Published in final edited form as:

J Stat Plan Inference. 2010 July ; 140(7): 1834–1843. doi:10.1016/j.jspi.2010.01.003.

Bounding the Resampling Risk for Sequential Monte Carlo Implementation of Hypothesis Tests

Hyune-Ju Kim

Department of Mathematics, Syracuse University, Syracuse, NY 13244

Abstract

Sequential designs can be used to save computation time in implementing Monte Carlo hypothesis tests. The motivation is to stop resampling if the early resamples provide enough information on the significance of the p-value of the original Monte Carlo test. In this paper, we consider a sequential design called the B-value design proposed by Lan and Wittes and construct the sequential design bounding the resampling risk, the probability that the accept/reject decision is different from the decision from complete enumeration. For the B-value design whose exact implementation can be done by using the algorithm proposed in Fay, Kim and Hachey, we first compare the expected resample size for different designs with comparable resampling risk. We show that the B-value design has considerable savings in expected resample size compared to a fixed resample or simple curtailed design, and comparable expected resample size to the iterative push out design of Fay and Follmann. The B-value design is more practical than the iterative push out design in that it is tractable even for small values of resampling risk, which was a challenge with the iterative push out design. We also propose an approximate B-value design that can be constructed without using a specially developed software and provides analytic insights on the choice of parameter values in constructing the exact B-value design.

Keywords

B-Value; Bootstrap; Permutation; Sequential Design; Approximation

1 Introduction

When we implement Monte Carlo hypotheses tests (MC tests) such as bootstrap or permutation tests, computation time can be often saved by early stopping. The main idea is to stop early and reject/accept the null hypothesis if the early replications provide enough evidence for a significant/insignificant p-value of the original MC test. A simple sequential design is a curtailed design where we stop and reject or accept the null hypothesis when we have the replications enough to ensure the rejection or acceptance of the null hypothesis with a full enumeration of the MC test. Fay et al. (2007) proposed using a truncated sequential probability ratio (tSPRT) boundary which is minimax with respect to the resampling risk, and provided an algorithm to calculate a valid p-value after reaching the stopping boundary. As discussed in Fay et al. (2007), there is another way to determine an MC boundary: by minimizing the resampling risk over a class of possible distributions for the p-value. Fay and Follmann (2002) used the class of distributions for the p-values generated by location shifts for a standard

normal test statistic and approximated the associated distributions for the p-values within this class using beta distributions. Within this class of distributions, Fay and Follmann (2002) found by numerical search the parameter values which gave the largest resampling risk for fixed boundaries, defining a “worst case” distribution. Their approach was then to determine the design for the MC test which bounded the resampling risk for that worst case distribution and would therefore bound the resampling risk for the entire class of distributions. Fay and Follmann (2002) used this approach in motivating the iterative push out (IPO) design. In Fay and Follmann (2002), the IPO boundary was shown to provide savings in the expected resample size compared to the fixed resample size design and a simple curtailed design. Its implementation, however, is not practical for small values of resampling risk.

Fay et al. (2007) did not fully explore using the Fay and Follmann (2002) approach for bounding the resampling risk within a class of distributions for the p-value, but only plotted resampling risk as a function of fixed p-values. In this paper, we explore the Fay and Follmann (2002) approach and focus on sequential designs which bound or approximately bound the resampling risk within the previously mentioned class of distributions for the p-value. In doing so, we use the B-value boundary, which is a tSPRT boundary written in terms of a B-value proposed by Lan and Witts (1988). The B-value is a statistic which produces the expected value of the test statistic at the end of study, conditional on information up to some point, and we make a decision based on this projected value. We construct the B-value boundary either by using the numerical algorithm introduced in Fay et al. (2007), FKH algorithm, or by using approximations introduced in this paper. We call the latter design the approximate B-value design, and it is much easier to construct, but it only provides a decision rule, without estimating the p-value. In order to numerically construct the B-value boundary, we implement the algorithm developed in Fay et al. (2007), which unlike the IPO design, provides a tractable design to bound the resampling risk below small values (e.g., below 0.01). Considering the class of the p-value distributions used in Fay and Follmann (2002), we empirically show that the B-value design can provide expected resample sizes significantly smaller than those of the fixed resample size and the simple curtailed sequential designs, and the expected resample size comparable to the IPO design. As another way of implementing the B-value design, we propose a method to construct an approximate B-value boundary by using results developed in sequential analysis. The approximate B-value design asymptotically bounds the resampling risk and the associated test asymptotically maintains its size. These approximations help one to understand analytic characteristics of the sequential MC test, and the approximate B-value design can be implemented without an aid of a special algorithm.

This paper is organized as follows. In Section 2, we introduce the notation and define the B-value design. We also summarize some exact properties of the B-value design, compared to the fixed resample size, curtailed and IPO designs. In Section 3, we propose the approximate B-value design, which only requires a simple numerical integration in order to make an accept/reject decision early, subject to a desirable resampling risk. Section 4 includes an example and further remarks.

2 B-value Design

2.1 Overview

Let T be a test statistic, whose large values lead to the rejection of the null hypothesis, H_0 , and $T_0 = T(\mathbf{d}_0)$ denote the value of T observed for the original sample. When independent replications of data, say $\mathbf{d}_1^*, \mathbf{d}_2^*, \dots$, are taken and we obtain $T_1 = T(\mathbf{d}_1^*), T_2 = T(\mathbf{d}_2^*), \dots$, the p-value of the MC test is defined as $p(\mathbf{d}_0) = P(T_i \geq T_0 | \mathbf{d}_0)$ for all possible replications of T . For given $\mathbf{d}_0$, consider $p(\mathbf{d}_0) = p$ as fixed. With such p , our goal is to stop taking Monte Carlo samples when we are reasonably sure that either $p \leq \alpha$ (i.e., the decision after an infinite resample is to reject the original null hypothesis) or $p > \alpha$ (i.e., the decision after an infinite

resample is to fail to reject the original null hypothesis). One way to think about the Monte Carlo problem is that we have independent and identically distributed random variables, $X_i = I(T_i \geq T_0)$, each distributed as a Bernoulli distribution with parameter p , and we want to design a study to decide between the two possibilities for the parameter p , either $p \leq \alpha$ or $p > \alpha$.

Suppose that $X_1, X_2, \dots$, are the independent and identically distributed Bernoulli random variables with parameter p , and consider the problem of testing $H_0^*: p \leq \alpha$ versus $H_1^*: p > \alpha$. Various sequential procedures can be applied to conduct the testing of H_0^* against H_1^* , and in this paper we use the B-value proposed by Lan and Wittes (1988) to project the current trend to the end of a study. For our purposes, we can calculate the B-value after every Monte Carlo sample, and use these B-values to efficiently conduct a sequential testing.

Consider the statistic Z_n to test whether $p = \alpha$ or not with n observations,

$$Z_n = \frac{S_n - n\alpha}{\sqrt{n\alpha(1-\alpha)}},$$

where $S_n = \sum_{i=1}^n X_i$. For the pre-determined maximum number of resamples, m , the B-value at $t = \sqrt{n/m}$ is defined as

$$B\left(\sqrt{\frac{n}{m}}\right) = \sqrt{\frac{n}{m}} Z_n = \frac{S_n - n\alpha}{\sqrt{m\alpha(1-\alpha)}}.$$

Consider a sequential test where we stop resampling at the n -th sample and reject $H_0^*: p \leq \alpha$ if $B\left(\sqrt{n/m}\right) \geq c_1$ for some $n \leq m$ (and thus do not reject the original null hypothesis H_0), or stop and fail to reject H_0^* if $B\left(\sqrt{n/m}\right) \leq c_2$ for some $n \leq m$ (and thus reject the original null hypothesis H_0) where c_1 and c_2 are appropriately chosen constants. This sequential test produces parallel boundaries on plots of (n, S_n) . We modify it by incorporating the idea of a simple curtailed test, which stops at the n -th sample and rejects H_0^* if $S_n \geq \lfloor \alpha(m+1) \rfloor = r_1$, and stops and accepts H_0^* if $n - S_n \geq \lceil (1-\alpha)(m+1) \rceil = r_0$. Let

$$\begin{aligned} U(n) &= \min \left\{ s \geq 0 : s - n\alpha \geq c_1 \sqrt{m\alpha(1-\alpha)} \right\}, \\ L(n) &= \max \left\{ s \geq 0 : s - n\alpha \leq c_2 \sqrt{m\alpha(1-\alpha)} \right\}, \end{aligned}$$

and define

$$\begin{aligned} B_{\text{Upper}} &= \left\{ (n, s) = (n, \min[U(n), r_1]) : n = n_0^+, n_0^+ + 1, \dots, m \right\} \\ B_{\text{Lower}} &= \left\{ (n, s) = (n, \max[L(n), n - r_0]) : n = n_0^-, n_0^- + 1, \dots, m \right\}, \end{aligned}$$

where n_0^+ is the smallest value of n such that $U(n) \leq n$ and n_0^- is the smallest n for which $L(n) \geq 0$. Also, denote the smallest value of n such that $L(n) \leq n - r_0$ as n_1^- and the smallest value of n such that $U(n) \geq r_1$ as n_1^+ . Then, the stopping boundary is formed by $\mathbf{B} = \mathbf{B}_{\text{Upper}} \cup \mathbf{B}_{\text{Lower}}$.

See Figure 1 for the boundary with $c_1 = -c_2 = 2.58281$, $m = 3,620$, and $\alpha = 0.05$, for which $n_0^+ = 36$, $n_0^- = 678$, $n_1^+ = 2,923$, and $n_1^- = 3,585$.

2.2 The B-value Design Bounding the Resampling Risk

Let the stopping rule for the design be denoted by a $b \times 2$ matrix, $\mathbf{B}$, such that the j -th row denotes values of (n, s) where sampling stops, and consider a sequential MC test φ defined on the B-value boundary as follows: $\varphi(n, s) = 0$ would lead us not to reject the original null hypothesis (i.e. reject $H_0: p \leq \alpha$) and $\varphi(n, s) = 1$ rejects the original null hypothesis (i.e. do not reject $H_0: p \leq \alpha$). Then, the size of the sequential MC test, φ , is estimated as $\alpha^*(\varphi, \mathbf{B}) = E_U E_{\mathbf{B}}(\varphi|P)$, where $E_{\mathbf{B}}$ is the expectation over the Monte Carlo sampling design after replacing p with the random variable P , and E_U represents expectation with respect to P under the continuous uniform distribution. For the B-value design with φ defined this way, we require the design to bound the resampling risk under the distribution for P , both under the null and a class of alternative distributions, where the resampling risk,

$$\gamma(\varphi, F, \mathbf{B}) = E_F E_{\mathbf{B}} \{I(P \leq \alpha)(1 - \varphi)\} + E_F E_{\mathbf{B}} \{I(P > \alpha)\varphi\},$$

is the expected probability that the accept/reject decision is different from complete enumeration for some distribution of $P \sim F$. In other words, we require $\gamma(\varphi, F, \mathbf{B}) \leq \gamma^*$ for all $F \in \mathcal{F}$, where $\mathcal{F}$ is the class of distributions which includes the null (i.e., the uniform) and many alternative distributions, and γ^* is some pre-specified bound on the resampling risk.

Fay and Follmann (2002) proposed the design to start with a simple design and to iteratively “push out” the boundary at the point where not stopping will decrease the resampling risk the greatest per added expected sample size. The pushing out algorithm will not be reviewed here, but we note that it is computationally expensive to carry out, so that IPO designs that bound the resampling risk less than or equal to $\gamma^* = 0.01$ are intractable. Fay and Follmann (2002) found distributions that met the resampling criterion by first defining a class of designs which may be expanded in some systematic way such that γ gets smaller as the designs get larger, then finding the distribution $\hat{F}_P \in \mathcal{F}$ which appears to give the largest values of $\gamma(\varphi, \hat{F}_P, \mathbf{B})$ for each member of the class of designs, and finally pick the smallest design in the class such that $\gamma(\varphi, \hat{F}_P, \mathbf{B}) \leq \gamma^*$. They estimated F_P with beta distributions and then searched for the $\hat{F}_P$ which gave the largest resampling risk for fixed boundaries of different sizes over all possible values of β , the probability of Type II error, and the numerical search showed that $1 - \beta \approx .47$ gave the largest resampling risk for fixed boundaries with $\alpha = 0.05$. As discussed in Fay et al. (2007), the tSPRT boundary or equivalently the B-value boundary is then can be determined by searching the values of c_1 and c_2 for fixed m such that $\gamma(\varphi, \hat{F}^*, \mathbf{B}) < \gamma^*$, where $\hat{F}^*$ is the “worst case” distribution chosen as in Fay and Follmann (2002).

In the rest of this section, we consider the B-value boundary of Section 2.1 and the test based on the valid p-value proposed in Fay et al. (2007), φ_{FKH} . Their valid p-value when (n, S_n) is a boundary point is defined as $\hat{p}_v(S_n, n) = F_{\hat{p}_{\text{MLE}}}(S_n/n)$, where $\hat{p}_{\text{MLE}}$ is the maximum likelihood estimator of p , $\hat{p}_{\text{MLE}}(S_n, n) = S_n/n$, for any closed boundary and $F_{\hat{p}}$ is a cumulative distribution function of $\hat{p}$. $F_{\hat{p}}$ is defined in (5.2) of Fay et al. (2007) and can be computed by using the FKH algorithm. For the B-value boundary obtained by the FKH algorithm and the test function defined as

$$\varphi_{\text{FKH}}(n, s) = \begin{cases} 1 & \text{if } \widehat{p}_v(s, n) \leq \alpha; \\ 0 & \text{if } \widehat{p}_v(s, n) > \alpha, \end{cases}$$

the following tables indicate the efficiency of the B-value design. In Table 1 we compare the B-value design with various values of $c = c_1 = -c_2$ to the IPO design of Fay and Follmann (2002) for $\alpha = .05$ and $\gamma^* = .025$. The distributions of the p-value considered in the tables are $\hat{F}^*$ chosen as the “worst case” beta distribution, a uniform distribution, and two point mass distributions. Following the notation of Fay and Follmann (2002), we let $H_{\alpha, 1-\beta}$ be a distribution such that $H_{\alpha, 1-\beta}(p) = 1 - \Phi\{\Phi^{-1}(1-p) - \Phi^{-1}(1-\alpha) + \Phi^{-1}(\beta)\}$ where Φ is a cumulative distribution function of a standard normal distribution, and $\hat{H}_{\alpha, 1-\beta}$ denote the beta distribution whose mean is same as the mean of $H_{\alpha, 1-\beta}$ and that $\hat{H}_{\alpha, 1-\beta}(\alpha) = H_{\alpha, 1-\beta}(\alpha) = 1 - \beta$. In this numerical experiment, we use $\hat{F}^* = \hat{H}_{0.05, 0.47}$ following Fay and Follmann (2002), and consider the case of $c = c_1 = -c_2$ and $\varepsilon = 2(1 - \Phi(c))$. The choices of ε (or c) are rather arbitrary in Tables 1 and 2, while m is chosen to be the smallest sample size to bound the resampling risk for c given. These values of m in Tables 1 and 2 are chosen by using the FKH algorithm, which required to repeatedly construct the B value boundary and to compute the associate risk, and some insights on such choices are provided in Section 3 with the approximate B-value design. We see that the expected resample size, $E(N)$, for the B-value design with $\varepsilon = 0.05$ is close to the $E(N)$ from the IPO design. The $E(N)$ for the B-value designs decrease with increasing ε (or decreasing c), and some B-value designs have lower $E(N)$ than the IPO design for most distributions tried. Note that changing the value of ε (or c) does not change the resampling risk within the range of values tried, which will be discussed further with the approximate B-value design in the next section. In Table 2, we compare the B-value design to the fixed resample size design and the curtailed design at $\gamma^* = 0.01$, a value for which the IPO design is intractable. We see that the B-value designs can give much lower values of $E(N)$ than the other designs, indicating that the B-value design is more efficient than the IPO and simple curtailed designs. In general, we see that for larger values of ε (i.e., smaller values of c), the B-value design shows smaller $E(N)$. Thus, at any m predetermined, it would be desirable to use the design with the smallest value of c , while bounding the resampling risk.

3 Approximate B-value Design

Although much more tractable than the IPO design, the numerical calculation of the B-value design in Section 2.2 is still computationally expensive and requires a specially developed software. We now consider an alternative design which we call the approximate B-value design. This design does not require a special program to construct the boundary, and this approximate B-value design would also provide some analytic insights on the properties of the B-value design illustrated in Section 2.2.

In this section, we use the B-value boundary defined in Section 2.1, but define the test function φ as follows: $\varphi(n, s) = 1$ if $(n, s) \in \mathbf{B}_{\text{Lower}}$ without reaching $\mathbf{B}_{\text{Upper}}$, and $\varphi(n, s) = 0$ if $(n, s) \in \mathbf{B}_{\text{Upper}}$ without reaching $\mathbf{B}_{\text{Lower}}$. Then (n, s) such that $\varphi(n, s) = 0$ would lead us not to reject the original null hypothesis and (n, s) such that $\varphi(n, s) = 1$ is the value to reject the original null hypothesis. We also let N denote the resample size at stopping. We note that φ of this section, the one based only on which boundary is crossed, is not exactly equivalent to φ_{FKH} of Section 2.2, which is based on the valid p-value estimated at each stopping point. Although not reported here, we observed in cases we tried that φ_{FKH} and φ are different only in a very small portion of the boundary, typically in the curtailed part, at which the valid p-value is close to α . In this sense, we call the test proposed in this section an approximate B-value test. Now, aiming to construct an approximate B-value design bounding the resampling risk, we discuss how to determine c_1 , c_2 , and m , and thus $\mathbf{B}$ such that $\gamma(\varphi, F, \mathbf{B}) \leq \gamma^*$ for a given choice of γ^* .

Since the approximate design does not automatically achieve the size of α , we also examine the size of the test following the approximate B-value design. Finally, we review how to estimate the expected resample size for the B-value design.

3.1 Approximate Upperbound of the Resampling Risk

In this section we show how to determine approximate bounds on the resampling risk given a B-value stopping boundary and the test function φ defined above. Suppose that a Monte Carlo sample forms a discrete stochastic process $\{X_1, X_2, \dots\}$, where X_i has a Bernoulli distribution with $E(X_i) = p$, and let $S_n = \sum_{i=1}^n X_i$. For the approximate B-value test associated with the B-value boundary of Section 2.1, we can summarize the decision rule using the following two random variables denoting the minimum values of n at which each boundary is crossed:

$$\begin{aligned} N^- &= \min \{n: S_n \text{ hits the lower boundary, } \mathbf{B}_{\text{Lower}}\}, \\ N^+ &= \min \{n: S_n \text{ hits the upper boundary, } \mathbf{B}_{\text{Upper}}\}. \end{aligned}$$

Recall that the resampling risk is $\gamma(\varphi, F, \mathbf{B}) = E_F E(\varphi I(P > \alpha) | P) + E_F E((1 - \varphi) I(P \leq \alpha) | P)$ where $P \sim F$. Then for $\alpha < p < 1$,

$$\begin{aligned} E(\varphi|p) &= \Pr(N^- \leq m \text{ and } N^- < N^+ | p) \\ &= \Pr(N^- \leq m | p) - \Pr(N^- \leq m \text{ and } N^- \geq N^+ | p) \\ &= h_1(p) - h_2(p) \end{aligned}$$

and for $0 < p \leq \alpha$,

$$\begin{aligned} E(1 - \varphi|p) &= \Pr(N^+ \leq m \text{ and } N^+ < N^- | p) \\ &= \Pr(N^+ \leq m | p) - \Pr(N^+ \leq m \text{ and } N^+ \geq N^- | p) \\ &= h_3(p) - h_4(p). \end{aligned}$$

In order to use asymptotic results developed for the B-values, consider the stopping boundaries formed only by the B-values without considering the truncation conditions: for each p , let

$$\begin{aligned} \tau^+ &= \min\{n \geq n_0^+: B(\sqrt{n/m}) \geq c_1\} \\ &= \min\{n \geq n_0^+: S_n - n\alpha \geq c_1 \sqrt{m\alpha(1-\alpha)}\} \\ &= \min\{n \geq n_0^+: W_n \geq b_1\} \end{aligned}$$

and

$$\begin{aligned} \tau^- &= \min\{n \geq n_0^-: B(\sqrt{n/m}) \leq c_2\} \\ &= \min\{n \geq n_0^-: S_n - n\alpha \leq c_2 \sqrt{m\alpha(1-\alpha)}\} \\ &= \min\{n \geq n_0^-: W_n \leq b_2\}, \end{aligned}$$

where $b_1 = c_1 \sqrt{m\alpha(1-\alpha)/(p(1-p))}$, $b_2 = c_2 \sqrt{m\alpha(1-\alpha)/(p(1-p))}$, and $W_n = \sum_{i=1}^n (X_i - \alpha) / \sqrt{p(1-p)}$ asymptotically behaves like a Brownian motion with drift $\mu = (p - \alpha) / \sqrt{p(1-p)}$. Then,

$$\begin{aligned} h_1(p) &= \Pr(N^- \leq m|p) \\ &= \Pr(\tau^- < n_1^- | p) + \Pr(\tau^- \geq n_1^- \text{ and } n - S_n \geq r_0 \text{ for some } n = n_1^-, \dots, m|p) \\ &\leq \Pr(W_m \leq 0|p) + \Pr(\tau^- < m \text{ and } W_m > 0|p) \end{aligned}$$

and

$$\begin{aligned} h_3(p) &= \Pr(N^+ \leq m|p) \\ &= \Pr(\tau^+ < n_1^+ | p) + \Pr(\tau^+ \geq n_1^+ \text{ and } S_n \geq r_1 \text{ for some } n = n_1^+, \dots, m|p) \\ &\leq \Pr(W_m \geq 0|p) + \Pr(\tau^+ < m \text{ and } W_m < 0|p). \end{aligned}$$

By using some asymptotic results developed in the context of sequential analysis, we obtain

$$\begin{aligned} \gamma(\varphi, F, \mathbf{B}) &\leq \int_{\alpha}^1 h_1(p) f_{p(X)}(p) dp + \int_0^{\alpha} h_3(p) f_{p(X)}(p) dp \\ &\leq \int_{\alpha}^1 \Phi(-\mu \sqrt{m}) f_{p(X)}(p) dp + R_1 + \int_0^{\alpha} 1 - \Phi(-\mu \sqrt{m}) f_{p(X)}(p) dp + R_2 \\ &= I_1 + R_1 + I_2 + R_2 \equiv \tilde{\gamma}, \end{aligned} \tag{1}$$

where f is a density function of P , and

$$\begin{aligned} I_1 &= \int_{\alpha}^1 \Phi(-\mu \sqrt{m}) f_{p(X)}(p) dp \cong \int_{\alpha}^1 \Pr(W_m \leq 0) f_{p(X)}(p) dp, \\ I_2 &= \int_0^{\alpha} (1 - \Phi(-\mu \sqrt{m})) f_{p(X)}(p) dp \cong \int_0^{\alpha} \Pr(W_m \geq 0) f_{p(X)}(p) dp, \\ R_1 &= \int_{\alpha}^1 e^{-2\mu(b_2+\rho)} \Phi\left(-\frac{2(b_2+\rho)}{\sqrt{m}} + \mu \sqrt{m}\right) f_{p(X)}(p) dp \cong \int_{\alpha}^1 \Pr(\tau^- < m \text{ and } W_m > 0|p) f_{p(X)}(p) dp, \\ R_2 &= \int_0^{\alpha} e^{2\mu(b_1+\rho)} \Phi\left(-\frac{2(b_1+\rho)}{\sqrt{m}} - \mu \sqrt{m}\right) f_{p(X)}(p) dp \cong \int_0^{\alpha} \Pr(\tau^+ < m \text{ and } W_m < 0|p) f_{p(X)}(p) dp, \end{aligned}$$

with $\rho \cong 0.583$, a correction factor for a discrete process.

See Appendix for the derivation of the upperbound in (1). Note that I_1 and I_2 depend only on α and m , not on c_1 and c_2 , and the major contribution to $I_1 + I_2$ comes from those p close to α . Also $R_1 + R_2$ is negligible compared to $I_1 + I_2$. These are expected from the formulations of I_1 , I_2 , R_1 and R_2 as shown on the right sides of the expressions above. Since I_1 works as an approximate bound of the resampling risk associated with stopping at the lower boundary between n_1^- and m , and I_2 serves as an approximate bound of the resampling risk associated with stopping at the upper boundary between n_1^+ and m , we note that the major contribution of the resampling risk comes from decisions made on these parts of the boundary, $\mathbf{B}_{\text{Upper}}$ for $n = n_1^+, \dots, m$ and $\mathbf{B}_{\text{Lower}}$ for $n = n_1^-, \dots, m$, rather than those on the parallel boundary points, $\mathbf{B}_{\text{Upper}}$ for $n = n_0^+, \dots, n_1^+$ and $\mathbf{B}_{\text{Lower}}$ for $n = n_0^-, \dots, n_1^-$.

Since I_1 and I_2 decrease as m increases at a given α , we can choose m to make $I_1 + I_2$ slightly less than γ^* . Then any choice of (c_1, c_2) to make $R_1 + R_2 \leq \gamma^* - I_1 - I_2$ would asymptotically

guarantee the design to bound the resampling risk. However, the larger the values of c_1 and $|c_2|$ are chosen, the larger resample size is expected, and thus, at m given or predetermined, it would be desirable to choose c_1 and $|c_2|$ as small as possible while bounding the resampling risk.

The approximate B-value design can be implemented as follows. For a given choice of α and γ^* such that $\gamma(\varphi, \hat{H}_{\alpha,47}, \mathbf{B}) \leq \gamma^*$, we first choose m such that $I_1 + I_2$, for $f \sim \hat{H}_{\alpha,47}$, is slightly smaller than γ^* . Then for such m , we find $c = c_1 = -c_2$ such that $g(c) = R_1 + R_2 \leq \gamma^* - I_1 - I_2$. For example, if we want to achieve $\alpha = 0.05$ and $\gamma(\varphi, \hat{H}_{0.05,0.47}, \mathbf{B}) \leq 0.01$, we can use a simple integration to find $I_1 + I_2 = 0.005046 + 0.004930 = 0.009976$ for the choice of $m = 3,620$ and $f \sim \text{Beta}(a = 0.3889, b = 2.5234) \approx \hat{H}_{0.05,0.47}$ of Table 2. Finding c such that $g(c) = R_1 + R_2 \leq \gamma^* - I_1 - I_2 = 0.000024$ would then bound the resampling risk to be under 0.01. We now note that for the values of c in Table 2, 2.241, 1.960, and 1.645 corresponding to $\varepsilon = 2(1 - \Phi(c)) = 0.025, 0.05$, and 0.1 , respectively, $g(c) = R_1 + R_2$ increases as c decreases and $R_1 + R_2 < 1.9 \times 10^{-7}, < 2.1 \times 10^{-6}$, and $< 2.3 \times 10^{-5}$, respectively. Thus the smallest value of c among these three, asymptotically bounding the resampling risk under $\gamma^* = 0.01$, is $c = 1.645$, which matches with Table 2, and the numerical search over all c less than or equal to 2.241 resulted in $c = 1.635$ for $\alpha = 0.05, \gamma^* = 0.01$ and $m = 3,620$. This indicates that for any m given or predetermined, we can first compute $I_1 + I_2$ to get an idea on γ^* and then can easily find the smallest design (i.e. c) satisfying the constraint on the resampling risk, while the construction of the exact boundary and the choice of m and c associated with φ_{FKH} would require rather time consuming searches.

Remark 1. The first upperbound of the above expression can be improved by incorporating $\int_{\alpha}^1 h_2(p) f_{p(X)}(p) dp$ and $\int_0^{\alpha} h_4(p) f_{p(X)}(p) dp$, but these probabilities are typically negligible compared to $\int_{\alpha}^1 h_1(p) f_{p(X)}(p) dp$ and $\int_0^{\alpha} h_3(p) f_{p(X)}(p) dp$, respectively.

Remark 2. When the distribution of P is a point mass at α , the resampling risk is greater than 0.5 since

$$\begin{aligned} h_1(\alpha) &\geq \Pr(m - S_m \geq r_0 | \alpha) \cong P_{\mu=0}(W_m \leq 0) = 0.5, \\ h_3(\alpha) &\geq \Pr(S_m \geq r_1 | \alpha) \cong P_{\mu=0}(W_m \geq 0) = 0.5, \end{aligned}$$

and $\int_{\alpha}^1 h_2(p) f_{p(X)}(p) dp$ and $\int_0^{\alpha} h_4(p) f_{p(X)}(p) dp$ are negligible. This explains the values of the resampling risk in Tables 1 and 2 obtained for the point mass at α .

3.2 Size of the Approximate B-value Design

Based on the similar arguments used in Section 3.1, for α given and large m ,

$$\begin{aligned} E(\varphi|p) &= h_1(p) - h_2(p) \\ &\leq h_1(p) \\ &\leq \Pr(W_m \leq 0|p) + \Pr(\tau^- < m \text{ and } W_m > 0|p) \\ &\cong \Phi(-\mu\sqrt{m}) + e^{-2\mu(b+\rho)} \Phi\left(-\frac{2(b+\rho)}{\sqrt{m}} + \mu\sqrt{m}\right) \\ &= g_{11}(p) + g_{12}(p, c), \end{aligned}$$

where $W_n = \sum_{i=1}^n (X_i - \alpha) / \sqrt{p(1-p)}$, $\rho \cong 0.583$, and $g_{12}(p, c)$ is obtained with $\eta = 0$, $\mu = \frac{p - \alpha}{\sqrt{p(1-p)}}$, and $b = -c_2 \sqrt{m_f \alpha(1-\alpha)/(p(1-p))}$ in Equation 3.28 of Siegmund (1985).

Then, for large m , an upperbound for the size of the test, $\alpha^*(\varphi, \mathbf{B})$, is obtained as

$$\begin{aligned} \alpha^*(\varphi, \mathbf{B}) &= E_v E(\varphi|P) = \int_0^1 E(\varphi|p) dp \\ &\leq \int_0^\alpha (g_{11}(p) + g_{12}(p, c)) dp + \int_\alpha^1 (g_{11}(p) + g_{12}(p, c)) dp \\ &= \alpha_1 + \alpha_2(c) + \alpha_3 + \alpha_4(c) \\ &= \tilde{\alpha}, \end{aligned}$$

where $\alpha_1 = \int_0^\alpha g_{11}(p) dp$, $\alpha_2(c) = \int_0^\alpha g_{12}(p, c) dp$, $\alpha_3 = \int_\alpha^1 g_{11}(p) dp$, and $\alpha_4(c) = \int_\alpha^1 g_{12}(p, c) dp$.

We note that α_2 and α_4 are negligible compared to α_1 and α_3 , and also that α_3 is relatively smaller than α_1 . The former is because $\int_0^1 \Pr(\tau^- < m \text{ and } W_m > 0 | p) dp$ is negligible with $\Pr(\tau^- < m)$ being small for $p > \alpha$ and $\Pr(W_m > 0)$ being small for $p \leq \alpha$, and the latter is expected from a positive drift of W_m when $p > \alpha$. This is illustrated by numerical integration results summarized in Table 3. In Table 3, we observe that the major contribution to $\tilde{\alpha}$ comes from $\alpha_1 + \alpha_3$, which only depends on m , not on c , and $\tilde{\alpha}$ is usually slightly over α . Considering that $\alpha^*(\varphi, \mathbf{B}) \cong \tilde{\alpha} - \int_0^1 h_2(p) dp = \tilde{\alpha} - \int_0^1 \Pr(N^- \leq m \text{ and } N^- \geq N^+ | p) dp$ and that $\int_0^1 h_2(p) dp$ is negligible, the approximate B-value design is expected to retain the size, at least approximately. Note that $c = 0.908$ in the last case was obtained as a smallest value of c bounding the approximate upperbound of the resampling risk, γ , under 0.01 at $m = 4,999$.

3.3 Expected Resample Size

As discussed in Section 3.1, the approximate B-value design can be determined by choosing m to make $I_1 + I_2$ smaller than the desirable resampling risk bound of γ^* and then by choosing the smallest c such that $R_1 + R_2 \leq \gamma^* - I_1 - I_2$. We now introduce approximations for $E(N)$, which would help us to study asymptotic efficiencies of B-value designs chosen to bound the resampling risk. As discussed in Sections 3.1 and 3.2, the B-value stopping time can be approximated as a stopping time of a Brownian motion process observed in discrete time, W_n , associated with a truncated parallel boundary, and thus we can use (1.15) of Samuel-Cahn (1974) or (3.17) of Siegmund (1985) with the correction suggested for the discrete time scale. Using these arguments, for $T_b = \min \{n : W_n \geq b\}$, we obtain

$$\begin{aligned} &E_p[\min(T_b, m)] \\ &\sim \frac{b+\rho}{\mu} + \left(m - \frac{b+\rho}{\mu}\right) \Phi(-\mu \sqrt{m} + b/\sqrt{m}) - \left(m + \frac{b+\rho}{\mu}\right) e^{2\mu(b+\rho)} \Phi(-\mu \sqrt{m} - b/\sqrt{m}) \\ &= \frac{b+\rho}{\mu} \left\{1 - \Phi(-\mu \sqrt{m} + b/\sqrt{m}) - e^{2\mu(b+\rho)} \Phi(-\mu \sqrt{m} - b/\sqrt{m})\right\} \\ &\quad + m \left\{\Phi(-\mu \sqrt{m} + b/\sqrt{m}) - e^{2\mu(b+\rho)} \Phi(-\mu \sqrt{m} - b/\sqrt{m})\right\}, \end{aligned}$$

where $\mu = (p - \alpha) / \sqrt{p(1-p)}$, $b = c \sqrt{m \alpha(1-\alpha)/p(1-p)}$, and $\rho \cong 0.583$. Using similar arguments in (3.37) of Siegmund (1985), we then may use

$$l(p) = E_p[\min(T_b, m)] + E_p[\min(T_{-b}, m)] - m \quad (2)$$

as an approximate lower bound of $E_p[\min(T, m)]$ for $T = \min(T_b, T_{-b})$. For the settings in Tables 1 and 2 with $\hat{H}_{\alpha, .47}$ and $\varepsilon = 0.005, 0.025, 0.05$, and 0.1 (or $c = 2.807, 2.241, 1.960$, and 1.645), we find $\int_0^1 l(p) f_{p(X)}(p) dp$ as 291.836, 249.616, 226.213, and 197.680 for $m = 580$, and 965.547, 810.270, 727.441, and 629.124 for $m = 3,620$, respectively. These bounds based on (2) are only approximate and also $E_p(N) \leq E_p(\min(T, m))$, but as discussed in Section 3.6 of Siegmund (1985), we observe that these approximations are close to $E(N)$ in Tables 1 and 2. This indicates that one can use (2) to get a quick and reasonably accurate estimate of the expected resample size without using the FKH algorithm to numerically construct the boundary and computing $E(N)$ for each choice of m and c , which helps one to efficiently choose m and c comparing approximate values of $E(N)$.

4 Example and Discussion

Now, we illustrate the application of the B-value design to the permutation test of Kim et al. (2000) developed to determine the number of change-points in a segmented line regression model. They noted that the test statistic to test the null hypothesis that there are k_0 change-points against the alternative that there are k_1 change-points for $0 \leq k_0 < k_1$, does not follow the classical asymptotic theory, so suggested to use a permutation distribution of the test statistic to estimate the p-value. Joinpoint, software developed by Kim et al. (2000) and available at <http://srab.cancer.gov/joinpoint>, sequentially conducts the permutation tests to select the number of change-points, and the sequential Monte Carlo tests proposed in Fay et al. (2007) and also in this paper are expected to improve the computational efficiency of Joinpoint. Figure 2 includes cancer incidence rates for (a) Hodgkin lymphoma and (b) Liver and bile duct cancer. These incidence rates are the age adjusted rates for all races and both sexes combined for the SEER-9 population, which includes about 9.5 % of the U.S. population (Surveillance, Epidemiology and End Results Program at National Cancer Institute). The permutation tests of no change-point versus one change-point implemented in R with 3,620 permuted data sets produced the p-values of 0.3944 and 0.0011 for Hodgkin Lymphoma and Liver and bile duct cancer, respectively, which led us to choose the straight line regression model for Hodgkin lymphoma and the one change-point model for Liver and bile duct cancer. Two plots in the lower part of Figure 2 show the trajectories of the B-value test with $m = 3,620$, $\alpha = 0.05$ and $c = 2.58281$ for these two cancer sites, where we plotted

$S_n = \sum_{i=1}^n I(T_i \geq T_0)$, marked as a triangle at each n , and a portion of the parallel boundary in Figure 1 up to $n = 1000$. S_n crossed the upper boundary when $n = 93$ for Hodgkin lymphoma, implying to reject $H_0^*: p \leq 0.05$, while S_n crossed the lower boundary at $n = 698$ for Liver and bile duct cancer, leading us not to reject $H_0^*: p \leq 0.05$. From the discussion in Sections 2 and 3, we know that the resampling risk of this B-value test is less than or equal to 0.01 and it is much more efficient than the simple curtailed test with $r_0 = 3,440$ and $r_1 = 181$.

In this paper, we considered the B-value design to conduct MC tests with a less number of replications while we control the resampling risk. To control the resampling risks, we considered the class of p-value distributions as in Fay and Follmann (2002), and compared the performance of the B-value design with those of a simple curtailed design and the IPO design proposed in Fay and Follmann (2002). The exact B-value design which can be obtained by using the FKH algorithm is shown to be more efficient than the curtailed design in terms of the number of resamples, and than the IPO design in terms of the computational complexity. We then proposed the approximate B-value design which provides some analytic insights on

the choice of parameters in numerically constructing the B-value design and which also serves as an approximate alternative that does not require a special algorithm.

There are some open problems that we leave for future research. Although the method of this paper works for a general choice of c_1 and c_2 , we considered only the symmetric choice of c_1 and c_2 in numerical examples. Such symmetric choice of c_1 and c_2 is equivalent to the truncated SPRT boundary of Fay et al. (2007) with the Type I and II errors set equally, and it might be of interest to consider asymmetric choices of c_1 and c_2 . Exploring how to construct more general designs aiming to minimize the expected resample size involves more details and requires further research. Also, the resampling risk defined in this paper gives the equal weight on Type I and II errors, and one may want to consider the resampling risk differently weighing the two types of errors. For a specific choice of weight, the method of Section 3 can be easily applied to approximate the resampling risk, size, and the expected resample size, but how to choose weights needs to be studied.

In this paper, we discussed how to construct the best design at m given or predetermined to have $I_1 + I_2 < \gamma^*$, empirically showing that $E(N)$ decreases as $c = c(m)$ decreases for m given. As indicated in numerical examples, however, there are many designs, that is, many choices of (m, c) , that satisfy the same constraints on α and γ^* , so the problem of determining an optimal design over all (m, c) is an interesting problem to pursue.

In order to conduct a sequential test for $H_0^*: p \leq \alpha$ versus $H_1^*: p > \alpha$, there are other sequential procedures proposed in the literature. For example, one may consider the likelihood ratio

boundary as follows. For $S_k = \sum_{i=1}^k I(T_i \geq T_0)$, the sequential likelihood ratio test rejects $H_0^*: p \leq \alpha$ if

$$\log[L(S_k; \hat{p}_{MLE})/L(S_k; \alpha)] \geq a$$

for some $m_0 \leq k \leq m_f$ and $a > 0$, where L denotes the likelihood function. This condition is equivalent to

$$(k+1) \left\{ \frac{S_k+1}{k+1} \left[\log \left(\frac{S_k+1}{k+1} \right) - \log \alpha \right] + \frac{k-S_k}{k+1} \left[\log \left(\frac{k-S_k}{k+1} \right) - \log (1-\alpha) \right] \right\} \geq a,$$

which can be written in terms of the following stopping time:

$$T = \min \{k: kH(S_k/k) \geq a\},$$

where $H(x) = x \ln x + (1-x) \ln (1-x) - x \ln(\alpha) - (1-x) \ln(1-\alpha)$. Then, some asymptotic results, as in Siegmund (1985, Chapter 5), can be used to construct an asymptotic sequential design, but it requires details more complicated than those in Section 3. We plan to pursue it in our future research to compare the performance of the B-value design with other sequential boundaries such as the likelihood ratio boundary.

Acknowledgments

Kim's research was partially supported by NIH Contract HHSN 261200700273P. The author thanks Dr. Michael P. Fay for very helpful discussions and extensive assistance to improve the presentation of the paper and Mark Hachey for the initial construction of Tables 1 and 2.

Appendix: Approximate Upperbound for $\gamma(\varphi, F, B)$

Note that

$$\begin{aligned}
 h_1(p) &= \Pr(N^- \leq m|p) \\
 &= \Pr(\tau^- < n_1^-|p) + \Pr(\tau^- \geq n_1^- \text{ and } n - S_n \geq r_0 \text{ for some } n=n_1^-, \dots, m|p) \\
 &\leq \Pr(\tau^- < n_1^-|p) + \Pr(\tau^- \geq n_1^- \text{ and } S_m - m\alpha \leq 0|p) \\
 &\leq \Pr(\tau^- < n_1^-|p) + \Pr(\tau^- \geq n_1^- \text{ and } W_m \leq 0|p) \\
 &= \Pr(\tau^- < n_1^-|p) + \Pr(W_m \leq 0|p) - \Pr(\tau^- < n_1^- \text{ and } W_m \leq 0|p) \\
 &= \Pr(W_m \leq 0|p) + \Pr(\tau^- < n_1^- \text{ and } W_m > 0|p) \\
 &\leq \Pr(W_m \leq 0|p) + \Pr(\tau^- < m \text{ and } W_m > 0|p) \\
 &\cong \Phi(-\mu \sqrt{m}) + e^{-2\mu(b+\rho)} \Phi\left(-\frac{2(b+\rho)}{\sqrt{m}} + \mu \sqrt{m}\right) \\
 &= g_{11}(p) + g_{12}(p, c)
 \end{aligned}$$

and

$$\begin{aligned}
 h_3(p) &= \Pr(N^+ \leq m|p) \\
 &= \Pr(\tau^+ < n_1^+|p) + \Pr(\tau^+ \geq n_1^+ \text{ and } S_n \geq r_1 \text{ for some } n=n_1^+, \dots, m|p) \\
 &= \Pr(\tau^+ < n_1^+|p) + \Pr(\tau^+ \geq n_1^+ \text{ and } S_m \geq r_1|p) \\
 &\leq \Pr(\tau^+ < n_1^+|p) + \Pr(\tau^+ \geq n_1^+ \text{ and } W_m \geq 0|p) \\
 &= \Pr(\tau^+ < n_1^+|p) + \Pr(W_m \geq 0|p) - \Pr(W_m \geq 0 \text{ and } \tau^+ < n_1^+) \\
 &= \Pr(W_m \geq 0|p) + \Pr(\tau^+ < n_1^+ \text{ and } W_m < 0|p) \\
 &\leq \Pr(W_m \geq 0|p) + \Pr(\tau^+ < m+ \text{ and } W_m < 0|p) \\
 &\cong 1 - \Phi(-\mu \sqrt{m}) + e^{2(b+\rho)\mu} \cdot \Phi\left(-\frac{2(b+\rho)}{\sqrt{m}} - \mu \sqrt{m}\right) \\
 &= g_{31}(p) + g_{32}(p, c)
 \end{aligned}$$

where $W_m = \sum_{i=1}^m (Y_i - \alpha) / \sqrt{p(1-p)}$, $b=c \sqrt{m\alpha(1-\alpha)/p(1-p)}$ and $\rho \cong 0.583$. Thus,

$$\int_{\alpha}^1 h_1(p) f_{p(X)}(p) dp \leq \int_{\alpha}^1 g_{11}(p) f_{p(X)}(p) dp + \int_{\alpha}^1 g_{12}(p) f_{p(X)}(p) dp = I_1 + R_1 = \gamma_1$$

and

$$\int_0^{\alpha} h_3(p) f_{p(X)}(p) dp \leq \int_0^{\alpha} g_{31}(p) f_{p(X)}(p) dp + \int_0^{\alpha} g_{32}(p) f_{p(X)}(p) dp = I_2 + R_2 = \gamma_2.$$

Therefore, when $c_1 = -c_2 = c$, for large m , we have

$$\gamma(\varphi, F, B) \leq I_1 + I_2 + R_1 + R_2.$$

Note that $I_1 = \int_{\alpha}^1 \Phi(-\mu \sqrt{m}) f_{p(X)}(p) dp$ works as an upperbound of $\int_{\alpha}^1 \Pr(\tau^- \geq n_1^- \text{ and } W_m \leq 0 | p) f_{p(X)}(p) dp$, which makes a most contribution to γ_1 with $R_1 = \int_{\alpha}^1 g_{12}(p, c) f_{p(X)}(p) dp = \int_{\alpha}^1 \Pr(\tau^- < m \text{ and } W_m > 0 | p) f_{p(X)}(p) dp$ negligible. Similarly, the major contribution to γ_2 comes from $I_2 = \int_0^{\alpha} [1 - \Phi(-\mu \sqrt{m})] f_{p(X)}(p) dp$ and $R_2 = \int_0^{\alpha} g_{32}(p) f_{p(X)}(p) dp$ is negligible.

References

- Fay MP, Follmann DA. Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests. *American Statistician* 2002;56:63–70.
- Fay MP, Kim HJ, Hachey M. On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests. *Journal of Computational and Graphical Statistics* 2007;16:946–967. [PubMed: 18633453]
- Kim HJ, Fay M, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine* 2000;19:335–351. [PubMed: 10649300]
- Lan KKG, Wittes J. The B-Value: A tool for monitoring data. *Biometrics* 1988;44:579–585. [PubMed: 3390511]
- Samuel-Cahn E. Repeated significance test II, for hypotheses about the normal distribution. *Communications in Statistics* 1974;3(8):711–733.
- Siegmund, D. *Sequential Analysis*. New York: Springer-Verlag; 1985.

Figure 1.
A B-value Boundary with $m = 3, 620$, $\alpha = 0.05$ and $c_1 = -c_2 = 2.58281$

Figure 2.
Hodgkin lymphoma and Liver and bile duct cancer

Table 1

$E(N|B)$ with $\alpha = 0.05$, $P \sim B$, and $\gamma(\varphi, \hat{H}_{\alpha, 47}, B) = 0.025$

	$\hat{H}_{\alpha, 47}$			Uniform		Point mass at α		Point mass at 0.001	
	E(N)	Risk		E(N)	Risk	E(N)	Risk	E(N)	Risk
B-value Design									
$(\epsilon, m) = (0.005, 580)$	279.637	0.02499	87.812	0.00719	536.429	0.52660	301.020	0	
$(\epsilon, m) = (0.025, 580)$	242.704	0.02499	76.056	0.00719	534.795	0.52660	240.816	0	
$(\epsilon, m) = (0.05, 580)$	221.092	0.02499	69.231	0.00719	530.807	0.52657	210.204	0	
$(\epsilon, m) = (0.1, 580)$	194.643	0.02500	61.072	0.00720	518.236	0.52627	176.531	0	
$(\epsilon, m) = (0.2, 600)$	163.118	0.02475	51.169	0.00713	494.378	0.52310	139.796	0	
IPO Design									
$m=576$	213.508	0.0250	62.85	0.0072	459.6047	0.522	251.7032	0	

Table 2

$E(N|B)$ with $\alpha = 0.05$, $P \sim B$, and $\gamma(\varphi, \hat{H}_{\alpha,47}, B) = 0.01$

	$\hat{H}_{\alpha,47}$		Uniform		Point mass at α		Point mass at 0.06	
	E(N)	Risk	E(N)	Risk	E(N)	Risk	E(N)	Risk
B-value Design								
$(\epsilon, m) = (0.005, 3620)$	954.475	0.00999	292.708	0.00289	3507.646	0.51065	2861.852	0.00426
$(\epsilon, m) = (0.025, 3620)$	803.470	0.00999	245.940	0.00289	3485.032	0.51065	2575.559	0.00426
$(\epsilon, m) = (0.05, 3620)$	723.584	0.00999	221.320	0.00289	3444.662	0.51063	2360.361	0.00426
$(\epsilon, m) = (0.1, 3620)$	626.848	0.00999	191.519	0.00289	3334.562	0.51049	2060.366	0.00430
$(\epsilon, m) = (0.2, 3681)$	510.256	0.01000	155.840	0.00289	3076.974	0.51058	1671.372	0.00452
Curtailed Design								
$m=3620$	2371.276	0.00999	722.753	0.00289	3515.422	0.50913	3016.334	0.00431
Fixed Design								
$m=3620$	3620	0.00999	3620	0.00289	3620	0.51065	3620	0.00426

Table 3

Approximate Size

c	$\varepsilon = 2(1 - \Phi(c))$	$\tilde{\gamma} = I_1 + I_2 + R_1 + R_2$	α	m	a_1	a_3	$\tilde{\alpha}$
2.241	0.025	0.024842	0.05	580	0.046756	0.004016	0.050772
2.241	0.025	0.009976	0.05	3620	0.048616	0.001509	0.050124
2.241	0.025	0.012958	0.01	3620	0.009404	0.000732	0.010135
0.908	0.364	0.009994	0.05	4999	0.048814	0.001276	0.050451