Implementing Monte Carlo Tests with P-value Buckets

AXEL GANDY, GEORG HAHN and DONG DING

Department of Mathematics, Imperial College London,
London SW7 2AZ, U.K.

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 (***, **, *). 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-value thresholds occurring in multiple testing settings, and that they can be computationally more efficient than standard implementations.

Keywords: algorithm, bootstrap, hypothesis testing, p-value, resampling, sampling.

1 Introduction

Software packages usually report the significance of statistical tests using p-values. The result of the test will often be interpreted by comparing those p-values to 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 ***. As pointed out in the literature, such levels of significance are sensible since they capture the magnitude of a p-value rather than its precise value, which in contrast to the magnitude is usually not reliably estimated (Boos & Stefanski, 2011).

In this article, we are concerned with statistical tests whose p-value p can only be approximated by sequentially drawn Monte Carlo samples. Among others, this scenario arises in bootstrap or permutation tests (Lourenco & Pires, 2014; Martínez-Camblor, 2014; Liu et al., 2013; Wu et al., 2013; Asomaning & Archer, 2012; Dazard & 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 & Clifford, 1991; Silva et al., 2009).

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

There are also approaches that aim to bound an integrated resampling risk for a single threshold (Fay & Follmann, 2002; Kim, 2010; Silva & Assunção, 2013). Such an error criterion is weaker than 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 resampling risk and, under conditions, have a finite runtime. 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, i.e. a sequence of random intervals with a joint coverage probability. 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 and 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. We have also implemented the method in the function *mctest* of the R-package *simctest*, which is available on CRAN.

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 computed as the probability of observing a statistic at least as extreme as t, i.e.

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

where $\mathbb{P}$ is a probability measure under the null hypothesis. For our purposes, we assume that $\mathbb{P}$ is either the true distribution of T under H_0 or an estimate of it, e.g. the distribution implied via bootstrapping.

We assume that the p-value p is not available analytically but can be approximated

using Monte Carlo simulation by drawing independent realizations of the test statistic T under $\mathbb{P}$. We will assume that we can generate a sequence $X_i, i \in \mathbb{N}$, of those draws and we let $X_i = 1$ if the ith replicate is greater or equal than t and $X_i = 0$ otherwise. As a consequence, X_i have a Bernoulli(p) distribution.

The algorithms we consider aim to return an interval containing p from a given set $\mathcal{J}$ of possibly overlapping sub-intervals of $[0, 1]$. The algorithms are sequential; for $n = 1, 2, \ldots$, based on $X_{1:n} = (X_1, \ldots, X_n)$, they will decide if they can stop and return an interval or whether they need to observe more X_i.

We use A to denote a generic algorithm of this type, I_A (or simply I) to denote the interval returned by A, and τ_A for the stopping time of A, i.e. the number of X_i that the algorithm observes before returning I_A. We use J for generic elements of $\mathcal{J}$. Our algorithms are built on the sequence (X_i), which have p as the unknown parameter. To emphasize that the underlying distribution is determined by p, we will use it as a subscript in the notation of probabilities and expected values, writing $\mathbb{P}_p$ and $\mathbb{E}_p$.

Formally, we require $\mathcal{J}$ to be a *set of p-value buckets*, which we define to be a set of sub-intervals of $[0, 1]$ of positive length that cover $[0, 1]$, i.e. $\bigcup_{J\in\mathcal{J}} J = [0, 1]$.

For example,

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

is a set of p-value buckets, which we will refer to in the remainder of the article as *classical buckets*. Deciding which of those buckets p falls into is equivalent to deciding where p lies in relation to the three traditional thresholds 0.001, 0.01 and 0.05.

A natural error criterion for an algorithm A is the risk of a *wrong* decision, defined as $\mathrm{RR}_p(A) = \mathbb{P}_p(p \notin I_A)$, and which we call the *resampling risk*. $\mathrm{RR}_p(A)$ is a function of the p-value p.

The algorithms A that we propose in this article bound the resampling risk uniformly in p at a given $\epsilon \in (0, 0.5)$, i.e.

$$\mathrm{RR}_p(A) \leq \epsilon \quad \text{for all } p \in [0, 1]. \tag{3}$$

2.2 Overlapping buckets

We say that the buckets $\mathcal{J}$ are *overlapping* if for all $p \in (0, 1)$ there exists $J \in \mathcal{J}$ such that p is contained in the interior of J. The following theorem shows that overlapping buckets are both a necessary and sufficient condition for a finite time algorithm A satisfying (3) to exist, where the effort is measured in terms of the stopping time τ_A.

Theorem 1. *The following statements are equivalent:*

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

All proofs can be found in Appendix A. A consequence of the theorem is that there is no algorithm A with finite expected effort (i.e., $\mathbb{E}_p(\tau_A) < \infty$ for all $p \in [0, 1]$) that achieves (3) for $\mathcal{J}^0$.

To turn the classical buckets $\mathcal{J}^0$ into a set of overlapping p-value buckets, we can add intervals that contain the classical thresholds in their interior. As a specific choice, we recommend

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

We will use $\mathcal{J}^*$ throughout the article and refer to them as the *extended buckets*. We recommend $\mathcal{J}^*$ for three reasons: First, this choice results in roughly an equal maximal effort when p is close to all three classical thresholds (see Example 2). Second, the maximal effort and the expected effort under the null are reasonable in practical applications. Third, the interval limits in $\mathcal{J}^*$ have only few decimal places and can thus be easily written down. Section 5 discusses additional (heuristic) ways of choosing buckets.

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 manual of the American Psychological Association (American Psychological Association, 2010, page 139), is the de facto standard for reporting significance.

We propose to extend the star rating system for the overlapping buckets in $\mathcal{J}^*$ in the manner given in Table 1 (referred to as the *extended star rating system*). The same coding could be used for other p-value buckets that contain $\mathcal{J}^0$. If the p-value bucket I returned by our algorithm allows for a clear decision with respect to the classical thresholds (first row of Table 1), we report the classical star rating. Otherwise, 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 (second row of Table 1).

For instance, suppose an algorithm returns the bucket $I = (0.05\%, 0.2\%]$ for p upon stopping. This implies $p \leq 1\%$ and thus we can safely report a ** significance.

However, as p could either be smaller or larger than the next classical threshold 0.1%, we report **~ to indicate the possibility of a higher significance.

2.4 The general construction

We suppose that we can compute a confidence sequence $C(X_{1:n})$, $n \in \mathbb{N}$, for p, i.e. a sequence of intervals $C(X_{1:n})$ such that its joint coverage probability is at least $1-\epsilon$, where $\epsilon > 0$ is the desired uniform bound on the resampling risk. Formally, we require

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

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

The generic algorithm we propose will depend on the choice of p-value buckets $\mathcal{J}$ and the method C for computing a confidence sequence. We will denote the algorithm by $A(\mathcal{J}, C)$. We define the stopping time

$$\tau_{A(\mathcal{J},C)} = \inf\{n \in \mathbb{N} : \text{there exists } J \in \mathcal{J} \text{ such that } C(X_{1:n}) \subseteq J\} \tag{5}$$

which denotes the minimal number of samples n needed until a confidence interval $C(X_{1:n})$ is fully contained in a bucket $J \in \mathcal{J}$. If $\tau_{A(\mathcal{J},C)} < \infty$, the result of our algorithm is a bucket $I \in \mathcal{J}$ such that $C(X_{1:n}) \subseteq I$. If multiple buckets exist with this property then an arbitrary one is chosen. If $\tau_{A(\mathcal{J},C)} = \infty$, our algorithm returns an arbitrary element $I \in \mathcal{J}$ such that $\lim_{n\to\infty} \frac{S_n}{n} \in I$, where $S_n = \sum_{i=1}^n X_i$. The limit exists by the law of large numbers.

Theorem 2. *If* (4) *holds then* $A = A(\mathcal{J}, C)$ *satisfies* (3).

This is an immediate consequence of the construction and the strong law of large numbers.

The confidence interval $C(X_{1:n})$ for p and the bucket $I \in \mathcal{J}$ that our algorithm returns are related but not equivalent. Following Boos & Stefanski (2011), we are ultimately only interested in reporting one of the pre-specified p-value buckets that p falls in; a more precise confidence statement on p is not required. The confidence interval $C(X_{1:n})$ for p serves to quantify the uncertainty in the estimation of p, and since $C(X_{1:n}) \subseteq I$ it ensures that the bucket I we report satisfies (3).

Lastly, if there exists $N \in \mathbb{N}$ such that $\tau_{A(\mathcal{J},C)} < N$, we can relax (4) to

$$\mathbb{P}_p(p \in C(X_{1:n}) \text{ for all } n < N) \geq 1-\epsilon \quad \text{for all } p \in [0,1]. \tag{6}$$

Example 1. *Suppose we are solely interested in the 5% threshold. Testing at 5% corresponds to the two classical buckets* $\mathcal{J}^e = \{[0, 0.05], (0.05, 1]\}$. *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 displayed in Figure 1 (left). We define the non-stopping region as the region in which sampling progresses until the sampling path (n, S_n) hits either its lower or upper boundary. 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 $\mathcal{J}^e$ 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 $\mathcal{J}^*$. The stopping region is infinite for the non-overlapping $\mathcal{J}^0$ and finite for the overlapping buckets $\mathcal{J}^*$.*

How likely is it to observe the different decisions which can occur 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.*

3 Construction of confidence sequences

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

3.1 The Robbins-Lai approach

Robbins (1970) showed that the sequence of sets

$$C_{\mathrm{RL}}(X_{1:n}) = \{p \in [0,1] : (n+1)b(n, p, S_n) > \epsilon\}$$

satisfies (4), where $b(n, p, s) = \binom{n}{s} p^s (1-p)^{n-s}$ (see eq. (14)). Lai (1976) showed that $C_{\mathrm{RL}}(X_{1:n})$ are intervals. Using these intervals with overlapping buckets leads to a bounded effort:

Lemma 1. *If $\mathcal{J}$ are overlapping buckets then the stopping time $\tau_{A(\mathcal{J},C_{RL})}$ can be bounded by a deterministic positive constant.*

The intervals $C_{\mathrm{RL}}(X_{1:n})$ need not be computed explicitly in order to check (5). Appendix B gives a simple criterion to check if $C_{\mathrm{RL}}(X_{1:n}) \subseteq J$ for $J \in \mathcal{J}$.

3.2 The Simctest approach

Gandy (2009) provides a method to compute a decision for H_0 with respect to a single threshold in the same Monte Carlo setting as the one of Section 2.1. This approach can also be used to construct confidence sequences for multiple thresholds.

For the purposes of this article, it suffices to mention that for a given threshold $\alpha \in [0,1]$, Gandy (2009) constructs two integer valued stopping boundaries $(L_{n,\alpha})_{n\in\mathbb{N}}$ and $(U_{n,\alpha})_{n\in\mathbb{N}}$, and defines a stopping time

$$\tau_\alpha = \inf\{k \in \mathbb{N} : S_k \geq U_{k,\alpha} \text{ or } S_k \leq L_{k,\alpha}\}.$$

The construction is parametrized by a spending sequence $(\epsilon_n)_{n\in\mathbb{N}}$ that is nonnegative, nondecreasing and converges to some $0 < \rho < 1$. (Gandy, 2009, Theorem 1) shows that, under conditions, $\mathbb{E}_p(\tau_\alpha) < \infty$ for $p \neq \alpha$ and that the probability of hitting the wrong boundary is bounded by ρ, i.e. $\mathbb{P}_p(S_{\tau_\alpha} \geq U_{\tau_\alpha,\alpha}) < \rho$ for $p < \alpha$, and similarly for $p > \alpha$.

In order to extend this approach to multiple thresholds, we first define the set of boundaries of intervals in $\mathcal{J}$ that are in the interior of $[0,1]$:

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

Then, for each $\alpha \in B_{\mathcal{J}}$ we construct the stopping boundaries $L_{n,\alpha}$ and $U_{n,\alpha}$ using the same ρ. We define

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

We define the confidence sequence of the Simctest approach as $C_{\mathrm{S}}(X_{1:n}) = \bigcap_{\alpha\in B_{\mathcal{J}}} I_{n,\alpha}$.

The following theorem shows that $C_{\mathrm{S}}(X_{1:n})$ has the desired joint coverage probability given in (4) (or (6) for overlapping buckets) when setting $\rho = \epsilon/2$. Moreover, the theorem shows that the algorithm $A(\mathcal{J}, C_{\mathrm{S}})$ has a bounded stopping time if $\mathcal{J}$ is a finite set of overlapping buckets.

Theorem 3. *Let $\epsilon \in (0,1)$. For each $\alpha \in B_{\mathcal{J}}$, construct $L_{n,\alpha}$ and $U_{n,\alpha}$ with error probability $\rho = \epsilon/2$. 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 $\alpha, \alpha' \in B_{\mathcal{J}}$, $\alpha < \alpha'$, *and* $n < N$.

1. *Then* $\mathbb{P}_p(p \in C_S(X_{1:n})$ *for all* $n < N) \geq 1 - \epsilon \quad$ *for all* $p \in [0, 1]$.

2. *Suppose* $N = \infty$, $\rho \leq 1/4$ *and* $\log(\epsilon_n - \epsilon_{n-1}) = o(n)$ *as* $n \to \infty$. *If* $\mathcal{J}$ *is a finite set of overlapping p-value buckets then there exists* $c < \infty$ *such that* $\tau_{A(\mathcal{J},C_S)} \leq c$.

Allowing $N < \infty$ in Theorem 3 is useful for stopping boundaries constructed to yield a finite runtime (see (6)).

The condition on the spending sequence in part 2 of Theorem 3 is identical to the condition imposed in Theorem 1 of Gandy (2009). It is satisfied by the *default spending sequence* defined in Gandy (2009) as $\epsilon_n = \rho n/(n + k)$ with $k = 1000$, which is also employed in the remainder of this article.

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 $\alpha, \alpha' \in B_{\mathcal{J}}$ with $\alpha < \alpha'$) can be checked for a fixed spending sequence $(\epsilon_n)_{n\in\mathbb{N}}$ 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 (13) given in the proof of Lemma 2 in Appendix A.

Lemma 2. *Suppose* $\rho \leq 1/4$ *and* $\log(\epsilon_n - \epsilon_{n-1}) = o(n)$ *as* $n \to \infty$. *Let* $\alpha, \alpha' \in B_{\mathcal{J}}$ *with* $\alpha < \alpha'$. *Then there exists* $n_0 \in \mathbb{N}$ *such that for all* $n \geq n_0$,

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

For $n < n_0$, the inequalities again have to be checked manually.

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 us 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)).

Theorem 4. *Let τ be the number of steps taken by a sequential procedure returning $I \in \mathcal{J}$ which respects* (3)*. Then, for every $\tilde{p} \in [0,1]$,*

$$\mathbb{E}_{\tilde{p}}(\tau) \geq \sup_{q \notin \tilde{J}} e(\tilde{p}, q, \epsilon, \epsilon), \tag{7}$$

where $\tilde{J} = \bigcup_{J \in \mathcal{J}, \tilde{p} \in J} J$ is the union of all buckets containing $\tilde{p}$ and

$$e(p, q, \alpha, \beta) = \frac{(1-\alpha)\log(\beta/(1-\alpha)) + \alpha \log((1-\beta)/\alpha)}{p \log(q/p) + (1-p)\log((1-q)/(1-p))}.$$

Furthermore, if $\tilde{p} \in [0,1]$ is such that exactly two elements of $\mathcal{J}$ contain $\tilde{p}$, say J_1 and J_2, then

$$\mathbb{E}_{\tilde{p}}(\tau) \geq \min_{\eta \in [0,1]} \max \left\{ \sup_{q \notin J_1} e(\tilde{p}, q, 1-\eta, \epsilon), \sup_{q \notin J_2} e(\tilde{p}, q, \min(\eta+\epsilon, 1), \epsilon) \right\}. \tag{8}$$

We call the bound given by (7) the basic lower bound and the bound given by the maximum of (7) and (8) the improved lower bound. The suprema in (7) and (8) 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 by conservatively replacing $e(\tilde{p}, q, \eta+\epsilon, \epsilon)$ by $e(\tilde{p}, q, \eta+\epsilon+\delta, \epsilon)$, where δ is the grid width. This is because e is decreasing in its third argument.

Figure 4 gives an example of both the basic and the improved lower bounds on $\mathbb{E}_{\tilde{p}}(\tau)$ for the extended 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 classical buckets $\mathcal{J}$ as well as the extended 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 $\mathcal{J}^*$ (right) 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 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. For both the Robbins-Lai and the Simctest approach, the files *RL.cpp* and *simctest.cpp* included in the Supplementary Material contain an implementation that computes the effort for a fixed set of p-value buckets.

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 $\mathcal{J}$ or $\mathcal{J}^*$ 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. Using Simctest, the effort of our algorithm of Section 2.4 differs from the theoretical lower bound by only a small factor.

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 a fixed p (see Figure 5, right) with respect to certain p-value distributions.

Here, we consider using the extended 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.

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 for each hypothesis.

We assume we want to test $n = 10^4$ hypotheses using the Bonferroni (1936) correction to correct for multiplicity. In order to be able to compute numbers of false classifications, we assign $n_{\text{alt}} = 100$ hypotheses to the alternative, the remaining $n - n_{\text{alt}} = 9900$ 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 uniformly in $[0, 1]$.

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

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

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

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

We now report the results from a single run of this setup. Our algorithm draws $N = 3.2 \times 10^5$ 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 & Hinkley, 1997). Due to this pseudo-count, this naïve approach is incapable of observing p-values below $(N+1)^{-1} = 3.125 \times 10^{-6}$ (see also Gandy & Hahn (2017)), and in particular incapable of observing any p-values in the two lowest buckets.

5 Discussion

The overlapping p-value buckets presented in Section 2.2 were chosen to be easily written down and to yield an equal maximal effort for all classical thresholds as well as a reasonable expected effort. However, these criteria are essentially arbitrary. A variety of further (heuristic) criteria can be used to obtain overlapping buckets from traditional testing thresholds $T = \{t_0, \ldots, t_m\}$. These include:

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

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

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 $S_n \in \{0, \ldots, n\}$) 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 of the resulting overlapping buckets.

The article leaves scope for a variety of future research directions. For instance, how can the 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) is capable of meeting the effort of the lower bound?

Acknowledgements

The second author was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. The third author was funded by a Presidents Ph.D. Scholarship of Imperial College London.

Supplementary material

The Supplementary Material includes *R* code (R Development Core Team, 2008) to implement the algorithms as well as to reproduce all figures and tables.

References

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

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

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

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

Besag, J. & Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika* **78**, 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.

Boos, D. & Stefanski, L. (2011). P-Value Precision and Reproducibility. *The American Statistician* **65**, 213–221.

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

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

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

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

Ding, D., Gandy, A. & Hahn, G. (2016). A simple method for implementing monte carlo tests. *arXiv:1611.01675* .

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

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

Gandy, A. & Hahn, G. (2014). MMCTest – A Safe Algorithm for Implementing Multiple Monte Carlo Tests. *Scand J Stat* **41**, 1083–1101.

Gandy, A. & Hahn, G. (2017). QuickMMCTest: quick multiple Monte Carlo testing. *Stat Comput* **27**, 823–832.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J Am Stat Assoc* **58**, 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**, 1834–1843.

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

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

Lourenco, V. & 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**, 1397–1409.

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

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

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

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

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

Address of corresponding author

Georg Hahn, Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, 677 Huntington Ave, Boston, MA 02115, USA.
E-mail: ghahn@hsph.harvard.edu

Appendix

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 $\alpha \in (0,1)$ which is not contained in the interior of any $J \in \mathcal{J}$. Let $I \in \mathcal{J}$ be the (random) interval reported by algorithm A which satisfies (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, eq. (4.81)), see also (Gandy, 2009, section 3.1), we have

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

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

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

Let $a_0 < a_1 < \cdots < a_k$ be the ordered boundaries of the 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 & Pearson (1934) confidence interval with coverage probability $1-\epsilon$ for p, where $n \in \mathbb{N}$ is the number of samples and S is the number of exceedances observed among those n samples. Let n be such that the length of all Clopper-Pearson intervals is less than Δ, i.e. $n = \min\{m \in \mathbb{N} : |I(S,m)| < \Delta \text{ for all } S \in \{0, \ldots, m\}\}$. 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 & Hahn (2014)).

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.) $\Rightarrow$ (1.): Since finite effort implies expected finite effort, (1.) follows immediately. □

Proof of Lemma 1. We first prove that the length of $C_{\mathrm{RL}}(X_{1:n})$ uniformly goes to zero. The bounded stopping time then follows after proving that once an interval is below a certain length, it is guaranteed to be contained in one of the buckets.

If $0 \le p \le S_n/n - [\log((n+1)/\epsilon)/(2n)]^{1/2}$ then, by Hoeffding's inequality (Hoeffding, 1963),

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

where $X \sim \mathrm{Binomial}(n,p)$. Hence, $p \notin C_{\mathrm{RL}}(X_{1:n})$. A similar argument shows that $b(n,p,S_n) \le \epsilon/(n+1)$ for $S_n/n + [\log((n+1)/\epsilon)/(2n)]^{1/2} \le p \le 1$. Thus, $|C_{\mathrm{RL}}(X_{1:n})| \le [2\log((n+1)/\epsilon)/n]^{1/2}$.

Now assume no $c > 0$ exists such that any interval $I \subseteq [0,1]$ with length less than c is contained in a $J \in \mathcal{J}$. Then for all $n \in \mathbb{N}$ there exists an interval $C_{\mathrm{RL}}(X_{1:n}) \subset [0,1]$ with $0 < |C_{\mathrm{RL}}(X_{1:n})| < 1/n$ such that $C_{\mathrm{RL}}(X_{1:n}) \not\subseteq J$ for all $J \in \mathcal{J}$. Let a_n be the mid point of $C_{\mathrm{RL}}(X_{1:n})$. 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 $J \in \mathcal{J}$ such that $(b-\epsilon, b+\epsilon) \subseteq J$. For large enough k we have $C_{\mathrm{RL}}(X_{1:n_k}) \subseteq (b-\epsilon, b+\epsilon)$, contradicting $C_{\mathrm{RL}}(X_{1:n_k}) \not\subseteq J$.

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

Proof of Theorem 3. 1. For threshold $\alpha \in B_{\mathcal{J}}$, let $\overline{E}_\alpha^N = \{S_{\tau_\alpha} \ge U_{\tau_\alpha,\alpha}, \tau_\alpha < N\}$ be the event that the upper boundary is hit first before time N and likewise let $\underline{E}_\alpha^N = \{S_{\tau_\alpha} \le L_{\tau_\alpha,\alpha}, \tau_\alpha < N\}$ be the event that the lower boundary is hit first. Then, for all $\alpha, \alpha' \in B_{\mathcal{J}}$ with $\alpha < \alpha'$,

$$\overline{E}_\alpha^N \supseteq \overline{E}_{\alpha'}^N \quad \text{and} \quad \underline{E}_\alpha^N \subseteq \underline{E}_{\alpha'}^N. \tag{9}$$

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 $U_{n,\alpha} \le U_{n,\alpha'}$ 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 $\overline{E}^N_{\alpha'}$. Hence, we have $\overline{E}^N_\alpha \supseteq \overline{E}^N_{\alpha'}$. The proof of $\underline{E}^N_\alpha \subseteq \underline{E}^N_{\alpha'}$ is similar.

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

$$\begin{aligned}\mathbb{P}_p(\text{there exists } n < N : p \notin C_{\mathrm{S}}(X_{1:n})) &\leq \mathbb{P}_p(\text{there exist } n < N, \alpha \in B_{\mathcal{J}} : p \notin I_{n,\alpha}) \\ &= \mathbb{P}_p\left(\bigcup_{\alpha \in B_{\mathcal{J}} : \alpha \leq p} \underline{E}^N_\alpha \cup \bigcup_{\alpha \in B_{\mathcal{J}} : \alpha \geq p} \overline{E}^N_\alpha\right) \\ &\leq \mathbb{P}_p\left(\bigcup_{\alpha \in B_{\mathcal{J}} : \alpha \leq p} \underline{E}^N_\alpha\right) + \mathbb{P}_p\left(\bigcup_{\alpha \in B_{\mathcal{J}} : \alpha \geq p} \overline{E}^N_\alpha\right). \end{aligned} \tag{10}$$

If $p < \min B_{\mathcal{J}}$, the first term is equal to 0. Otherwise, let $\alpha' = \max\{\alpha \in B_{\mathcal{J}} : \alpha \leq p\}$. Then, by (9),

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

The second term on the right hand side of (10) can be dealt with similarly.

2. By (12) and as $\Delta_n = o(n)$ there exists $n_0 \in \mathbb{N}$ such that

$$|\{\alpha \in B_{\mathcal{J}} : \tau_\alpha > n_0\}| \leq 1. \tag{11}$$

We will show that $\tau_{A(\mathcal{J}, C_{\mathrm{S}})} \leq n_0$. First, the assumption on the ordering of L_n and U_n excludes the possibility that $C_{\mathrm{S}}(X_{1:n_0}) = \emptyset$. Second, (11) implies $|C_{\mathrm{S}}(X_{1:n_0}) \cap B_{\mathcal{J}}| \leq 1$.

If $|C_{\mathrm{S}}(X_{1:n_0}) \cap B_{\mathcal{J}}| = 1$ then let $\alpha \in B_{\mathcal{J}}$ be such that $\alpha \in C_{\mathrm{S}}(X_{1:n_0})$. 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 $C_{\mathrm{S}}(X_{1:n_0}) \subseteq J$ due to $|C_{\mathrm{S}}(X_{1:n_0}) \cap B_{\mathcal{J}}| = 1$, thus showing $\tau_{A(\mathcal{J}, C_{\mathrm{S}})} \leq n_0$.

If $|C_{\mathrm{S}}(X_{1:n_0}) \cap B_{\mathcal{J}}| = 0$ then let β be in the interior of $C_{\mathrm{S}}(X_{1:n_0})$. As $\mathcal{J}$ is overlapping, there exists $J \in \mathcal{J}$ such that $\beta \in J$. As $C_{\mathrm{S}}(X_{1:n_0}) \cap B_{\mathcal{J}} = \emptyset$ this implies $C_{\mathrm{S}}(X_{1:n_0}) \subseteq J$, thus showing $\tau_{A(\mathcal{J}, C_{\mathrm{S}})} \leq n_0$. □

Proof of Lemma 2. By arguments in (Gandy, 2009, Proof of Theorem 1), we have

$$\frac{U_{n,\alpha} - n\alpha}{n} \leq \frac{\Delta_n + 1}{n} \to 0, \qquad \frac{L_{n,\alpha'} - n\alpha'}{n} \geq -\frac{\Delta_n + 1}{n} \to 0, \tag{12}$$

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

$$2\left(\frac{\Delta_n}{n} + \frac{1}{n}\right) \leq \alpha' - \alpha \text{ for all } n \geq n_0. \tag{13}$$

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 (12).

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 $L_{n,\alpha} \leq L_{n,\alpha'}$ and $U_{n,\alpha} \leq U_{n,\alpha'}$ for all $n \geq n_0$ as desired. □

Proof of Theorem 4. We suppose that $I \in \mathcal{J}$ is the (random) bucket reported by a sequential algorithm that respects (3). Let $\tilde{p} \in [0,1]$. For any $q \in [0,1] \setminus \tilde{J}$, we can consider the hypotheses $H_0 : p = \tilde{p}$ against $H_1 : p = q$ and the test that rejects H_0 if and only if $\tilde{p} \notin I$. By (3), the type I error of such a test is at most ϵ. Also, the type II error is at most ϵ, as $q \notin \tilde{J}$ implies $\mathbb{P}_q(\tilde{p} \in I) \leq \mathbb{P}_q(q \notin I) \leq \epsilon$. Hence, using the lower bound in (Wald, 1945, eq. (4.80)), we get (7).

To see (8): For any $q \in [0,1] \setminus J_1$ consider the hypotheses $H_0 : p = \tilde{p}$ and $H_1 : p = q$ and the test that rejects H_0 if and only if $I \neq J_1$. This test has type I error $1-\eta$, where $\eta = \mathbb{P}_{\tilde{p}}(I = J_1)$, and type II error of at most ϵ. Using (Wald, 1945, eq. (4.80)) we get $\mathbb{E}_{\tilde{p}}(\tau) \geq e(\tilde{p}, q, 1-\eta, \epsilon)$. Similarly, for any $q \in [0,1] \setminus J_2$, we can test the hypotheses $H_0 : p = \tilde{p}$ and $H_1 : p = q$ by rejecting H_0 if and only if $I \neq J_2$. This test has type I error of at most $\min(\eta+\epsilon, 1)$ and type II error of at most ϵ. Again, using (Wald, 1945, eq. (4.80)) we get $\mathbb{E}_{\tilde{p}}(\tau) \geq e(\tilde{p}, q, \min(\eta+\epsilon, 1), \epsilon)$. Eq. (8) follows as these inequalities hold for all q and due to the fact that we can account for the unknown η by minimizing over it. □

B A simple stopping criterion for Robbins-Lai

The following describes a simple criterion to determine whether a confidence interval computed via the Robbins-Lai approach of Section 3.1 is fully contained in a bucket. For a single threshold this approach has been suggested in Ding et al. (2016). Let interval $C_{\mathrm{RL}}(X_{1:n})$ and bucket $J \in \mathcal{J}$ as well as n, S_n and ϵ be as in Sections 2.4 and 3.1. Then $C_{\mathrm{RL}}(X_{1:n}) \subseteq J$ if and only if for $p \in \{\min J, \max J\}$,

$$(n+1)b(n, S_n, p) = (n+1)\binom{n}{S_n} p^{S_n}(1-p)^{n-S_n} \leq \epsilon. \tag{14}$$

As (14) is also satisfied if $C_{\mathrm{RL}}(X_{1:n})$ 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.

After applying a (monotonic) log transformation to (14), taking the derivative with respect to p yields

$$\frac{S_n}{p} - \frac{n - S_n}{1-p} \begin{cases} \geq 0 & p = \min J, \\ \leq 0 & p = \max J. \end{cases} \tag{15}$$

If (14) and (15) are satisfied, then $C_{\mathrm{RL}}(X_{1:n}) \subseteq J$.

Bucket	$[0, 0.1\%]$	$(0.1\%, 1\%]$	$(1\%, 5\%]$	$(5\%, 1]$
Code	***	**	*	
Bucket	$(0.05\%, 0.2\%]$	$(0.8\%, 1.2\%]$	$(4.5\%, 5.5\%]$	
Code	**~	*~	~	

Table 1: Extended star rating system for $\mathcal{J}^*$.

	Robbins-Lai	Simctest	Lower bound
H_0	2228	1853	975
H_{1a}	16878	13837	7126
H_{1b}	40059	30896	15885

Table 2: Expected (integrated) effort for both Robbins-Lai and Simctest applied to $\mathcal{J}^*$.

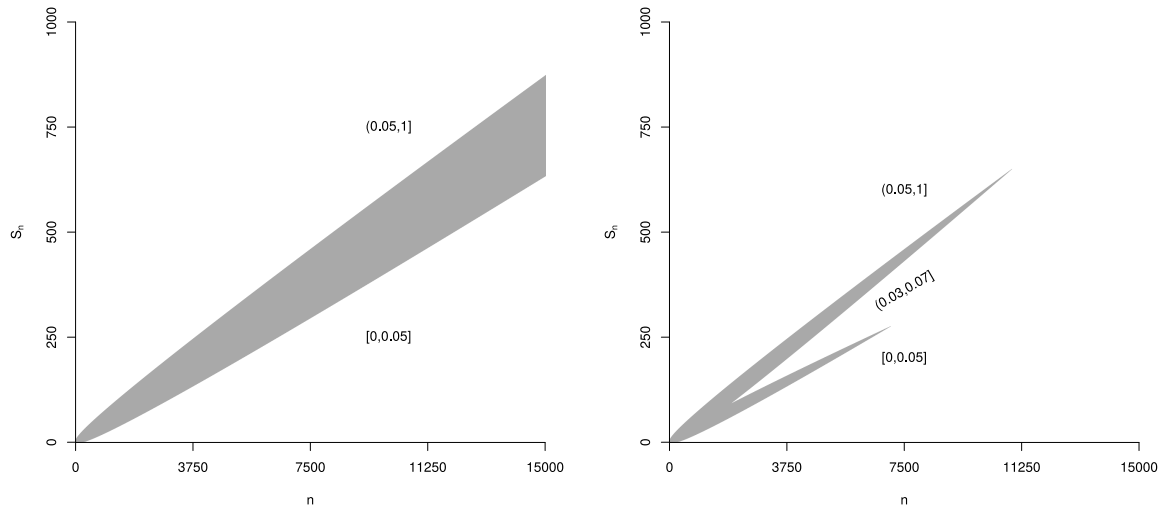

Figure 1: Non-stopping region (grey) to compute a decision on p with respect to J^e (left), which corresponds to a 5% threshold, and with respect to the overlapping buckets $J^e \cup \{(0.03, 0.07]\}$ (right).

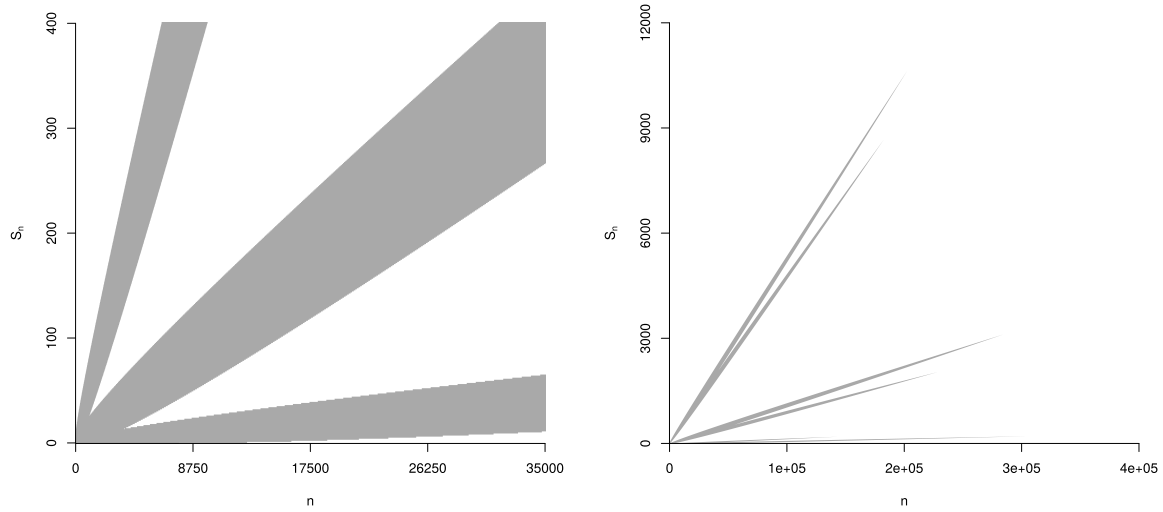

Figure 2: Non-stopping region (gray) for $\mathcal{J}^0$ (left) and $\mathcal{J}^*$ (right).

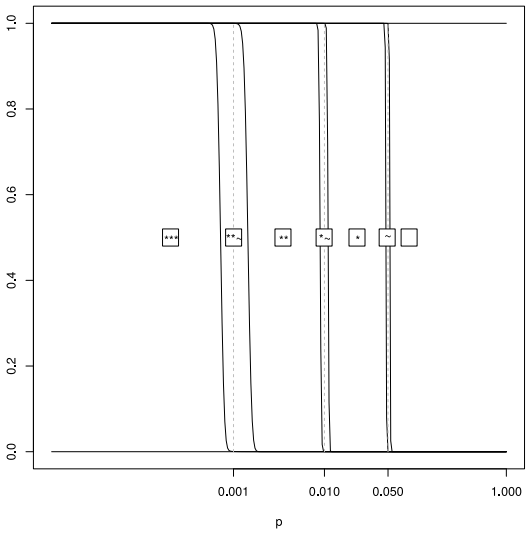

Figure 3: Probabilities of observing each possible decision for $\mathcal{J}^*$ as a function of p.

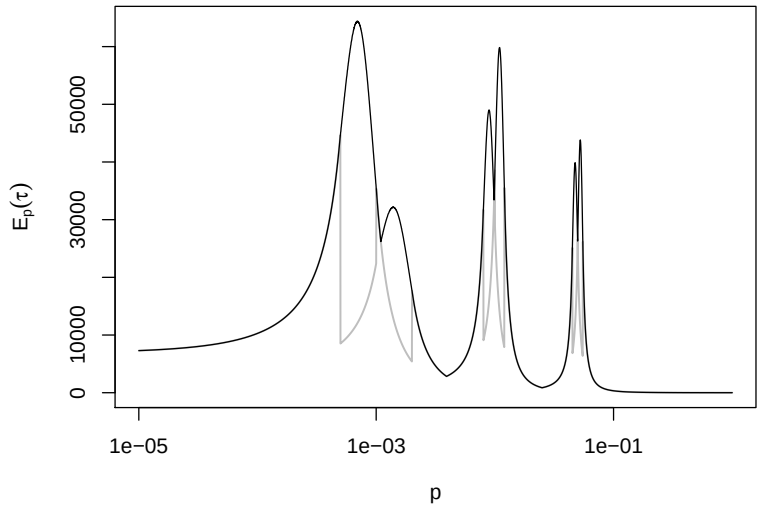

Figure 4: Basic (grey) and improved (black) lower bounds on the effort $\mathbb{E}_p(\tau)$ for $\mathcal{J}^*$.

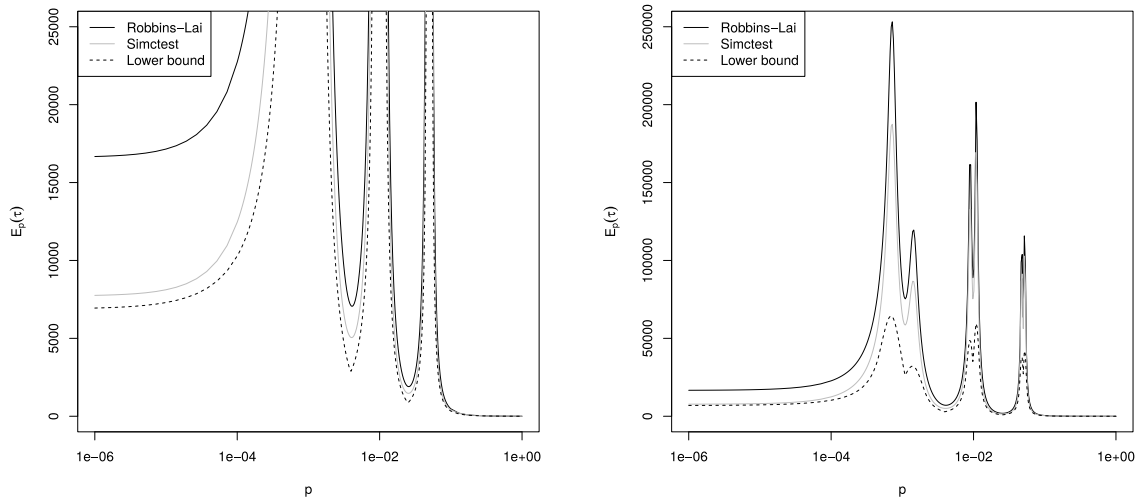

Figure 5: Expected effort to compute a decision with respect to $\mathcal{J}^0$ (left) and $\mathcal{J}^*$ (right) as a function of p. Confidence sequences computed with both Simctest (grey) and Robbins-Lai (black). Lower bound on the effort indicated with a dashed line.