

Bayesian Adaptive Designs for Clinical Trials

BY YI CHENG

*Department of Mathematical Sciences, Indiana University,
South Bend, Indiana 46634, USA*

ycheng@iusb.edu

AND YU SHEN

*Department of Biostatistics, M. D. Anderson Cancer Center,
Houston, Texas 77030, USA*

yshen@mdanderson.org

SUMMARY

A Bayesian adaptive design is proposed for a comparative two-armed clinical trial using decision theoretical approaches. A loss function is specified to consider the cost for each patient, and the costs of making incorrect decisions at the end of a trial. At each interim analysis, the decision to terminate or to continue the trial is based on the expected loss function while concurrently incorporating efficacy, futility and cost. The maximum number of interim analyses is not pre-fixed but decided adaptively by the observed data. We derive explicit connections between the loss function and the frequentist error rates, so that the desired frequentist properties can be maintained for regulatory settings. The operating characteristics of the design are able to be evaluated on frequentist grounds. Extensive simulations are carried out to compare the proposed design with existing ones. The design is general enough to accommodate both continuous and discrete types of data. We illustrate the methods with an animal study evaluating a medical treatment for cardiac arrest.

Some key words: Adaptive designs; Decision theory; Group sequential clinical trials; Loss function; Martingale convergence theorem.

1 INTRODUCTION

When conducting a large-scale efficacy trial, it is important to sequentially monitor the trial for ethical, scientific and economic concerns. Using information from the interim analyses to adaptively make decisions during the trial is a dynamic process. Recently, there have been substantial interest and development of statistical methods in the adaptive sample size re-estimation design in the frequentist framework (Proschan & Hunsberger, 1995; Fisher, 1998; Shen & Fisher, 1999; Lehmacher & Wassmer, 1999; Liu & Chi, 2001; and Müller & Schäfer, 2001, among others). Tsiatis & Mehta (2003), on the other hand, compare the efficiency of some adaptive designs with the standard group-sequential design. An important feature of an adaptive clinical trial is a continual updating of the design using accumulated information from prior experience and interim data. Such designs carry the Bayesian spirit and naturally fit the Bayesian paradigm. However, there has been limited literature in this area using Bayesian decision theoretical approaches.

A variety of Bayesian approaches have been proposed in group sequential trials for different purposes. Bayesian designs are compared with frequentist group sequential designs using decision theoretical approaches in Berry & Ho (1988) and Lewis & Berry (1994). Studies by Eales & Jennison (1992), Cressie & Biele (1994) and Barber & Jennison (2002), among others, search optimal group sequential designs under various settings using Bayesian decision theoretical approaches. The maximum sample size/block size is pre-determined for all these methods. By using the non-stationary Markov approach, Lai (1973) formulates a hypotheses testing problem that minimizes the expected sample sizes.

In this study, we generalize the Bayesian decision theoretical approach by allowing the maximum sample size to be random and sequentially determined by the observed data. One practical concern with the use of Bayesian methods in the design and conduct of clinical trials has been the control of the false-positive error rate in the regulatory setting. We use loss functions that explicitly quantify the costs caused by false-positive and false-

negative decisions. An asymptotic functional relationship can be derived between the design parameters of the loss function and the frequentist error rate. We can thus maintain the desired frequentist properties, such as type I and II error rates, for the designs by choosing an appropriate loss function. Another advantage of such a design is that we are able to simultaneously integrate considerations of efficacy, futility, and cost in the decision making, whereas the self-designing trials (Shen & Fisher, 1999 and Thach & Fisher, 2002) and similar adaptive designs require a separate futility stopping rule. It is, therefore, feasible to identify optimal designs with minimum expected sample sizes while achieving the specified power. The method is general enough to be applied to both continuous and discrete data.

We formulate the design and inference strategy in Section 2, and explore two typical types of outcomes in detail. The extensive simulations that we carry out are presented in Section 3, and the methods are illustrated by an example in Section 4.

2 BAYESIAN ADAPTIVE DESIGN WITH ONE-STEP BACKWARD INDUCTION

Consider a clinical trial for comparing a treatment T and a control C , where the treatment response is X_T and the control response is X_C . The block size at each stage is denoted by $2B_i$, where B_i is the sample size for each treatment arm, and $i = 1, 2, \dots$. Let \bar{X}_{T_i} and \bar{X}_{C_i} be the observed mean of the i -th block for the two arms. Let θ be the parameter of interest,

$$X_i = \bar{X}_{T_i} - \bar{X}_{C_i} \sim F(\cdot|\theta), \quad \text{and} \quad \int_{-\infty}^{\infty} x dF(x|\theta) = \theta,$$

where $F(x|\theta)$ is the cumulative density function of x given θ . The prior distribution for θ is denoted by $\pi(\theta)$ with a prior mean $E(\theta|\pi) = \delta$. For instance, with a normal prior density of $\phi(\theta|\delta, \sigma^2/B_0)$, it is equivalent to a normal likelihood arising from a previous trial of B_0 patients with an observed or hypothetical mean δ for the treatment difference (Spiegelhalter et al., 1994). In the frequentist's hypothesis testing framework, the one-sided hypotheses are

$$H_0: \theta \leq \theta_0 \quad \text{versus} \quad H_1: \theta > 0.$$

The hypotheses to be tested include two scenarios: with ($\theta_0 > 0$) and without ($\theta_0 = 0$) the range of equivalence (Freedman, 1987). If θ is within the equivalence range, there is insufficient information to indicate a preference for any one of the treatments. In Sections 2.1 and 2.2, we describe the general results and decision rules for the proposed design, while we focus on normally distributed data and binary outcomes to illustrate the methods in detail in Sections 2.3 and 2.4, respectively.

2.1 Loss function and decision rules

Let A and R represent the actions of accepting and rejecting the null hypothesis, respectively. The loss function for each action is defined by

$$L(\theta, A) = \begin{cases} 0, & \text{if } \theta \leq \theta_0; \\ K_1, & \text{if } \theta > \theta_0; \end{cases}$$

$$L(\theta, R) = \begin{cases} K_0, & \text{if } \theta \leq 0; \\ 0, & \text{if } \theta > 0. \end{cases}$$

In particular, K_0 and K_1 are the losses for making the type I and type II errors, respectively. The conventional Bayesian $0 - K_i$ loss, with $i = 0, 1$, (Berger, 1985) is a special case when $\theta_0 = 0$.

The stopping rule is devised to minimize the loss. The information at each interim stage is updated by Bayes theory. Let $\mathcal{X}_j = \{X_1, \dots, X_j\}$ define the cumulated data up to step j ; the corresponding information set at that time can be denoted by σ -algebra, $\mathcal{F}_j = \sigma(\mathcal{X}_j)$. The total cost/loss of terminating the trial at the j -th step is the cost of patients, plus the loss of taking one of the two actions, A or R , whichever is smaller. It can be expressed as

$$L_{stop}(\mathcal{X}_j) = 2K_2 \sum_{i=1}^j B_i + \min \left[E\{L(\theta, A)|\mathcal{F}_j\}, E\{L(\theta, R)|\mathcal{F}_j\} \right], \quad (2.1)$$

where K_2 is the unit cost of enrolling a patient into the trial and

$$E\{L(\theta, A)|\mathcal{F}_j\} = K_1 \text{pr}(\theta > \theta_0|\mathcal{F}_j); \quad E\{L(\theta, R)|\mathcal{F}_j\} = K_0 \text{pr}(\theta \leq 0|\mathcal{F}_j).$$

At each interim analysis, we can use cumulated data up to that stage and estimate the expected loss of continuing the trial by observing one more block of data. Specifically, the expected loss of continuing the trial to the $(j + 1)$ -th stage can be expressed as the total cost for sampling up to the $(j + 1)$ -th step plus the expected minimum loss of accepting or rejecting the null hypothesis, given the data observed up to the j -th step,

$$L_{cont}(\mathcal{X}_j) = 2K_2 \sum_{i=1}^{j+1} B_i + E\left(\min \left[E\{L(\theta, A)|\mathcal{F}_{(j+1)}\}, E\{L(\theta, R)|\mathcal{F}_{(j+1)}\} \right] \middle| \mathcal{F}_j \right), \quad (2.2)$$

The outside expectation in (2.2) is with respect to the predictive distribution of the treatment difference. Specifically, the predictive density of X can be expressed as

$$g(x|\mathcal{X}_j) = \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta|\mathcal{X}_j)d\theta,$$

where $f(x|\theta)$ is the probability density of X . Given the observed data up to the j -th stage, a critical region, denoted by R_j , is derived from the loss function,

$$R_j = \left\{ \mathcal{X}_j : \frac{\text{pr}(\theta \leq 0|\mathcal{F}_j)}{\text{pr}(\theta > \theta_0|\mathcal{F}_j)} \leq \frac{K_1}{K_0} \right\}, \quad j = 1, 2, \dots \quad (2.3)$$

The acceptance region at each stage, A_j , is the complement of R_j . To search for the optimal adaptive design to minimize the expected loss, we use the following two-step strategies starting from the first block of data (with $j = 1$).

- (1) If $L_{stop}(\mathcal{X}_j) \leq L_{cont}(\mathcal{X}_j)$, we terminate the trial, and the maximum block size is j . Then, if cumulative data \mathcal{X}_j is in the rejection region R_j , we conclude that the new treatment is more effective than the control. Otherwise, the new treatment is no more effective than the control.
- (2) If $L_{stop}(\mathcal{X}_j) > L_{cont}(\mathcal{X}_j)$, we continue to observe the $(j + 1)$ -th block and repeat steps (1) and (2).

We use a one-step backward induction algorithm for this decision problem (DeGroot 1970); the algorithm is illustrated in detail in Section 2.3. The total number of blocks to be observed in the trial, denoted by M , is a stopping time, since $\{M \geq j\}$ only depends

on \mathcal{X}_j . In contrast to the sequential designs with a fixed maximum number of blocks, M is not pre-fixed but is a random integer. Thus, it is critical to ensure that the trial will be terminated with a finite number of interim analyses with the given loss function. The following theorem proves that the design will lead to the termination of trial with a finite number of blocks. The proof of Theorem 1 is given in the Appendix.

THEOREM 1. *If the unit cost for each sample, $K_2 > 0$, then M is a stopping time satisfying $P(M < +\infty \mid \theta) = 1$.*

2.2 Connections to the frequentist designs

Bayesian decision-theoretical designs have been compared with the frequentist group sequential designs in terms of their frequentist operating characteristics (Heitjan et al, 1992; and Lewis & Berry, 1994). Bayesian designs often use loss functions, whereas frequentist methods use pre-fixed error rates, to determine the sample sizes (Lindley, 1997). However, the comparisons are mainly empirical, and few theoretical connections have been established between these designs and the frequentist properties. We will describe an explicit relationship between the design parameters in the loss functions of the proposed design and the frequentist type I error rate.

The design parameters, K_i ($i = 0, 1, 2$) in the proposed loss function, along with the stopping rules, allow us to control the probabilities of type I and type II errors. With the specified loss function applied here, it is clear that a design with a set of parameters (aK_0, aK_1, aK_2) is equivalent to the design with parameters (K_0, K_1, K_2) for any positive value a . Therefore, the ratios of these parameters, $K_0 : K_1 : K_2$, determine the operating characteristics of a design. We first discuss the case with $\theta_0 = 0$.

In a clinical trial, if the loss function indicates that the trial should be terminated at step j , there are two possible actions to take: A or R . Based on the Likelihood Principle (Berger, 1985), the probability of making a false-positive conclusion at step j

is $P(R_j|\theta = 0)$, where

$$R_j = \left\{ \mathcal{X}_j : \frac{\text{pr}(\theta \leq 0|\mathcal{X}_j)}{\text{pr}(\theta > 0|\mathcal{X}_j)} \leq \frac{K_1}{K_0} \right\} = \left\{ \mathcal{X}_j : \text{pr}(\theta \leq 0|\mathcal{X}_j) \leq \frac{K_1}{K_0 + K_1} \right\}.$$

If all related density functions satisfy the regularity conditions, the posterior distribution of θ given \mathcal{X}_j is asymptotically normal with the mean and variance equal to their posterior mean, δ_j , and posterior variance, s_j^2 , respectively (Hartigan, Ch. 11, 1983). Thus, under H_0 , $\text{pr}(\theta \leq 0|\mathcal{X}_j)$ is asymptotically distributed as $\Phi(-\delta_j/s_j)$, where Φ is the standard normal cumulative distribution function. Under H_0 , δ_j/s_j converges in distribution to Z , where Z follows a standard normal distribution. Therefore, we have

$$\text{pr}(\theta \leq 0|\mathcal{X}_j) \rightarrow \Phi(Z), \quad \text{in distribution.}$$

Note that $\Phi(Z)$ follows a uniform distribution on $(0, 1)$. Thus, recall the rejection region R_j with the above property,

$$\text{pr}(R_j|\theta = 0) \rightarrow \text{pr} \left\{ \Phi(Z) \leq \frac{K_1}{K_0 + K_1} \right\} = \frac{K_1}{K_0 + K_1}, \quad \text{in distribution.}$$

For $\theta_0 > 0$, R_j shrinks as θ_0 increases. Therefore, we have

$$\limsup_{j \rightarrow \infty} \text{pr}(R_j|\theta = 0) \leq \frac{K_1}{K_0 + K_1}.$$

For a given type I error rate, α , we may use the above inequality as a guideline to choose the value of $K_1/(K_0 + K_1)$ to control the false-positive conclusion probability, if the overall sample size is sufficiently large. From equation (2.3), it is clear that R_j depends on K_0 and K_1 only through their ratio, K_0/K_1 . Specifically, we have

$$K_0/K_1 = (1 - \alpha)/\alpha, \tag{2.4}$$

by letting $K_1/(K_0 + K_1) = \alpha$.

In decision analyses, values in the loss function are often chosen to reflect as closely as possible the actual costs incurred in the trial (Gittins & Pezeshk, 2000 and Stallard et al., 1999). Here, we may use this as a guideline to decide the costs for each patient in the trial, K_2 . K_0 and K_1 are often much larger in magnitude relative to K_2 , implying

a great loss for making an incorrect decision at the end of the trial when comparing with the unit cost for one sample. For instance, if $K_2 = \$1000$ is the cost to enroll and treat one patient in the trial, then K_1 can be on the magnitude of \$100,000,000 by considering that hundreds of thousands of future patients would not benefit from this effective treatment if the trial concludes inefficacy of the treatment at the end. Moreover, the design parameters of the loss function play an important role to evaluate and control the loss at each interim analysis. Reflected from the specified loss function, the trial should continue if the posterior distribution at the interim analysis does not have a clear indication for the value of θ one way or the other. The cost of additional patients, by the unit of K_2 , may be paid in order to gain more information about the treatment under investigation. If the posterior distribution indicates a clear trend for the mass of θ , the terminal decision to identify a better treatment should be made to minimize the loss. This is also the time when the cost of enrolling more patients dominates the total cost. As pointed by one reviewer, a high value of K_1 implies that future patients might benefit from a new effective treatment. However, the new treatment may be superseded within a few years, which would reduce the “value” of the treatment. We acknowledge such a possibility, but it is often difficult to explicitly build this concern prospectively into a trial design (Anscombe, 1963 and Eales & Jennison, 1992).

2.3 *Special Case I: Normal responses*

In this section, we focus on continuous outcomes with a normal distribution to elaborate the details in the decision making rules and computations. In addition to the results in Section 2.2 for general distributions, we will derive a strict upper boundary to control the false-positive probability for outcomes with a normal distribution. Let

$$X_i = \bar{X}_{T_i} - \bar{X}_{C_i} \sim N(\theta, \sigma^2/B_i),$$

where σ^2 is known. It is not conceptually different with an unknown variance. The prior distribution for θ is assumed to be $N(\delta, \sigma^2/B_0)$, where B_0 can be interpreted as the “sample size” for the prior information on the treatment (Spiegelhalter et al, 1994).

Thus, we can denote $X_0 = \delta$ to keep the notation for prior information coherent with that from the interim analysis. After data from block j are observed, the posterior distribution of θ is $N(\delta_j, s_j^2)$, where

$$\delta_j = \frac{\sum_{i=0}^j B_i X_i}{\sum_{i=0}^j B_i}, \quad s_j^2 = \frac{\sigma^2}{\sum_{i=0}^j B_i}$$

are the posterior mean and variance, respectively. The rejection region, R_j , then follows

$$R_j = \left\{ \mathcal{X}_j : \frac{\Phi(-\delta_j/s_j)}{1 - \Phi\{(\theta_0 - \delta_j)/s_j\}} \leq \frac{K_1}{K_0} \right\}.$$

Note that

$$\frac{\Phi(-\delta_j/s_j)}{1 - \Phi\{(\theta_0 - \delta_j)/s_j\}}$$

is a decreasing function of δ_j , and

$$\sup_{\delta_j} \frac{\Phi(-\delta_j/s_j)}{1 - \Phi\{(\theta_0 - \delta_j)/s_j\}} = \infty, \quad \inf_{\delta_j} \frac{\Phi(-\delta_j/s_j)}{1 - \Phi\{(\theta_0 - \delta_j)/s_j\}} = 0.$$

Therefore, there exists a unique critical value, c_j , such that $R_j = \{\mathcal{X}_j : \delta_j \geq c_j\}$ or equivalently,

$$c_j = \arg \left\{ x : \frac{\Phi(-x/s_j)}{1 - \Phi\{(\theta_0 - x)/s_j\}} - \frac{K_1}{K_0} = 0 \right\}.$$

For the special case without an equivalence range, i.e. $\theta_0 = 0$, we can solve c_j explicitly. To control the frequentist properties, we like to solve the ratio of K_0/K_1 for any specified type I error rate, α . To simplify the notation, let

$$h = \Phi^{-1}\{K_1/(K_0 + K_1)\}.$$

It is of interest to solve h corresponding to a given α . Note that $h \leq 0$ based on the general rule for the selection of K_0 and K_1 . Specifically, the assumption that the type I error rate is smaller than the type II error rate leads to $K_0 \geq K_1$. For $j = 0, 1, \dots$, let $n_j = \sum_{i=0}^j B_i$ be the cumulated sample sizes at stage j including the ‘‘sample size’’ from the prior information. Under the null hypothesis of $\theta = 0$,

$$-\frac{\delta_j}{s_j} \sim N \left\{ -\frac{n_0 \delta}{\sigma \sqrt{n_j}}, \frac{n_j - n_0}{n_j} \right\},$$

then the probability of rejecting the null hypothesis at the j -th interim analysis follows as

$$\text{pr}(R_j|\theta = 0) = \text{pr}(\delta_j/s_j > h|\theta = 0) = \Phi \left\{ \frac{h\sigma\sqrt{n_j} + n_0\delta}{\sigma\sqrt{(n_j - n_0)}} \right\}. \quad (2.5)$$

To solve h for all $j \geq 1$, we need to find an upper boundary for (2.5). Note that the right-hand side of (2.5) is a unimodal function of n_j . By taking the first derivative for n_j , it is clear that the function reaches its maximum at $\sqrt{n_j} = -h\sigma/\delta$, and the function increases when $\sqrt{n_j} \leq -h\sigma/\delta$. Therefore,

$$\sup_{n_j} \Phi \left\{ \frac{h\sigma\sqrt{n_j} + n_0\delta}{\sigma\sqrt{(n_j - n_0)}} \right\} \leq \Phi \left\{ -\frac{\sqrt{(h^2\sigma^2 - n_0\delta^2)}}{\sigma} \right\}. \quad (2.6)$$

It is worth noting that the $h^2\sigma^2 - n_0\delta^2 \geq 0$ as long as $n_0 \leq n_j$, which is always the case. We set the right-hand side of (2.6) equal to α , and we solve for h , which follows

$$h_1 = - \left(z_\alpha^2 + \frac{n_0\delta^2}{\sigma^2} \right)^{1/2},$$

where $\Phi(z_\alpha) = 1 - \alpha$. In a similar vein, we can solve the corresponding h when $\sqrt{n_j} > -h\sigma/\delta$. The total number of blocks is at least one, thus $n_j \geq n_1$ for $j \geq 1$. When $\sqrt{n_1} > -h\sigma/\delta$,

$$\sup_{n_j} \Phi \left\{ \frac{h\sigma\sqrt{n_j} + n_0\delta}{\sigma\sqrt{(n_j - n_0)}} \right\} \leq \Phi \left\{ \frac{h\sigma\sqrt{n_1} + n_0\delta}{\sigma\sqrt{(n_1 - n_0)}} \right\}. \quad (2.7)$$

Let the right-hand side of (2.7) equal α , and solve for h . The solution is

$$h_2 = \frac{z_{1-\alpha}\sigma\sqrt{(n_1 - n_0)} - n_0\delta}{\sigma\sqrt{n_1}}.$$

Now, for any given significance level α , we can determine the ratio of K_0/K_1 based on this upper bound:

$$\frac{K_0}{K_1} = \begin{cases} \{1 - \Phi(h_1)\}/\Phi(h_1), & \text{if } \sqrt{(n_1)} \leq \sqrt{\{(\frac{\sigma}{\delta})^2 z_\alpha^2 + n_0\}}; \\ \{1 - \Phi(h_2)\}/\Phi(h_2), & \text{if } \sqrt{n_1} > \sqrt{\{(\frac{\sigma}{\delta})^2 z_\alpha^2 + n_0\}}. \end{cases} \quad (2.8)$$

With K_0/K_1 defined by (2.8), we have

$$\sup_j \text{pr}(R_j|\theta = 0) \leq \alpha.$$

Note that using the ratio K_0/K_1 derived from (2.8) leads to a more conservative design compared with that using the ratio of $K_0/K_1 = (1 - \alpha)/\alpha$ from Section 2.2, since the former is an absolute upper boundary.

The one-step backward induction algorithm is used for the evaluation of loss functions at each interim analysis. After observing data up to the j -th interim analysis, the loss to terminate the trial at the j -th step follows

$$L_{stop}(\mathcal{X}_j) = 2K_2 \sum_{i=1}^j B_i + \min \left[K_1 \left\{ 1 - \Phi \left(-\frac{\delta_j}{s_j} \right) \right\}, K_0 \Phi \left(-\frac{\delta_j}{s_j} \right) \right].$$

We also need to evaluate the predictive loss to continue the trial to the $(j + 1)$ -th block, based on data up to the j -th step. The predictive distribution of X_{j+1} given observed data up to the j -th step is

$$X_{j+1} | \mathcal{X}_j \sim N \left(\delta_j, s_j^2 + \frac{\sigma^2}{B_{j+1}} \right).$$

For each possible value of X_{j+1} , say x_{j+1} , we can compute the posterior mean and posterior variance recursively as the following:

$$\delta_{j+1} = \frac{n_j \delta_j + B_{j+1} x_{j+1}}{n_j + B_{j+1}}, \quad s_{j+1}^2 = \frac{\sigma^2}{n_j + B_{j+1}}.$$

The predicted loss of continuing and observing one more block is then

$$L_{cont}(\mathcal{X}_j) = 2K_2 \sum_{i=1}^{j+1} B_i + \int_{-\infty}^{+\infty} \min \left[K_1 \left\{ 1 - \Phi \left(-\frac{\delta_{j+1}}{s_{j+1}} \right) \right\}, K_0 \Phi \left(-\frac{\delta_{j+1}}{s_{j+1}} \right) \right] d\Phi \left\{ \frac{x_{j+1} - \delta_j}{(s_j^2 + \sigma^2/B_{j+1})^{1/2}} \right\},$$

where the integral is to the variable x_{j+1} . If $L_{stop} < L_{cont}$, the trial stops at the j -th block. Otherwise, we observe the $(j + 1)$ -th block data and the algorithm repeats.

2.4 Special case II: Binary responses

In this section we consider another important special case with binary outcomes. Consider a clinical trial comparing two treatments for a binary outcome. Assume that the

true rates of success are denoted by p_t and p_c for the new treatment and control group, respectively. Then

$$X_{T_i}|p_t \sim \text{Binomial}(B_i, p_t), \quad X_{C_i}|p_c \sim \text{Binomial}(B_i, p_c),$$

The difference in efficacy is $\theta = p_t - p_c$. We assume the following prior distributions for p_t and p_c :

$$p_t \sim \text{Be}(a_t, b_t), \quad p_c \sim \text{Be}(a_c, b_c).$$

The density function for θ is

$$\pi(\theta|a_t, b_t, a_c, b_c) = \begin{cases} \int_0^1 \text{dbeta}(\theta + x, a_t, b_t) * \text{dbeta}(x, a_c, b_c) dx, & \text{if } -1 < \theta < 0; \\ \int_0^{1-\theta} \text{dbeta}(\theta + x, a_t, b_t) * \text{dbeta}(x, a_c, b_c) dx, & \text{if } 0 < \theta < 1 \end{cases}$$

where $\text{dbeta}(x, a, b)$ is the density function of the beta distribution with parameters a and b . Because of the conjugate nature of the beta distribution, the posterior distributions continue to have beta distributions. The sufficient statistic by the end of the j -th stage is denoted by

$$(s_{t_j}, f_{t_j}, s_{c_j}, f_{c_j}), \quad \text{where} \quad s_{t_j} + f_{t_j} = \sum_{i=1}^j B_i, \quad s_{c_j} + f_{c_j} = \sum_{i=1}^j B_i.$$

s_{t_j} and f_{t_j} are the cumulated number of successes and failures observed on the treatment arm up to stage j , and the same notations, s_{c_j} and f_{c_j} , are used for the control arm. After observing data from block j and terminating the trial, the expected losses for the two decisions, A and R , follow

$$\begin{aligned} E\{L(\theta, A)|\mathcal{X}_j\} &= K_1 \int_{\theta_0}^1 \pi(\theta|a_{t_j}, b_{t_j}, a_{c_j}, b_{c_j}) d\theta, \\ E\{L(\theta, R)|\mathcal{X}_j\} &= K_0 \int_{-1}^0 \pi(\theta|a_{t_j}, b_{t_j}, a_{c_j}, b_{c_j}) d\theta. \end{aligned} \tag{2.9}$$

where

$$a_{t_j} = a_t + s_{t_j}, \quad b_{t_j} = b_t + f_{t_j}, \quad a_{c_j} = a_c + s_{c_j}, \quad b_{c_j} = b_c + f_{c_j}.$$

Following a transformation to the integrals in (2.9), we have

$$\begin{aligned} E\{L(\theta, A)|\mathcal{X}_j\} &= K_1 \int_0^{1-\theta_0} \text{dbeta}(x, a_{c_j}, b_{c_j}) \{1 - \text{pbeta}(\theta_0 + x, a_{t_j}, b_{t_j})\} dx \\ E\{L(\theta, R)|\mathcal{X}_j\} &= K_0 \int_0^1 \text{dbeta}(x, a_{c_j}, b_{c_j}) \text{pbeta}(x, a_{t_j}, b_{t_j}) dx \end{aligned}$$

where $pbeta(\cdot, a, b)$ is the cumulative distribution function of $Be(a, b)$.

The loss of stopping the trial at the j -th block or observing one more block follows (2.1) and (2.2), respectively. The expected loss to observe one more block refers to the predictive distribution of $s_{t_{j+1}}, s_{c_{j+1}}$ given (s_{t_j}, s_{c_j}) in block $(j + 1)$, which is given by

$$\text{pr}(s_{t_{j+1}}, s_{c_{j+1}} | s_{t_j}, s_{c_j}) = \binom{B_{j+1}}{s_{t_{j+1}} - s_{t_j}} \binom{B_{j+1}}{s_{c_{j+1}} - s_{c_j}} \frac{\beta(a_{t_{j+1}}, b_{t_{j+1}})}{\beta(a_{t_j}, b_{t_j})} \frac{\beta(a_{c_{j+1}}, b_{c_{j+1}})}{\beta(a_{c_j}, b_{c_j})},$$

where $\beta(\cdot, \cdot)$ is a beta function.

Even though it is possible to derive an absolute upper boundary for binary outcomes to control the type I error rate as for normal outcomes, the derivation is much more tedious compared to that for the normal distribution. In contrast, we can use the asymptotic boundary derived in Section 2.2 to choose the ratio of K_0/K_1 in order to control the false-positive rate. An alternative way is to use the normal approximation for the binary outcome to apply the boundary in §2.3.

3 NUMERICAL RESULTS

We evaluate the operating characteristics of the proposed designs by extensive Monte Carlo simulations, and compare them with the performance of the existing group sequential designs, including the frequentist designs of Pocock (1977), O'Brien-Fleming (1979), and the adaptive self-designing trial (Shen & Fisher, 1999). For a direct comparison with the frequentist designs, the true values of θ are given as fixed, which can be considered to have a prior with all mass on θ . The block sizes are pre-fixed, by letting semi-block sizes $B_0 = 1$, $B_1 = \max\{B, 10\}$, and $B_i = B$ for $i \geq 2$. For each scenario, the same block sizes are used among the three adaptive designs; and the same equal block size, $2B$, are used for the Pocock and O'Brien-Fleming, OBF, designs under comparison. All simulations are repeated 10,000 times. The rates of type I and type II errors and the average total sample number (ASN) for each design are estimated and compared among different designs.

It is worth noting that the proposed Bayesian design and self-designing trial do not enforce a maximum sample size, while the Pocock and O'Brien-Fleming trials have maximum sample sizes. We consider one-sided hypothesis testing. The modified one-sided Pocock and O'Brien-Fleming designs are used to ensure comparability among the designs. The maximum number of groups, N , in the designs of Pocock and O'Brien-Fleming is obtained from Table 2 in Pocock (1977) with $\Delta = \sqrt{B}\delta/\sigma$, where the block size, $2B$, is a constant for each stage. For a given N , Pocock's interim test boundaries are of the form

$$-z < \sqrt{(jB)} \sum_{i=1}^j X_i / (j\sigma) < z,$$

where the z value is obtained from Table 1 in Pocock (1977). The asymptotic boundaries for O'Brien-Fleming's one-sided test have the form

$$-z_l / \{\Delta\sqrt{(j)}\} + 0.5\Delta\sqrt{(j)} < \sqrt{(jB)} \sum_{i=1}^j X_i / (j\sigma) < z_u \sqrt{(N/j)},$$

where $z_l = \log\{\beta/(1 - \alpha)\}$ and z_u are obtained from DeMets and Ware (1982).

We first generate normal outcomes with mean $\theta = 0.5$ and variance $\sigma = 1$. The value of δ is set at 0.4, 0.5, 0.6, and 0.7. Using the guidelines in Sections 2.2 and 2.3, K_0/K_1 is determined by (2.8) for "Bayes Adapt I" or by (2.4) for "Bayes Adapt II" in Table 1 for the given type I error rate of 0.025. The value of K_2/K_1 is searched to yield the frequentist type II error rate empirically for given δ and B .

Table 1 shows that the frequentist type I error rates are strictly maintained under the specified level for the proposed Bayesian designs with either boundary. Because of the upper boundary being used to control the type I error rate, the proposed design is conservative in terms of the type I error. The magnitude is similar to that of the self-designing trial, but there is no additional futility stopping rule required for the proposed Bayesian designs. It is not surprising that the boundary taking (2.8) is more conservative than the other from (2.4), since the former is a strict upper boundary. When the prior mean, δ , is over-stated relative to the true treatment efficacy, θ , the frequentist group sequential designs with the fixed maximum sample sizes lead to a substantial loss of

power. On the other hand, the proposed Bayesian designs and the self-designing trial can update the prior information using the observed data through interim analyses.

It is interesting to note the advantages of the proposed Bayesian designs over the self-designing trial in terms of both power and average sample number, ASN, when the type I error rates for both designs are maintained. For instance, when the expected treatment effect is 40% over-estimated to the true effect, the Bayesian design can still maintain almost 90% power, whereas the self-designing trial achieves 77% power. One possible reason is that the design based on the decision theoretical approach uses sufficient statistics from interim data, while the self-designing trial does not. Of course, it is not surprising that the frequentist group sequential designs have power that is only between 60-67% with fixed maximum sample sizes. Figure 1 shows a histogram for the number of blocks based on 100 simulated trials. More than 75% of the trials are terminated with the number of interim analyses being four or less.

For binary outcomes, first we are interested in comparisons with the usual fixed sample design. We use constant $Be(1, 1)$ and $Be(2, 2)$ priors with block sizes 32 or 48. The empirical type I error rates, denoted by $\hat{\alpha}$, are determined when taking $p_t = p_c = 0.5$. The frequentist type II error rates are estimated with $p_t = 0.5 + \delta/2$ and $p_c = 0.5 - \delta/2$. Using normal approximation to the binary outcomes, the ratio of K_0/K_1 in the loss function is determined by (2.4).

Tables 2 shows the frequentist error rates for the proposed Bayesian designs. As expected, the proposed Bayesian designs have false-positive probabilities around the specified level under the null hypothesis. When the true difference, θ , is smaller than the prior mean, δ , the proposed designs can extend the trial and achieve adequate power using interim data. In contrast, the usual fixed sample designs do not have the flexibility to adjust for the total sample size to achieve the desired power.

Secondly, we are particularly interested in comparisons with the designs of Lewis & Berry (1994), which were proposed for binary outcomes only. To directly compare the proposed design with that of Lewis & Berry (1994), we use the same set-up as in their simulations, including the same block sizes and priors as in Table 4 of Lewis and Berry

and assuming $\delta = \theta$. When $\delta = \theta$, the proposed design has power similar to that of Lewis & Berry's design, but the average sample number is slightly increased (less than 5%) under the alternative. Under the null hypothesis, the average sample number is smaller for the proposed design than for the design of Lewis and Berry. However, the computation of the proposed design is much less intensive compared to that of Lewis and Berry's design, and the implementation is straightforward with one-step backward induction. A major difference between the two designs is that the design of Lewis and Berry has a pre-fixed maximum number of blocks, while our proposed design does not have such a restriction. Thus, the power for the Lewis and Berry design will be reduced if $\theta < \delta$.

4 EXAMPLE

To illustrate the proposed design, we apply the design to a completed randomized animal experiment that evaluated epinephrine in cardiac arrest (Niemann et al., 1992; and Lewis & Berry, 1994). The experiment compared cardiac resuscitation outcomes in a canine model using the standard therapy, immediately delivering an electric shock, versus an alternative therapy, high-dose epinephrine therapy plus conventional CPR, before a countershock of prolonged ventricular fibrillation. The study was originally designed as a frequentist fully sequential trial to detect an increase in the proportion of animals successfully resuscitated from 20% to 60% with the new therapy. The trial was terminated after observing outcomes from 28 animals. The observed data are presented in the following table, where S_c and S_t are the cumulative number of successes for each block in the control arm and treatment arm, respectively.

Block	$2B_i$	S_c	S_t
1	20	3	6
2	8	3	9

Using the proposed Bayesian design with 20 animals in the first block and 8 animals in the following blocks, $K_0/K_1 = 19$, $K_2 = 0.005$, and flat priors of $Be(1,1)$ for p_c

and p_t , the frequentist type I and type II error rates are 0.035 and 0.834, from Monte Carlo simulations. Based on the decision rules from our design, the experiment should be terminated at the end of the second block with a total sample size of 28. The null hypothesis is then rejected, and the posterior probability that $\theta > 0$ is 0.987.

5 DISCUSSION

One new feature of the proposed Bayesian adaptive designs is that the total sample size is adaptively determined by prior information and cumulated data, rather than prefixed as in the designs described by Berry & Ho (1988) and Lewis & Berry (1994). The design based on a decision theoretical approach, on the other hand, is fundamentally different from the adaptive designs developed in the frequentist framework. With the flexible loss functions, the proposed designs can be terminated at any stage through the loss functions for either futility or superiority of the treatment arm. Thus, we are able to simultaneously integrate efficacy, futility, and cost in decision making, whereas the self-designing trial and other existing adaptive designs often require dealing with these factors separately. Such a unified loss function makes it possible to find optimal strategies under a variety of circumstances.

As elaborated in Lewis & Berry (1994), the use of traditional Bayesian sequential designs in the regulatory setting is often hampered by concerns of violating frequentist properties. To be consistent with regulatory standards, we propose Bayesian designs that are derived and checked to have acceptable frequentist properties. An asymptotic upper boundary to control the false-positive error rate through the loss function is derived for general distributions, and an absolute boundary is obtained for normal distributions. Note that the derivation of (2.4) is based on the Likelihood Principle (Berger, 1985), therefore the stopping rules are not taken into account in the calculation of the probability to reject the null hypothesis at a given stopping time. In contrast, from the frequentist view point, the terminal decision probability depends on what stopping rules are used, and not just on the observed data (Schervish, 1995).

The proposed designs are general enough to accommodate various types of outcomes, because the use of one-step backward induction makes it possible to estimate the predicted losses of different decisions in the monitoring process. In contrast, the M -step backward induction used in Lewis & Berry (1994) can be difficult and extremely time consuming computationally for continuous outcomes with even moderate M , and is only feasible for binary outcomes. Regardless of the computation intensity for M -step backward induction, it is interesting to know how much efficiency may be gained by using M -step backward induction compared to one-step backward induction. Intuitively, for fully sequential designs, there can be a significant gain when using M -step backward induction over one-step backward induction, because only one patient is observed each time. On the other hand, for group sequential designs that are commonly used in clinical trials, the data from each block provide adequate information for the prediction. As a result, the gain by performing M -step backward induction is minor compared to the use of one-step backward induction in group sequential designs, as we found in the simulations. However, compared to M -step backward induction, the computation for one-step backward induction is much simpler and the algorithm is easier to implement for various distributions. In many clinical trials, the primary outcomes of interest often need long-term follow-up with potential censoring (Shen and Cai, 2003). To implement this design to censored survival data requires further technical modifications.

ACKNOWLEDGEMENT

The authors thank the editor and two referees for their insightful comments. This research was done in part while the first author was visiting the Biostatistics Department at The University of Texas, M. D. Anderson Cancer Center. This research was partially supported by a grant from the National Cancer Institute.

APPENDIX

Proof of Theorem 1

For $j = 1, 2, \dots$, define

$$Y_j = \min \left[E\{L(\theta, A)|\mathcal{F}_j\}, E\{L(\theta, R)|\mathcal{F}_j\} \right].$$

By Jensen's inequality, we have

$$E(Y_{j+1}|\mathcal{F}_j) \leq \min \left(E\left[E\{L(\theta, A)|\mathcal{F}_{j+1}\} \mid \mathcal{F}_j \right], E\left[E\{L(\theta, R)|\mathcal{F}_{j+1}\} \mid \mathcal{F}_j \right] \right) = Y_j.$$

The stochastic process $\{(Y_j, \mathcal{F}_j); j \geq 1\}$ is a nonnegative uniformly bounded supermartingale process. According to the martingale convergence theorem, Y_j converges almost surely to a bounded random variable Y_∞ (Chung, 1974). Therefore,

$$0 \leq Y_j - E(Y_{j+1}|\mathcal{F}_j) \rightarrow Y_\infty - Y_\infty = 0 \quad \text{a.s. as } j \rightarrow \infty.$$

Let $D_j = \{\mathcal{X}_j : L_{stop}(\mathcal{X}_j) \leq L_{cont}(\mathcal{X}_j)\}$, where $L_{stop}(\mathcal{X}_j)$ and $L_{cont}(\mathcal{X}_j)$ are given in (2.1) and (2.2), respectively. The above arguments together with the unit cost, $K_2 > 0$, imply that $\text{pr}(D_j)$ converges to 1. It is equivalent to $\text{pr}(M < \infty) = 1$. \square

REFERENCES

- Anscombe, F. J. (1963). Sequential medical trials. *J. Am. Statist. Assoc.* **58**, 365-83.
- Barber, S. & Jennison, C. (2002). Optimal asymmetric one-sided group sequential tests. *Biometrika* **89**, 49-60.
- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.
- Berry, A.D. & Ho, C. (1988). One-sided sequential stopping boundaries for clinical trials: a decision-theoretical approach. *Biometrics* **44**, 219-27.
- Chung, K.L. (1974). *A Course in Probability Theory*. New York: Academic Press.
- Cressie, N. & Biele, J. (1994). A sample-size-optimal Bayesian procedure for sequential pharmaceutical trials. *Biometrics* **50**, 700-11.

- DeGroot, M.H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.
- DeMets, D.L. & Ware, J.H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **69**, 661-63.
- Eales, J.D. & Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13-24.
- Fisher, L. (1998). Self-designing clinical trials. *Statist. in Med.* **17**, 1551-62.
- Freedman, B. (1987). Equipoise and the ethics of clinical research. *New Engl. J. Med.* **317**, 141-45.
- Gittins, J & Pezeshk H. (2000). How large should a clinical trial be? *The Statistician* **49**, 177-87.
- Hartigan, J.A. (1983). *Bayes Theory*. New York: Springer-Verlag.
- Heitjan, D.F., Houts, P.S., & Harvey, H.A. (1992) . A decision-theoretic evaluation of early stopping rules. *Statist. in Med.* **11**, 673-83.
- Lai, T.L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *Ann. Statist.* **1**, 659-73.
- Lehmacher, W. & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286-90.
- Lewis, R.J. & Berry, D.A. (1994). Group sequential clinical trials: A classical evaluation of Bayesian decision-theoretic designs. *J. Am. Statist. Assoc.* **89**, 1528-34.
- Lindley, D. (1997). The choice of sample size. *The Statistician* **46**, 129-38.
- Liu, Q., and Chi, G.Y.H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* **57**, 172-7.

- Müller, H-H & Schäfer, H. (2001). Adaptive group sequential design for clinical trials: combining the advantages of adaptive and of classical group sequential procedure. *Biometrics* **57**, 886-91.
- Niemann, J.T., Cairns, C.B., Sharma, J., & Lewis, R.J. (1992). Treatment of prolonged ventricular fibrillation: immediate countershock versus high-dose epinephrine and CPR preceding countershock. *Circulation* **85**, 281-7.
- O'Brien, P.C., & Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-56.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-9.
- Proschan, M. & Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315-24.
- Schervish, M.J. (1995). *Theory of Statistics*. New York: Springer-Verlag.
- Shen, Y. & Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190-7.
- Shen, Y. & Cai, J. (2003). Sample size reestimation for clinical trials with censored survival data. *J. Am. Statist. Assoc.* **98**, 418-26.
- Spiegelhalter, D.J., Freedman, L.S., & Parmar, M.K.B.,(1994). Bayesian approaches to randomized trials (with discussion). *J. R. Statist. Soc. B* **157** 357-416.
- Stallard, N., Thall, P.F., & Whitehead J. (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* **55**, 971-7.
- Thach, C. & Fisher, L.D. (2002). Self-designing two-stage trials to minimize expected costs. *Biometrics* **58**, 432-8.
- Tsiatis, A.A. & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**, 367-78.

Table 1. *The comparison of power and average sample number (ASN) between the Bayesian designs and other group sequential designs with one-sided $\alpha = 0.025$, and true $\theta = 0$ at null and $\theta = 0.5$ under the alternative*

Design	δ	$B = 6$				$B = 8$			
		$\hat{\alpha}$	ASN_{α}	$1-\hat{\beta}$	ASN_{β}	$\hat{\alpha}$	ASN_{α}	$1-\hat{\beta}$	ASN_{β}
Pocock	0.4	0.026	174.7	0.985	74.9	0.025	171.3	0.982	78.9
OBF	0.4	0.022	68.6	0.985	84.4	0.025	70.9	0.984	83.7
Self-designing	0.4	0.012	84.4	0.911	87.0	0.010	92.9	0.931	90.0
Bayes Adapt I	0.4	0.014	49.7	0.934	72.7	0.012	59.3	0.959	79.5
Bayes Adapt II	0.4	0.016	51.1	0.938	72.0	0.017	60.4	0.961	78.3
Pocock	0.5	0.023	117.2	0.904	68.6	0.024	124.9	0.923	74.0
OBF	0.5	0.024	47.7	0.929	65.5	0.024	49.7	0.937	69.1
Self-designing	0.5	0.013	62.5	0.888	78.5	0.014	71.4	0.918	83.9
Bayes Adapt I	0.5	0.012	45.8	0.921	69.4	0.016	54.4	0.946	75.5
Bayes Adapt II	0.5	0.018	47.8	0.930	68.2	0.017	54.9	0.951	73.7
Pocock	0.6	0.025	82.0	0.760	59.7	0.026	94.0	0.811	67.5
OBF	0.6	0.025	35.0	0.810	50.8	0.214	37.3	0.848	55.9
Self-designing	0.6	0.013	51.9	0.836	70.3	0.014	59.9	0.869	74.6
Bayes Adapt I	0.6	0.013	42.0	0.905	66.6	0.013	49.2	0.928	71.7
Bayes Adapt II	0.6	0.020	45.3	0.914	64.6	0.020	51.3	0.942	69.7
Pocock	0.7	0.025	58.9	0.595	49.3	0.022	63.1	0.616	54.2
OBF	0.7	0.024	26.7	0.670	38.1	0.023	28.9	0.668	39.9
Self-designing	0.7	0.014	43.9	0.761	61.3	0.014	49.3	0.772	65.1
Bayes Adapt I	0.7	0.015	39.8	0.889	63.9	0.015	45.9	0.920	68.9
Bayes Adapt II	0.7	0.022	42.7	0.907	61.6	0.021	48.9	0.932	66.1

Bayes Adapt I: K_0/K_1 is determined by formula (2.8). Bayes Adapt II: K_0/K_1 satisfies the equation $K_1/(K_1 + K_0) = \alpha$. For both designs, $K_2/K_1 = 0.1^4 B * \delta^3$. ASN_{α} and ASN_{β} are average sample sizes under $\theta = 0$ and $\theta = 0.5$, respectively.

Table 2. The comparison of power and average sample number (ASN) between the proposed Bayesian optimal design and fixed sample design for binary responses; priors are $Be(1, 1)$; $\delta = 0.4$; $p_t = p_c = 0.5$ for H_0 ; $p_t = 0.5 + \theta/2$ and $p_c = 0.5 - \theta/2$ for H_1

θ	Proposed Bayesian design II				Fixed sample design	
	B=16		B=24		$P(\text{reject } H_0)$	ASN
	$P(\text{reject } H_0)$	ASN	$P(\text{reject } H_0)$	ASN		
0.40	0.921	46.0	0.973	55.0	0.906	46
0.36	0.874	50.4	0.945	57.3	0.835	46
0.32	0.801	52.3	0.875	60.9	0.741	46
0.28	0.710	54.0	0.812	64.4	0.631	46
0.00	0.047	40.2	0.047	56.2	0.050	46

$K_0/K_1 = 19$ which satisfies the equation $K_1/(K_1 + K_0) = \alpha$ for $\alpha = 0.05$. $K_2/K_1 = 0.005$.

Table 3. The comparison of power and average sample number (ASN) between the proposed Bayesian optimal design and Lewis and Berry Bayesian design for binary responses; $p_t = p_c = 0.5$ for H_0 ; $p_t = 0.5 + \delta/2$ and $p_c = 0.5 - \delta/2$ for H_1

Priors	δ	B	Proposed Bayesian design II				Lewis-Berry's design			
			$\hat{\alpha}$	ASN_α	$1 - \hat{\beta}$	ASN_β	$\hat{\alpha}$	ASN_α	$1 - \hat{\beta}$	ASN_β
$B(1, 1)$	0.4	16	0.047	40.2	0.921	46.0	0.039	42.1	0.946	44.3
	0.2	16	0.030	131.4	0.926	171.7	0.035	155.9	0.960	161.4
$B(2, 2)$	0.4	16	0.030	40.6	0.942	48.6	0.027	38.3	0.907	46.2
	0.2	16	0.026	125.5	0.917	171.9	0.034	152.3	0.958	162.1

$K_0/K_1 = 19$ which satisfies the equation $K_1/(K_1 + K_0) = \alpha$ for $\alpha = 0.05$. $K_2/K_1 = 0.005$ for $\delta = 0.4$. $K_2/K_1 = 0.00003$ for $\delta = 0.2$.

Histogram of the Number of Blocks

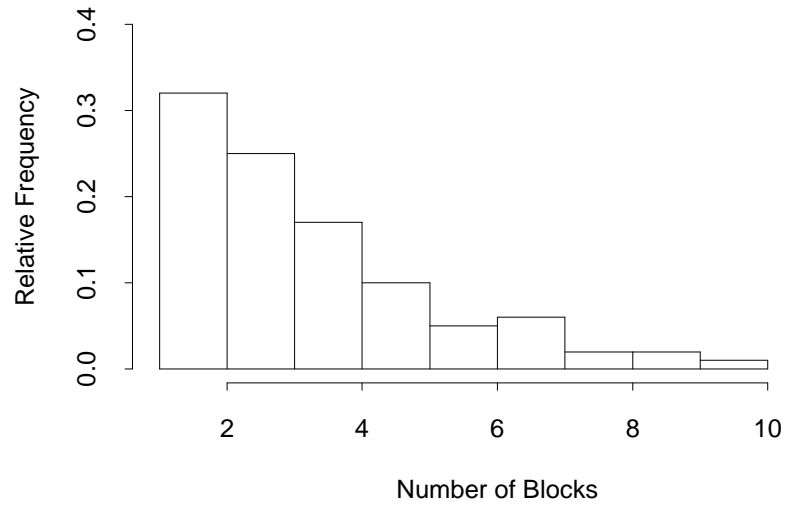


Figure 1: The histogram of the number of blocks in relative frequency, where $\theta = 0.6$, $\delta = 0.6$, and $B = 12$.