Brazilian Journal of Probability and Statistics
2018, Vol. 32, No. 2, 215–238
https://doi.org/10.1214/16-BJPS339

Truncated sequential Monte Carlo test with exact power

Ivair Silva[a,b,1] **and Renato Assunção**[c,2]

[a]*Harvard Medical School and Harvard Pilgrim Healph Care Institute*
[b]*Federal University of Ouro Preto*
[c]*Federal University of Minas Gerais*

Abstract. Monte Carlo hypothesis testing is extensively used for statistical inference. Surprisingly, despite the many theoretical advances in the field, statistical power performance of Monte Carlo tests remains an open question. Because the last assertion may sound questionable for some, the first goal in this paper is to show that the power performance of truncated Monte Carlo tests is still an unsolved question. The second goal here is to present a solution for this issue, that is, we introduce a truncated sequential Monte Carlo procedure with statistical power arbitrarily close to the power of the theoretical exact test. The most significant contribution of this work is the validity of our method for the general case of any test statistic.

1 Introduction

When the true distribution of a test statistic is unknown, the exact hypothesis test is intractable. The Monte Carlo hypothesis test provides an effective solution when it is feasible to generate values from the test statistic under the null hypothesis (H_0). By means of this simulation, a reference empirical distribution for U can be obtained and used to make a decision about whether to accept/reject H_0.

To illustrate the applicability of Monte Carlo tests, consider the classic problem of testing the independence between rows and columns in contingency tables with fixed marginals a priori. If the table has many columns and rows, or the totals of the marginals are large, the Fisher's exact test is impracticable. If the expected counts are small, the Pearson's Chi-square test can give spurious results. However, given that all marginals totals are fixed, tables can be generated under the null hypothesis from a hypergeometric distribution, and then the Monte Carlo test is applicable.

Monte Carlo simulations are used to find efficient solutions in a wide diversity of fields (Li and Kulldorff (2009), Smith, Forster and McDonald (1996), Gates (1991), Smith (1996)). A modern and well-known problem solved by a

[1]Supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and by Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG).
[2]Supported by CNPq and Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG).
Key words and phrases. Exact hypothesis testing, power loss upper bounds, p-value, sequential Monte Carlo design.
Received March 2015; accepted October 2016.

Monte Carlo test approach is the spatial scan test for detection of spatial clusters (Kulldorff (2001)). Although the analytical distribution of the scan statistic remains unknown, it is simple to generate values from it under the hypothesis of non-existence of spatial clusters in the map.

Monte Carlo tests can be broadly separated into conventional and sequential procedures. The former is based on a fixed and pre-determined number $(m-1)$ of simulations (Hope (1968), Barnard (1963), Birnbaum (1974), Dwass (1957)). Alternatively, in the later the total number of simulations is random, and the user proceeds with the simulation until there is enough evidence to make a decision (Besag and Clifford (1991)).

For a given data set, let u_0 be an observed value of the test statistic, and let U to denote the test statistic as a random variable. Denote the probability distribution of U by $F(u)$, and let $u_1, \ldots, u_{m-1}$ be a sequence of Monte Carlo copies of U generated under the null hypothesis. Under the conventional Monte Carlo test, H_0 is rejected if $x < (\alpha_{\text{mc}} m)$, where x is the number of u_i's that are greater than or equal to u_0, and $\alpha_{\text{mc}} \in (0, 1)$ is a desired significance level for the Monte Carlo test. A well-known property is that, for any $F(u)$, the Monte Carlo test is a level α_{mc} test. When $F(u)$ is a continuous distribution, and m is proportional to $1/\alpha_{\text{mc}}$, with rational α_{mc}, the Monte Carlo test is of size α_{mc}. One way of proving these properties was presented by Silva, Assunção and Costa (2009).

Under the sequential Monte Carlo test approach, the simulations are interrupted when x_l intersects a lower or upper stopping boundary, where x_l is the number of simulated values greater than or equal to u_0 at the lth simulation. In practice, and depending on the nature of the problem, the boundaries may not be intersected in a viable time. Even worse, the simulation can last indefinitely (Gandy (2009)). Thus, it becomes necessary to pre-define a maximum number of simulations, a truncation rule, to stop such simulations. These sequential procedures are called "truncated sequential Monte Carlo tests". Although it is somewhat uncertain when the term "truncated" was used by the first time in sequential analysis, because the term is found in the work of Armitage (1958), it is safe to assert that the first use certainly dates back to fifty decades. Truncated sequential procedures are more realistic than open-ended approaches; this is so because, in practice, the simulations must be interrupted at some point in order to favor a decision about H_0.

Surprisingly, although Monte Carlo simulation is used in many fields of science, such as numerical analysis, applied mathematics and statistical inference, the power performance of the Monte Carlo test, in comparison to the exact test, remains an open question in practical terms. After showing, using a real example, that the conventional Monte Carlo test can lead to elevated power losses such as, for example, 20% of the exact test power, the second goal of this manuscript is to present a general solution for this problem.

Generally, a sequential approach is adopted in order to handle at least one of the two following challenges: (i) saving execution time for computational intensive

test statistics; and (ii) bounding the power loss in small values, like in at most 1% of the exact test, for example.

As stressed earlier, the first objective of this paper is to show that item (ii) is still an open question. The second goal is to provide a general solution for the problem. We are not strongly concerned with saving computational execution time. Instead, our proposal is meant to ensure a true control of the power loss in small values, such as 1%, given pre-fixed and finite maximum number of simulations. Unlike other truncated methods, our results are valid for the general case of any test statistic, that is, the method is free from assumptions. Thus, in this sense, we say that our proposal is a test with exact power. Additionally, a valid p-value, P_{mc}, is deduced in order to support the decision about accept/reject H_0, that is, for the proposed Monte Carlo p-value holds that $\Pr(P_{\text{mc}} \leq \alpha | H_0) \leq \alpha$, for all $0 \leq \alpha \leq 1$.

Showing that item (ii) has been an open challenge is our first goal. However, the solution of the problem is certainly of major interest for some readers, especially for the experts in the field of Monte Carlo testing for whom the reported problem is well known. Hence, this material is organized in a way that the reader interested in seeing beforehand the solution can just jump from here to read the three first paragraphs of Section 3, then read Section 3.1, and then jump to read the solution presented in Section 4, which can be understood without having to read other parts of the manuscript. But, those readers unfamiliar with the subject and interested to see some properties and limitations of conventional procedures, can just read each section following the natural order at which they appear, which is organized in the following way: next section presents an overview of the main proposals for sequential Monte Carlo test designs. Section 3 shows that the theoretical power performance of Monte Carlo tests is an open question, and this is reinforced by presenting a numerical counter-example for testing the mean of a Poisson distribution. Section 4 introduces our truncated sequential Monte Carlo test with exact power. Section 5 presents a brief discussion on some implications of the main results.

2 Sequential Monte Carlo test designs: Current proposals and their limitations

This section presents a brief description of the main advantages and disadvantages of some prominent sequential Monte Carlo procedures found in the literature.

Under an exact hypothesis test criteria, H_0 is rejected if $p \leq \alpha$, where p is the p-value and α is a desired significance level. For the continuous case, the probability distribution function of the p-value, $F_P(p) = \Pr(P \leq p)$, can be written in the following way:

$$F_P(p) = \begin{cases} 1 - F_A\big(F_0^{-1}(1-p)\big), & \text{for right-hand tests,} \\ F_A\big(F_0^{-1}(p)\big), & \text{for left-hand tests,} \\ 1 - F_A\big(F_0^{-1}(1-p)\big) + F_A\big(F_0^{-1}(p)\big), & \text{two-sided,} \end{cases} \quad (2.1)$$

where P is the p-value as a random variable, and F_A and F_0 denote the probability distribution functions of U under H_A and H_0, respectively. $F_0^{-1}(p)$ is such that $\Pr(U \leq F_0^{-1}(p)|H_0) = p$.

The study of the Monte Carlo test power, π_m, can be done by evaluating the probability of rejecting the null hypothesis as a function of p. Because the p-value is a random variable, P, π_m can be seen as the expectation $\int_0^1 \pi_m(\alpha_{\text{mc}}, p) F_P(dp)$, where $\pi_m(\alpha_{\text{mc}}, p)$ is the probability of rejecting H_0 with the Monte Carlo test for a fixed p, and F_P is a probability measure defined according to (2.1). Here we emphasize that, for a more general discussion, the significance level of the Monte Carlo test, α_{mc}, is not required to be equal to the significance level, α, of the exact test. This is arbitrary for the user in practice.

For cases where F_P is continuous, Hope (1968) studied the behavior of the conventional Monte Carlo test power with respect to m and proved that it converges to the power of the uniformly more powerful test. He restricted the evaluations to a class of probability density functions for P, given by all monotonic densities with respect to p. After taking the derivative of (2.1) with respect to p, we can see that this assumption is the same as assuming a monotonic behavior for the likelihood ratio with respect to the argument $F_0^{-1}(p)$.

Also limited to the continuous case, Jockel (1986) explored the composed tests with hypothesis $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$, where $\theta \in \mathbb{R}$. By assuming that $F_P(p)$ is concave with respect to p, Jockel (1986) derived an expression to bound the power loss of the conventional Monte Carlo test into arbitrarily small values. Jockel (1986) argues that likelihood ratio tests usually satisfy the concavity assumption over $F_P(p)$. We can figure out the reasoning behind this argument by noting that, for one-sided tests, a concavity of $F_P(p)$ implies in a monotonicity of the likelihood ratio just as assumed by Hope (1968).

Let τ be the class of concave p-value distributions. In practice, we have to verify the validity of these assumptions by analyzing the third line in (2.1). But this presupposes some familiarity with the behavior of F_0 and F_A with respect to θ. Therefore, it is not practical to check whether concavity is a reasonable assumption. This is so because F_0 and F_A are unknown when Monte Carlo tests are in use. Thus, this sort of assumption does not represent a realistic solution.

Fay and Follmann (2002) proposed the Iterative Push Out procedure (IPO), which is designed to save execution time in the simulations. To bound the resampling risk (RR), the probability of disagreement about the accept/reject decision between the Monte Carlo test procedure and the exact test, Fay and Follmann (2002) considered a rather restrictive class, $\Im$, for the p-value distribution, with cumulative distribution function given by:

$$H_{\alpha,1-\beta}(p) = 1 - \Phi\{\Phi^{-1}(1-p) - \Phi^{-1}(1-\alpha) + \Phi^{-1}(\beta)\}, \qquad (2.2)$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Normal distribution, α is the desired significance level and β is the Type II error probability. Their

approach consisted on finding the worst distribution in $\Im$ in the sense of having the largest RR. The class $\Im$ is implied by a test statistic U that either follows the standard normal distribution under the null hypothesis and a $N(\mu, 1)$ under the alternative, or it follows a central and a non-central $\chi_1^{(2)}$ under the null and alternative hypotheses, respectively. It is important to note that RR is an upper bound for the power loss, therefore an upper bound for the former is also a bound for the last.

Klein et al. (2010) noted that the IPO procedure is not applicable when one wants to bound RR in arbitrary small values (e.g., 1%), and proposed a truncated sequential procedure to save execution time and to guarantee upper bounds for RR. This was possible by restricting their analysis to the distribution class $\Im$.

Fay, Kim and Hachey (2007) also proposed an algorithm to implement a truncated Sequential Probability Ratio Test (tSPRT) to save execution time in the Monte Carlo test. They studied the behavior of RR as a function of p, emphasizing the fact that RR is close to 0.5 for $p = \alpha_{\text{mc}}$.

In order to save computational effort in sequential Monte Carlo tests without power losses with respect to the conventional Monte Carlo test, Silva and Assunção (2013) introduced an optimal generalized truncated sequential Monte Carlo test. Their proposal provided a theoretical expected number of simulations considerably smaller than the predecessors proposals, but the investigations were not devoted to treat the Monte Carlo power losses with respect to the exact test.

Gandy (2009) proposed an open-ended sequential procedure that uniformly bounds the resampling risk in arbitrarily small values for any test statistic. We call this method by "risk spending approach". This RR control is possible because the procedure considered by Gandy (2009) is not truncated. We stress that, in practice, it is necessary to establish a maximum number of simulations. The reason is that, if $p = \alpha_{\text{mc}}$, Gandy (2009) shows that the expected number of simulations is infinite. Since all his results are based on asymptotic arguments and valid only when the open-ended strategy is adopted, the effective control of RR is an open problem in practical terms.

In summary, the existing tentative of avoiding power losses in Monte Carlo tests are valid only under limited circumstances, like assuming a specific shape for the p-value distribution, or taking the risk of having simulations running forever. Actually, both limitations are inconvenient for the practice.

The next section presents the response for the first goal of this paper, that is, we show how and why bounding the power losses of Monte Carlo tests remains an open problem. The explanation involves: (i) that cases where $p = \alpha$ are pathological for the Monte Carlo theory in such a way that can lead to power losses of 50% even for very large values of m, (ii) to illustrate that such elevated power losses can really happen in practice like for the problem of testing the mean of a Poisson sample, and (iii) to explain why one of the most prominent methods, the risk spending approach of Gandy (2009) also fails to solve the problem.

3 The challenge of bounding power losses from Monte Carlo tests

Let $u_1, u_2, \ldots, u_{m-1}$ be a sample generated from the distribution of U under H_0, and let u_0 be the observed value of the test statistic for a fixed data set. Under right-hand tests, that is, if H_0 is rejected for large values of U, the Monte Carlo test statistic is X_{m-1}, which counts the number of u_i's greater than or equal to u_0. If H_0 is rejected for small values of U, left-hand tests, X_{m-1} is defined as the number of u_i's smaller than or equal to u_0. A conventional Monte Carlo test rejects H_0 if $X_{m-1} < C$. For $\alpha_{\text{mc}} \in (0, 1)$, this test criterion has significance level equal to α_{mc} if $C = \lfloor \alpha_{\text{mc}} m \rfloor$ (Silva, Assunção and Costa (2009)). The value $\lfloor y \rfloor$ is the greatest integer smaller than y. With two-sided tests, H_0 is rejected with level α_{mc} if $X_{m-1} < C_1$ or if $X_{m-1} > C_2$, with $C_1 = \lfloor \alpha_{\text{mc}} m \rfloor / 2$, and $C_2 = m - (\lfloor \alpha_{\text{mc}} m \rfloor / 2)$. A valid p-value for the conventional procedure can be calculated by $P_m = (X_{m-1} + 1)/(m + 1)$. Denote this conventional Monte Carlo test by MC_m. We shall work only with one-sided tests from now on, but all results in this paper can be extended for two-sided tests by using similar reasonings. The Monte Carlo test statistic, X_{m-1}, follows a binomial distribution with $(m - 1)$ essays and success probability equal to the observed p. Thus, MC_m rejects H_0 with probability:

$$\begin{aligned} \pi_m(\alpha_{\text{mc}}, p) &= \Pr(X_{m-1} \leq C - 1 | P = p) \\ &= \sum_{x=0}^{C-1} c_x^{m-1} p^x (1 - p)^{m-1-x}, \end{aligned} \tag{3.1}$$

where $c_b^a = a!/[b!(a - b)!]$, with a and b positive integers.

An interesting MC_m property, which has not been pointed out in the literature yet, is the fact of being necessary to set m as a multiple of $\lfloor 1/\alpha_{\text{mc}} \rfloor$. This is needed in order to avoid power losses in comparison to another design, say MC_{m_1}, which by its turn is based on a smaller number of simulations, m_1, where m_1 is the greatest multiple of $\lfloor 1/\alpha_{\text{mc}} \rfloor$ smaller than m. After restricting the choice of m to the set of the multiples of $\lfloor 1/\alpha_{\text{mc}} \rfloor$, Hope (1968) and Jockel (1986) assumed some conditions over the shape of $F(u)$ to prove that the Monte Carlo test power increases monotonously with m. But here we emphasize that, if the power is monotonously increasing in a sequence $m_1 < m_2 < \cdots$, then the terms of this sequence are multiples of $\lfloor 1/\alpha_{\text{mc}} \rfloor$. The last assertion holds for any shape of $F(u)$. Thus, for the conventional Monte Carlo test, a rule of thumb is: m must be always selected as a multiple of $1/\alpha_{\text{mc}}$.

Theorem 3.1. *The power of the conventional Monte Carlo test is non-increasing with m for* $\lfloor j/\alpha_{\text{mc}} \rfloor \leq m < \lfloor (j + 1)/\alpha_{\text{mc}} \rfloor$, *where j is a positive integer. Then, if power is the only concern, m must be chosen as a multiple of* $\lfloor 1/\alpha_{\text{mc}} \rfloor$.

The proof is left to the Appendix. As an example, take $\alpha_{\text{mc}} = 0.01$, then a Monte Carlo test with $m_1 = 1050$ is less powerful than a second design with $m_2 = 1000$. A potential increasing in power, with respect to m_2, occurs for m_1 starting from 1100.

3.1 A trap near α

The power $\pi_m(\alpha_{\text{mc}})$ of the MC_m test is obtained by integrating expression (3.1), $\pi_m(\alpha_{\text{mc}}, p)$, with respect to the distribution of the p-value, F_P, as follows:

$$\pi_m(\alpha_{\text{mc}}) = \int_0^1 \pi_m(\alpha_{\text{mc}}, p) F_P(dp). \tag{3.2}$$

Let α be the significance level of the exact test. For $P = p$, that is, given a realized p-value after an observed data set, under the exact hypothesis test, H_0 is rejected if $p \leq \alpha$, and it is not rejected if $p > \alpha$. Let $\pi(\alpha, p)$ be the probability of rejecting H_0 under the exact test, which, for a given p, can be expressed by:

$$\pi(\alpha, p) = \begin{cases} 1, & \text{if } p \leq \alpha, \\ 0, & \text{if } p > \alpha. \end{cases}$$

Thus, the power of the exact test is calculated as following:

$$\pi(\alpha) = \int_0^1 \pi(\alpha, p) F_P(dp) = \int_0^\alpha F_P(dp). \tag{3.3}$$

A well-known fact is the convergence of the MC_m power to the exact power. To state this we use the notation "$\overset{\text{a.e.}}{\to}$" for "almost everywhere" convergence.

Theorem 3.2. *Let $\pi_m(\alpha_{\text{mc}})$ and $\pi(\alpha)$ denote the statistical power of the conventional Monte Carlo test and of the exact test, respectively. Thus, $\pi_m(\alpha_{\text{mc}}) \overset{\text{a.e.}}{\to} \pi(\alpha)$ as $m \to \infty$.*

Proof. Take $\alpha_{\text{mc}} = \alpha$. As $X_{m-1}/m \overset{\text{a.e.}}{\to} p$ when $m \to \infty$, then, for $p < \alpha$, $\Pr(X_{m-1} \leq m\alpha - 1 | P = p) \overset{\text{a.e.}}{\to} 1$, and if $p \geq \alpha$, $\Pr(X_{m-1} \leq m\alpha - 1 | P = p) \overset{\text{a.e.}}{\to} 0$. Thus, according to the dominated convergence theorem, $\lim_{m\to\infty} \pi_m(\alpha) = \pi(\alpha)$. □

Although the convergence stated in Theorem 3.2 is clearly a very important property of the conventional Monte Carlo test, under a practical point of view, it is more useful to understand the relationship between $\pi_m(\alpha_{\text{mc}})$ and $\pi(\alpha)$ for finite m. Let

$$D(\alpha, \alpha_{\text{mc}}, m, p) = \pi(\alpha, p) - \pi_m(\alpha_{\text{mc}}, p).$$

In words, $D(\alpha, \alpha_{\text{mc}}, m, p)$ is the difference between the exact test power and the MC_m power for a fixed $P = p$. Note that $\pi_m(\alpha_{\text{mc}}, p)$ is decreasing with p, and this is so because it is the cumulative binomial distribution evaluated at $C - 1$.

The function $\pi(\alpha, p)$, by its turn, is a step function assuming 1 for $p \in (0, \alpha]$ and 0 otherwise. Then, the difference $D(\alpha, \alpha_{\text{mc}}, m, p)$ is positive and increasing for $p \in (0, \alpha]$, and it is negative for $p \in (\alpha, 1)$. Consequently, the maximum power difference occurs at $p = \alpha$, that is:

$$D(\alpha, \alpha_{\text{mc}}, m, \alpha) = \max_{p \in (0,1)} \left\{ D(\alpha, \alpha_{\text{mc}}, m, p) \right\} \tag{3.4}$$

$$= 1 - \sum_{x=0}^{\lfloor m\alpha_{\text{mc}} \rfloor - 1} c_x^{m-1} \alpha^x (1-\alpha)^{m-x-1}. \tag{3.5}$$

Figure 1 illustrates this behavior for values of $m = 1000,\ 10{,}000$, $\alpha = 0.05$ and $\alpha_{\text{mc}} = 0.05, 0.06$. The step function in each graph of Figure 1 represents

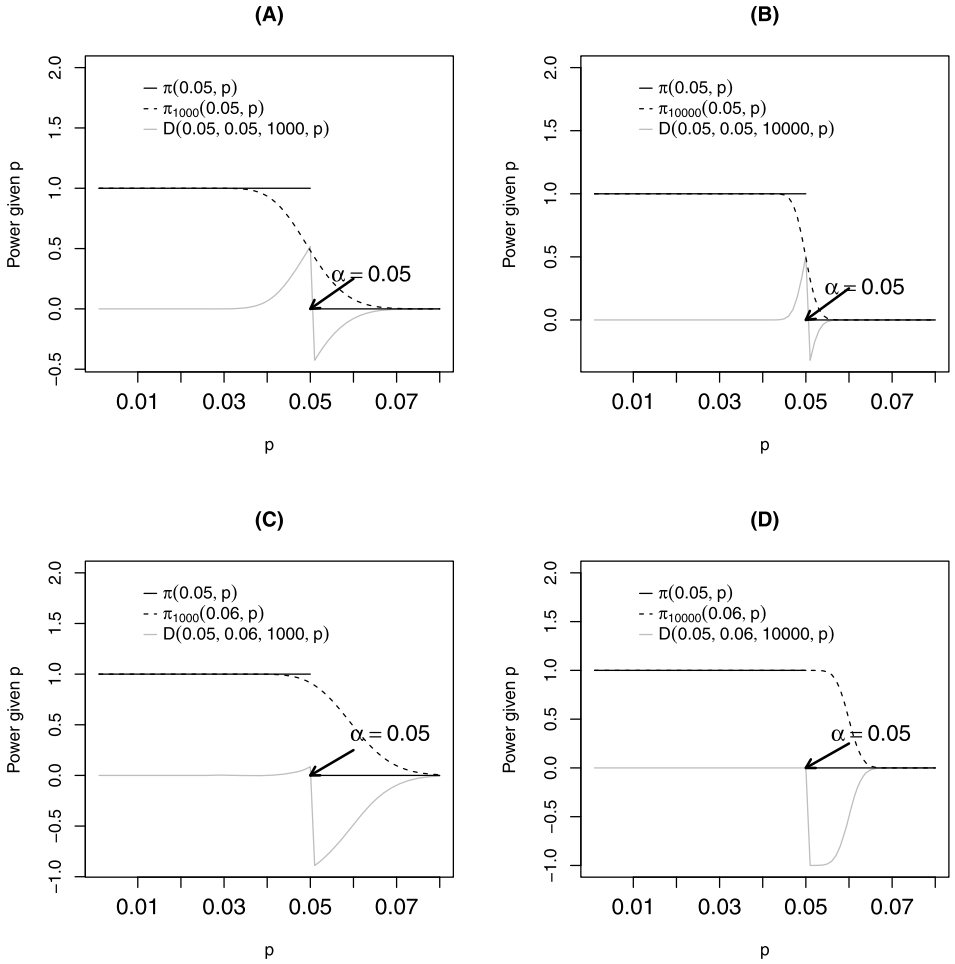

Figure 1 *Power difference between the exact test and* MC_m *for fixed p-values.*

$\pi(\alpha = 0.05, p)$, the dashed curve represents $\pi_m(\alpha_{\text{mc}}, p)$, and the dotted line is the difference between them. The maximum difference occurs at $p = \alpha$ because the exact power function is zero after α. For the case $m = 1000$ and $\alpha_{\text{mc}} = 0.05$, Figure 1(A), the maximum difference is $D(0.05, 0.05, 1000, 0.05) = 0.517$. Figure 1(B) has the case $m = 10{,}000$, which is close to 1 for a larger interval of p than in (A), but also tumbles to 0.51 for $p = \alpha$. The maximum difference keeps close to 0.5 even for astronomically large values of m, like 10^{10}, for example. Then, the power loss is elevated when F_P is concentrated near α. This problem explains the utility of assuming restricted classes of distributions to study the MC_m power, like $\Im$ and τ, since these assumptions restrict the amount of mass that F_P allocates around α.

At first sight, the discussion about the behaviour of the Monte Carlo power function around α may sound like technicality without practical relevance. An analyst could argue that the situations where the distribution of P is substantially high around α are rare, then this problem can be overlooked. But this is not true. In fact, F_P does not belong to distribution families like $\Im$ or τ even for classically common applications like, for example, the problem of testing the mean of a Poisson distribution.

3.2 Power losses from the conventional Monte Carlo test for Poisson data

This section is meant to show that, for some applications, the actual power loss from a conventional Monte Carlo test can be excessively large even when using a very large number $(m - 1)$ of simulations, that is, the guarantee of a satisfactory statistical power for procedures based on the conventional Monte Carlo test remains an open question under a general perspective.

For a counter-example against the performance of previous Monte Carlo test procedures, our discussion uses an important test statistic for the sequential analysis field, the maximized sequential probability ratio test statistic (MaxSPRT). MaxSPRT was developed for the prospective rapid cycle vaccine safety surveillance, and implemented by the Centers for Disease Control and Prevention (CDC) sponsored Vaccine Safety Datalink (VSD) (Kulldorff et al. (2011)). MaxSPRT is an extension of the well-known Sequential Probability Ratio Test (SPRT) (Wald (1945)). Unlike SPRT, MaxSPRT is defined for composite alternative hypothesis rather than simple. Currently, MaxSPRT is widely used for monitoring increased risks of adverse events in post-market safety surveillance (Yih et al. (2009), Belongia et al. (2010), Klein et al. (2010)).

In the vaccine safety surveillance context, C_t is the random variable that counts the number of adverse events in a known risk window, from 1 to W days, after a vaccination that was administrated in a period $[0, t]$. Commonly, under the null hypothesis, C_t is supposed to have a Poisson distribution with mean μ_t, where μ_t is a known function of the population at risk, adjusted for age, gender and any other covariates of interest. Under H_A, C_t is Poisson with mean $R\mu_t$, where R

is the unknown increased relative risk due to the vaccine. The MaxSPRT statistic, denoted here by LLR_t, is given by: $LLR_t = (\mu_t - c_t) + c_t \log(c_t/\mu_t)$, if $c_t \geq \mu_t$, and $LLR_t = 0$, otherwise.

The monitoring can be done by a continuous or group sequential fashion. In the continuous case, a test is performed as soon as a chunk of events arrives, and LLR_t is confronted against a critical value, CV, to decide about the H_0 acceptance/rejection. Basically, the surveillance is interrupted for the rejection of H_0 at the first t for which $LLR_t \geq \text{CV}$, with $t \in (0, T]$. T is a pre-specified maximum length of surveillance, and defined in terms of the expected number of events under H_0. In the group approach, there are G pre-defined moments of testing, $t_1, , \ldots, t_G$, for which, if $LLR_{t_i} \geq \text{CV}$, for some $i = 1, \ldots, G$, the null hypothesis is rejected. In both approaches, continuous or group sequential, the surveillance is interrupted for the acceptance of H_0 if $t = T$ happens before $LLR_t \geq \text{CV}$. For the present example, we shall consider the group sequential approach.

Based on the exact distribution of LLR_t, we have calculated the true power under the exact MaxSPRT test. It was done for R in the range $[1, 1.1, \ldots, 2]$, a significance level of 0.05, and four different designs for the number of group sequential tests, $G = 1, 2, 4, 10$. The case $G = 1$ is the classical non-sequential test for the parameter of a Poisson-based distribution, and represents the design for an unique test at time T. Here, and accordingly to a common practice in the sequential analyses field, the moments of testing were distributed in equally spaced sizes. Thus, we fixed the interval between two sequential tests by $10/G$, except for the last moment, T, which was taken in a way to satisfy the desired alpha level of 0.05. For $G = 1, 2, 4, 10$, the solutions were $T = 10.0359566, 6.65359, 8.17171, 32.18292$, respectively. We used this large number of decimal places because of the discrete nature of C_t, which provided a precision of 10^{-8} for type I error probabilities around $\alpha = 0.05$. Maximum surveillance sizes around 10 and 30 are likely common when monitoring rare adverse events.

Assuming μ_t known, for each fixed G, we used a numerical procedure to run a Markov chain of a Poisson process in order to calculate the exact distribution of LLR_t. Then, using expression (3.3), we have calculated the MaxSPRT power for different values of R. In parallel, the MC_m power, expression (3.2), was also obtained for each scenario. Because we want to stress the fact that the power losses from Monte Carlo tests can be expressively large even if we use a very large m value, here the MC_m power calculations were made by fixing $m = 10^{10}$.

The y-axis of Figure 2 presents the effective ratio between the MC_m power and the exact MaxSPRT power. The x-axis brings the values of the exact MaxSPRT powers. Each line represents a different scheme associated to a specific value of G. At first, for fixed G and varying R, observe the inconvenient characteristic of having higher losses for smaller exact powers. It is also an undesirable fact that, for fixed R and varying G, the power losses are more expressive for the more powerful designs, which are the less frequent tests ($G = 1$ and $G = 2$). The effective MC_m power losses reach expressive values as, for example: 20% ($G = 1$, $R = 1.1$); 10%

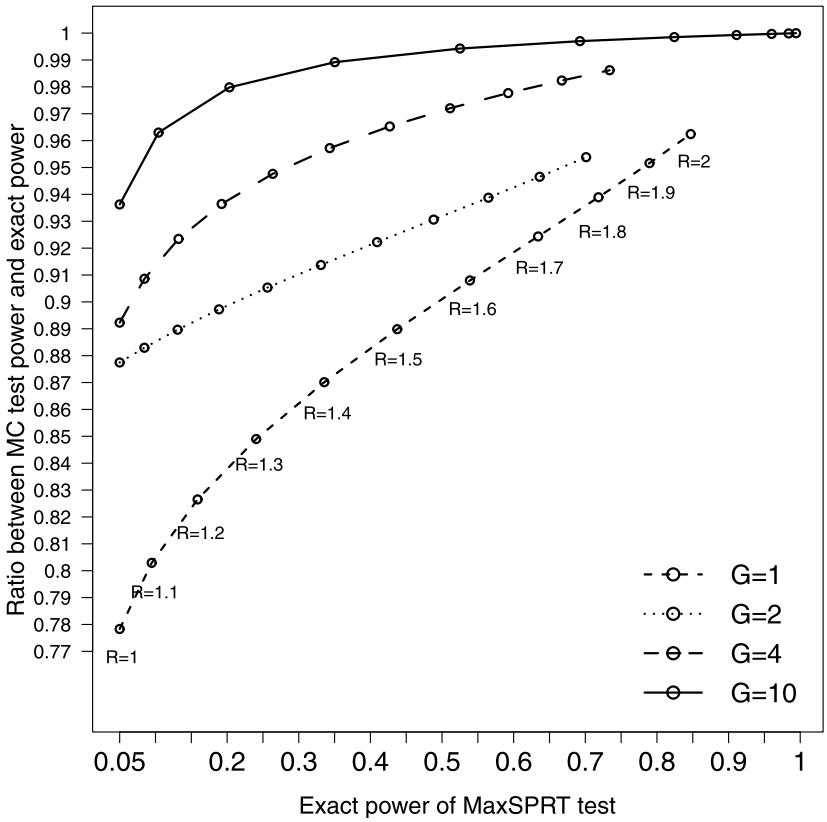

Figure 2 *Ratio between the* MC_m *power and the exact test power for the MaxSPRT statistic with relative risks* (R) *in the range* $[1, 1.1, \ldots, 2]$.

$(G = 2, R = 1.4)$; or 5% $(G = 4, R = 1.4)$. This occurs because the values of c_t, leading to the rejection of H_0 at the fringe of CV, have an expressive mass of probability due to its discrete nature, and, as already discussed, the MC_m power is around 0.5 in this situation $(p = \alpha = 0.05)$.

To clarify the calculations used to access the MC_m power losses, consider the case $G = 1$, where the critical value, in the scale of the number of cases, is $c_T = 16$, that is, $\Pr(X_T \geq 16|T = 10.0359566, R = 1) = 0.05$, where $X_T \sim \text{Poisson}(T)$. Now, if $R = 1.5$, the exact MaxSPRT power is $\Pr(X_T \geq 16|T = 10.0359566, R = 1.5) = 0.4374351$. We point out that the single probability at $p = \alpha$ (i.e., $c_T = 16$) is responsible for a large amount of the overall power: $\Pr(X_T = 16|T = 10.0359566, R = 1.5) = 0.102426$. From (3.1), we have $\pi_{10^{10}}(\alpha_{\text{mc}} = 0.05, p = 0.05) = 0.4999945$. Thus, the power loss, given $c_T = 16$, is equal to $(1 - 0.4999945) \times 0.102426 = 0.05121356$, about 11.7% of 0.4374351, the exact MaxSPRT power. A similar reasoning was applied to calculate the power losses from different values of G.

Now, it is opportune to mention that the MC_m power, for fixed p, is not substantially affected for p close and suitably greater than α (see Figure 1). But, for the example of $G = 1$ above, because the distribution of X_T is discrete, the smallest p, greater than 0.05, is 0.08535, which is associated to $c_T = 15$. This event occurs with probability approximately equal to zero for any $m > 1000$. Fortunately, the critical values for the MaxSPRT statistic can be exactly calculated by a numerical Markov chain algorithm (Kulldorff et al. (2011)), that is, the Monte Carlo approach is not needed because the exact test can be naturally applied.

Elevated power losses, such as the experienced with the numerical example above, cannot be figured out in real applications since Monte Carlo tests are required just because of the inability of performing similar analytical studies. When Monte Carlo tests are needed, the exact test statistic distribution is unknown. But, these power losses are potentially present in many hypotheses testing situations. For example, when the maximum length of surveillance, T, is impressively large, for example, $T = 10{,}000$, and it is allowed to look to the data as many times as the user wants (continuous sequential approach), the exact critical value, for performing the MaxSPRT test, is computationally intensive to be calculated. Then, the Monte Carlo test is a convenient option. Since this last scenario is directed to treat, in essence, the same type of data as earlier, it is evident that an application of the Monte Carlo test can lead to elevated power losses such as the illustrated above.

3.3 Power losses from the risk spending approach

The method of Gandy (2009), referred here by "risk spending approach", is an open-ended procedure directed to bound the resampling risk of Monte Carlo tests. Note that a bound for the resampling risk is also a bound for the power loss. Let S_l and L_l denote upper and lower thresholds such that H_0 is rejected at the lth Monte Carlo iteration if $X_l \leq L_l$, and H_0 is not rejected if $X_l \geq S_l$. If neither S_l nor L_l are hit at iteration l, the simulations are continued. Let ε_l denote a nondecreasing sequence such that $\varepsilon_l \to \varepsilon$ as $l \to \infty$, where ε (<0.5) is a desired upper bound for the resampling risk, and then it is to be settled satisfactorily small like for example, $\varepsilon = 0.01$. The thresholds are constructed in a way that:

$$\begin{aligned} \Pr[\text{hit } S_l \text{ until } l \,|\, p = \alpha] &\leq \varepsilon_l, \\ \Pr[\text{hit } L_l \text{ until } l \,|\, p = \alpha] &\leq \varepsilon_l. \end{aligned} \tag{3.6}$$

The main fact to be emphasized is that, in practice, the user has to establish a truncation on the number of simulations. Otherwise, as a matter of chance, the number of simulations can continue for a long time for a p very close to α, and worse, the expected number of simulations is infinite for $p = \alpha$ (Gandy (2009)). But, as stated by Gandy (2009), a truncated version of the method no longer ensures a true control of possible power losses.

To see why the truncated version of the risk spending approach can lead to elevated power losses, let m denote the maximum number of simulations at which the procedure is interrupted if neither S_l nor L_l are hit until $l = m$. In the indecision situation, that is, the cases where neither S_l nor L_l are hit until $l = m$, the decision is based on the Monte Carlo p-value, which, as suggested by Gandy (2009), is given by the maximum likelihood estimator of p, $\hat{p}_{\text{naive}} = X_m/m$.

By the construction in (3.6), the probability of rejecting H_0 until the mth simulation is at most ε_m. Thus, in the indecision situation, the probability of rejecting H_0 is solely guided by the Monte Carlo p-value $\hat{p}_{\text{naive}}$. Therefore, in the indecision situation H_0 is rejected if $\hat{p}_{\text{naive}} \leq \alpha$, and H_0 is not rejected if $\hat{p}_{\text{naive}} > \alpha$. Hence,

$$\begin{aligned}\Pr[\text{rejecting } H_0 | p = \alpha] &\leq \Pr[\text{hit } L_l \text{ until } l | p = \alpha] + \Pr[\hat{p}_{\text{naive}} \leq \alpha | p = \alpha] \\ &\leq \varepsilon + \Pr[X_m \leq \alpha m | p = \alpha].\end{aligned}$$

Remind from Section 3.1 that the probability $\Pr[X_m \leq \alpha m | p = \alpha]$ is close to 0.5 even for m values in the magnitude of thousands of billions. Therefore, if $p = \alpha$, the power loss from the risk spending approach is of at least $100 \times (1 - \Pr[X_m \leq \alpha m | p = \alpha] - \varepsilon)\%$. For example, for $\alpha = 0.05$ and $m = 10^{10}$, we have $\Pr[X_m \leq \alpha m | p = \alpha] \approx 0.5000119$. Using $\varepsilon = 0.01$, the power loss is of at least 48%. The power loss can be reduced by increasing ε, but this would simultaneously lead to an overshoot in the overall test size to $\alpha + \varepsilon$.

3.4 On a general alpha-liberal solution with exact power

Previous sections showed that power losses of former Monte Carlo tests can be large in practical contexts even if we use a very large number of simulations. In order to bound the power losses in arbitrarily small values, here we offer an alpha-liberal Monte Carlo procedure that is valid for the general case of any test statistic.

By taking $\alpha_{\text{mc}} > \alpha$, the power function of MC_m given p, evaluated for $p = \alpha$, that is, $\pi_m(\alpha_{\text{mc}}, p = \alpha)$, becomes more and more near to 1 as α_{mc} moves away from α. The combination of Figure 1(A) and (C) illustrates this behavior. Formally, from expressions (3.2) and (3.3), and for any $\alpha, \alpha_{\text{mc}} \in (0, 1)$, the power difference between the exact test and the Monte Carlo test can be expressed with the following expectation:

$$D(\alpha, \alpha_{\text{mc}}, m) = \int_0^1 \big(1_{(0,\alpha]}(p) - \pi_m(\alpha_{\text{mc}}, p)\big) F_P(dp), \tag{3.7}$$

where $1_{(0,\alpha]}(p)$ is the step function of $p \in (0, \alpha]$. Because $1_{(0,\alpha]}(p)$ is greater than or equal to $\pi_m(\alpha_{\text{mc}}, p)$ only for $p \leq \alpha$, from expression (3.3):

$$\begin{aligned}D(\alpha, \alpha_{\text{mc}}, m) &\leq \int_0^\alpha \big(1 - \pi_m(\alpha_{\text{mc}}, p)\big) F_P(dp) \\ &= \pi(\alpha) - \int_0^\alpha \pi_m(\alpha_{\text{mc}}, p) F_P(dp) \\ &\leq \pi(\alpha) - \pi(\alpha)\pi_m(\alpha_{\text{mc}}, \alpha).\end{aligned} \tag{3.8}$$

The last inequality holds because $\pi_m(\alpha_{\text{mc}}, p)$ is decreasing with p. Therefore, $D(\alpha, \alpha_{\text{mc}}, m)$ reaches its maximum value when the mass $\pi(\alpha)$ is concentrated at the point $p = \alpha$. For $\pi(\alpha) > 0$, the upper bound, $b(\alpha, \alpha_{\text{mc}}, m)$, for the relative power loss, is obtained by observing that:

$$D(\alpha, \alpha_{\text{mc}}, m)/\pi(\alpha) \leq 1 - \pi_m(\alpha_{\text{mc}}, \alpha) = b(\alpha, \alpha_{\text{mc}}, m). \tag{3.9}$$

The right-hand side of inequality (3.9) is an upper bound for $D(\alpha, \alpha_{\text{mc}}, m)$, that is, this inequality can be used to bound the potential power losses in arbitrary small values. For example, for $m = 3000$, $\alpha = 0.05$ and $\alpha_{\text{mc}} = 0.06$, from (3.9), the power loss is at most 0.783%. For $m = 10{,}000$, and $\alpha_{\text{mc}} = 0.055$, the upper bound is 1.234%. The choice of α_{mc} can be conveniently calibrated in order to use a minimum m needed to attempt an arbitrary bound. This naive approach results in a stronger statement: the power loss of conventional Monte Carlo tests with finite m are uniformly bounded by using $\alpha_{\text{mc}} = \alpha + \delta$, $\delta > 0$.

Theorem 3.3. *For any $\varepsilon, \alpha \in (0, 1)$, and $\delta \in (0, \alpha/2)$, it is always possible to find $m < \infty$ such that $D(\alpha, \alpha + \delta, m) \leq \varepsilon$.*

Proof. Let $X \sim \text{Bin}(m - 1, \alpha)$ and $Y \sim \text{Bin}(m, \alpha)$. From expression (3.4), a crude upper bound for the MC test power loss is the upper tail $\Pr(X \geq \lfloor \alpha_{\text{mc}} m \rfloor)$. Observe that, for all $a \in [0, 1, \ldots, m - 1]$, the following holds:

$$\Pr(X \geq a) \leq \Pr(Y \geq a). \tag{3.10}$$

Krafft (1969) established that:

$$\Pr\left(\left|\frac{Y}{m - \alpha}\right| \geq \delta\right) < (2/m)^{1/2} \exp\{-2m\delta^2 - 4m\delta^4/3\}/\delta,$$

where δ is a constant, $m > 2$ and $\alpha, (1 - \alpha) \geq \max\{4/m, 2\delta\}$. We can dispense the refinement of the last boundary in order to obtain a simplified inequality:

$$\Pr(Y/m - \alpha \geq \delta) < 2^{1/2} \exp\{-2m\delta^2\}/\delta.$$

Replacing $(\delta + \alpha)$ by α_{mc}, we have:

$$\Pr(Y \geq \alpha_{\text{mc}} m) \leq 2^{1/2} \exp\{-2m\delta^2\}/\delta. \tag{3.11}$$

Without loss of generality, since δ and m are arbitrary, the truncation $\lfloor \alpha_{\text{mc}} m \rfloor$ in the power loss expression, (3.4), can be ignored because δ can conveniently be replaced by some smaller rational value. Thus, from (3.10) and (3.11),

$$D(\alpha, \alpha_{\text{mc}}, m, \alpha) \leq 2^{1/2} \exp\{-2m\delta^2\}/\delta = C(\delta, m). \tag{3.12}$$

To complete the reasoning, note that $C(\delta, m)$ is decreasing with m. Thus, for $\varepsilon > 0$ and $0 < \delta \leq \alpha/2$, $C(\delta, m) \leq \varepsilon$ for $m \geq m_0(\varepsilon, \delta) = -\log(\varepsilon\delta/\sqrt{2})\delta^{-2}/2$. □

Table 1 offers the minimum m values to ensure upper bounds, ε, for the potential relative power losses of MC_m. This table contains three different scenarios. The

Table 1 *Minimum m values to bound the relative power loss of* MC_m *in comparison to the exact test of* $\alpha = 0.05$ *level*

$\varepsilon(100)\%$	$H_{0.05,0.5}(p)$	$F_w(p)$	MC_m $(\alpha_{\mathrm{mc}} = 0.06)$
5%	60	1220	1434
4.5%	60	1500	1517
4%	60	1900	1617
3.5%	80	2480	1717
3%	80	3360	1850
2.5%	100	4840	2000
2%	120	7560	2200
1.5%	160	13,440	2450
1%	240	30,240	2800
0.5%	460	120,960	3417
0.1%	2300	3,019,500	4884

first scenario is based on the worst distribution case, $H_{0.05,0.5}(p)$, from the distribution class $\Im$. The second scenario considers the concavity assumption over F_P, and it is based on the worst case, $F_w(p)$, from the class τ. The last line of the table presents the minimum m values required when the alpha-liberal MC_m approach is adopted. In this case, the values are obtained by manipulating inequality (3.9).

We see that the approach using the class $\Im$ gives the smallest m values. According to Fay and Follmann (2002), this approach is valid when U is likely to follow a standard normal distribution under H_0 and a $N(\mu, 1)$ under H_A, or a central and a non-central $\chi_1^{(2)}$ under the null and alternative hypothesis, respectively. The case $F_w(p)$ requires larger m values than the former because the concavity assumption is actually more general. Silva and Assunção (2013) offered an analytical proof showing that $\Im \subset \tau$. But, the concavity assumption can be equally problematic to check since F_P, in practice, is unknown. Further, classes $\Im$ and τ are limited to continuous cases. The last line of Table 1 presents the intermediate minimum m values. This line is based on the approach that does not require some knowledge about F_P, and holds for continuous, discrete, or mixed distributions. Additionally, we point out that the liberality from equation (3.9) can be freely calibrated in satisfactorily small values (Theorem 3.1). For example, the less liberal alpha level of 0.051 presents a bound magnitude smaller than 1% for any m greater than 260,000.

Assumptions over the shape of F_P, or usage of liberal criterions, are solutions that require small to moderate m magnitudes. But, when usage of large m is feasible due to simplicity in the simulation process, like, for instance, the conditional MaxSPRT statistics introduced by Li and Kulldorff (2009), a truncated Monte Carlo design with bounded power losses can always be provided. In the next section, we introduce our truncated sequential Monte Carlo design which, by construction, ensures small power loss upper bounds.

4 Truncated sequential Monte Carlo test with exact power: Our proposal

Sequential designs for Monte Carlo testing are usually adopted to save execution time in computational intensive applications. But, like proposed by Gandy (2009), a sequential approach can also be used in order to avoid power losses. Focused on this last goal, we introduce a truncated procedure with power satisfactorily close to the exact test. The key for the solution was finding a stopping criterion for the simulations that mimics the step function in $(0, \alpha]$ more efficiently than the conventional Monte Carlo test.

The procedure is described as following. Recall that X_l is the number of simulated values greater than or equal to the observed statistic u_0 at lth simulation. For given constants m, s, t_1 and C_e, such that $s \leq t_1 \leq m$, the simulation is interrupted if $X_l \geq s$ for some $l \leq t_1$, or if $l = m - 1$. The null hypothesis is rejected if $\psi_e = 1$, where:

$$\psi_e = \begin{cases} 0, & \text{if } X_{t_1} \geq s \text{ or } (X_{t_1} < s, X_{m-1} \geq C_e), \\ 1, & \text{if } X_{t_1} < s \text{ and } X_{m-1} < C_e. \end{cases} \tag{4.1}$$

We denote this scheme by $\mathrm{MC}_e(\alpha, s, C_e)$, and sometimes simply by MC_e.

The method is quite simple. Basically, H_0 is not rejected if X_l is greater than s until time t_1. Otherwise, not further looks at the process X_l are necessary until time $l = (m - 1)$, at which the decision rule is exactly equal to the conventional Monte Carlo test, but having C_e as critical value.

Observe that the conventional Monte Carlo test is a particular case of MC_e derived by setting $s = t_1 = m - 1$ and $C_e = \lfloor \alpha_{\mathrm{mc}} m \rfloor$.

Let t^* denote the number of simulations generated until the interruption moment. If the tuning parameters m, s, t_1 and C_e are settled according to Theorem 4.1, a valid p-value following the MC_e procedure, denoted by P_e, can be calculate as follows:

$$P_e = \begin{cases} X_{t^*}/t^*, & \text{if } t^* \leq t_1, \text{ or if } t^* = m - 1 \text{ and } X_{m-1} \geq C_e, \\ \hat{P}, & \text{if } t^* = m - 1 \text{ and } X_{m-1} < C_e, \end{cases} \tag{4.2}$$

where,

$$\hat{P} = \sum_{x=0}^{s-1} \sum_{y=0}^{y_0} c_y^{m-t_1-1} c_x^{t_1} \left[m \;\; c_{x+y}^{m-1} \right]^{-1}, \tag{4.3}$$

where $y_0 = \min\{m - t_1, X_{m-1} - 1 - x\}$.

Theorem 4.1. *The Monte Carlo p-value P_e given in (4.2) is valid if m, s, t_1 and C_e are such that $s/t_1 > \alpha$, $C_e/m > \alpha$, and*

$$\int_0^1 \Pr(\psi_e = 1 | P = p)\, dp \leq \alpha. \tag{4.4}$$

Proof. We want to prove that $\Pr(P_e \leq \alpha | H_0) \leq \alpha$. First, observe that, in general, $\Pr(\psi_e = 1|H_0) \leq \int_0^1 \Pr(\psi_e = 1|P = p)\, dp$. Thus, from (4.4), it holds that $\Pr(\psi_e = 1|H_0) \leq \alpha$. Because the choice of the tuning parameters is restricted to satisfy $s/t_1 > \alpha$ and $C_e/m > \alpha$, the event $\{P_e \leq \alpha\}$ can only happen when H_0 is rejected. But, because the probability of rejecting H_0, under H_0, is smaller than or equal to α, we conclude: $\Pr(P_e \leq \alpha | H_0) \leq \alpha$. □

Because the probability $\Pr(\psi_e = 1|P = p)$ is downward monotone with p, the power loss of MC_e can be bounded at a small constant ε by finding a parameterization that simultaneously satisfies $\Pr(\psi_e = 1|P = \alpha) \leq \varepsilon$ and $\Pr(\psi_e = 1|H_0) \leq \alpha$.

Theorem 4.2. *Consider a Monte Carlo test procedure with decision rule in the form of* (4.1). *If* m, C_e, s, *and* t_1 *are such that*:

$$\begin{cases} \Pr(\psi_e = 0|P = \alpha) \leq \varepsilon, \\ \Pr(\psi_e = 1|H_0) \leq \alpha, \end{cases} \tag{4.5}$$

then the test is of α*-level and the power loss with respect to the exact test is of at most* $(100 \times \varepsilon)\%$.

Proof. The test is of α-level, and this is obvious due to condition $\Pr(\psi_e = 1|H_0) \leq \alpha$. From (4.1), and for fixed $P = p$, the probability of rejecting H_0 with MC_e is given by:

$$\Pr(\psi_e = 1|P = p) = \Pr(X_{t_1} < s, X_{m-1} < C_e | P = p). \tag{4.6}$$

It merits remark that this probability is downward monotone with p. It is also worth noting that the exact power is the probability $\Pr(P \leq \alpha)$.

For fixed $P = p$, the difference $D_e(p)$ between the probabilities of rejecting H_0 with the exact and with MC_e, in this order, is equal to

$$\begin{aligned} D_e(p) &= \Pr(P \leq \alpha | P = p) - \Pr(\psi_e = 1|P = p) \\ &= \begin{cases} 1 - \Pr(\psi_e = 1|P = p), & \text{if } p \leq \alpha \\ -\Pr(\psi_e = 1|P = p), & \text{if } p > \alpha. \end{cases} \end{aligned} \tag{4.7}$$

Hence, the actual difference D_e between the power of exact and MC_e tests is given by the following expectation:

$$\begin{aligned} D_e &= \int_0^1 D_e(p) F_P(dp) = \int_0^1 \big[1_{(0,\alpha]} - \Pr(\psi_e = 1|P = p)\big] F_P(dp) \\ &\quad \big[\text{because } 1_{(0,\alpha]} \geq \Pr(\psi_e = 1|P = p) \text{ only if } p \leq \alpha\big] \\ &\leq \int_0^\alpha \big[1 - \Pr(\psi_e = 1|P = p)\big] F_P(dp) \end{aligned}$$

$$
\begin{aligned}
&= \Pr(P \le \alpha) - \int_0^{\alpha} \Pr(\psi_e = 1 | P = p) F_P(dp) \qquad (4.8)\\
&\quad \big[\text{because } \Pr(\psi_e = 1 | P = p) \text{ is decreasing with } p\big]\\
&\le \Pr(P \le \alpha) - \Pr(\psi_e = 1 | P = \alpha) \int_0^{\alpha} F_P(dp)\\
&= \Pr(P \le \alpha)\big[1 - \Pr(\psi_e = 1 | P = \alpha)\big].
\end{aligned}
$$

Then, the last term can be use to construct an upper bound for the percentual power difference, i.e. $100 \times \frac{D_e}{\Pr(P \le \alpha)}\%$, as follows:

$$
\begin{aligned}
\frac{100 \times D_e}{\Pr(P \le \alpha)}\% &\le \frac{100 \times \Pr(P \le \alpha)[1 - \Pr(\psi_e = 1 | P = \alpha)]}{\Pr(P \le \alpha)}\%\\
&= 100 \times \big[1 - \Pr(\psi_e = 1 | P = \alpha)\big]\%\\
&= \big[100 \times \Pr(\psi_e = 0 | P = \alpha)\big]\% \qquad (4.9)\\
\big[\text{from condition (4.5)}\big] &\le (100 \times \varepsilon)\%.
\end{aligned}
$$

In conclusion, (4.5) implies that ε is a crude upper bound for the relative power loss $\frac{D_e}{\Pr(P \le \alpha)}$ of MC_e with respect to the exact test. □

To find solutions for system (4.5) one needs to have computable expressions. For this, note that the probability of rejecting the null, for fixed $P = p$, is given by:

$$
\begin{aligned}
&\Pr(\psi_e = 1 | P = p)\\
&\quad = \Pr(X_{t_1} < s, X_{m-1} < C_e | P = p)\\
&\quad = \sum_{x=0}^{s-1} \Pr(X_{m-1} < C_e | P = p, X_{t_1} = x) \Pr(X_{t_1} = x | P = p) \qquad (4.10)\\
&\quad = \sum_{x=0}^{s-1} \sum_{y=0}^{C_e - 1 - x} c_y^{m-t_1-1} c_x^{t_1} p^{y+x} (1-p)^{m-1-y-x}.
\end{aligned}
$$

Using the fact that $\Pr(\psi_e = 0 | P = \alpha) = 1 - \Pr(\psi_e = 1 | P = \alpha)$, and using (4.10), one can easily compute the left-hand side of the first line of system (4.5) for fixed tuning parameters.

For manipulation of the second line in (4.5) note that:

$$
\begin{aligned}
\Pr(\psi_e = 1 | H_0) &\le \int_0^1 \Pr(\psi_e = 1 | P = p)\, dp\\
&= \sum_{x=0}^{s-1} \sum_{y=0}^{C_e - 1 - x} c_y^{m-t_1-1} c_x^{t_1} \big[m c_{x+y}^{m-1}\big]^{-1}. \qquad (4.11)
\end{aligned}
$$

Table 2 *Tuning parameters to use* MC_e *with guaranteed power loss upper bounds* $\varepsilon = 0.02, 0.015, 0.01, 0.005$. *It was done for significance levels* $\alpha = 0.01, 0.025, 0.05, 0.1$. *We used a precision tolerance of* 10^{-5}. *The m values are divided by* 10^6

		$(100 \times \varepsilon)\%$			
α		2%	1.5%	1%	0.5%
0.1	m	3.6	6	15	30
	s	3	2	3	3
	t_1	6	2	4	4
	C_e	361,537	601,903	1,502,900	3,004,991
0.05	m	7	20	25	30
	s	3	3	2	2
	t_1	12	10	3	2
	C_e	351,969	1,003,060	1,253,025	1,503,450
0.025	m	6.7	8.65	31	200
	s	2	2	2	3
	t_1	7	7	6	10
	C_e	168,481	217,516	777,648	5,006,000
0.01	m	9	40	100	200
	s	1	4	4	4
	t_1	2	48	40	35
	C_e	91,150	401,720	1,002,929	2,005,859

Numerical solutions for system (4.5) can be reached by fixing two of the four tuning parameters (m, s, t_1, C_e), and optimizing over the other two. Here we provide a set of solutions for meaningful values of α and ε, which are available in Table 2. An efficient strategy to find solutions for system (4.5) is based on searching for values of C_e near to $\lfloor \alpha(m+1) \rfloor$. The intuition is that the solution cannot differ much from the critical value of the conventional Monte Carlo test. Solutions will generally be found for small values of s and t_1, and for intuitive reasons, thinner bounds are related to larger m values. Parameterizations for $\alpha = 0.01, 0.025, 0.05, 0.1$ and $\varepsilon = 0.03, 0.02, 0.015, 0.01, 0.005$, are offered in Table 2. Observe the million scale of m. For example, for $\alpha = 0.05$, it is guaranteed at least 99% of the exact test power if $s = 2$, $t_1 = 3$, $m = 25 \times 10^6$, and $C_e = 1{,}253{,}025$. It merits mention that all solutions shown in this table attend the conditions of Theorem 4.1, hence, the p-value P_e is valid to be use with any of these tuning parameterizations.

4.1 A refinement for the stopping boundaries of MC_e

A refinement for MC_e, in the spirit of saving computation time, can be obtained by generalizing system (4.5). To do so, consider to perform additional tests while the simulation runs. For given constants s_l, i_l, $l \in \{l_2, \ldots, l_h\}$, where h represents the total number of sequential tests to be performed, the simulation process is interrupted as soon as $X_l \geq s_l$ or $X_l < i_l$. Denote the p-value followed by this

refined procedure with P_r, and let l^* be the value of l in the interruption moment. The null hypothesis is rejected if $\psi_l = 1$, where:

$$\psi_l = \begin{cases} 0, & \text{if } P_r \geq \alpha, \\ 1, & \text{if } P_r \leq \alpha, \end{cases}$$

and P_r is defined as:

$$P_r = \begin{cases} X_{l^*}/l^*, & \text{if } X_{l^*} \geq s_l, \\ \hat{P}_i, & \text{if } X_{l^*} < i_{l^*}, \end{cases} \tag{4.12}$$

where $\hat{P}_i = \sum_{x=0}^{s_{l_1}-1} \sum_{y=0}^{y_0} \frac{c_y^{l^*-l_1} c_x^{l_1}}{c_{x+y}^{l^*}(l^*-x-y)}$, and $y_0 = \min\{l^* - l_1, X_{l^*} - 1 - x\}$. The term l_1 represents the minimum l for which $\Pr(X_l \geq s_l, X_{m-1} < C_e | P = \alpha) > 0$.

We call this procedure MC_r. Recall the notation for MC_e. By setting $l_1 = t_1$ and $s_{l_1} = s$, a valid MC_r parameterization to bound the overall power loss at ε_2 can be obtained by solving the following system:

$$\begin{cases} \sum_{l=1}^{m-1} \Pr(X_l < i_l, X_{m-1} \geq C_e | H_0) \leq \varepsilon_1, \\ \sum_{l=2}^{m-1} \Pr(X_l \geq s_l, X_{m-1} < C_e | P = \alpha) \leq \varepsilon_2, \end{cases} \tag{4.13}$$

where ε_1 is $\Pr(X_{l_1} \geq s_{l_1}, X_{m-1} < C_e | H_0)$.

In practice, there is no need to solve system (4.13) in order to find s_l and i_l. For an observed realization of X_l, say x_l, the procedure has exactly the same effect over the errors of Type I and II if we track the measures:

$$\begin{cases} \mathbb{P}_s(x_l) = \Pr(X_l \geq x_l, X_{m-1} < C_e | P = \alpha), \text{ and} \\ \mathbb{P}_i(x_l) = \Pr(X_l \leq x_l, X_{m-1} \geq C_e | H_0), \end{cases} \tag{4.14}$$

using two flat thresholds, δ and η, where $\delta = (\varepsilon_2 - \varepsilon)/h$, and $\eta = \varepsilon_1/h$, with $l = l_2, \ldots, l_h$. Then, the simulations are interrupted, and H_0 is not rejected, for the first l such that $\mathbb{P}_s(x_l) < \delta$, situation where the p-value is x_l/l, or, the simulations are stopped for the rejection of the null if $\mathbb{P}_i(x_l) < \eta$ occurs before $\mathbb{P}_s(x_l) < \delta$, where the p-value is $\hat{P}_i$. The parameters ε, l_1 and s_{l_1} can be fixed directly from the MC_e procedure by setting $\varepsilon = \Pr(X_{l_1} \geq s_{l_1}, X_{m-1} < C_e | P = \alpha)$, $l_1 = t_1$, $s_{l_1} = s$, and $\varepsilon_1 = \varepsilon$.

The gain with this refinement is more substantial when H_0 is true. In cases where p is close to α, $(m-1)$ simulations will typically be required. For simplicity, we suggest usage of small number of interim sequential Monte Carlo tests, such as h equal to 5 or 10. To help with this decision, for all the designs of Table 2, use $h = 5$, $l_1 = t_1$, $l_2 = 2000$, $l_3 = 5000$, $l_4 = 20{,}000$, $l_5 = 1{,}000{,}000$. Evidently, some designs of Table 2 have m smaller than 1,000,000, i.e, for those cases the number of sequential tests is 4.

Table 3 *Tuning parameters to use the refinement*, MC_r, *with guaranteed power loss upper bounds* $\varepsilon = 0.02, 0.015, 0.01, 0.005$, *and significance levels* $\alpha = 0.01, 0.025, 0.05, 0.1$

		$\varepsilon(100)\%$			
α		2%	1.5%	1%	0.5%
0.1	δ	0.00018	0.00005	5.2×10^{-6}	0.00003
	η	0.00482	0.00369	0.00249	0.00122
0.05	δ	0.00009	0.00066	0.00002	0.00014
	η	0.00497	0.00309	0.00250	0.00111
0.025	δ	0.00009	4.67×10^{-6}	0.00002	4.67×10^{-5}
	η	0.00491	0.00374	0.00248	0.00123
0.01	δ	10^{-5}	0.00262	0.00192	0.00114
	η	0.00499	0.00113	0.00058	0.00011

Example 1. Accordingly to Table 2, set $\alpha = 0.05$, $m = 30 \times 10^6$, $s = 2$, $t_1 = 2$, and $C_e = 1{,}503{,}450$. This specific MC_e design leads to a power loss upper bound equal to $\varepsilon = \Pr(X_2 \geq 2, X_{m-1} < 1{,}503{,}450 | P = 0.05) = 0.00443$. Using the refinement through MC_r in order to save computation time, from Table 3, a global power loss bound of 0.5% ($\varepsilon_2 = 0.005$), from MC_r, is obtained by setting $\delta = (0.005 - 0.00443)/4 = 0.0001425$, and $\eta = 0.00443/4 = 0.0011075$.

Here follows a friendly way of expressing the monitoring measures $\mathbb{P}_s(x_l)$ and $\mathbb{P}_i(x_l)$:

$$\begin{cases} \mathbb{P}_s(x_l) = 1 - \displaystyle\sum_{x=0}^{C_e-1} \big[\Pr(W \leq x_l - 1)\Pr(Y = x)\big], \\ \mathbb{P}_i(x_l) = \displaystyle\sum_{x=0}^{C_e-1} \Pr(W \leq x_l) \Big/ m, \end{cases} \tag{4.15}$$

where W is hypergeometric-distributed with parameters x, $m - x$, and l ($W \sim \mathrm{Hyp}(x, m - x, l)$), i.e., $\Pr(W = w) = \frac{c_w^x c_{l-w}^{m-x}}{c_l^m}$, and $Y \sim \mathrm{Bin}(m, \alpha)$.

For the cases where H_0 is rejected, here follows a simplified expression for calculating the p-value. If l^* is the value of l at which the simulation process is interrupted, then $\hat{P}_i = \sum_{x=0}^{x_{l^*}} \Pr(W^* \leq s - 1)/(l^* + 1)$, with $W^* \sim \mathrm{Hyp}(x, l^* - x, l_1)$.

5 Last comments

There is no doubt that Monte Carlo simulation technics are extensively used for statistical research and inferential analyses. Its status as a competitor for asymptotic tests was anticipated by Jockel (1984). This opinion is also supported by the

results in the present work, which basically offers a Monte Carlo test procedure with controlled power losses relatively to the exact test. This power loss control is obtained for arbitrary and valid significance levels, what is not always possible with the asymptotic treatment. Therefore, when simulating the test statistic under the null hypothesis is a feasible option, there is no reason to deliberately use asymptotic approximations in place of a Monte Carlo approach to get a valid p-value. This result is especially important in the cases where sample sizes are small, that is, cases where asymptotic treatments are certainly biased.

Appendix: Rule of thumb for the number of simulations

Here we prove Theorem 3.1, that is, for the conventional Monte Carlo test, and if statistical power is of meaningful concern, the pre-defined number of simulations, m, must be a multiple of $\lfloor 1/\alpha_{\text{mc}} \rfloor$. To demonstrate this, it is sufficient to show that the Monte Carlo test power, $\pi_m(\alpha_{\text{mc}})$, is non-increasing with m for $\lfloor (j-1)/\alpha_{\text{mc}} \rfloor < m < \lfloor j/\alpha_{\text{mc}} \rfloor$, with $j > 1$ an integer.

Proof of Theorem 3.1. When applying the conventional Monte Carlo test, one rejects H_0 if, among $(m-1)$ simulated u_i's, the number of values greater than or equal to u_0, X_{m-1}, is not greater than $\lfloor \alpha_{\text{mc}} m \rfloor - 1$. Obviously, that requires $m \geq 1/\alpha_{\text{mc}}$, because, otherwise, H_0 is never rejected. Consider two Monte Carlo tests which differ from the number of simulations, $(m-1)$ and $(m+k-1)$, $k > 0$. For the first design, based on m, we reject H_0 only if X_{m-1} is at most $\lfloor \alpha_{\text{mc}} m \rfloor - 1$, whereas, for the second test, with $(m+k-1)$ simulations, H_0 is rejected if such number is at most $\lfloor \alpha_{\text{mc}}(m+k) \rfloor - 1$. According to (3.1), for any observed $P = p$, the power of the second test is greater than that from the first test if

$$\sum_{y=0}^{h_1} c_y^{m+k-1} p^y (1-p)^{(m+k-1)-y} > \sum_{y=0}^{h_2} c_y^{m-1} p^y (1-p)^{(m-1)-y}, \tag{A.1}$$

where $h_1 = \lfloor \alpha_{\text{mc}}(m+k) \rfloor - 1$, and $h_2 = \lfloor \alpha_{\text{mc}} m \rfloor - 1$.

Observe that inequality (A.1) is equivalent to

$$\Pr\big(X \leq \lfloor \alpha_{\text{mc}}(m+k) \rfloor - 1\big) > \Pr\big(Y \leq \lfloor \alpha_{\text{mc}} m \rfloor - 1\big),$$

where $X \sim \text{Bin}(m+k-1, p)$ and $Y \sim \text{Bin}(m-1, p)$. If $0 < k < 1/\alpha$, then

$$\Pr\big(X \leq \lfloor \alpha_{\text{mc}}(m+k) \rfloor - 1\big) = \Pr\big(X \leq \lfloor \alpha_{\text{mc}} m \rfloor - 1\big),$$

and the inequality (A.1) becomes

$$\Pr\big(X \leq \lfloor \alpha_{\text{mc}} m \rfloor - 1\big) > \Pr\big(Y \leq \lfloor \alpha_{\text{mc}} m \rfloor - 1\big),$$

which is not valid since X is binomial with larger number of trials, $(m+k-1)$, than Y, with $(m-1)$. Thus, because these arguments hold for any $p \in (0,1)$, if $0 < k < 1/\alpha$, then $\pi_{m+k}(\alpha_{\text{mc}}) \leq \pi_m(\alpha_{\text{mc}})$. □

Acknowledgments

We are very grateful for the rich writing suggestions from Anandra Santos Ribeiro de Oliveira.

References

Armitage, P. (1958). Numerical studies in the sequential estimation of a binomial parameter. *Biometrika* **45**, 1–15. MR0092321

Barnard, G. A. (1963). Discussion of professor Bartlett's paper. *Journal of the Royal Statistical Society* **25B**.

Belongia, E. A., Irving, S. A., Shui, I. M., Kulldorff, M., Lewis, E., Yin, R., Lieu, T. A., Weintraub, E., Yih, W. K., Li, R., Baggs, J. and the Vaccine Safety Datalink Investigation Group (2010). Real-time surveillance to assess risk of intussusception and other adverse events after pentavalente, bovine-derived rotavirus vaccine. *Pediatric Infectious Disease Journal* **29**, 1–5.

Besag, J. and Clifford, P. (1991). Sequential Monte Carlo p-value. *Biometrika* **78**, 301–304. MR1131163

Birnbaum, Z. W. (1974). Computers and unconventional test-statistics. In *Reliability and Biometry* (F. Proschan and R. J. Serfling, eds.) 441–458. MR0362638

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* **28**, 181–187. MR0087280

Fay, M. P. and Follmann, D. A. (2002). Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests. *The American Statistician* **56**, 63–70. MR1944904

Fay, M. P., Kim, H.-J. and Hachey, M. (2007). On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests. *Journal of Computational and Graphical Statistics* **16**, 946–967. MR2412490

Gandy, A. (2009). Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association* **104**, 1504–1511. MR2750575

Gates, J. (1991). Exact Monte Carlo tests using several statistics. *Journal of Statistical Computation and Simulation* **38**, 211–218.

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society* **30B**, 582–598.

Jockel, K. H. (1984). Application of Monte-Carlo tests—some considerations. *Biometrics* **40**, 263.

Jockel, K. H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *The Annals of Statistics* **14**, 336–347. MR0829573

Klein, N. P., Fireman, B., Yih, W. K., Lewis, E., Kulldorff, M., Ray, P., Baxter, R., Hambidge, S., Nordin, J., Naleway, A., Belongia, E. A., Lieu, T., Baggs, J., Weintraub, E. and the Vaccine Safety Datalink (2010). Measles-mumps-rubella-varicella combination vaccine and risk of febrile seizures. *Pediatrics* **126**, e1–e8.

Krafft, O. (1969). A note on exponential bounds for binomial probabilities. *Annals of the Institute of Statistical Mathematics, Tokyo* **21**, 219–220.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of Royal Statistical Society* **164A**, 61–72. MR1819022

Kulldorff, M., Davis, R. L., Margarette, K., Lewis, E., Lieu, T. and Platt, R. (2011). A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential Analysis* **30**, 58–78. MR2770706

Li, L. and Kulldorff, M. (2009). A conditional maximized sequential probability ratio test for pharmacovigilance. *Statistics in Medice* **29**, 284–295. MR2750517

Silva, I. and Assunção, R. (2013). Optimal generalized truncated sequential Monte Carlo test. *Journal of Multivariate Analysis* **121**, 33–49. MR3090467

Silva, I., Assunção, R. and Costa, M. (2009). Power of the sequential Monte Carlo test. *Sequential Analysis* **28**, 163–174. MR2518828

Smith, P. (1996). Monte Carlo exact tests for log-linear and logistic models. *Journal of the Royal Statistical Society B*. To appear.

Smith, P. W. F., Forster, J. J. and McDonald, J. W. (1996). Monte Carlo exact tests for contingency tables. *Journal of the Royal Statistical Society A* **159**, 309–321.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16**, 117–186. MR0013275

Yih, W. K., Nordin, J. D., Kulldorff, M., Lewisc, E., Lieua, T. A., Shia, P. and Weintraube, E. S. (2009). An assessment of the safety datalink of adolescent and adult tetanus-diphteria-acellular pertussis (tdap) vaccine, using active surveillance for adverse events in the vaccine safety datalink. *Vaccine* **27**, 4257–4262.

Department of Statistics
Federal University of Ouro Preto
Campos Universitário Morro do Cruzeiro
Ouro Preto 35400 000
MG, Brazil
E-mail: ivairest@gmail.com

Departament of Computational Science
Federal University of Minas Gerais
Belo Horizonte
MG, Brazil
E-mail: assuncao@dcc.ufmg.br