On the expected runtime of multiple testing algorithms with bounded error

Georg Hahn

Abstract

Consider the testing of multiple hypotheses in the setting where the p-values of all hypotheses are unknown and thus have to be approximated using Monte Carlo simulations. One class of algorithms published in the literature for this scenario ensures guarantees on the correctness of their testing result (for instance, a bound on the resampling risk) through the computation of confidence statements on all approximated p-values. This article focuses on the expected runtime of those algorithms and shows the following four main results. Computing a decision on a single hypothesis tested at a fixed threshold requires an infinite expected runtime. In applications relying on the decisions of multiple hypotheses computed with a Bonferroni-type threshold, all but two hypotheses can be decided in finite expected runtime. This result does not extend to applications which require full knowledge of all individual decisions (for instance, step-up or step-down procedures), in which case no algorithm can guarantee even a single decision in finite expected runtime. Nevertheless, simulations show that in practice, the number of pending decisions typically remains low.

Keywords: algorithm, bounded error, computational effort, finite expected runtime, multiple testing

1 Introduction

Consider the testing of \( m \in \mathbb{N} \) hypotheses \( H_{01}, \ldots, H_{0m} \) in a scenario in which the p-values \( p_1, \ldots, p_m \) corresponding to the \( m \) hypotheses are unknown and thus have to be approximated using Monte Carlo simulations, for instance through bootstrap or permutation tests. Several algorithms published in the literature are designed for this scenario, either using a truncation rule to reach fast decisions (Besag and Clifford, 1991; Davidson and MacKinnon, 2000; Andrews and Buchinsky, 2000, 2003; van Wieringen et al., 2008; Sandve et al., 2011; Silva and Assunção, 2013, 2018) or using a heuristic approach to minimize the computational effort without truncation (Lin, 2005; Silva et al., 2009; Gandy and Hahn, 2017).

This article focuses on the expected runtime of methods which aim to provide a guarantee of correctness on the decision of each hypothesis through the computation of a sequence of confidence intervals on each p-value. Under a simple and weak asymptotic condition on the length of the intervals produced by the confidence sequence, the article shows the following four main results: (1.) Algorithms which rely on the decision (rejection or non-rejection) of a single hypothesis with respect to a fixed threshold have an infinite expected runtime (section 2). (2.) For applications relying on independent testing, that is multiple hypotheses tested at a constant (Bonferroni-type) threshold, all but two hypotheses can be decided in finite expected runtime (section 3). (3.) This result does not extend to applications which require full knowledge of all individual decisions, for instance step-up or step-down procedures, in which case no algorithm can guarantee even a single decision in finite expected runtime (section 4). (4.) In practice, however, the number of pending decisions is typically low as demonstrated in the conclusions (section 5).

Although unconsidered in their original publications, the expected runtimes derived in this article apply to, for instance, the algorithms of Guo and Peddada (2008) and Gandy and Hahn.
which provide a guarantee on the correctness of the decision on each hypothesis through the computation of Clopper and Pearson [1934] confidence intervals. Expected runtimes also hold true for the methods of Gandy and Hahn [2016] and Ding et al. [2018] which use the confidence sequences of Robbins [1970] and Lai [1976], as they do for the confidence sequences of Darling and Robbins [1967a,b] and the binomial confidence intervals of Armitage [1958] employed in Fay et al. [2007] and Gandy [2009]. The aforementioned algorithms are routinely used in many applications (Roth et al., 2016; Termenon et al., 2016; Jin et al., 2016; Reiss et al., 2015; Rubin-Delanchy et al., 2015), albeit in a rather heuristic fashion, therefore making it important to rigorously characterize their expected runtimes. The results of this article do not apply to the push-out design of Fay and Follmann (2002) and the B-value design of Kim (2010), which both achieve a bounded resampling risk without confidence statements on the p-values.

Supplementary material containing R code (R Development Core Team, 2011) to reproduce fig. 1 is provided.

2 A single decision requires infinite expected runtime

Following an argument similar to (Gandy, 2009, section 3.1), computing the decision of a single hypothesis \( H_{01} \) with random p-value \( p_1 \) requires an infinite expected runtime. Assume a sequential algorithm \( A \) tests \( H_{01} \) at some given threshold \( \alpha \) by approximating \( p_1 \) through the drawing of Monte Carlo samples and gives a guarantee of \( 1 - \epsilon \) on the correctness of its decision, where \( \epsilon \in (0, 1) \).

Computing a decision on \( H_{01} \) is equivalent to deciding whether \( p_1 \leq \alpha \) or \( p_1 > \alpha \). For some \( p_0 > \alpha \), consider testing \( H_0 : p_1 \leq \alpha \) against \( H_1 : p_1 = p_0 \). A test can be constructed by rejecting \( H_0 \) if and only if \( A \) does not reject \( H_{01} \). Due to the guarantee of algorithm \( A \), both the type 1 and type 2 errors of this test are \( \epsilon \). For such a sequential test, a lower bound on the runtime \( \tau \) (the expected number of steps) is given by (Wald, 1945, equation (4.81)) as

\[
E(\tau | p = p_1) \geq \frac{\epsilon \log \left( \frac{1}{1-p_1} \right) + (1 - \epsilon) \log \left( \frac{1-\epsilon}{\epsilon} \right)}{p_1 \log \left( \frac{p_1}{\alpha} \right) + (1 - p_1) \log \left( \frac{1-p_1}{1-\alpha} \right)}. \tag{1}
\]

The same bound (1) can be derived for the case \( p_0 \leq \alpha \). Abbreviate the numerator of (1) by \( C \) and consider \( p_1 \) in a Bayesian setup such that the following condition is satisfied.

**Condition 1.** Assume \( p_1 \) has a distribution function \( F(p_1) \) with derivative \( F'(\alpha) > 0 \), and that for a suitable \( \gamma > 0 \), there exists a constant \( d > 0 \) such that \( F'(p_1) \geq d \) in \((\alpha, \alpha + \gamma)\).

Amongst others, condition 1 is satisfied for the distributions of the exponential family. Under condition 1, \( E(\tau) \) can be bounded by

\[
E(\tau) = \int_0^1 E(\tau | p = p_1) dF(p_1) \geq \int_0^{\alpha+\gamma} E(\tau | p = p_1) dF(p_1) \\
\geq C \cdot d \cdot \int_\alpha^{\alpha+\gamma} \left( p \log \left( \frac{p}{\alpha} \right) + (1 - p) \log \left( \frac{1-p}{1-\alpha} \right) \right)^{-1} dp = \infty,
\]

as the integrand is proportional to \((p-\alpha)^{-2}\) as \( p \to \alpha \). This proves an infinite expected runtime for the sequential test of \( H_0 \) and, equivalently, for algorithm \( A \).

3 Bonferroni-type multiple testing in expected finite time

Assume the testing of \( H_{01}, \ldots, H_{0m} \) is carried out by comparing each \( p_i \) to a threshold value \( \alpha_i \in (0, 1) \), \( i \in \{1, \ldots, m\} \), as done in, for instance, step-up and step-down procedures (Gandy
For this, without loss of generality, assume that $p_1 \leq \cdots \leq p_m$ and $\alpha_1 \leq \cdots \leq \alpha_m$. In applications which rely on multiple testing at a constant (Bonferroni-type) threshold with a guarantee of correctness through confidence statements on the p-values, it will be shown that decisions on all but two hypotheses can be computed in expected finite time. Similarly to section 2, the guarantee of correctness on all decisions is assumed to hold simultaneously for all hypotheses at $1 - \epsilon$ for some pre-specified $\epsilon \in (0, 1)$.

To be precise, a stronger statement is proven. For all but two hypotheses, it can be decided in expected finite time which of the intervals

$$I = \{[\alpha_i, \alpha_{i+1}) : \alpha_i < \alpha_{i+1}, i = 0, \ldots, m - 1\} \cup \{[0, \alpha_1), [\alpha_m, 1]\}$$

their p-values fall into.

As the p-values are unknown, they are approximated through Monte Carlo samples, and confidence statements are provided via confidence intervals. Let $g(n)$ be the length of the confidence interval for a p-value after drawing $n$ (Monte Carlo) samples, and $\hat{p}_n$ be the maximum likelihood estimate of $p$ based on $n$ samples.

**Condition 2.** Any confidence interval for $p$ contains $\hat{p}_n$. Moreover, $g(n) \in o(n^\gamma)$ for some $-\frac{1}{2} < \gamma < -\frac{1}{3}$.

Define $D = \min_{I \in \mathcal{I}} d(p, \partial I)$ for a (random) $p$, where $d(p, \partial I)$ is the distance of $p$ to the boundary of $I$. If the distance of $p$ to $\hat{p}_n$ is less than $D/2$ and the confidence interval for $p$ has length $g(n) < D/2$, the confidence interval for $p$ will be entirely contained in some $I \in \mathcal{I}$ (assuming it contains $\hat{p}_n$ as required by condition 2). Therefore, a decision on which interval $I \in \mathcal{I}$ contains $p$ is obtained on reaching the stopping time $\tau = \inf\{n : |p_n - p| < D/2, g(n) < D/2\}$.

Let $\tau_1, \ldots, \tau_m$ be the stopping times of the $m$ p-values and $\tau_{(1)} \leq \cdots \leq \tau_{(m)}$ be their order statistic.

**Theorem 1.** Let the density of $p$ be bounded above by some finite constant. Assume $m \geq 3$. Under condition 3, $\mathbb{E}(\tau_{(m-s)}) < \infty$ for $2 \leq s < m$.

The proof of theorem 1 is found in appendix B. For the Bonferroni (1936) correction, determining intervals $I \in \mathcal{I}$ containing the confidence interval of each hypothesis is equivalent to determining if the p-value of a hypothesis is above or below the constant testing threshold (subject to the overall $1 - \epsilon$ error probability), thus giving a decision on all but two hypotheses in finite expected time by theorem 1. This result does not extend to multiple testing applications which depend on all individual decisions, for instance step-up or step-down procedures, in which case no algorithm can guarantee even a single decision in finite expected time (section 4).

Condition 2 is satisfied for a variety of confidence sequences and the algorithms they use.

1. The length of Clopper and Pearson (1934) confidence intervals is $g(n) \propto n^{-1/2}(-\log \rho_n)^{1/2}$ (see lemma 1 in appendix A), where $(\rho_n)_{n \in \mathbb{N}}$ is a non-negative sequence controlling how the overall error $\epsilon$ is spent (that is, $\sum_{n=1}^\infty \rho_n = \epsilon$). If $-\log(\rho_n) \propto \log(n)$ as in Gandy (2009) and Gandy and Hahn (2014), then $g(n) \in o(n^\gamma)$ for any $-\frac{1}{2} < \gamma < -\frac{1}{3}$. Since the Clopper and Pearson (1934) intervals contain $\hat{p}_n$, condition 2 is satisfied. The intervals are employed in the algorithms of Guo and Peddada (2008) and Gandy and Hahn (2014).

2. Confidence intervals produced by the binomial confidence sequences of Robbins (1970) and Lai (1976) satisfy $g(n) \propto n^{-1/2}\{\log(n \log n)\}^{1/2}$ (see lemma 2 in appendix A). In Lai (1976, section 3(A)) it is shown that the intervals contain $\hat{p}_n$, thus satisfying condition 2. These confidence sequences are employed in the methods of Gandy and Hahn (2016) and Ding et al. (2018).
3. Intervals produced by the confidence sequences of [Darling and Robbins (1967a,b)] have length $g(n) \propto n^{-1/2} \{\log(\log n)\}^{1/2}$ [Darling and Robbins (1967a), section 1] and are based around the empirical mean, thus satisfying condition $2$.

4. Binomial confidence intervals of [Armitage (1958)] are employed in the methods of [Fay et al. (2007)] and [Gandy (2009)]. Being binomial exact intervals as the ones of [Clopper and Pearson (1934)], they satisfy condition $2$.

4 Extension to infinite expected runtime for multiple testing

For multiple testing applications having the property that with non-zero probability, no decision on any hypothesis can be made until it is known whether a certain hypothesis is rejected or non-rejected (as it is the case, for instance, for step-up and step-down procedures), under reasonable assumptions on the distribution of $p$-values, an algorithm with both a guarantee of correctness and a finite expected runtime for any number of decisions cannot exist.

Consider testing the $m$ hypotheses $H_{01}, \ldots, H_{0m}$ under the following condition.

**Condition 3.** The $p$-values $p = (p_1, \ldots, p_m)$ corresponding to $H_{01}, \ldots, H_{0m}$ have a joint distribution with support $[0, 1]^m$ and multivariate density $f_p(p)$. For all $\delta > 0$ there exists a $\kappa > 0$ such that $f_p(p) > \kappa$ on $[\delta, 1 - \delta]^m$.

This condition is satisfied for many common densities. Assume testing is carried out with a step-up procedure that compares each $p_i$ to a threshold value $\alpha_i$, $i \in \{1, \ldots, m\}$, where $\alpha_1 \leq \cdots \leq \alpha_{m-1} < \alpha_m$ (cf. section $3$). For any $0 < \eta < (\alpha_m - \alpha_{m-1})/4$, define $A = [\alpha_{m-1} + 2\eta, \alpha_m - 2\eta]^{m-1} \times [\alpha_m - \eta, \alpha_m + \eta]$. Let $B = [\delta_B, 1 - \delta_B]^m$ and choose $\delta_B \in (0, 1)$ in such a way that $A \cap B \neq \emptyset$.

Assuming the distribution of $p = (p_1, \ldots, p_m)$ satisfies condition $3$ with $\delta = \delta_B$, draw a vector $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_m)$ from $f_p$ conditional on being in $A$. By definition of $A$, $\tilde{p}_m$ is the largest value in $\tilde{p}$ and will thus be compared to the threshold value $\alpha_m$. By properties of a step-up procedure, as $\alpha_{m-1} < \tilde{p}_i < \tilde{p}_m$ for all $i \neq m$, no decision on any hypothesis can be made unless it is known whether $\tilde{p}_m$ lies below or above $\alpha_m$. In the former case, all hypotheses are rejected, in the latter case, all hypotheses are non-rejected. Under the additional assumption that the marginal distribution of $p_m$ satisfies condition $1$ by section $2$ deciding $H_{0m}$ requires an infinite expected time. Thus on the event $\{p \in A\}$, the expected time to decide any number of hypotheses is also infinite, $\mathbb{E}(\tau|p \in A) = \infty$, where $\tau$ denotes the number of samples.

By the law of total expectation, the unconditional expected time can be bounded as $\mathbb{E}(\tau) \geq \mathbb{E}(\tau|p \in A) \cdot \mathbb{P}(p \in A) = \infty$, where it was used that $\mathbb{E}(\tau|p \in A) = \infty$ and that $f_p(p) > \kappa > 0$ on $B$ implies $\mathbb{P}(p \in A) \geq \mathbb{P}(p \in A \cap B) > 0$.

The above consideration proves an infinite expected runtime to obtain any number of decisions for the widely used class of step-up procedures ([Simes, 1986] [Hochberg, 1988] [Rom, 1990] [Benjamini and Hochberg, 1995] [Benjamini and Yekutieli, 2001]). A similar construction proves the same result for step-down procedures ([Sidak, 1967] [Holm, 1979] [Shaffer, 1986]). Extensions to other testing applications in which obtaining any decision can be made dependent on the decision of a single hypothesis are possible but application-specific.

5 Conclusions

Theorem $1$ shows that the expected time to determine intervals $I \in I$ (defined in $(2)$) containing the $p$-values of all but two hypotheses is finite. For independent (Bonferroni-type) testing, this immediately implies at most two pending decisions in finite expected time.

For step-up and step-down procedures, the situation is more involved. This is because the individual decision on a hypothesis need not matter in step-up (step-down) procedures so long as
Figure 1: Survival function of the number of undecided hypotheses tested with the Benjamini and Hochberg (1995) step-up procedure with threshold $\alpha = 0.1$. $R = 10^4$ repetitions. Number of hypotheses $m$ ranging from $10^1$ (solid), $10^2$ (dashed), $10^3$ (dotted), to $10^4$ (dash-dotted).

there exists another hypothesis with a larger (smaller) p-value for which a decision is available. As a consequence, the rejection or non-rejection of the two remaining (undecided) hypotheses in the sense of theorem 1 may well be determined by the other hypotheses whose decision can be computed in expected finite time, meaning that theorem 1 is not always representative of the actual number of decisions. Although in section 4 the same observation was used to show that under conditions, the expected time to compute any decision can be infinite, in practice the decisions on a large number of hypotheses can often be computed quickly. For the Benjamini and Hochberg (1995) step-up procedure with threshold $\alpha = 0.1$, fig. 1 displays the survival function of the number of undecided hypotheses which remain if, out of $m \in \{10^1, 10^2, 10^3, 10^4\}$ hypotheses, for only $m - 2$ hypotheses it can be decided which $I \in I$ their p-values are contained in. The figure is based on $R = 10^4$ repetitions using p-values generated from the mixture distribution of Sandve et al. (2011), consisting of a proportion $\pi_0 = 0.8$ drawn from a uniform distribution in $[0, 1]$ and the remaining proportion $1 - \pi_0$ drawn from a beta(0.5, 25) distribution. As can be seen, often only a handful of hypotheses remain without decision.

A Auxiliary lemmas

Lemma 1. The two-sided Clopper and Pearson (1934) confidence interval $I_n$ with coverage probability $1 - \rho_n$ based on $n$ samples has length $|I_n| \leq 2(2n)^{-1/2}(-\log \rho_n)^{1/2}$.

Proof. Suppose $S < n$ exceedances are observed among $n$ samples. Let $\xi = (2n)^{-1/2}(-\log \rho_n)^{1/2}$ and regard the following probabilities conditional on $S$ and $n$. The upper limit $p_u$ of the interval $I_n$ is the solution to $\mathbb{P}(X \leq S|p = p_u) = \rho_n$, where $X \sim \text{Binomial}(n, p)$. If $p > S/n + \xi$, by Hoeffding’s inequality (Hoeffding 1963),

$$\mathbb{P}(X \leq S) = \mathbb{P}\left(\frac{X}{n} - \mathbb{E}\left(\frac{X}{n}\right) \leq \frac{S}{n} - \mathbb{E}\left(\frac{X}{n}\right)\right) \leq \exp\left(-\frac{2(S/n - p)^2n^2}{n}\right) < \rho_n.$$ 

Thus $p_u \leq S/n + \xi$. If $S = n$ then $p_u = 1$, implying $p_u \leq S/n + \xi$. Similarly, the lower limit $p_l$ of $I_n$ satisfies $p_l \geq S/n - \xi$. Together, $|I_n| = p_u - p_l \leq 2\xi$. \qed
The proof of lemma 1 is analogous to the one of (Gandy and Hahn 2014, Lemma 2).

**Lemma 2.** The confidence interval $I_n$ produced by the Robbins (1970) and Lai (1976) binomial confidence sequence based on $n$ samples has length $|I_n| \leq n^{-1/2}(\log(4n \log n))^{1/2}$ for large enough $n$.

**Proof.** In (Lai 1976 equation (3)) it is shown that $\mathbb{P}(p \notin I_n) \propto (n \log n)^{-1/2}$. Since $I_n$ need not be centered around the maximum likelihood estimate $\hat{p}_n$, let $I_n^* \supseteq I_n$ be the smallest symmetric interval around $\hat{p}_n$ containing $I_n$. Denote $I_n^* = [\hat{p}_n - t_n, \hat{p}_n + t_n]$ for some $t_n > 0$. By Hoeffding’s inequality (Hoeffding 1963), $\mathbb{P}(p \notin I_n^*) \leq 2 \exp(-2nt_n^2)$. Since $\mathbb{P}(p \notin I_n^*) \leq \mathbb{P}(p \notin I_n)$, $\mathbb{P}(p \notin I_n^*)$ is at least of order $(n \log n)^{-1/2}$. Thus, $t_n \leq (4n)^{-1/2}(\log(4n \log n))^{1/2}$ for large enough $n$. □

**B Proof of theorem 1**

**Proof.** The cdf of $D = \min_{t \in I} d(p, \partial I)$ is bounded above by

$$\mathbb{P}(D \leq t) = \mathbb{P}(\exists I \in \mathcal{I} : d(p, \partial I) \leq t) \leq \sum_{I \in \mathcal{I}} \int_0^t d(p, \partial I)f_p(p)dp \leq (m + 1) \int_0^t Ud\tau \leq (m + 1)Ut,$$

where it was used that the cardinality of $\mathcal{I}$ is at most $m + 1$, that the distance function $d$ is bounded above by 1, and that the density $f_p$ of the p-values is bounded above by a constant $U$.

In the following, an upper bound on the survival function $\mathbb{P}(\tau > t)$ will be derived. First, $\mathbb{P}(|p_t - p| \geq D/2) \leq 2 \exp(-D^2t/2)$ by Hoeffding’s inequality (Hoeffding 1963). Second, using $g(t) \in o(t^\gamma)$ by condition 2, there exists a $t_0 > 0$ such that the event $\{g(t) < D/2\}$ is implied by $\{D > t^\gamma\}$ for all $t > t_0$. Indeed, $g(t) \in o(t^\gamma)$ and $Dt^{-\gamma} > 1$ imply the existence of a $t_0 > 0$ such that $g(t) < g(t)Dt^{-\gamma} < D/2$ for all $t > t_0$.

The survival function $\mathbb{P}(\tau > t)$ can now be bounded above by conditioning on $D$. For $t > t_0$,

$$\mathbb{P}(\tau > t) = \mathbb{P}(\tau > t|D \leq t^\gamma)\mathbb{P}(D \leq t^\gamma) + \mathbb{P}(\tau > t|D > t^\gamma)\mathbb{P}(D > t^\gamma) \leq (m + 1)Ut^\gamma + \mathbb{P}(|p_t - p| \geq D/2|D > t^\gamma) \leq (m + 1)Ut^\gamma + 2 \exp(-\gamma/2t^{2\gamma+1}) ,$$

where it was used that $\mathbb{P}(\tau > t|D \leq t^\gamma) \leq 1$, $\mathbb{P}(D > t^\gamma) \leq 1$ and that $\tau > t$ if either $|p_t - p| \geq D/2$ or $g(t) \geq D/2$ by the definition of $\tau$ in section 3. The latter can be omitted as $\{D > t^\gamma\}$ implies $\{g(t) < D/2\}$ for $t > t_0$. Using $\gamma > -1/2$ by condition 2 $t^{2\gamma+1} \to \infty$ in the argument of the exponential function as $t \to \infty$ and thus $\mathbb{P}(\tau > t) \in O(t^\gamma)$.

Using the fact that the cumulative distribution function (cdf) of the $r$th order statistic $X_{(r)}$ of $n \in \mathbb{N}$ independent and identically distributed random variables $X_1, \ldots, X_n$ with cdf $F_X$ can be expressed as $F_{X_{(r)}}(x) = \sum_{i=r}^n \binom{n}{i} F_X(x) (1 - F_X(x))^{n-i}$ for $r \in \{1, \ldots, n\}$ (David and Nagaraja 2003), the expectation of $\tau_{(m-s)}$ can be bounded as

$$\mathbb{E}(\tau_{(m-s)}) = \int_0^\infty 1 - F_{\tau_{(m-s)}}(t)dt = \int_0^\infty \sum_{i=0}^{m-s-1} \binom{m}{i} F_{\tau}(t) (1 - F_{\tau}(t))^{m-i} dt \leq \sum_{i=0}^{m-3} \binom{m}{i} \int_0^\infty \mathbb{P}(\tau > t)^{m-i}dt,$$

where it was used that the stopping times are non-negative, that $F_{\tau}(t) \leq 1$ and that $2 \leq s < m$ implies $m - s - 1 \leq m - 3$. Using $\mathbb{P}(\tau > t) \in O(t^\gamma)$, the integrals $\int_0^\infty \mathbb{P}(\tau > t)^{m-i}dt$ behave like $\int_0^\infty t^{\gamma(m-i)}dt$ and converge as $\gamma < -1/3$ (cf. condition 2) and $m \geq 3$ imply $\gamma(m - i) < -1$ for all $i \in \{0, \ldots, m - 3\}$. □
References

Andrews, D. and Buchinsky, M. (2000). A Threestep Method for Choosing the Number of Bootstrap Repetitions. *Econometrica*, 68(1):23–51.

Andrews, D. and Buchinsky, M. (2003). Evaluation of a three-step method for choosing the number of bootstrap repetitions. *J Econometrics*, 103(1-2):345–386.

Armitage, P. (1958). Numerical Studies in the Sequential Estimation of a Binomial Parameter. *Biometrika*, 45(1-2):1–15.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc B Met*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, 29(4):1165–1188.

Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika*, 78(2):301–304.

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. and Pearson, E. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413.

Darling, D. and Robbins, H. (1967a). Confidence sequences for mean, variance, and median. *P Natl Acad Sci USA*, 58(1):66–68.

Darling, D. and Robbins, H. (1967b). Iterated logarithm inequalities. *P Natl Acad Sci USA*, 57(5):1188–92.

David, N. and Nagaraja, H. (2003). *Order Statistics*. Wiley.

Davidson, R. and MacKinnon, J. (2000). Bootstrap tests: How many bootstraps? *Economet Rev*, 19(1):55–68.

Ding, D., Gandy, A., and Hahn, G. (2018). A simple method for implementing Monte Carlo tests. *arXiv:1611.01675*, pages 1–17.

Fay, M. and Follmann, D. (2002). Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests. *Am Stat*, 56(1):63–70.

Fay, M., Kim, H.-J., and Hachey, M. (2007). On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests. *J Comput Graph Stat*, 16(4):946–967.

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). MMCTest – A Safe Algorithm for Implementing Multiple Monte Carlo Tests. *Scand J Stat*, 41(4):1083–1101.

Gandy, A. and Hahn, G. (2016). A Framework for Monte Carlo based Multiple Testing. *Scand J Stat*, 43(4):1046–1063.

Gandy, A. and Hahn, G. (2017). QuickMMCTest: quick multiple Monte Carlo testing. *Stat Comput*, 27(3):823–832.
Guo, W. and Peddada, S. (2008). Adaptive Choice of the Number of Bootstrap Samples in Large Scale Multiple Testing. *Stat Appl Genet Mol Biol*, 7(1):1–16.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

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

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Stat*, 6(2):65–70.

Jin, H., Rube, H., and Song, J. (2016). Categorical spectral analysis of periodicity in nucleosomal DNA. *Nucleic Acids Res*, 44(5):2047–2057.

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.

Lin, D. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reiss, P., Huo, L., Zhao, Y., Kelly, C., and Ogden, R. (2015). Wavelet-Domain Regression And Predictive Inference In Psychiatric Neuroimaging. *Ann Appl Stat*, 9(2):1076–1101.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann Math Statist*, 41(5):1397–1409.

Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77(3):663–665.

Roth, M., Dahmen, J., Muir, D., Imhof, F., Martini, F., and Hofer, S. (2016). Thalamic nuclei convey diverse contextual information to layer 1 of visual cortex. *Nat Neurosci*, 19:299–307.

Rubin-Delanchy, P., Burn, G., Griffé, J., Williamson, D., Heard, N., Cope, A., and Owen, D. (2015). Bayesian cluster identification in single-molecule localization microscopy data. *Nat Methods*, 12:1072–1076.

Sandve, G., Ferkingstad, E., and Nygård, S. (2011). Sequential Monte Carlo multiple testing. *Bioinformatics*, 27(23):3235–3241.

Shaffer, J. (1986). Modified Sequentially Rejective Multiple Test Procedures. *J Am Stat Assoc*, 81(395):826–831.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*, 62(318):626–633.

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

Silva, I. and Assunção, R. (2018). Truncated sequential Monte Carlo test with exact power. *Braz J Probab Stat*, 32(2):215–238.

Silva, I., Assunção, R., and Costa, M. (2009). Power of the sequential Monte Carlo test. *Sequential Anal*, 28(2):163–174.
Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Termenon, M., Jaillard, A., Delon-Martin, C., and Achard, S. (2016). Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project. *Neuroimage*, 142:172–187.

van Wieringen, W., van de Wiel, M., and van der Vaart, A. (2008). A Test for Partial Differential Expression. *J Am Stat Assoc*, 103(483):1039–1049.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Ann Math Statist*, 16(2):117–186.