Implementing Monte Carlo Tests with P-value Buckets

Axel Gandy, Georg Hahn and Dong Ding
Department of Mathematics, Imperial College London

Abstract

Software packages usually report the results of statistical tests using p-values. Users often interpret these by comparing them to standard thresholds, e.g. 0.1%, 1% and 5%, which is sometimes reinforced by a star rating (***, **, *). In this article, we consider an arbitrary statistical test whose p-value p is not available explicitly, but can be approximated by Monte Carlo samples, e.g. by bootstrap or permutation tests. The standard implementation of such tests usually draws a fixed number of samples to approximate p. However, the probability that the exact and the approximated p-value lie on different sides of a threshold (the resampling risk) can be high, particularly for p-values close to a threshold. We present a method to overcome this. We consider a finite set of user-specified intervals which cover $[0,1]$ and which can be overlapping. We call these p-value buckets. We present algorithms that, with arbitrarily high probability, return a p-value bucket containing p. We prove that for both a bounded resampling risk and a finite runtime, overlapping buckets need to be employed, and that our methods both bound the resampling risk and guarantee a finite runtime for such overlapping buckets. To interpret decisions with overlapping buckets, we propose an extension of the star rating system. We demonstrate that our methods are suitable for use in standard software, including for low p-values occurring in multiple testing settings, and that they can be computationally more efficient than standard implementations.

Keywords: Algorithms, Bootstrap/resampling, Hypothesis Testing, Sampling

1 Introduction

Software packages usually report the significance of statistical tests using p-values. Most users will base further steps of their analyses on where those p-values lie with respect to certain thresholds. To facilitate this, many tests in statistical software such as *R* (R Development Core Team, 2008), *SAS* (SAS Institute Inc., 2011) or *SPSS* (IBM Corp., 2013) translate the significance to a star rating system, in which typically $p \in (0.01, 0.05]$ is denoted by *, $p \in (0.001, 0.01]$ is denoted by ** and $p \leq 0.001$ is denoted by ***.

In this article, we consider a statistical test whose p-value p can only be approximated by sequentially drawn Monte Carlo samples. Among others, this scenario arises in bootstrap or permutation tests, see e.g. Lourenco and Pires (2014); Martínez-Camblor (2014); Liu et al. (2013); Wu et al. (2013); Asomaning and Archer (2012); Dazard and Rao (2012).

Standard implementations of Monte Carlo tests in software packages usually take a fixed number of samples and estimate p as the proportion of exceedances over the observed value of the test statistic. Examples of this include the computation of a bootstrap p-value inside the function *chisq.test* in *R* or the function *t-test* in *SPSS*. However, there is no control of the *resampling risk*, the probability that the exact and the approximated p-value lie on two opposite sides of a testing threshold (usually 0.1%, 1% or 5%).

Sequential methods to approximate p-values have been studied in the literature. Early works provided ad hoc attempts to reduce the computational effort without focusing on a specific error criterion (Besag and Clifford, 1991; Silva et al., 2009).

Further developments aimed at a uniform bound on the resampling risk for a single threshold (Davidson and MacKinnon, 2000; Andrews and Buchinsky, 2000, 2001; Gandy, 2009). Gandy (2009) shows that such a uniform bound necessarily results in an infinte running time.

There are also approaches that aim to bound an integrated resampling risk for a single threshold (Fay and Follmann, 2002; Kim, 2010; Silva and Assunção, 2013). Such an error criterion is weaker that a uniform bound on the resampling risk and can be achieved with finite effort.

In this article, we present algorithms that work with multiple thresholds, aim for uniform bounds on the error and, under conditions, have finite running time. We first generalize testing thresholds to a finite set of user-specified intervals (called "p-value buckets") which cover $[0, 1]$ and which can be overlapping. Our algorithms return one of those p-value buckets which is guaranteed to contain the unknown (true) p up to a uniformly bounded error.

We prove that methods achieving both a finite runtime and a bounded resampling risk need to operate on overlapping p-value buckets. In order to report decisions computed with overlapping buckets, we propose to use an extension of the classical star rating system (*, **, ***) used to indicate the significance of a hypothesis.

Our methods rely on the computation of a confidence sequence for p. We present two approaches to compute such a confidence sequence, prove that both approaches indeed bound the resampling risk and achieve a finite runtime for overlapping buckets. We compare both approaches in a simulation section and demonstrate that they achieve a competitive computational effort which is close to a theoretical lower bound on the effort we derive.

The article is structured as follows. Section 2 introduces the mathematical setting of our article (Section 2.1), the rationale behind overlapping p-value buckets (Section 2.2), our proposed extension of the traditional star rating system (Section 2.3) and a general algorithm to compute a decision for p with respect to a set of p-value buckets (Section 2.4). The general algorithm relies on the construction of certain confidence sequences for p for which we present two approaches: one based on likelihood martingales (Robbins, 1970; Lai, 1976) in Section 3.1 and one based on the *Simctest* algorithm (Gandy, 2009) in Section 3.2. In Section 4 we first derive a theoretical lower bound on the expected effort (Section 4.1) and demonstrate that our methods achieve a computational effort which stays within a multiple of the optimal effort (Sections 4.2, 4.3). An application to multiple testing is considered in Section 4.4. The article concludes with a discussion in Section 5. All proofs can be found in Appendix A. The Supplementary Material includes R-code to implement the algorithms as well as to reproduce all figures and tables.

2 General algorithm

2.1 Setting

We consider one hypothesis H_0 which we would like to test with a given statistical test. Let T denote the test statistic and let t be the evaluation of T on some given data. For simplicity, we assume that H_0 should be rejected for large values of t. In this case the p-value is commonly defined as the probability of observing a statistic at least as extreme as t, i.e.

$$p = \mathcal{P}(T \geq t), \tag{1}$$

where $\mathcal{P}$ is a probability measure under the null hypothesis.

We assume that the p-value p is not available analytically but can be approximated using Monte Carlo simulation, by drawing independent data under H_0 and evaluating the

<table>
<tr><td>Bucket
Code</td><td colspan="2">$[0, 0.1\%]$
***</td><td colspan="2">$(0.1\%, 1\%]$
**</td><td colspan="2">$(1\%, 5\%]$
*</td><td colspan="2">$(5\%, 1]$</td></tr>
<tr><td>Bucket
Code</td><td></td><td colspan="2">$(0.05\%, 0.2\%]$
**∼</td><td colspan="2">$(0.8\%, 1.2\%]$
*∼</td><td colspan="2">$(4.5\%, 5.5\%]$
∼</td><td></td></tr>
</table>

Table 1: Extended star rating system for the p-value buckets $\mathcal{J}^*$.

test statistic T on them. Comparing the result to the observed realization of T then allows to approximate the p-value as $\hat{p} = \mathcal{P}_n(T \geq t)$, where $\mathcal{P}_n$ is the estimated null-distribution based on n samples (for instance, using bootstrap tests). The exceedances over the observed realization of T can equivalently be modeled using a stream of independent random variables X_i, $i \in \mathbb{N}$, having a Bernoulli(p) distribution.

Let $\mathcal{J}$ be a set of sub-intervals of $[0, 1]$ of positive length that cover $[0, 1]$, i.e. $\bigcup_{J \in \mathcal{J}} J = [0, 1]$. We call any such $\mathcal{J}$ a set of *p-value buckets*. For example,

$$\mathcal{J} = \mathcal{J}^0 := \{[0, 10^{-3}], (10^{-3}, 0.01], (0.01, 0.05], (0.05, 1]\} \tag{2}$$

is a set of p-value buckets. Finding a bucket $I \in \mathcal{J}^0$ such that $p \in I$ is equivalent to deciding where p lies in relation to the traditional levels 0.001, 0.01 and 0.05.

The goal of our algorithms is to find a bucket $I \in \mathcal{J}$ containing p. A natural error criterion is the risk of a *wrong* decision $\mathcal{P}_p(p \notin I)$, which we call the *resampling risk*. Our methods bound the resampling risk uniformly in p at a given $\epsilon \in (0, 1)$, i.e.

$$\mathcal{P}_p(p \notin I) \leq \epsilon \text{ for all } p \in [0, 1]. \tag{3}$$

We will show that there is no algorithm achieving this for $\mathcal{J}^0$ with finite expected effort – for finite effort we need *overlapping* p-value buckets, which we discuss in the next section.

2.2 Overlapping buckets

We say that the buckets $\mathcal{J}$ are overlapping if any $p \in (0, 1)$ is contained in the interior of a $J \in \mathcal{J}$. For instance, the buckets $\mathcal{J}^0$ in (2) are not overlapping, whereas the buckets

$$\mathcal{J}^* = \mathcal{J}^0 \cup \left\{(5 \cdot 10^{-4}, 2 \cdot 10^{-3}], (0.008, 0.012], (0.045, 0.055]\right\},$$

employed in the remainder of this article, are overlapping.

We let τ be the random effort of an algorithm, defined as the number of exceedance indicators X_i used. The following theorem shows that overlapping buckets are both a necessary and sufficient prerequisite for a finite time algorithm satisfying (3) to exist.

Theorem 1. *The following statements are equivalent:*

1. *There exists an algorithm satisfying* (3) *with* $\mathbb{E}_p[\tau] < \infty$ *for all* $p \in [0, 1]$.
2. *The p-value buckets* $\mathcal{J}$ *are overlapping.*
3. *There exists an algorithm satisfying* (3) *with* $\tau < C$ *for some deterministic* $C > 0$.

2.3 Extended Star Rating System

It is commonplace to report the significance of a hypothesis using a star rating system: strong significance is encoded as *** ($p < 0.1\%$), significance at 1% is encoded as ** and weak significance ($p < 5\%$) as a single star. This classification, recommended in the publication

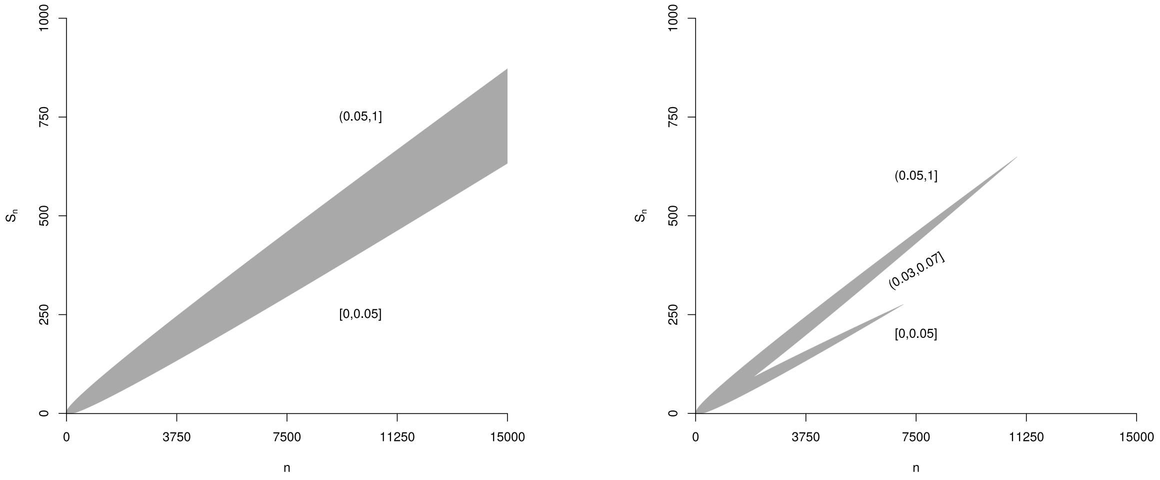

Figure 1: Left: Non-stopping region (gray) to decide p with respect to J^e (corresponding to a 5% threshold). Right: Non-stopping region for the overlapping buckets $J^e \cup \{(0.03, 0.07]\}$.

manual of the American Psychological Association (American Psychological Association, 2010, page 139), is the de facto standard for reporting significance. As we have seen in Theorem 1, it is impossible to produce such a star rating for Monte Carlo tests with finite effort, but it is possible to report results with overlapping buckets with finite effort.

We propose to extend the star rating system to overlapping buckets as described in Table 1, using the p-value buckets $\mathcal{J}^*$ as example. If the algorithm gives a clear decision with respect to the classical thresholds, we report the classical star rating. Otherwise, i.e. if the reported bucket I equals $(0.05\%, 0.2\%]$, $(0.8\%, 1.2\%]$ or $(4.5\%, 5.5\%]$, we propose to report significance with respect to the smallest classical threshold larger than $\max I$ and to indicate the possibility of a higher significance with a tilde symbol.

For instance, suppose an algorithm returns the bucket $I = (0.05\%, 0.2\%]$ for p upon stopping. Since in this case a decision with respect to all classical thresholds larger than $\max I = 0.2\%$ is available, we know that $p \leq 1\%$ and can safely report a ** significance. However, as p could either be smaller or larger than the next smaller classical threshold $0.1\% \in J$, we report **$^\sim$ to indicate the possibility of a higher significance.

2.4 The general construction

We suppose that for each $n \in \mathbb{N}$, we can compute a confidence interval I_n for p based on $X_1, \dots, X_n$ such that the joint coverage probability of the sequence I_n, $n \in \mathbb{N}$, is at least $1 - \epsilon$, where $\epsilon > 0$ is the desired uniform bound on the resampling risk, i.e. we require

$$\mathcal{P}_p(p \in I_n \ \forall n \in \mathbb{N}) \geq 1 - \epsilon \quad \text{for all } p \in [0, 1]. \tag{4}$$

In Sections 3.1 and 3.2 we consider two constructions satisfying (4).

For a given set $\mathcal{J}$ of p-value buckets, we define the stopping time

$$\tau_{\mathcal{J}} = \inf \{n \in \mathbb{N} : \exists I \in \mathcal{J} : I_n \subseteq I\} \tag{5}$$

which denotes the minimal number of samples n needed until a confidence interval I_n is fully contained in a bucket $I \in \mathcal{J}$. The result of our algorithm is this bucket I if $\tau_{\mathcal{J}} < \infty$. If $\tau_{\mathcal{J}} = \infty$ we let I be an arbitrary element of $\mathcal{J}$ such that $\lim_{n\to\infty} \frac{S_n}{n} \in I$, where $S_n = \sum_{i=1}^n X_i$ denotes the cumulative sum of exceedances observed among the first n samples.

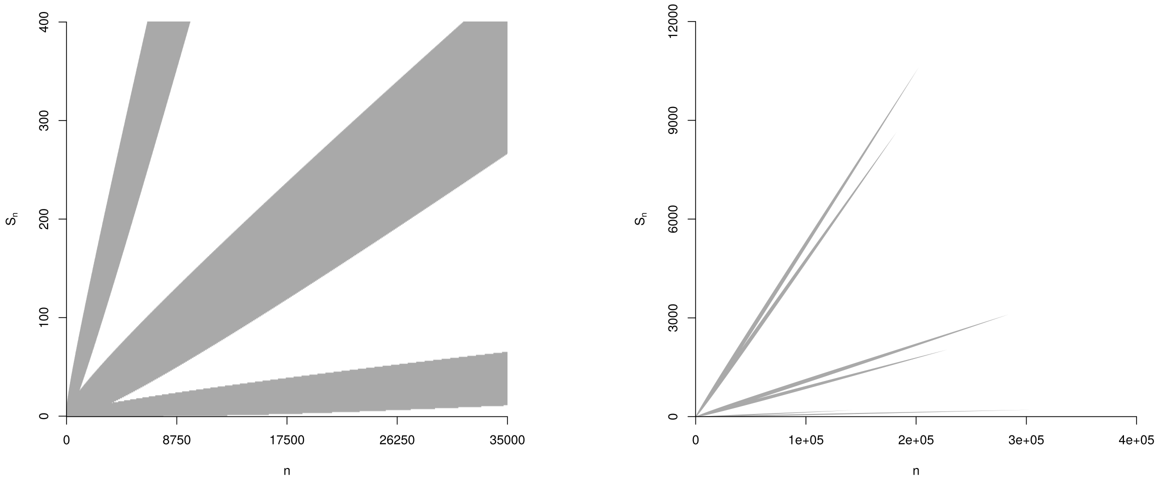

Figure 2: Non-stopping region for $\mathcal{J}^0$ (left) and $\mathcal{J}^*$ (right).

The random interval I constructed in this way satisfies the uniform bound on the resampling risk (3) due to the strong law of large numbers and due to (4).

If τ_J is bounded, meaning if there exists $N \in \mathbb{N}$ such that $\tau_{\mathcal{J}} < N$, we can relax (4) to

$$\mathcal{P}_p(p \in I_n \; \forall n < N) \geq 1 - \epsilon \quad \text{for all } p \in [0,1]. \tag{6}$$

Example 1. *Suppose we are solely interested in the 5% threshold. Testing at 5% corresponds to the two classical buckets $J^e = \{[0, 0.05], (0.05, 1]\}$. Using the approach of Section 3.2 with $\epsilon = 10^{-3}$ to compute a confidence sequence for p, we arrive at the non-stopping region (gray) displayed in Figure 1 (left).*

Sampling progresses until the sampling path (n, S_n) hits either (lower or upper) boundary of the non-stopping region. As displayed in Figure 1 (left), we report the interval $[0, 0.05]$ ($(0.05, 1]$) upon hitting the lower (upper) boundary first.

Adding the bucket $(0.03, 0.07]$ to J^e results in overlapping buckets with a finite non-stopping region displayed in Figure 1 (right). In Figure 1 (right), the sample path can leave the non-stopping region in three ways: Either to the top via the former upper boundary of Figure 1 (left), in which case we report the classic interval $(0.05, 1]$, to the bottom via the former lower boundary corresponding to the bucket $[0, 0.05]$, or to the middle corresponding to the added bucket $(0.03, 0.07]$.

Example 2. *Similarly to Example 1, Figure 2 shows the non-stopping region for $\mathcal{J}^0$ and $\mathcal{J}^*$. Again, the stopping region is infinite for the non-overlapping $\mathcal{J}^0$ and finite for $\mathcal{J}^*$ consisting of overlapping p-value buckets.*

How likely is it to observe the different decisions possible when testing with $\mathcal{J}^*$? Figure 3 shows the probability of obtaining each decision in the extended star rating system for $\mathcal{J}^*$ as a function of p. These probabilities are computed as follows: For a given p, we iteratively (over n) compute the distribution of S_n conditional on not stopping. This allows us to compute the probability of stopping and the resulting decision.

*Figure 3 shows that intermediate decisions (~, *~, **~) only occur with appreciable probability for a narrow range of p-values. For most p-values, a decision in the sense of the classical star rating system is reached.*

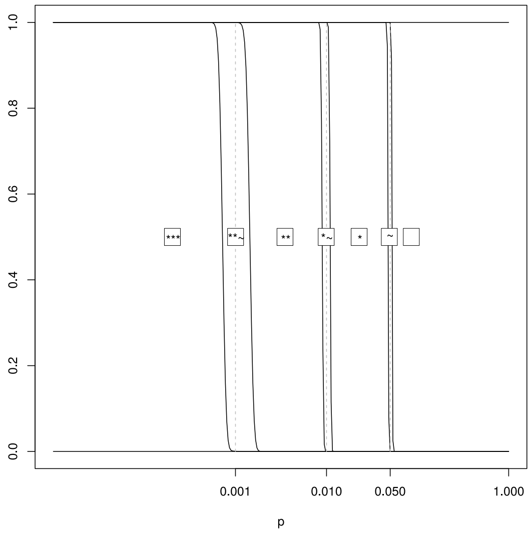

Figure 3: Probabilities of observing each possible decision with $\mathcal{J}^*$ as a function of p.

3 Construction of Confidence Sequences

We now present two approaches to computing confidences sequences and show that, for overlapping buckets, the resulting stopping times are bounded.

3.1 The Robbins-Lai approach

Confidence sequences can be constructed from likelihood martingale inequalities (Robbins, 1970; Lai, 1976). To be precise, Robbins (1970) proves that the following inequality

$$\mathcal{P}_p\left(\exists n \in \mathbb{N} : b(n, p, S_n) \leq \frac{\epsilon}{n+1}\right) \leq \epsilon \tag{7}$$

holds true for all $p \in (0, 1)$ and $\epsilon \in (0, 1)$, where $b(n, p, x) = \binom{n}{x} p^x (1-p)^{n-x}$. The statement (7) is trivially true for $p \in \{0, 1\}$. Therefore, $I_n = \{p \in [0, 1] : (n + 1)b(n, p, S_n) > \epsilon\}$ is a sequence of confidence sets for p with the desired coverage probability of $1 - \epsilon$.

Lai (1976) further shows that I_n are intervals. Indeed, if $0 < S_n < n$ we have $I_n = (g_n(S_n), f_n(S_n))$, where $g_n(x) < f_n(x)$ are the two distinct roots of $(n + 1)b(n, p, x) = \epsilon$. In the case $S_n = 0$, the equation $(n+1)b(n, p, x) = \epsilon$ has only one root r_n, leading to $I_n = [0, r_n)$. Likewise for the case $S_n = n$, which leads to $I_n = (r_n, 1]$.

For overlapping buckets $\mathcal{J}$, the stopping time $\tau_{\mathcal{J}}$ can always be bounded by a deterministic positive constant. Indeed, the following two lemmas show that the length of I_n uniformly goes to zero and, moreover, that once an interval is below a certain length, it is guaranteed to be contained in one of the buckets, ensuring that the general algorithm stops.

Lemma 1. *Let $n \in \mathbb{N}$ and $|I_n|$ be the length of the interval I_n. Then* $|I_n| \leq \left[\frac{2}{n} \log\left(\frac{n+1}{\epsilon}\right)\right]^{1/2}$.

Lemma 2. *If $\mathcal{J}$ is an overlapping set of p-value buckets then there exists $c > 0$ s.t. for all intervals $I \subseteq [0, 1]$ with length less than c there exists $J \in \mathcal{J}$ such that $I \subseteq J$.*

The two roots $g_n(S_n)$ and $f_n(S_n)$ need not be computed explicitly to determine if $I_n \subseteq J$. Indeed, for every $J \in \mathcal{J}$, it suffices to first check if $(n+1)b(n, \alpha, S_n) > \epsilon$ at $\alpha \in \{\min J, \max J\}$ to see if any boundary of J is contained in I_n. If this is not the case, one can use the derivative of $(n + 1)b(n, \alpha, S_n)$ with respect to α at $\alpha \in \{\min J, \max J\}$ together with the unimodality of $(n + 1)b(n, \alpha, S_n)$ in α to check if $I_n \subseteq J$. Details are given in Appendix B. For a single threshold this approach has been suggested in Ding et al. (2016).

3.2 The Simctest approach

Gandy (2009) constructed stopping boundaries to compute a decision for a p-value with respect to a single threshold $\alpha \in [0,1]$. We revisit this method before showing how it can be extended to p-value buckets.

Before observing Monte Carlo samples, two integer sequences $(L_i)_{i\in\mathbb{N}}$ and $(U_i)_{i\in\mathbb{N}}$ serving as lower and upper stopping boundaries are computed. The algorithm then proceeds to draw samples $(X_i)_{i\in\mathbb{N}}$ until the trajectory (n, S_n) hits either boundary. The stopping time for this method is thus $\tau = \inf\{k \in \mathbb{N} : S_k \geq U_k \text{ or } S_k \leq L_k\}$.

The two boundaries $(L_i)_{i\in\mathbb{N}}$ and $(U_i)_{i\in\mathbb{N}}$ are a function of both the threshold α and some bound on the resampling risk ρ. They are computed recursively in such a way that, given $p \leq \alpha$ ($p > \alpha$), the probability of hitting the upper (lower) boundary is less than ρ. Starting with $U_1 = 2$, $L_1 = -1$, the two sequences are recursively defined as

$$\begin{aligned} U_n &= \min\left\{j \in \mathbb{N} : \mathcal{P}_\alpha(\tau \geq n, S_n \geq j) + \mathcal{P}_\alpha(\tau < n, S_\tau \geq U_\tau) \leq \epsilon_n\right\}, \\ L_n &= \max\left\{j \in \mathbb{Z} : \mathcal{P}_\alpha(\tau \geq n, S_n \leq j) + \mathcal{P}_\alpha(\tau < n, S_\tau \leq L_\tau) \leq \epsilon_n\right\}, \end{aligned} \tag{8}$$

where $(\epsilon_n)_{n\in\mathbb{N}}$ is a non-decreasing sequence satisfying $\epsilon_n \to \rho$ as $n \to \infty$ and $0 \leq \epsilon_n \leq \rho$. It controls how the overall error ρ is spent over all iterations of the algorithm (called a *spending sequence* in Gandy (2009)). In the remaining sections of this article we use

$$\epsilon_n = \rho \frac{n}{n+k} \tag{9}$$

with $k = 1000$, which is the *default spending sequence* suggested in Gandy (2009).

The aforementioned method has a finite expected stopping time (for $p \neq \alpha$) and the probability of hitting the *wrong* boundary (leading to a decision not equal to the one obtained based on the unknown p) is bounded by ρ (under the conditions $\rho \leq 1/4$ and $\log(\epsilon_n - \epsilon_{n-1}) = o(n)$ as $n \to \infty$, see (Gandy, 2009, Theorem 1)). Thus, upon stopping we define $I = [0, \alpha]$ in case of hitting the lower boundary ($S_\tau \leq L_\tau$) and $I = (\alpha, 1]$ in case of hitting the upper boundary ($S_\tau \geq U_\tau$). By construction, the interval I has a coverage probability of $1 - \rho$.

To extend the approach of Gandy (2009) to multiple thresholds we proceed as follows. We first define the set of boundaries of the intervals in $\mathcal{J}$ that are in the interior of $[0,1]$:

$$A_\mathcal{J} := \{\min J,\ \max J :\ J \in \mathcal{J}\} \setminus \{0, 1\},$$

where $\min J$ ($\max J$) denote the lower (upper) limit of the interval J, respectively. Then we construct the above stopping boundaries for each $\alpha \in A_\mathcal{J}$, denoted as $L_{n,\alpha}$ and $U_{n,\alpha}$, using the same ρ.

We define corresponding stopping times $\sigma_\alpha = \inf\{k \in \mathbb{N} : S_k \geq U_{k,\alpha} \text{ or } S_k \leq L_{k,\alpha}\}$ (based on the same sequence $X_j, j \in \mathbb{N}$, see Section 2.1). We then define

$$I_{n,\alpha} = \begin{cases} [0,1] & \text{if } n < \sigma_\alpha, \\ [0,\alpha) & \text{if } n \geq \sigma_\alpha, S_{\sigma_\alpha} \leq L_{\sigma_\alpha,\alpha}, \\ (\alpha, 1] & \text{if } n \geq \sigma_\alpha, S_{\sigma_\alpha} \geq U_{\sigma_\alpha,\alpha}, \end{cases}$$

and let $I_n = \bigcap_{\alpha \in A_\mathcal{J}} I_{n,\alpha}$.

The following theorem shows that I_n has the desired joint coverage probability given in (4) or (6) when setting $\rho = \epsilon/2$.

Theorem 2. *Let $N \in \mathbb{N} \cup \{\infty\}$. Suppose that $U_{n,\alpha} \leq U_{n,\alpha'}$ and $L_{n,\alpha} \leq L_{n,\alpha'}$ for all $\alpha, \alpha' \in A_\mathcal{J}$, $\alpha < \alpha'$, and $n < N$ (computed as in (8) with overall error ρ for each $\alpha \in A_\mathcal{J}$). Then for all $p \in [0,1]$,*

$$\mathcal{P}_p(p \in I_n\ \forall n < N) \geq 1 - 2\rho.$$

Allowing $N < \infty$ is useful for stopping boundaries constructed to yield a finite runtime (see (6)).

The condition on the monotonicity of the boundaries ($U_{n,\alpha} \leq U_{n,\alpha'}$ and $L_{n,\alpha} \leq L_{n,\alpha'}$ for all $n \in \mathbb{N}$ and $\alpha, \alpha' \in \mathcal{J}$ with $\alpha < \alpha'$) can be checked for a fixed spending sequence ϵ_n in two ways: For finite N, the two inequalities can be checked manually after constructing the boundaries. For $N = \infty$, the following lemma shows that under conditions, the monotonicity of the boundaries holds true for all $n \geq n_0$, where $n_0 \in \mathbb{N}$ can be computed as a solution to inequality (15), given in the proof of Lemma 3 in Appendix A. For $n < n_0$, the inequalities again have to be checked manually.

Lemma 3. *Suppose $\rho \leq 1/4$ and $\log(\epsilon_n - \epsilon_{n-1}) = o(n)$ as $n \to \infty$. Let $\alpha, \alpha' \in A_{\mathcal{J}}$ with $\alpha < \alpha'$. Then there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,*

$$L_{n,\alpha} \leq L_{n,\alpha'}, U_{n,\alpha} \leq U_{n,\alpha'}.$$

The condition on the spending sequence in Lemma 3 is identical to the condition imposed in Theorem 1 of Gandy (2009) and is satisfied by the default spending sequence (9). Therefore, our default spending sequence (9) with the p-value buckets used in this article ($\mathcal{J}^0$ and $\mathcal{J}^*$) satisfies the boundary conditions of Theorem 2.

The following theorem shows that for overlapping buckets the algorithm has a bounded stopping time.

Theorem 3. *Suppose the conditions of Theorem 2 and Lemma 3 hold true with $N = \infty$. If $\mathcal{J}$ is a finite set of overlapping p-value buckets then the general construction of Section 2.4 has a bounded stopping time $\tau_{\mathcal{J}}$, i.e. there exists $c < \infty$ s.t. $\tau_{\mathcal{J}} \leq c$.*

4 Computational effort

This section investigates the expected computational effort of the algorithm of Section 2.4. We start by deriving a theoretical lower bound on the expected effort in Section 4.1. We then compare both the Simctest and Robbins-Lai approach of Section 3 in terms of their expected effort as a function of p (Section 4.2). Integrating this effort for certain p-value distributions of practical interest allows to compare both approaches in practical situations (Section 4.3). Section 4.4 shows that the algorithm can be used for small p-values arising in multiple testing settings.

4.1 Lower bounds on the expected effort

In this section we construct lower bounds on the expected number of steps of sequential procedures satisfying (3). The key idea is to consider hypothesis tests implied by (3) and then to use the lower bounds for the expected effort of sequential tests (Wald, 1945, eq. (4.80)).

We suppose that I is the (random) bucket reported by a sequential procedure that respects (3). Let $\tilde{p} \in [0, 1]$. We give a basic and an improved lower bound on $\mathbb{E}_{\tilde{p}}[\tau]$.

First, let $\tilde{J} = \bigcup_{J \in \mathcal{J}, \tilde{p} \in J} J$ be the union of all buckets that $\tilde{p}$ is contained in. For any $q \in [0, 1] \setminus \tilde{I}$, we can consider the hypotheses $H_0 : p = \tilde{p}$ against $H_1 : p = q$ and the test that rejects H_0 iff $p \notin I$. By (3), the type I error of such a test is at most ϵ. Also, the type II error is at most ϵ, as $\mathcal{P}_q(p \in I) \leq \mathcal{P}_q(q \notin I) \leq \epsilon$. Hence, $\mathbb{E}_{\tilde{p}}[\tau]$ is bounded from below by the lower bound in (Wald, 1945, eq. (4.80)), which we call $a(q)$. Thus we get

$$\mathbb{E}_{\tilde{p}}[\tau] \geq \max_{q \notin \tilde{J}} a(q). \tag{10}$$

The maximum in (10) can be evaluated by looking at the boundary points of $\tilde{J}$.

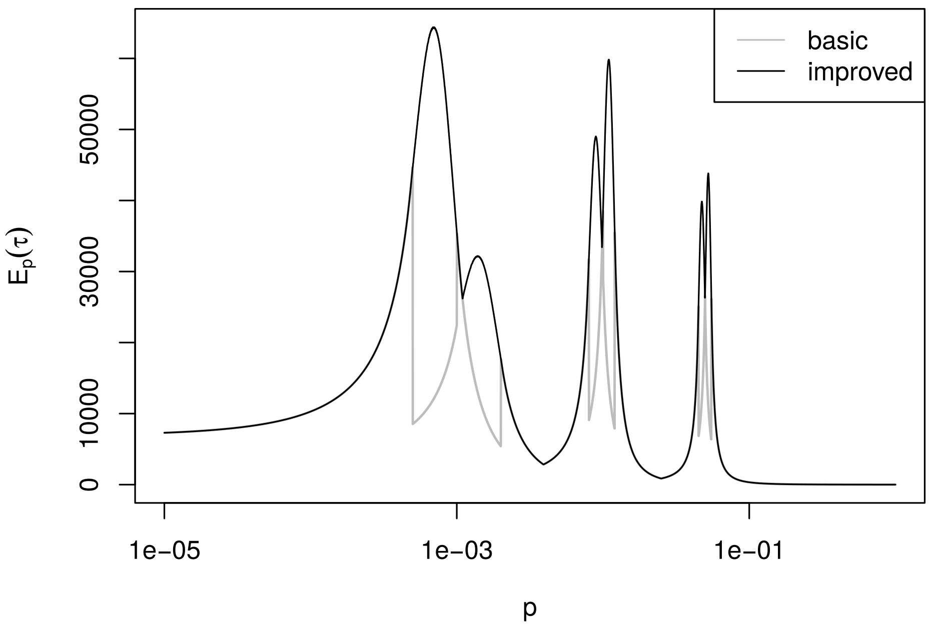

Figure 4: Basic (gray) and improved (black) lower bounds on the effort $\mathbb{E}_p[\tau]$ for the buckets $\mathcal{J}^*$.

The bound (10) can be improved if the number of elements of $\mathcal{J}$ containing $\tilde{p}$ is exactly two, say J_1 and J_2. Suppose that for a given sequential procedure, $\eta = \mathcal{P}_{\tilde{p}}(I_1)$. Let $q_1 \in [0,1] \setminus J_1$. Consider the hypotheses $H_0 : p = \tilde{p}$ and $H_1 : p = q_1$ and the corresponding test that rejects H_0 iff $I \neq J_1$. This test has type I error $1 - \eta$ and type II error ϵ. Again, using (Wald, 1945, eq. (4.80)) we get a lower bound on $\mathbb{E}_{\tilde{p}}[\tau]$, which we call $b_1(q, \eta)$. Similarly, for any $q_2 \in [0,1] \setminus J_2$, we can test the hypotheses $H_0 : p = \tilde{p}$ and $H_1 : p = q_2$ by rejecting H_0 iff $I \neq J_2$. This test has type I error of at most $\eta + \epsilon$ and type II error of at most ϵ. Again, using (Wald, 1945, eq. (4.80)) we get a lower bound on $\mathbb{E}_{\tilde{p}}[\tau]$, which we call $b_2(q, \eta)$.

As η is dependent on the specific procedure, we can get a universal lower bound on $\mathbb{E}_{\tilde{p}}[\tau]$ by minimizing over η, thus

$$\mathbb{E}_{\tilde{p}}[\tau] \geq \max\left(\max_{q \notin \tilde{J}} a(q), \min_{\eta \in [0,1]} \max\left\{\max_{q \notin J_1} b_1(q, \eta), \max_{q \notin J_2} b_2(q, \eta)\right\}\right). \tag{11}$$

The maxima in (11) can be evaluated by looking at the boundary points of $\tilde{J}$, J_1 and J_2. The minimum can be bounded from below by looking at a grid of values for η and conservatively replacing $b_2(q, \eta)$ by $b_2(q, \eta + \delta)$, where δ is the grid width.

Figure 4 gives an example of both the basic and the improved lower bound on $\mathbb{E}_{\tilde{p}}[\tau]$ for the buckets $\mathcal{J}^*$. The improved bound is much higher (and thus better) in the areas where there are overlapping buckets.

4.2 Expected effort for (non-)overlapping buckets

This section investigates both the non-overlapping p-value buckets $\mathcal{J}$ as well as the overlapping buckets $\mathcal{J}^*$ with respect to the implied expected effort as a function of p.

Using the non-stopping regions depicted in Figure 2, Figure 5 shows the expected effort (measured in terms of the number of samples drawn) to compute a decision with respect to $\mathcal{J}$ (left) and $\mathcal{J}^*$ as a function of $p \in [10^{-6}, 1]$. For any given p, the expected effort is computed by iteratively (over n) updating the distribution of S_n conditional on not having

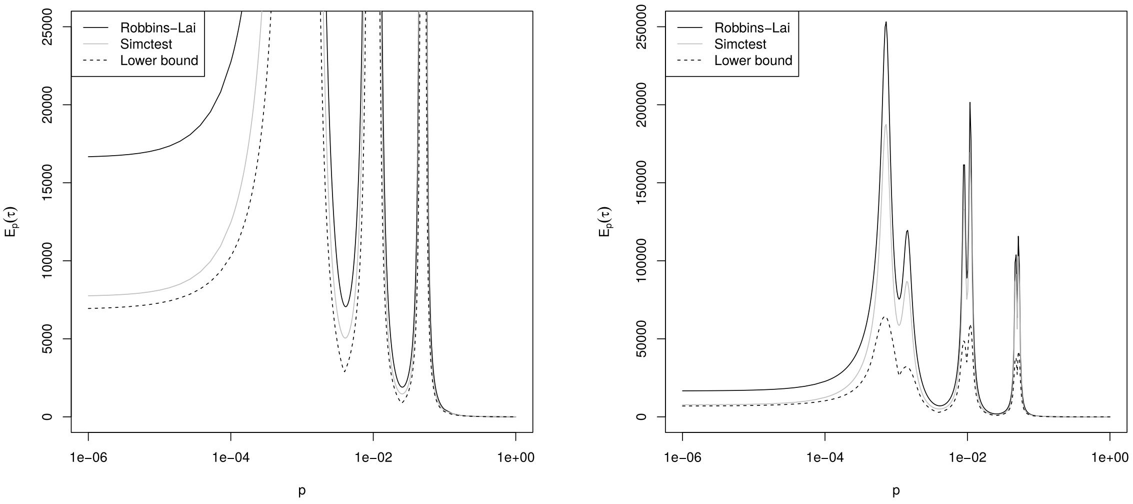

Figure 5: Expected effort to compute a decision with respect to $\mathcal{J}$ (left) and $\mathcal{J}^*$ (right) as a function of p. Both Simctest and Robbins-Lai are used to compute confidence sequences. Lower bound on the effort given as dashed line.

stopped up to time n. Using this distribution, we work out the probability of stopping at step n and add the appropriate contribution to the overall effort.

The effort diverges as p approaches any of the thresholds in $\mathcal{J}$. For $\mathcal{J}^*$ the effort stays finite even in the case that p coincides with one of the thresholds (Figure 5, right). The effort is maximal in a neighborhood around each threshold, while in-between thresholds, the effort slightly decreases. For p-values larger than the maximal threshold in $\mathcal{J}^*$, the effort decreases to zero. The effort for Simctest seems to be uniformly smaller than the one for Robbins-Lai for both $\mathcal{J}$ and $\mathcal{J}^*$.

Figure 5 also shows the lower bound (dashed line) on the effort derived in Section 4.1. In connection with Simctest, the effort of our algorithm of Section 2.4 differs from the theoretical lower bound by roughly a factor of three.

4.3 Expected effort for three specific p-value distributions

The expected effort of the proposed methods for repeated use can be obtained by integrating the expected effort for fixed p (Figure 5, right) with respect to certain p-value distributions.

Here, we consider using the overlapping buckets $\mathcal{J}^*$ with three different p-value distributions. These are a uniform distribution in the interval $[0, 1]$ (H_0), as well as two alternatives given by the density $\frac{1}{2} + 10\mathbb{I}(x \leq 0.05)$ (H_{1a}) and by a Beta$(0.5, 25)$ distribution (H_{1b}), where $\mathbb{I}$ denotes the indicator function.

Table 2 shows the expected effort as well as the lower bound on the expected effort. The Simctest approach (Section 3.2) dominates the one of Robbins-Lai (Section 3.1) for this specific choice of distributions. As expected, the effort is lowest for a uniform p-value distribution, and more extreme for the alternatives having higher probability mass on low p-values. Using Simctest, the expected effort stays within roughly a factor of two of the theoretical lower bound derived in Section 4.1.

	Robbins-Lai	Simctest	Lower bound
H0	2228	1853	975
H1a	16878	13837	7126
H1b	40059	30896	15885

Table 2: Expected (integrated) effort for both Robbins-Lai and Simctest applied to $\mathcal{J}^*$. Expectations are taken over three different p-value distributions: a uniform $U[0,1]$ distribution (H_0) and two mixture distributions H_{1a} and H_{1b}.

4.4 Application to multiple testing

We consider the applicability of our algorithm of Section 2.4 to the (lower) testing thresholds occurring in multiple testing scenarios. In the following example, we demonstrate that our algorithm is well suited as a screening procedure for the most significant hypotheses: Even for small threshold values, it is capable of detecting more rejections than a naïve sampling procedure that uses an equal number of samples.

We assume we want to test $n = 10^4$ hypotheses using the Bonferroni (1936) procedure to correct for multiplicity. In order to be able to compute numbers of false classifications, we assign $n_A = 100$ hypotheses to the alternative, the remaining $n - n_A = 9900$ hypotheses are from the null. The p-values of the alternative are then set to $1 - F(X)$, where F is the cumulative distribution function of a Student's t-distribution with 100 degrees of freedom and X is a random variable sampled from a t-distribution with 100 degrees of freedom and noncentrality parameter uniformly chosen in $[2, 6]$. The p-values of the null are sampled from a uniform distribution in $[0, 1]$.

We apply our algorithm from Section 2.4 with $\epsilon = 10^{-3}$ and confidence sequences computed with the Simctest approach (Section 3.2). To speed up the Monte Carlo sampling, we sample in batches of geometrically increasing size $a^i b$ in each iteration i, where $b = 10$ and $a = 1.1$. Likewise, both the stopping boundaries and the stopping condition (hitting of either boundary) in Simctest are updated (checked) in batches of the same size.

In order to screen hypotheses, we aim to group them by the order of magnitude of their p-values. For this we employ the overlapping buckets

$$\mathcal{J}^s = \left\{ \left[0, 10^{-7}\right] \right\} \cup \left\{ \left(10^{i-2}, 10^{i}\right] : i = -6, \dots, 0 \right\}$$

which group the p-values in buckets spanning two orders of magnitude each (and $\left[0, 10^{-7}\right]$).

We now report the results from a single run of this setup. Our algorithm draws $N = 3.2 \cdot 10^5$ samples per hypothesis. Of the 10^4 hypotheses, 28 are correctly allocated to the two lowest buckets. As expected, the p-values from the null are all allocated to larger buckets (covering values from 10^{-4} onwards).

An alternative approach would be to draw an equal number of N samples per hypothesis and to compute a p-value using a pseudo-count (Davison and Hinkley, 1997). Due to this pseudo-count, this naïve approach is incapable of observing p-values below $N^{-1} = 3.125 \cdot 10^{-6}$, and in particular incapable of observing any p-values in the lower bucket.

5 Discussion

In this article we investigate methods capable of computing a decision for a single hypothesis H_0 with unknown p-value p (approximated via Monte Carlo sampling) that achieve both a bounded resampling risk and a finite runtime. We first generalize testing thresholds to p-value buckets and prove that methods having both aforementioned properties necessarily need to operate on overlapping p-value buckets (Section 2).

In order to report decisions when testing with overlapping buckets, we propose to use an extension of the traditional star rating system used to report the significance of a hypothesis.

Our algorithms rely on the computation of a confidence sequence for the unknown p. We give two constructions of such confidence sequences (Section 3), prove that both approaches indeed satisfy the bound on the resampling risk and yield a finite runtime for overlapping buckets. We (empirically) demonstrate that our methods achieve a competitive computational effort that is close to a theoretical lower bound on the effort (Section 4).

The choice of (overlapping) p-value buckets we employ in our article is arbitrary. However, a variety of (heuristic) techniques can be used to obtain overlapping buckets from traditional thresholds $T = \{t_0, \dots, t_m\}$. These include:

1. The bucket overlapping each threshold $t \in T$ can be chosen as a fixed proportion $\rho \in (0, 1)$, leading to the interval $[\rho t, \rho^{-1} t]$.

2. Since the length of a confidence interval for a binomial quantity p behaves proportionally to $\sqrt{p(1-p)} \in O\left(\sqrt{p}\right)$ as $p \to 0$, a bucket around any $t \in T$ can be chosen as $J := [t - \rho\sqrt{t}, t + \rho\sqrt{t}]$, where $\rho > 0$ is such that $0 \notin J$.

3. The buckets can be chosen to match the precision of a naïve sampling method which draws a fixed number of samples $n \in \mathbb{N}$ per hypothesis. For this we compute all $n + 1$ possible confidence intervals (one for each possible $S_n \in \{0, \dots, n\}$) for each threshold $t \in T$ and record all confidence intervals which cover t. The union of those intervals can then be used as a bucket for t.

The tuning parameter ρ can be chosen, for instance, to minimize the maximal (worst case) effort for the resulting overlapping buckets.

The present article leaves scope for a variety of future research directions. For instance, how can overlapping p-value buckets be chosen to maximize the probability of obtaining a classical decision (*, ** or ***), subject to a suitable optimization criterion? How can the lower bound on the computational effort derived in Section 4.1 be improved? Which algorithm (possibly based on our generic algorithm in connection with Simctest) is capable of meeting the lower bound effort?

A Proofs

Proof of Theorem 1. We prove a circular equivalence of the three statements.

(1.) $\Rightarrow$ (2.): Suppose the buckets $\mathcal{J}$ are not overlapping. This implies that there exists $\alpha \in (0, 1)$ that is not contained in the interior of any $J \in \mathcal{J}$.

Let $I \in \mathcal{J}$ be the (random) interval reported by an algorithm satisfying (3).

Let $n \in \mathbb{N}$ such that $\alpha - 1/n \geq 0$ and $\alpha + 1/n \leq 1$.

Consider the hypotheses $H_0 : p = \alpha - 1/n$ and $H_1 : p = \alpha + 1/n$ and the test that rejects H_0 iff $\alpha - 1/n \notin I$. As I cannot contain both $\alpha - 1/n$ and $\alpha + 1/n$ (otherwise α would be in the interior of the interval I) and because of (3), this test has type I and type II error of at most ϵ. Hence, by the lower bound on the expected number of steps of a sequential test given in (Wald, 1945, equation (4.81)) (see also (Gandy, 2009, section 3.1)), we have

$$\mathbb{E}_{\alpha+1/n}[\tau] \geq \frac{\epsilon \log(\frac{\epsilon}{1-\epsilon}) + (1-\epsilon)\log(\frac{1-\epsilon}{\epsilon})}{(\alpha + \frac{1}{n})\log(\frac{\alpha+1/n}{\alpha-1/n}) + (1 - \alpha - \frac{1}{n})\log(\frac{1-\alpha-1/n}{1-\alpha+1/n})}.$$

As $n \to \infty$, the right hand side converges to ∞, contradicting (1.).

(2.) ⇒ (3.): We construct an explicit (but not very efficient) algorithm for this.

Let $a_0 < a_1 < \ldots < a_k$ be the set of boundaries of buckets in $\mathcal{J}$, i.e. $\{a_0, \ldots, a_k\} = \{\max J : J \in \mathcal{J}\} \cup \{\min J : J \in \mathcal{J}\}$. Let $\Delta = \min\{a_i - a_{i-1} : i = 1, \ldots, k\}$ be the minimal gap between those boundaries.

Let $I(S, n)$ be the two-sided Clopper and Pearson (1934) confidence interval with coverage probability $1 - \epsilon$ for p when n is the number of samples and S is the number of exceedances. Let $n \in \mathbb{N}$ be such that the length of all Clopper-Pearson intervals is less than Δ, i.e. $n = \min\{m \in \mathbb{N} : \forall S \in \{0, \ldots, m\} : |I(S, m)| < \Delta\}$. This is well-defined as the length of the Clopper Pearson confidence interval $I(S, n)$ decreases to 0 uniformly in S as $n \to \infty$; see e.g. the proof of Condition 2 in Lemma 2 of Gandy and Hahn (2014) for this.

Consider the algorithm that takes n samples $X_1, \ldots, X_n$ and then returns an arbitrary interval $I \in \mathcal{J}$ that satisfies $I \supseteq I(\sum_{i=1}^n X_i, n)$ (to be definite, order all elements in $\mathcal{J}$ arbitrarily and return the first element satisfying the condition. Such an I always exists as the buckets are overlapping by (2.) and as $|I(\sum_{i=1}^n X_i, n)| < \Delta$, implying that it overlaps with at most one possible boundary. This algorithm satisfies (3) due to the coverage probability of $1 - \epsilon$ of the Clopper-Pearson interval.

(3.) ⇒ (1.): Since finite effort implies expected finite effort, (1.) follows immediately. □

Proof of Lemma 1. If $0 \leq p \leq \frac{S_n}{n} - \left[\frac{1}{2n} \log\left(\frac{n+1}{\epsilon}\right)\right]^{1/2}$ then, by Hoeffding's inequality (Hoeffding, 1963),

$$b(n, p, S_n) = \mathcal{P}(X = S_n) \leq \mathcal{P}\left(\frac{X}{n} - p \geq \frac{S_n}{n} - p\right) \leq \exp\left(\frac{-2(S_n - np)^2}{n}\right) \leq \frac{\epsilon}{n+1},$$

where $X \sim \text{Bin}(n, p)$ is a binomial random variable. Hence, $p \notin I_n$. A similar argument can be made for $\frac{S_n}{n} + \left[\frac{1}{2n} \log\left(\frac{n+1}{\epsilon}\right)\right]^{1/2} \leq p \leq 1$. Thus, $|I_n| \leq \left[\frac{2}{n} \log\left(\frac{n+1}{\epsilon}\right)\right]^{1/2}$. □

Proof of Lemma 2. Suppose this is not true. Then for all $n \in \mathbb{N}$ there exists an interval $I_n \subset [0, 1]$ with $0 < |I_n| < \frac{1}{n}$ s.t. $\forall J \in \mathcal{J}$: $I_n \not\subseteq J$. Let a_n be the mid points of I_n. As (a_n) is a bounded sequence, there exists a convergent subsequence (a_{n_k}). Let $b = \lim_{k \to \infty} a_{n_k}$.

If $b \in (0, 1)$ then, as $\mathcal{J}$ is overlapping, there exists $\epsilon > 0$ and $J \in \mathcal{J} : (b - \epsilon, b + \epsilon) \subseteq J$. For large enough k we have $I_{n_k} \subseteq (b - \epsilon, b + \epsilon)$, contradicting $I_{n_k} \not\subseteq J$.
If $b = 0$ then, as $\mathcal{J}$ is a covering of $[0, 1]$ consisting of intervals of positive length there exists $\epsilon > 0$ and $J \in \mathcal{J}$ s.t. $[0, \epsilon) \subseteq J$. For large enough k we have $I_{n_k} \subseteq [0, \epsilon)$, again contradicting $I_{n_k} \not\subseteq J$. If $b = 1$ a contradiction can be derived similarly. □

Proof of Theorem 2. For a given threshold $\alpha \in A_{\mathcal{J}}$, let

$$\overline{E}_\alpha^N = \{S_{\tau_\alpha} \geq U_{\tau_\alpha, \alpha}, \tau_\alpha < N\}$$

be the event that the upper boundary is hit first before time N and likewise let

$$\underline{E}_\alpha^N = \{S_{\tau_\alpha} \leq L_{\tau_\alpha, \alpha}, \tau_\alpha < N\}$$

be the event that the lower boundary is hit first. Then, for all $\alpha, \alpha' \in A_{\mathcal{J}}$ with $\alpha < \alpha'$ the following holds:

$$\overline{E}_\alpha^N \supseteq \overline{E}_{\alpha'}^N \quad \text{and} \quad \underline{E}_\alpha^N \subseteq \underline{E}_{\alpha'}^N. \tag{12}$$

Indeed, to see $\overline{E}_\alpha^N \supseteq \overline{E}_{\alpha'}^N$, we can argue as follows. On the event $\overline{E}_{\alpha'}^N$, as $U_{n,\alpha} \leq U_{n,\alpha'}$ for all $n \in \mathbb{N}$, the trajectory (n, S_n) must hit the upper boundary $U_{n,\alpha}$ of α no later than $\tau_{\alpha'}$, hence $\tau_\alpha \leq \tau_{\alpha'} < N$. It remains to prove that the trajectory does not first hit the lower boundary $L_{n,\alpha}$ of α. Indeed, if the trajectory does hit the lower boundary of α before hitting its upper boundary, it also hits the lower boundary of α' (as $L_{n,\alpha} \leq L_{n,\alpha'}$ for all $n < N$) before time $\tau_{\alpha'}$, thus contradicting being on the event $\overline{E}_{\alpha'}^N$. Hence, we have $\overline{E}_{\alpha}^N \supseteq \overline{E}_{\alpha'}^N$. The proof of $\underline{E}_{\alpha}^N \subseteq \underline{E}_{\alpha'}^N$ is similar.

Using this notation, for all $p \in [0,1]$,

$$\begin{aligned}
&\mathcal{P}_p(\exists n < N : p \notin I_n) \leq \mathcal{P}_p(\exists n < N, \alpha \in A_{\mathcal{J}} : p \notin I_{n,\alpha}) \\
=&\mathcal{P}_p\left(\bigcup_{\alpha \in A_{\mathcal{J}}:\alpha \leq p} \underline{E}_{\alpha}^N \cup \bigcup_{\alpha \in A_{\mathcal{J}}:\alpha \geq p} \overline{E}_{\alpha}^N \right) \leq \mathcal{P}_p\left(\bigcup_{\alpha \in A_{\mathcal{J}}:\alpha \leq p} \underline{E}_{\alpha}^N \right) + \mathcal{P}_p\left(\bigcup_{\alpha \in A_{\mathcal{J}}:\alpha \geq p} \overline{E}_{\alpha}^N \right).
\end{aligned} \tag{13}$$

If $p < \min A_{\mathcal{J}}$ then the first term is equal to 0. Otherwise, let $\alpha' = \max\{\alpha \in A_{\mathcal{J}} : \alpha \leq p\}$. Then, by (12),

$$\mathcal{P}_p\left(\bigcup_{\alpha \in A_{\mathcal{J}}:\alpha \leq p} \underline{E}_{\alpha}^N \right) = \mathcal{P}_p\left(\underline{E}_{\alpha'}^N\right) \leq \rho.$$

The second term on the right hand side of (13) can be dealt with similarly. □

Proof of Lemma 3. By arguments in (Gandy, 2009, Proof of Theorem 1), we have

$$\frac{U_{n,\alpha} - n\alpha}{n} \leq \frac{\Delta_n + 1}{n} \to 0, \qquad \frac{L_{n,\alpha'} - n\alpha'}{n} \geq -\frac{\Delta_n + 1}{n} \to 0 \tag{14}$$

as $n \to \infty$, where $\Delta_n = \sqrt{-n\log(\epsilon_n - \epsilon_{n-1})/2}$. Since $\Delta_n = o(n)$ there exists $n_0 \in \mathbb{N}$ such that

$$2\left(\frac{\Delta_n}{n} + \frac{1}{n}\right) \leq \alpha' - \alpha \text{ for all } n \geq n_0. \tag{15}$$

Splitting $\frac{2}{n} = \frac{1}{n} + \frac{1}{n}$ and multiplying by n yields $n\alpha + \Delta_n + 1 \leq n\alpha' - \Delta_n - 1$ from which $U_{n,\alpha} \leq L_{n,\alpha'}$ follows by (14).

By definition, we have $L_{n,\alpha} \leq U_{n,\alpha}$ and $L_{n,\alpha'} \leq U_{n,\alpha'}$ for all $n \in \mathbb{N}$, thus implying $L_{n,\alpha} \leq L_{n,\alpha'}$, $U_{n,\alpha} \leq U_{n,\alpha'}$ for all $n \geq n_0$ as desired. □

Proof of Theorem 3. By (14) and as $\Delta_n = o(n)$ there exists $n_0 \in \mathbb{N}$ such that

$$|\{\alpha \in A_{\mathcal{J}} : \sigma_\alpha > n_0\}| \leq 1. \tag{16}$$

We will show that $\tau_{\mathcal{J}} \leq n_0$.

First, the assumption on the ordering of L_n and U_n exclude the possibility of $I_{n_0} = \emptyset$. Second, (16) implies $|I_{n_0} \cap A_{\mathcal{J}}| \leq 1$.

If $|I_{n_0} \cap A_{\mathcal{J}}| = 1$ then let $\alpha \in A_{\mathcal{J}}$ be such that $\alpha \in I_{n_0}$. As $\mathcal{J}$ is overlapping, there exist $J \in \mathcal{J}$ such that α is in the interior of J. Hence, α cannot be a boundary of J, implying $I_{n_0} \subseteq J$ due to $|I_{n_0} \cap A_{\mathcal{J}}| = 1$, thus showing $\tau_{\mathcal{J}} \leq n_0$.

If $|I_{n_0} \cap A_{\mathcal{J}}| = 0$ then let β be in the interior of I_{n_0}. As $\mathcal{J}$ is overlapping, there exists $J \in \mathcal{J}$ such that $\beta \in J$. As $I_{n_0} \cap A_{\mathcal{J}} = \emptyset$ this implies $I_{n_0} \subseteq J$, showing $\tau_{\mathcal{J}} \leq n_0$. □

B A simple stopping criterion for Robbins-Lai

The following describes a simple criterion to determine whether a confidence interval computed via Robbins-Lai (Section 3.1) is fully contained in a bucket. Let interval I_n and bucket $J \in \mathcal{J}$ as well as n, S_n and ϵ be as in Section 3.1. Then $I_n \subseteq J$ if and only if

$$(n+1)b(n, S_n, p) = (n+1)\binom{n}{S_n} p^{S_n}(1-p)^{n-S_n} \leq \epsilon \tag{17}$$

for $p \in \{\min J, \max J\}$.

As (17) is also satisfied if I and J are simply disjoint, we verify that $(n+1)b(n, S_n, p)$ is indeed increasing at $\min J$ and decreasing at $\max J$ using the derivative of $(n+1)b(n, S_n, p)$ with respect to p. This then proves that the two limits of bucket J are indeed not both smaller than $\min I$ or larger than $\max I$. We first apply a (monotonic) log transformation,

$$\log\left[(n+1)b(n, S_n, p)\right] = \log(n+1) + \log\binom{n}{S_n} + S_n \log p + (n - S_n)\log(1-p),$$

and then take the derivative with respect to p:

$$\frac{S_n}{p} - \frac{n - S_n}{1-p} \begin{cases} \geq 0 & p = \min J, \\ \leq 0 & p = \max J. \end{cases} \tag{18}$$

If (17) and (18) are satisfied, then $I_n \subseteq J$.

References

American Psychological Association (2010). *Publication manual of the American Psychological Association (6th ed.).* American Psychological Association, Washington, DC.

Andrews, D. and Buchinsky, M. (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica*, 68(1):23–51.

Andrews, D. and Buchinsky, M. (2001). Evaluation of a three-step method for choosing the number of bootstrap repetitions. *J Econometrics*, 103(1-2):345–386.

Asomaning, N. and Archer, K. (2012). High-throughput DNA methylation datasets for evaluating false discovery rate methodologies. *Comput Stat Data An*, 56(6):1748–1756.

Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika*, 78(2):301–4.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Clopper, C. J. and Pearson, E. S. . (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413.

Davidson, R. and MacKinnon, J. (2000). Bootstrap Tests: How Many Bootstraps? *Economet Rev*, 19(1):55–68.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.

Dazard, J.-E. and Rao, J. (2012). Joint adaptive meanvariance regularization and variance stabilization of high dimensional data. *Comput Stat Data An*, 56(7):2317–2333.

Ding, D., Gandy, A., and Hahn, G. (2016). A simple method for implementing monte carlo tests. *arXiv:1611.01675*, pages 1–12.

Fay, M. and Follmann, D. (2002). Designing Monte Carlo Implementations of Permutation or Bootstrap Hypothesis Tests. *Am Stat*, 56(1):63–70.

Gandy, A. (2009). Sequential Implementation of Monte Carlo Tests With Uniformly Bounded Resampling Risk. *J Am Stat Assoc*, 104(488):1504–1511.

Gandy, A. and Hahn, G. (2014). Mmctesta safe algorithm for implementing multiple monte carlo tests. *Scand J Stat*, 41(4):1083–1101.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J Am Stat Assoc*, 58(301):13–30.

IBM Corp. (2013). *IBM SPSS Statistics for Windows*. IBM Corp., Armonk, NY.

Kim, H.-J. (2010). Bounding the Resampling Risk for Sequential Monte Carlo Implementation of Hypothesis Tests. *J Stat Plan Infer*, 140(7):1834–1843.

Lai, T. (1976). On Confidence Sequences. *Ann Stat*, 4(2):265–280.

Liu, J., Huang, J., Ma, S., and Wang, K. (2013). Incorporating group correlations in genome-wide association studies using smoothed group Lasso. *Biostatistics*, 14(2):205–219.

Lourenco, V. and Pires, A. (2014). M-regression, false discovery rates and outlier detection with application to genetic association studies. *Comput Stat Data An*, 78:33–42.

Martínez-Camblor, P. (2014). On correlated z-values distribution in hypothesis testing. *Comput Stat Data An*, 79:30–43.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robbins, H. (1970). Statistical Methods Related to the Law of the Iterated Logarithm. *Ann Math Stat*, 41(5):1397–1409.

SAS Institute Inc. (2011). *Base SAS 9.3 Procedures Guide*. SAS Institute Inc., Cary, NC.

Silva, I. and Assunção, R. (2013). Optimal generalized truncated sequential Monte Carlo test. *J Multivariate Anal*, 121:33–49.

Silva, I., Assunção, R., and Costa, M. (2009). Power of the Sequential Monte Carlo Test. *Sequential Analysis*, 28(2):163–174.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann Math Stat*, 16(2):117–186.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14(2):232–243.