. Prudent use of the available prior information is a crucial element of experimental planning. Most sample size formulae in current use employ this information only in the form of point estimates, even though it is usually more accurately expressed as a distribution over a range of values. In this paper, we review several Bayesian and mixed Bayesian/likelihood approaches to sample size calculations based on lengths and coverages of posterior credible intervals. We apply these approaches to the design of an experiment to estimate the difference between two binomial proportions, and we compare results to those derived from standard formulae. Consideration of several criteria can contribute to selection of a final sample size.

1. INTRODUCTION

Consideration of the optimal number of experimental units is well recognized as an essential step in the design of biomedical investigations. Sample sizes are usually determined either from power calculations or from formulae based on confidence interval widths. While the forms of standard sample size equations that arise from power and interval width considerations are similar, the objectives differ, and often lead to substantially different sample size requirements.[1–3] Since in the recent past there have been many articles in medical journals that encourage the use of interval estimation rather than hypothesis testing and p-values,[4–6] this paper focuses on sample sizes based on interval widths, although we provide references to related work based on power considerations.

Currently, the most frequently used sample size formulae arise from the relationship between the standard error of the estimator of the parameter of interest and the sample size.[7–10] In almost all cases of practical importance, the resulting formulae require as input, along with the desired confidence coefficient and interval length, a point estimate of one or more unknown parameters of the model. Since the formulae can be highly sensitive to the choice of inputs, careful selection of the parameter estimates and target criteria are essential steps in determining a final sample size. Consider the following prototypic example.

CCC 0277–6715/97/070769–13$17.50

Received January 1995
Revised May 1996

1.1. Example 1

Deep-vein thrombosis (DVT) is a common complication of major knee surgery. Without any prophylaxis, more than half of all patients will develop DVT. Consider the design of a randomized clinical trial to compare two different prophylactic drugs, warfarin and low-molecular weight heparin. What sample size is needed to provide sufficient information to specify the true difference in DVT rates to within a total interval width of 5 percentage points? The most common solution, as given in all four standard references above, is to find a sample size such that the $100(1-\alpha)$ per cent confidence interval (CI) would have total width $w = 0{\cdot}05$, that is, to calculate

$$n = \frac{4Z^2_{1-\alpha/2}[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{w^2} \tag{1}$$

where π_1 is the rate of DVT under warfarin, π_2 is the rate of DVT under low molecular weight heparin, $Z_{1-\alpha/2}$ is the standard normal $(1-\alpha/2)$ upper quantile, and n is the sample size for each treatment group. The parameters π_1 and π_2 are of course unknown. The sample size is maximized when $\pi_1 = \pi_2 = 0{\cdot}5$, giving $n = 3074$ if $\alpha = 0{\cdot}05$. This conservative estimate, however, can be much larger than that truly required. Suppose that there are two previous studies available at the time of planning. One study of warfarin found 3 out of 14 ($\hat{\pi}_1 = 0{\cdot}21$, 95 per cent CI = (0·06, 0·51)) patients developed DVT[11] while another study found that 11 out of 65 ($\hat{\pi}_2 = 0{\cdot}17$, 95 per cent CI = (0·09, 0·29)) patients given low molecular weight heparin developed DVT.[12] Using these point estimates in (1) gives $n = 1899$. This sample size, however, can differ substantially from that calculated using other reasonable values, such as the upper or lower 95 per cent CI limits, which suggest $n = 2801$ and $n = 850$, respectively. Choosing $n = 2801$ appears conservative, but may be wasteful of resources if the true rates are closer to the point estimates. Selecting $n = 1899$ appears risky, however, since the confidence intervals around the point estimates are wide. A compromise may be desirable, but the criterion for the choice of the final n remains unclear.

One reason for this uncertainty is that formulae such as (1) require point estimates of π_1 and π_2, while a better summary of the available information is a distribution over a range of values. Furthermore, regardless of the choice of n or the true values of π_1 and π_2, the lengths of intervals reported at the end of a trial depend on the data collected, which is of course unknown at the planning stage. This paper reviews a Bayesian approach to sample size determination, which makes full use of the available prior information. The prior distribution leads to a predictive (marginal) distribution for the data that includes the dependence of the final inferences on both the unknown parameter values and sampling variation. One can then define various sample size criteria in terms of the average coverage probability or the average length of intervals of posterior credible sets over all possible data sets, weighted by the predictive distribution. These criteria clearly expose the compromises and risks related to choice of particular sample sizes.

Several authors have recognized the utility of using prior distributions rather than point estimates in sample size calculations. Early work included that of Dudewicz,[13] who calculated confidence intervals for power for the Normal density. Goldstein[14] considered estimating the mean of an arbitrary distribution, suggesting a Bayesian criterion based on the expected change in the point estimate for the mean over future sample values. Gould[15] considered both frequentist confidence densities based on previous data as well as Bayesian prior densities in examining the relationship between power and sample size. In the presence of uncertain parameter values, sample sizes having specified power with high probability as well as expected power for a given sample size were calculated. Berger[16] discussed a decision theoretic approach to finding the optimal sample size for both fixed and sequential sampling. Although this approach may be

theoretically attractive, it adds another layer of complexity and it is often the case that different interested parties (patients, physicians, pharmaceutical companies) have very different loss functions, impeding its use in practice. Here we consider loss functions only implicitly, in balancing the precision of the estimate versus the costs implied by a larger sample. Spiegelhalter and Freedman[17] proposed a predictive approach based on power considerations, while Yateman and Skene[18] suggest simulating a large number of possible data sets to determine sample sizes for complex survival studies.

Adcock[19,20] and Pham-Gia and Turkkan[21] considered Bayesian sample size determination for multinomial, normal and single binomial experiments, respectively. Adcock[22] reviewed these and suggested criteria based on the average coverage of tolerance intervals from normal approximations to the true posterior densities[19,20] or average posterior variances.[21] Joseph *et al.*[23] proposed the use of highest posterior density (HPD) intervals from the exact posterior distributions in the context of estimating sample sizes for a single binomial parameter. HPD intervals are optimal in the sense that they lead to the smallest sample sizes for any given coverage. Gould[24] has investigated sample sizes for differences in binomial proportions from a Bayesian viewpoint for equivalence trials.

In the next section we review several Bayesian criteria for sample size selection and apply them to the case of the difference between two binomial proportions in Section 3. The methods discussed here apply to any medical experiment where it is desirable to report results in terms of credible intervals for the difference between two binomial proportions. This includes both equivalence and comparative clinical trials as well as many other types of experiments, although we do not consider specific issues for any one type in detail here. We also discuss mixed Bayesian/likelihood approaches that use the prior distribution to derive the predictive distribution of the data but assume that one uses only the likelihood for inference. These are intended to satisfy investigators who recognize that prior information is important for planning purposes, but prefer to base final inferences only on the data. For example, confidence intervals that do not utilize prior information are most often reported in the medical literature, so that one could use the methods to derive sample sizes that ensure sufficiently narrow CI widths. Section 4 provides examples, and the final section contains further discussion. We defer the practical implementation of the criteria to the Appendix.

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

2. BAYESIAN CRITERIA FOR SAMPLE SIZE DETERMINATION

Let θ denote the parameter under study, Θ the parameter space for θ, and $f(\theta)$ the prior distribution of θ. We assume that the experiment under consideration provides data $x = (x_1, x_2, \ldots, x_n)$, where n is the sample size, and the components of x are exchangeable and belong to the data space $\mathcal{X}$.

The predictive distribution of x, also known as the pre-posterior marginal distribution of the data, is

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

and the posterior distribution of θ given x is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_{\Theta} f(x|\theta)f(\theta)\,d\theta} \tag{3}$$

where $f(x|\theta)$ is the likelihood of the data x. If $f(\theta)$ is a (possibly improper) uniform density over Θ, then (3) is the normalized likelihood. In the criteria that follow, one can either let $f(\theta)$ represent the true prior information in both (2) and (3), that is, a fully Bayesian (FB) approach, or let $f(\theta)$ be the true prior information in (2) but substitute a uniform density for $f(\theta)$ in (3), which we call the mixed Bayesian/likelihood (MBL) approach.

Typically, we wish a highest posterior density (HPD) or other posterior credible interval of length l that covers θ with probability $(1-\alpha)$. The posterior distribution of θ, however, depends on the data x, which is of course unknown at the planning stages of the experiment. We can eliminate this uncertainty by several different methods, leading to the criteria listed in the following sections.

2.1. Average Coverage Criterion (ACC)

We can allow the coverage probability $1-\alpha$ to vary with x, while holding the HPD interval length, l, fixed. This leads to sample size defined by the minimum n satisfying

$$\int_{\mathcal{X}} \left\{ \int_{a(x,n)}^{a(x,n)+l} f(\theta|x)\,d\theta \right\} f(x)\,dx \geqslant 1-\alpha \tag{4}$$

where $f(x)$ is given by (2), $f(\theta|x)$ is given by (3), and $a(x,n)$ is the lower limit of the HPD interval of length l for the posterior density $f(\theta|x)$, which in general depends on both x and n. We can regard the left hand side of equation (4) as an average of posterior coverage probabilities of fixed length l, weighted by the predictive distribution $f(x)$.

2.2. Average Length Criterion (ALC)

Conversely, we can fix the coverage probability, and allow the HPD interval length to vary depending on the data. In this case, for each x in $\mathcal{X}$ we must first find the HPD length $l'(x,n)$ such that

$$\int_{a(x,n)}^{a(x,n)+l'(x,n)} f(\theta|x)\,d\theta = 1-\alpha \tag{5}$$

and the sample size is the minimum integer n that satisfies

$$\int_{\mathcal{X}} l'(x,n)f(x)\,dx \leqslant l \tag{6}$$

where l is the prespecified average length. The left hand side of equation (6) averages the lengths of fixed coverage HPD intervals, weighted by the predictive distribution $f(x)$. The ACC and the ALC often lead to substantially different sample sizes.

2.3. Worst Outcome Criterion (WOC)

The ACC and ALC are based on averages over all samples, but since inferences are conditional on the observed sample, they lead to larger than desirable coverages and lengths for some

samples. A conservative approach is to ensure a maximum length of l and a minimum coverage probability of $(1-\alpha)$, regardless of the data x that occur. Hence, l and $(1-\alpha)$ are both fixed in advance, and we choose the minimum n such that

$$\inf_{x \in \mathscr{X}} \left\{ \int_{a(x,n)}^{a(x,n)+l} f(\theta \mid x)\, \mathrm{d}\theta \right\} \geqslant 1-\alpha. \tag{7}$$

A slight modification of the WOC, which we refer to as the MWOC, is to take the infimum in (7) over a subset, $\mathscr{S}$, of $\mathscr{X}$. For example, we might choose the set $\mathscr{S}$ to be the 99 per cent HPD region according to the predictive distribution (2). Here we ensure a maximum length l with a minimum coverage $(1-\alpha)$ for 99 per cent of all data sets x most likely to occur according to the prior information. Thus we can avoid the situation of having to select an unnecessarily large sample size to guard against highly improbable data. See example 2 of Section 4 for an illustration of this phenomenon.

3. BAYESIAN SAMPLE SIZES FOR THE DIFFERENCE BETWEEN TWO BINOMIAL PROPORTIONS

As an application of the criteria given in Section 2, let π_1 and π_2 denote two independent binomial parameters. Suppose that $f(\pi_1, \pi_2)$ is the prior distribution that summarizes the pre-experimental information about π_1 and π_2. This can be derived from past data, expert knowledge, or a combination of both. Proposed techniques have included directly matching percentiles[25] or means and standard deviations[26,27] to a member of a predetermined family of distributions, methods based on mean deviations,[28] as well as methods that directly use the predictive distribution of the data.[29] Prior specifications in clinical trials have been discussed by Spiegelhalter *et al.*[30] and Hughes.[31] Below, we consider both independent and dependent prior distributions for π_1 and π_2. The former arise naturally if there are data from past studies or pilot data for each arm of the study, although even in that situation, it is reasonable to consider that knowledge of π_1 may influence opinions about likely values of π_2. Therefore, it is often desirable to allow dependence between π_1 and π_2. For example, if π_1 represents a baseline rate for the control group, it may be more natural to specify a joint prior distribution for π_1 and the difference π_2–π_1, that is, to specify $f(\pi_1, \pi_2 - \pi_1) = f(\pi_1) f(\pi_2 - \pi_1 \mid \pi_1)$. When the sample size is large relative to the prior information, the posterior distribution will be predominantly determined by the likelihood function. In this case, the phenomenon of 'stable estimation'[16] or robustness[32,33] of the posterior distribution to the specification of the prior distribution may mean that the sample sizes are also robust to such changes, so whether one utilizes independent or dependent prior distributions may be less important.

Let x_1 and x_2 be the total number of 'successes' out of n_1 and n_2 trials from independent binomial experiments with parameters π_1 and π_2, respectively. The posterior distribution of (π_1, π_2) is then

$$f(\pi_1, \pi_2 \mid x_1, x_2, n_1, n_2) = k\, f(\pi_1, \pi_2) \prod_{i=1}^{2} \pi_i^{x_i} (1-\pi_i)^{n_i - x_i} \tag{8}$$

where k is a normalizing constant. In general, the specific form of the posterior distribution (8) will depend on the prior distribution. If we can represent the prior information on π_1 and π_2 by independent beta distributions with parameters (c_1, d_1) and (c_2, d_2), respectively, then we can

write (8) as

$$f(\pi_1, \pi_2 | x_1, x_2, n_1, n_2) = k \prod_{i=1}^{2} \pi_i^{x_i + c_i - 1}(1 - \pi_i)^{n_i - x_i + d_i - 1} \tag{9}$$

where $k = [B(x_1 + c_1, n_1 - x_1 + d_1)B(x_2 + c_2, n_2 - x_2 + d_2)]^{-1}$, and $B(\cdot, \cdot)$ represents the beta function. The predictive distribution of (x_1, x_2) is

$$p(x_1, x_2) = \prod_{i=1}^{2} \frac{\binom{n_i}{x_i} B(x_i + c_i, n_i - x_i + d_i)}{B(c_i, d_i)}. \tag{10}$$

For notational simplicity, we suppress the dependence of $p(x_1, x_2)$ on n_1, n_2 and the prior parameters. While we cannot represent all prior distributions by a beta distribution, it is a very flexible family[34] for distributions with support on the interval [0, 1], and has the advantage of being the conjugate family for binomial likelihoods. Here interest focuses on the difference between π_1 and π_2 which we can introduce through the change of variable transformation $\theta = \pi_1 - \pi_2$, and $\pi_1 = \pi_1$ and which has unit Jacobean. The joint posterior distribution for (π_1, θ) is

$$f(\pi_1, \theta | x_1, x_2, n_1, n_2) = k\, \pi_1^{x_1 + c_1 - 1}(1 - \pi_1)^{n_1 - x_1 + d_1 - 1}(\pi_1 - \theta)^{x_2 + c_2 - 1}(1 - \pi_1 + \theta)^{n_2 - x_2 + d_2 - 1} \tag{11}$$

which is non-zero over the region in the plane bounded by the lines $\pi_1 = 0$, $\pi_1 = 1$, $\pi_1 = \theta$ and $\pi_1 = \theta + 1$. It follows that the marginal posterior distribution of θ is

$$f(\theta | x_1, x_2, n_1, n_2) = \int_{\max(0, \theta)}^{\min(\theta + 1, 1)} f(\pi_1, \theta | x_1, x_2, n_1, n_2) \mathrm{d}\pi_1. \tag{12}$$

We can solve this integral analytically, the integrand (11) being a polynomial in π_1. In practice this can be cumbersome, since the polynomial can be of very high degree. Here we approximated the posterior distribution (12) by the closest fitting beta density on the interval $[-1, 1]$. One could also use other more numerical techniques such as Gaussian quadrature.

We now apply the criteria of Section 2 to find sample sizes for θ. The sample space $\mathscr{X}$ is discrete, taking values in the set $(0, 1, \ldots, n_1) \times (0, 1, \ldots, n_2)$. Henceforth we set $n_1 = n_2 = n$, that is, we assume equal sample sizes from each distribution. Extension to unequal sample sizes, such as allowing n_1 to be a fixed multiple of n_2, is straightforward. One could also consider finding the minimum sum $n_1 + n_2$ that satisfies one of the criteria, which may be worthwhile, for example, when there is substantially more prior information on π_1 than on π_2.

3.1. ACC

When θ is the difference between two binomial parameters, we can specify ACC as the minimum n satisfying

$$\sum_{x_1 = 0}^{n} \sum_{x_2 = 0}^{n} \Pr\{\theta \in (a(x_1, x_2), a(x_1, x_2) + l)\} p(x_1, x_2) \geqslant 1 - \alpha \tag{13}$$

where

$$\Pr\{\theta \in (a(x_1, x_2), a(x_1, x_2) + l)\} = \int_{a(x_1, x_2)}^{a(x_1, x_2) + l} f(\theta | x_1, x_2, n)\, \mathrm{d}\theta$$

$f(\theta|x_1, x_2, n)$ being given by (12), $a(x_1, x_2)$ is the lower limit of the HPD interval given x_1, x_2, and n, l is the HPD interval length provided by the investigator, and $p(x_1, x_2)$ is given by (10). We suppress the dependence of $a(x_1, x_2)$ on n to ease the notation.

3.2. ALC

According to (6), we seek the minimum n satisfying

$$\sum_{x_1=0}^{n} \sum_{x_2=0}^{n} l'(x_1, x_2)\, p(x_1, x_2) \leqslant l \tag{14}$$

where $p(x_1, x_2)$ is given by (10). We find the lengths $l'(x_1, x_2)$ that correspond to the HPD intervals for each (x_1, x_2) pair by solving

$$\int_{a(x_1, x_2)}^{a(x_1, x_2) + l'(x_1, x_2)} f(\theta|x_1, x_2, n)\, \mathrm{d}\theta = 1 - \alpha \tag{15}$$

where $f(\theta|x_1, x_2, n)$ is given by (12), and $a(x_1, x_2)$ and $a(x_1, x_2) + l'(x_1, x_2)$ are the lower and upper HPD limits of this distribution, respectively. Again, for notational purposes, we have suppressed the dependence of l' on n.

3.3. WOC

For a given interval length l and coverage probability $1 - \alpha$, we define criterion WOC by the minimum n satisfying

$$\int_{a(x_1^*, x_2^*)}^{a(x_1^*, x_2^*) + l} f(\theta|x_1^*, x_2^*, n)\, \mathrm{d}\theta \geqslant (1 - \alpha) \tag{16}$$

where $a(\cdot, \cdot)$ is defined as above, and we define x_1^*, x_2^* as the numbers of successes that maximize the length of the HPD interval, $x_1^*, x_2^* \in [0, 1, \ldots, n]$. We conjecture that

$$x_i^* = \begin{cases} \dfrac{n + c_i + d_i + 1}{2} - c_i \text{ or } \dfrac{n + c_i + d_i - 1}{2} - c_i, & \text{if } n + c_i + d_i \text{ is odd, and } n \geqslant |c_i - d_i|, \\ \dfrac{n + c_i + d_i}{2} - c_i, & \text{if } n + c_i + d_i \text{ is even, and } n \geqslant |c_i - d_i|, \\ n, & \text{if } \quad 0 \leqslant n \leqslant |c_i - d_i|, \end{cases} \tag{17}$$

for $i = 1, 2$. This conjecture states that we maximize the length of the HPD interval for θ when the posterior beta parameters for each of π_1 and π_2 are as close as the sample size will allow. Although it appears difficult to prove analytically, this result is intuitively reasonable, since this choice of parameters maximizes the variance of each beta distribution over the set of all possible choices. Further, we verified x^* via exhaustive simulations for all $n \leqslant 1000$, and it is also true asymptotically as $n \to \infty$, since for a normal distribution, maximum variance implies minimum HPD coverage probability.

4. EXAMPLES

We now illustrate the criteria of Section 3 with three examples. In each case, the value given for the ACC or ALC is the average of 10 repetitions of the Monte Carlo procedure described in the

Table I. Sample sizes for example 1, using fully Bayesian, mixed Bayesian/likelihood, and standard frequentist criteria

	ACC	ALC	MWOC(95)	MWOC(99)	WOC
Full Bayes	1799	1763	2582	2687	3033
Mixed Bayes/likelihood	1840	1794	2625	2731	3070
Frequentist	1899		2825	2903	3074

Appendix. Monte Carlo methods are not needed for the WOC or MWOC. We calculated the frequentist sample sizes using (1), where to correspond to the ACC or ALC, we substituted means of the marginal prior densities for π_1 and π_2, while we substituted 0·5 to correspond to the WOC. To correspond to the MWOC sample sizes, we used the exact binomial method[34] to calculate joint 95 per cent or 99 per cent confidence sets for (π_1, π_2) from the prior information, and we used the values in each set closest to 0·5.

4.1. Example 1 (Revisited)

The results cited in Section 1 suggest beta prior distributions with parameters $c_1 = 3$ and $d_1 = 11$ for the probability of DVT for warfarin patients, and $c_2 = 11$ and $d_2 = 54$ for the low molecular weight heparin group. Table I summarizes the sample sizes for $1 - \alpha = 0{\cdot}95$ and $l = 0{\cdot}05$ as well as their closest corresponding frequentist estimates. The FB WOC sample size is only slightly less than the corresponding frequentist sample size, and in general, the difference should be close to $0{\cdot}5(c_1 + d_1 + c_2 + d_2)$. The MBL WOC sample size is very close to the corresponding frequentist sample size, as expected. All other comparisons reveal lower sample sizes from the FB and MBL approaches. The frequentist estimate of 1899 is 5·5 per cent to 7·7 per cent higher than the corresponding FB sample sizes from ACC or ALC, respectively, and is 3·2 per cent to 5·8 per cent higher than the MBL sample sizes. The frequentist estimate using values from the 95 per cent upper confidence limits is 8·5 per cent higher than that suggested by the FB MWOC(95), and is even higher than the MBL MWOC(99) size. In trials with high per subject costs, even a small percentage reduction in sample size can lead to substantial savings.

We can now base the ultimate sample size selected on the above information. When it is crucial that the total width of the posterior interval not exceed 5 percentage points, a sample size in the area of 2600 or 2700 should suffice. Otherwise, we may choose a sample size in the range 1750 to 1850, knowing that slightly higher or lower than the average length may result, depending on the data.

In the above example, we considered total interval widths of 5 percentage points. In general, the desired width depends on the prior information about treatment differences and the clinical range of equivalence. The latter may be a function of treatment costs, side-effects, and other considerations. If we anticipate a large difference or it becomes apparent in an interim analysis, it may be unethical to continue sampling simply to obtain a narrow interval. Data monitoring committees, therefore, should have the ability to modify the sample size requirements if necessary.

4.2. Example 2 (Rare Events)

Consider a clinical trial planned to study the rates of myocardial infraction (MI) for patients with acute unstable angina pectoris following two different study regimens. A previous study[35] showed that both aspirin and an aspirin and heparin combination had lower MI rates compared to placebo, with rates of 4/121, 2/122 and 14/118 events per total number of patients in each group,

Table II. Sample sizes for example 2, using fully Bayesian, mixed Bayesian/likelihood, and standard frequentist criteria

	ACC	ALC	MWOC(95)	MWOC(99)	WOC
Fully weighted prior distributions					
Full Bayes	726	674	1437	1608	8414
Mixed Bayes/likelihood	884	823	1630	1807	8534
Frequentist	822		2438	2902	8537
Down-weighted prior distributions					
Full Bayes	806	702	1810	2049	8475
Mixed Bayes/likelihood	896	793	1914	2154	8534
Frequentist	822		3456	4171	8537

respectively. The confidence interval for the difference in rates between the aspirin and aspirin and heparin regimens, however, ranges from -3 to 6 percentage points. Using the above prior information, what sample size do we need so that the 95 per cent HPD interval for the difference in rates between aspirin and aspirin with heparin has a total length of 3 percentage points? We can summarize the prior information as $c_1 = 4$, $d_1 = 117$, $c_2 = 2$ and $d_2 = 120$. Often, we need to downweight these prior distributions, to reflect design, population, or other differences between previous trials and the one planned.[32] Therefore, Table II presents sample size requirements for both fully weighted and down-weighted (each beta prior parameter set to half its former value) prior distributions. Weights other than one-half could also be considered. The MWOC values are less than one-quarter those given by the WOC, indicating that the worst possible outcome is highly unlikely to occur. Here the differences between Bayesian and frequentist sample sizes are more pronounced. For example, the 99 per cent frequentist estimate is 61 per cent higher than the MBL MWOC(99) sample size. Notice that the ACC and ALC adjust to the changing prior information, whereas the frequentist equivalent is fixed. We can see differences of over 90 per cent between the two approaches in the case of down-weighted prior information. The Bayesian approaches suggest that 2000 subjects is quite conservative, and one could consider sizes as low as 700, depending on the risk one is willing to take.

It can be observed for this example that while the FB sizes are uniformly lower, the MBL estimates from the ACC and ALC hover around the frequentist sample size suggested by using point estimates from the prior data. One should use caution, however, in employing the mixed Bayesian/likelihood estimates, as 'paradoxes' can arise. For example, if $c_1 = d_1 = c_2 = d_2 = 10$, $l = 0{\cdot}05$ and $\alpha = 0{\cdot}05$, then the FB ACC sample size equals 2910, while the MBL ACC size equals 2926. If, however, more prior information becomes available such that $c_1 = d_1 = c_2 = d_2 = 1000$, the FB ACC size reduces to 1072 while the MBL ACC size increases to 3068. Of course, this 'paradox' is explained by the convergence of the prior distributions around the 'worst case' values $\pi_1 = \pi_2 = 0{\cdot}5$, but this example illustrates the inefficiencies related to one's ignoring prior information in final inferences. The problem is especially important when the prior information forms a substantial part of the total available information.

4.3. Example 3 (Dependent Prior Distributions)

As discussed in Section 3, when the sample sizes are large compared to the information in the prior distribution, or when the prior distribution covers only a narrow portion of the possible range for both π_1 and π_2, over which the sample size requirements may not substantially vary,

Table III. Sample sizes for example 3 from dependent and independent prior distributions, using fully Bayesian, mixed Bayesian/likelihood, and standard frequentist criteria

	ACC	ALC	MWOC(95)	MWOC(99)	WOC
Dependent prior distributions					
Full Bayes	53	24	145	171	175
Mixed Bayes/likelihood	64	39	153	179	184
Independent prior distributions					
Full Bayes	59	44	164	176	187
Mixed Bayes/likelihood	67	55	168	180	189
Frequentist	70		193	193	193

the form of the prior distribution should not greatly affect the sample sizes. For small or moderate sample sizes, however, there may be some differences. The following example examines this issue by calculating sample sizes from a dependent prior distribution, and compares these to sample sizes obtained from an independent prior distribution whose marginal means and variances for π_1 and π_2 match those of the dependent prior distribution.

Consider the situation where we expect a control group rate to be about 10 per cent, but it could perhaps be lower or it could be as high as 30 per cent. We expect the rate in the treatment group to be similar to the control group rate, although this is not known with great certainty. Reasonable prior distributions may then be $\pi_1 \sim \text{beta}(0{\cdot}6, 5{\cdot}4)$ and $\pi_2 \mid \pi_1 \sim \text{beta}(6\pi_1, 6(1-\pi_1))$. The correlation between π_1 and π_2 is 0·74. The sample sizes that correspond to this prior information and that satisfy the various criteria with $1-\alpha = 0{\cdot}95$ and $l = 0{\cdot}2$ appear in Table III. For comparison purposes, the sample sizes starting from the matching independent prior distribution with $\pi_1 \sim \text{beta}(0{\cdot}6, 5{\cdot}4)$ and $\pi_2 \sim \text{beta}(0{\cdot}2769, 2{\cdot}492)$ also appear in Table III. Changing from dependent to independent prior distributions can lead to different sample size requirements, as seen here for the ACC and especially the ALC values. These differences are due in part to changes in the predictive distribution for (x_1, x_2), but also since there is additional prior information about π_2 when it is correlated to π_1. The MWOC sizes show less variation, since the prior information in both cases cover wide ranges, so that outcomes leading to large HPD intervals can occur. One can use other functions of π_1 for the beta coefficients of the conditional density of $\pi_2 \mid \pi_1$ when it is likely that $\pi_1 > \pi_2$ or $\pi_1 < \pi_2$.

5. DISCUSSION

A common problem encountered in the use of standard sample size formulae is their sensitivity to the values selected for unknown parameters. There is almost always at least some amount of prior information about the unknown parameters, but rarely to the degree that one can give a reliable point estimate. Therefore, in specifying a probability distribution over a range of values, the FB and MBL approaches allow for a more satisfactory formulation of the problem. As is evident from the examples provided in Section 4, the Bayesian estimates often suggest smaller sample sizes than the corresponding frequentist estimates. This is largely due to the efficient use of prior information provided by the Bayesian approach. The sample sizes discussed here may also be adequate if one will ultimately employ a multivariate analysis (for example, logistic regression) for inference.[36]

The choice between the ACC and the ALC appears somewhat arbitrary, even though the sample sizes differ substantially. One may consider the ALC more conventional, since fixed

coverage (usually 95 per cent) intervals are most often reported, regardless of length. Whether to use criteria that average over the predictive distribution of the data or to consider the worst possible outcome depends in part on the degree of risk one is willing to take. In enumerating this risk, it is useful to calculate the MWOC sample sizes for a range of subsets of the sample space. It is also important to realize that criteria based on averages will only attain their target values approximately half of the time. Consideration of all of the criteria can lead to a more informed choice.

The MWOC considers the set of most likely experimental outcomes, regardless of the associated interval coverages or lengths. It is also possible to base cut-offs directly on these coverages or lengths. For example, one can select a sample size such that the coverage will be greater or equal to that desired (for a specified fixed length) with sufficiently high probability. Conversely, one can consider fixed coverages, and ensure a maximum length with high probability. Such criteria differ from the MWOC when, for example, a given coverage can occur in both common and less common outcomes in the sample space $\mathscr{X}$. A sufficient condition for such criteria to coincide with the MWOC is that the relationship between the probability of the outcomes is monotone with respect to the lengths, for example, when more common outcomes are associated with longer lengths. Otherwise, the MWOC is more conservative than the above criteria.

The general Bayesian sample size criteria presented in Section 2 applies to virtually any sample size problem. While this paper focused on treatment differences, one could use other outcome summaries such as relative risks or odds ratios. One must, however, exercise caution in the interpretation of average ratios. This is especially true for rare events, which can lead to very large ratios for some elements in the sample space. The Monte Carlo approach described in the Appendix should make calculations feasible in cases where posterior distributions are difficult to calculate exactly. Further work is required to investigate whether the criteria are useful and worthwhile for more complex (for example, multivariate) problems.

APPENDIX

The sample size criteria given in Section 2 in general do not have closed form solutions, as is the case for the standard formula (1). Therefore, one must carry out a numerical search for the correct sample size. One way to formalize the process is to employ a bisectional search strategy, which stops when the relevant criterion is satisfied for n but not for $n-1$. For each possible value of n, one evaluates the relevant criteria, and chooses the next candidate depending on the result of the previous candidate. Below we briefly outline a generally applicable Monte Carlo algorithm for implementing the criteria described in Section 2.

Denote the parameter of interest (for example, treatment difference or risk ratio) by θ, and let $x \in \mathscr{X}$ be the data vector of length n. We can write the ACC as

$$\int_{\mathscr{X}} \left\{ \int_{R(x)} f(\theta \mid x)\, \mathrm{d}\theta \right\} f(x)\, \mathrm{d}x \geqslant 1-\alpha \tag{18}$$

where $R(x)$ represents a region of length l, and $f(x)$ is the predictive distribution. The algorithm proceeds by drawing a random sample of size M from $f(x)$, and calculating the inner integral for each of the M sampled data points. The average of these M integrals then approximates the left hand side of the ACC for the chosen value of n.

Similarly, we could use the minimum or relevant quantile of the set of integral values in determining whether we satisfy the WOC or MWOC. We can devise analogous Monte Carlo algorithms for the ALC. To ensure sufficiently small Monte Carlo error, we can run the

simulations several times, and calculate the variance of the computed sample sizes. For the examples in Section 4, we averaged ten repetitions with $M = 1000$ to form the final sample size estimate. This number of repetitions usually brought the Monte Carlo error to below 0·5 percent of the final sample size, and often much below. For the application to differences between binomial parameters, exact use of equation (17) and the predictive distribution (10) precludes Monte Carlo methods for the WOC and MWOC.

One might use several methods, including exact analytic computation, Gaussian quadrature,[37] normal or other approximations, sampling importance resampling (SIR),[38] or even the Gibbs sampler[39] to calculate the required integral for each sampled point. To obtain the results in Section 4, we used a beta approximation to the distribution of the difference of two proportions suggested by Springer.[40] For the ALC, we approximated the interval lengths using the method of Tanner.[41] To calculate the sample sizes when the prior information is such that π_1 and π_2 are correlated, we obtained a random sample from the posterior density using the SIR algorithm, and we obtained approximate HPD lengths and coverage probabilities from the sorted sample. We used values of x_1 and x_2 that provided the maximum marginal variances to calculate the WOC sample sizes.

Programs written in S-plus for all criteria (including the case of unequal n_1 and n_2) discussed in this paper are available from the authors, or send the e-mail message 'send samplesize-prop' to statlib@lib.stat.cmu.edu to receive the software by e-mail.

ACKNOWLEDGEMENTS

The authors would like to thank the referees and the deputy editor for their insightful comments which led to a much improved paper. The research in this paper was supported in part by the Natural Sciences and Engineering Research Council. Lawrence Joseph is a research scholar of the Medical Research Council of Quebec.

REFERENCES

1. Bristol, D. R., 'Sample sizes for constructing confidence intervals and testing hypotheses', *Statistics in Medicine*, **8**, 803–811 (1989).
2. Beal, S. L., Sample size determination for confidence intervals on the population mean and on the difference between two population means', *Biometrics*, **45**, 969–977 (1989).
3. Grieve, A. P. 'Confidence intervals and sample sizes', *Biometrics*, **47**, 1597–1603 (1991).
4. Gardner, M. J. and Altman, D. G. 'Confidence intervals rather than p-values: Estimation rather than hypothesis testing', *British Medical Journal*, **292**, 746–750 (1986).
5. Anonymous. 'Report with confidence', *Lancet*, **1**, 488 (1987).
6. Evans, S. J. W., Mills, P. and Dawson, J. 'The end of the *p*-value?', *British Heart Journal*, **60**, 177–180 (1988).
7. Lachin, J. M. 'Introduction to sample size determination and power analysis for clinical trials', *Controlled Clinical Trials*, **2**, 93–113 (1981).
8. Desu, M. M. and Raghavarao, D. *Sample Size Methodology*, Academic Press, Inc, Boston, 1990.
9. Lemeshow, S., Hosmer Jr., D. W., Klar, J. and Lwanga, S. K. *Adequacy of Sample size in Health Studies*, Wiley, Chichester, 1990.
10. Lipsey, M. W. *Design Sensitivity*, *Statistical Power for Experimental Research*, Sage Publications, Newbury Park, 1990.
11. Francis, C. W., Marder, V. J., Evarts, C. M. and Yaukoolbodi, S. 'Two-step warfarin therapy. Prevention of postoperative venous thrombosis without excessive bleeding', *Journal of the American Medical Association*, **249**, 374–378 (1983).
12. Leclerc, J. R., Geerts, W. H., Desjardins, L., Jobin, F., Laroche, F., Delorme, F., Haviernick, S., Atkinson, S. and Bourgoin, J. 'Prevention of deep vein thrombosis after major knee surgery – A randomized,

double-blind trial comparing a low molecular weight heparin fragment (enoxaparin) to placebo', *Thrombosis and Haemostasis*, **67**, 417–423 (1991).
13. Dudewicz, E. J. 'Confidence intervals for power with special reference to medical trials', *Australian Journal of Statistics*, **14**, 211–216 (1972).
14. Goldstein, M. 'A Bayesian criterion for sample size', *Annals of Statistics*, **9**, 670–672 (1981).
15. Gould, A. L. 'Sample sizes required for binomial trials when the true response rates are estimated', *Journal of Statistical Planning and Inference*, **8**, 51–58 (1983).
16. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985.
17. Spiegelhalter, D. J. and Freedman, L. S. 'A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion', *Statistics in Medicine*, **5**, 1–13 (1986).
18. Yateman, N. A. and Skene A. M. 'The use of simulation in the design of two cardiovascular survival studies', *Statistics in Medicine*, **12**, 1365–1372 (1993).
19. Adcock, C. J. 'A Bayesian approach to calculating sample sizes for multinomial sampling', *Statistician*, **36**, 155–159 (1987).
20. Adcock, C. J. 'A Bayesian approach to calculating sample sizes', *Statistician*, **37**, 433–439 (1988).
21. Pham-Gia, T. G. and Turkkan, N. 'Sample size determination in Bayesian analysis', *Statistician*, **41**, 389–397 (1992).
22. Adcock, C. J. '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**, 399–401 (1992).
23. Joseph, L., Wolfson, D., and Du Berger, R. 'Sample size calculations for binomial proportions via highest posterior density intervals', *Statistician*, **44**, 143–154 (1995).
24. Gould, A. L. 'Sample sizes for event rate equivalence trials using prior information', *Statistics in Medicine*, **12**, 2009–2023 (1993).
25. Press, S. J. *Bayesian Statistics*: *Principles*, *Models and Applications*, Wiley, New York, 1989.
26. Bunn, D. W. 'The estimation of a Dirichlet prior density', *Omega*, **6**, (4), 371–373 (1978).
27. Lee, P. M. *Bayesian Statistics*: *an Introduction*, 3rd edn, Halsted Press, New York, 1992.
28. Pham-Gia, T. G., Turkkan, N., and Duong, Q. P. 'Using the mean deviation in the elicitation of the prior distribution', *Statistics and Probability Letters*, **13**, 373–381 (1992).
29. Chaloner, K. M. and Duncan, G. T. 'Assessment of a beta prior distribution: PM elicitation', *Statistician*, **32**, 174–180 (1989).
30. Spiegelhalter, D. J., Freedman, L. S. and Parmer, M. K. B. 'Bayesian approaches to randomized trials', *Journal of the Royal Statistical Society*, *Series A*, **157**, 357–416 (1994).
31. Hughes, M. D. 'Reporting Bayesian analyses of clinical trials', *Statistics in Medicine*, **12**, 1561–1563 (1993).
32. Greenhouse, J. B. and Wasserman, L. 'Robust Bayesian methods for monitoring clinical trials', *Statistics in Medicine*, **14**, 1379–1391 (1995).
33. Kass, R. E. and Greenhouse, J. B. 'Comment: A Bayesian perspective', *Statistical Science*, **4**, 310–317 (1989).
34. Johnson, N. and Kotz, S. *Continuous Univariate Distributions-2*, chapter 24, Wiley, New York, 1970.
35. Theroux, P., Ouimet, H., McCans, J., Latour, J-G., Joly, P., Lévy, G., Pelletier, E., Juneau, M., Stasiak, J., DeGuise, P., Pelletier, G. B., Rinzler, D., Waters, D. D. 'Aspirin, heparin, or both to treat acute unstable angina', *New England Journal of Medicine*, **17**, 1105–1111 (1988).
36. Aickin, M. and Ritenbaugh, C. 'A criterion for the adequacy of a simple design when a complex model will be used for analysis', *Controlled Clinical Trials*, **12**, 560–565 (1991).
37. Thisted, R. *Elements of Statistical Computing*, Chapman and Hall, New York, (1988).
38. Rubin, D. B. Using the SIR algorithm to simulate posterior distributions, in *Bayesian statistics 3,* eds. J. M. Bernado, M. H. Degroot, D. V. Lindley, and A. F. M. Smith. New York: Oxford University Press, 395–402 (1988).
39. Gelfand, A. E., and Smith, A. F. M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistics Association*, **85**, 398–409 (1990).
40. Springer, M. *The Algebra of Random Variables*, Wiley, New York, 1979.
41. Tanner, M. *Tools for Statistical Inference*, Springer-Verlag, New York, 1991.

