Power of the Sequential Monte Carlo Test

Ivair Silva[1], Renato Assunção[1], and Marcelo Azevedo[1]

Departamento de Estatistica, Universidade Federal de Minas Gerais,
Av. Antônio Carlos, 6627 - Pampulha - Belo Horizonte - MG,Brazil
CEP 31270-901 Fone: +5531 3409.5000 Fax: +5531 3409.4188
irs@ufmg.br, assuncao@est.ufmg.br, azevedo@est.ufmg.br

Abstract. Many statistical tests obtain their p-value from a Monte Carlo sample of m values of the test statistic under the null hypothesis. The number m of simulations is fixed by the researcher prior to any analysis. In contrast, the sequential Monte Carlo test does not fix the number of simulations in advance. It keeps simulating the test statistics until it decides to stop based on a certain rule. The final number of simulations is a random number N. This sequential Monte Carlo procedure can decrease substantially the execution time in order to reach a decision. This paper has two aims concerning the sequential Monte Carlo tests: to minimize N without affecting its power; and to compare its power with that of the fixed-sample Monte Carlo test. We show that the power of the sequential Monte Carlo test is constant after a certain number of simulations and therefore, that there is a bound to N. We also show that the sequential test is always preferable to a fixed-sample test. That is, for every test with a fixed sample size m there is a sequential Monte Carlo test with equal power but with smaller number of simulations.

Keywords. Monte Carlo test; P-value; Sequential estimation; Sequential test; Significance test.

1 INTRODUCTION

To carry out hypothesis testing one must find the distribution of the test statistic U under the null hypothesis, from which the p-value is calculated. Either because it is too cumbersome or it is impossible to obtain this distribution analytically, Monte Carlo tests are used in many situations (Manly, 2006). In particular, areas such as spatial statistics (Assunção *et al.*, 2007; Diggle *et al.*, 2004; Kulldorff, 2001) and data mining (Kulldorff *et al.*, 2003; Rolka, 2007) rely heavily on Monte Carlo tests to draw inference. Other areas have situations in which Monte Carlo tests seems to be the best current approach such as the exact tests in categorical data analysis (Booth and Butler, 1999; Caffo and Booth, 2003), and some regression problems in econometrics (Khalaf and Kichian, 2005; Luger, 2006).

The conventional Monte Carlo test generates a large number of independent copies of U from the null distribution. Assuming that large values of U lead to the null hypothesis rejection, a Monte Carlo value is calculated based on the proportion of the simulated values that are larger or equal than the observed value of U.

As the statistics field evolves to deal with ever more complex models, Monte Carlo tests become costly. The simulation of each independent copy of U under the null hypothesis can take a long time.

In many applications, after a few simulations are carried out, it becomes intuitively clear that a large number of simulations is not necessary. For instance, suppose that after 100 simulations, the observed value is around the median of the generated values. It is not likely that the null hypothesis will be eventually rejected even if a much larger number of simulations (such as 9999) is carried out. Most researchers would be confident to stop at this point if a valid p-value could be provided.

Besag and Clifford (1991) introduced the idea of sequential Monte Carlo tests, an alternative way to obtain p-values without fixing the number of simulations previously. Their method makes a decision concerning the null hypothesis after each simulated value up to a maximum number of simulations. This approach can substantially shorten the number of simulations required to decide about the significance of the observed test statistic.

Although the proposal of Besag and Clifford (1991) stands as a major contribution to the practice of modern data analysis, it is under utilized and has some unanswered theoretical questions. One important aspect of the sequential Monte Carlo tests is the relative comparison of its power with that of the conventional Monte Carlo test. Based on the Besag and Clifford results, we can always obtain a sequential Monte Carlo with significance level α which does not require more simulations than a conventional Monte Carlo test at the same level. However, the relationship between the power functions of these tests is not clear. In terms of power, is there a cost when we apply the sequential test instead of the

conventional Monte Carlo test? The answer is no, and the first aim of this paper is to demonstrate this. The second objective of this work is to show how we can make the choice of the maximum number of simulations in the sequential Monte Carlo tests without losing power.

The next section contains a summary of the definitions and notation associated with the conventional and the sequential Monte Carlo tests. Section 3 discusses the power of the sequential procedure and Section 4 shows how to establish the parameters of the sequential test such that it has the same power as a given conventional Monte Carlo test. In Section 5, we develop bounds for the difference of power between a conventional and a sequential Monte Carlo tests. Section 6 closes the paper with a discussion of the implications of our results.

2 A SEQUENTIAL MONTE CARLO TEST

Let U be a test statistic with distribution F under the null hypothesis H_0. Suppose that large values of U leads to the rejection of the null hypothesis. When F can be evaluated explicitly, the p-value of the upper-tail test based on the observed value u_0 of U is given by $p = 1 - F(u_0)$. Let $P = 1 - F(U)$ be the random variable associated with the p-value. If F is a continuous function, P has a uniform distribution in $(0, 1)$ under the null hypothesis. When we can not evaluate F, we need to find other ways to calculate the p-value. The Monte Carlo test proposed by Dwass (1957) is an alternative if we can simulate samples from the null hypothesis.

The fixed-size or conventional Monte Carlo test generates a sample of size $m - 1$ of the test statistic U under the null hypothesis H_0. Denote each simulated value by u_i, $i = 1, \ldots, m - 1$. The Monte Carlo p-value p_{mc} is equal to r/m if the observed value u_0 is the r-th largest value among the m values $u_0, u_1, \ldots, u_{m-1}$. In this conventional Monte Carlo procedure, if the rank of u_0 is among the αm larger ranks of $u_0, u_1, \ldots, u_{m-1}$, we reject the null hypothesis at the α significance level. We denote this procedure by $MCconv(m, \alpha)$.

Let P_{mc} be the corresponding random variable associated with the realized Monte Carlo p-value p_{mc}. Under the null hypothesis, we have $\mathbb{P}(P_{mc} \leq a) = a$ if a is one of the values $1/m, 2/m, \ldots, 1$. In addition to that, irrespective of the validity of the null hypothesis, $P_{mc} \to P$ almost everywhere as m goes to infinity.

However, when early on there is little evidence against the null hypothesis, it is wasteful to run the procedure for large values of m such as, for example, $m = 10000$. This is the main motivation for Besag and Clifford to develop the sequential Monte Carlo test. In brief, the sequential version of the test selects a small integer h, such as $h = 10$ or $h = 20$. It keeps simulating by Monte Carlo from the null hypothesis distribution until h of the simulated values are larger than the observed value u_0. There is also an upper limit $n - 1$ for the total number of simulations. The p-value is based on the proportion of simulated values larger than u_0 at the stopping time.

In other words, simulate independently and sequentially the random values $U_1, U_2, \ldots, U_L$ from the same distribution as U under the null hypothesis. The random variable L has possible values h, $h + 1, \ldots, n - 1$ and its value is determined in the following way: L is the first time when there are h simulated values larger than u_0. If this has not occurred at step $n - 1$, then let $L = n - 1$. Let g be the number of simulated U_i's larger than u_0 at termination. If we denote by l the realized number of Monte Carlo withdrawals, then the sequential p-value is given by

$$p_s = \begin{cases} h/l, & \text{if } g = h, \\ (g + 1)/n, & \text{if } g < h \end{cases} \tag{2.1}$$

For example, if up to $n-1 = 999$ Monte Carlo withdrawals are considered and the sampling scheme stops as soon as $h = 10$ exceeding values of U occurs, then the possible values of the sequential p-value are $10/10$, $10/11$, $10/12, \ldots, 10/1000$, $9/1000$, $8/1000, \ldots, 1/1000$.

The most important random variable in our paper is L, the total number of simulations carried out, which has distribution under the null hypothesis given by

$$\mathbb{P}(L \leq l) = \begin{cases} 0, & \text{se } l \leq h-1 \\ 1 - h/(l+1), & \text{se } l = h, h+1, ..., n-1 \\ 1, & \text{if } l = n \end{cases}$$

Its expected value was found by Besag and Clifford(1991):

$$\mathbb{E}(L) = \sum_{l=1}^{n-1} P(L \geq l) = \sum_{l=h+1}^{n-1} l^{-1} \cong h + h \log\left(\frac{n-0.5}{h+0.5}\right) \tag{2.2}$$

To reach a decision with the sequential Monte Carlo test, it is necessary to fix the values of three tuning parameters, the significance level n, h e α, and hence we denote the test by $MCseq(n, h, \alpha)$. Typically, n is taken equal to the number m of simulations one would run if carrying out the conventional Monte Carlo test. If this typical choice is really necessary is one of the issues studied in this paper.

3 POWER OF THE SEQUENTIAL MONTE CARLO TEST

In this section we study the power of the sequential Monte Carlo procedure $MCseq(n, h, \alpha)$. Its behavior depends on the value of n with respect to $h/\alpha + 1$. We deal initially with the case $n \geq h/\alpha + 1$.

3.1 $MCseq(n, h, \alpha)$ with $n \geq h/\alpha + 1$

This constraint implies that $\alpha \geq h/(n-1)$. That is, α is not smaller than h divided by the maximum number of simulations. A typical choice found in practical analysis is $n - 1 = 999$ and $\alpha = 0.05$. Then, the condition $n \geq h/\alpha + 1$ is valid if $h \leq 49$. This is likely to cover most of the choices one would make for h in practice.

The power of the procedure $MCseq(n, h, \alpha)$ is constant for all $n \geq h/\alpha + 1$ and hence, taking n larger than $h/\alpha + 1$ is not worth in terms of power. In other words, $n = \lfloor h/\alpha \rfloor + 1$ is optimal in terms of number of simulations for a test with error type I probability α. The notation $\lfloor x \rfloor$ represents the ceiling of x, the smallest integer greater or equal to x.

To see this result, label the event $[U_i \geq u_0]$ as a success. Since U_i has c.d.f F, the probability $\mathbb{P}(U_i \geq u_0)$ is the observed p-value $p = 1 - F(u_0)$. The probability of carrying out L simulations until h successes is a probability function from a truncated multinomial variable on $n, n+1, ...$, then, we reject H_0 if, and only if, $h/\alpha \leq L \leq n-1$. This means that, in $\lfloor h/\alpha \rfloor - 1$ simulations we obtain at most $h - 1$ successes. Therefore, for an observed value u_0, the probability of rejecting H_0 in the sequential test is given by

$$\mathbb{P}(L \geq (h/\alpha) \mid P = p) = \sum_{x=0}^{h-1} \binom{h/\alpha - 1}{x} p^x (1-p)^{h/\alpha - x - 1} \tag{3.1}$$

Since the last expression does not involve n, the power of the sequential Monte Carlo test is constant as long as $n \geq (h/\alpha) + 1$. Since the error type I is fixed at α, $\lfloor h/\alpha \rfloor + 1$ is an upper bound for n.

For example, if $h = 5$ and $\alpha = 0.05$, then $n = 101$ minimize the sampling effort while holding constant the test power. It is not worth to select a larger sample size such as, for example $n = 1000$, expecting to have a better test. Using (2.2), we know that $\mathbb{E}(L) \approx 19$ if $n = 101$ under the null hypothesis. If one decides to use $n = 1000$ then $\mathbb{E}(L) \approx 31$, 50% larger compared with that associated with optimal n. However, the more substantial gain of using the optimal n is when the null hypothesis is false. In this situation, it is more probable that we need to run the sequential test up to the maximum number $n - 1$ of simulations and then choosing $n = 101$ will save many simulations compared with the larger sample size $n = 1000$, which does not increase the power.

3.2 $MCseq(n, h, \alpha)$ with $n < h/\alpha + 1$

The power of the procedure $MCseq(n, h, \alpha)$ do not have a monotone behavior with the increase of n when it is in the range $h + 1 < n < h/\alpha + 1$. In fact, at least in principle, the power can have a non-monotone behavior as n increases from $h + 1$ towards the $h/\alpha + 1$. However, the most usual behavior is that the power is an increasing function of n, for n in that range.

To understand this limitation of the analysis, let us assume that $n < h/\alpha + 1$. We have two possible evaluations of the sequential p-value depending on the value of g, according to (2.1). Hence, we reject the null hypothesis either when estimating the p-value by g/l or when estimating the p-value by $(g + 1)/n$.

However, we can never reject the null hypothesis if the p-value p_s is of the form g/l. The reason is that, if $p_s = g/l$, then we obtained h values exceeding u. The smallest value for g/l is $h/(n - 1)$. Since $n < h/\alpha + 1$, we have that $p_s \geq h/(n - 1) > \alpha$ and we can not reject the null hypothesis.

Therefore, the only other possibility to reject the null hypothesis when $n < h/\alpha + 1$ is when p_s is of the form $(g + 1)/n$. In this case, we need $(g + 1)/n \leq \alpha$, or $g \leq n\alpha - 1$. Given that $P = p$, the probability of rejecting H_0 is equal to

$$\mathbb{P}(G \leq n\alpha - 1 \mid P = p) = \sum_{x=0}^{\lfloor n\alpha \rfloor - 1} \binom{n-1}{x} p^x (1-p)^{n-1-x} \tag{3.2}$$

The power for $n < h/\alpha + 1$ is given by integrating out (3.2) with respect to the p-value probability distribution F_P:

$$\pi(n, h, \alpha, F_P) = \int_0^1 \sum_{x=0}^{\lfloor n\alpha \rfloor - 1} \binom{n-1}{x} p^x (1-p)^{n-1-x} F_P(dp) \tag{3.3}$$

Denote by $\pi(n, h, \alpha, F_P)$ the power function of teh sequential procedure. Depending on F_P, the power curve can be non-monotone. Therefore, for $n < h/\alpha + 1$, the sequential power behavior depends heavily on the shape of the P-value density.

4 A SEQUENTIAL MC TEST EQUIVALENT TO A FIXED-SIZE MC TEST

From now on, we consider only the case $n \geq h/\alpha + 1$. Given a conventional Monte Carlo test $MCconv(m, \alpha)$, we find in this section a sequential test $MCseq(n, h, \alpha)$ with the same power as the conventional one. For the fixed-size Monte Carlo test, let G be the random count of U_is that are greater or equal to u_0 among the $m - 1$ generated. The null hypothesis is rejected if $(G + 1)/m \leq \alpha$ or, equivalently, if $G \leq \alpha m - 1$. The random variable G has a binomial distribution with parameters $m - 1$ and success probability equal to the p-value p. Therefore, $MCconv(m, \alpha)$ rejects the null hypothesis with probability $P(G \leq \lfloor \alpha m \rfloor - 1 \mid P = p)$:

$$\mathbb{P}(\text{Reject } H_0 \mid P = p) = \sum_{y=0}^{\lfloor \alpha m \rfloor - 1} \binom{m-1}{y} p^y (1-p)^{m-y-1} \tag{4.1}$$

Then, the $MCconv(m, \alpha)$ power is

$$\pi(m, \alpha, F_P) = \int_0^1 \sum_{y=0}^{\lfloor \alpha m \rfloor - 1} \binom{m-1}{y} p^y (1-p)^{m-y-1} \, F_P(p) \, dp \tag{4.2}$$

while the $MCseq(n, h, \alpha)$ power for $n > h/\alpha + 1$ is given by integrating out (3.1) with respect to F_P:

$$\pi(n, h, \alpha, F_P) = \int_0^1 \sum_{x=0}^{h-1} \binom{h/\alpha - 1}{x} p^x (1-p)^{h/\alpha - x - 1} \, F_P(p) \, dp \tag{4.3}$$

As a result, the power (4.3) of $MCseq(n, h, \alpha)$ and the power (4.2) of $MCconv(m, \alpha)$ are equal if we take $h = \alpha m$. That is, given a conventional MC procedure $MCconv(m, \alpha)$, we have sequential MC procedure in $MCseq(n, \alpha m, \alpha)$ with equal power. This is valid for all $n > h/\alpha + 1$ and hence we take the minimum possible value $n = \lfloor h/\alpha \rfloor + 1$ to have the equivalent procedures $MCconv(m, \alpha)$ and $MCseq(m + 1, \alpha m, \alpha)$.

Under the null hypothesis or under an alternative not too far from the null, there will be considerable reduction in the number of simulations required to reach a decision if the sequential test is adopted holding fixed the main statistical characteristic (size and power) of the fixed-size MC tests. Therefore, we can have large gains if the sequential procedure is adopted.

We showed that, given a conventional MC test, there is a simple rule to find a sequential MC test with the same power but typically requiring a smaller number of simulations. However, one can trade a slight power loss in exchange for a smaller number of Monte Carlo simulations. If we want to adopt a general sequential MC test rather than the fixed-size MC test, it is important to have control over the power loss we are subjected. The next section establishes bounds for this loss.

5 BOUNDS ON THE POWER DIFFERENCES

Equations (3.1) and (4.1) give the null hypothesis rejection probability for $MCseq(n, h, \alpha)$ and $MCconv(m, \alpha)$ for a fixed realized p-value $P = p$. Since it is wasteful to take n larger than $h/\alpha + 1$, we assume that n is equal to $\lfloor h/\alpha \rfloor + 1$. To obtain the power, we need to integrate (3.1) and (4.1) with respect to the probability density $f_P(p)$ of P. Under the null hypothesis, $f_P(p)$ is the density of an uniform distribution in $(0, 1)$. Under an alternative hypothesis, $f_P(p)$ is concentrated towards the lower half of the interval $(0, 1)$.

Let $D(P)$ be the random variable

$$D(P) = \sum_{y=0}^{\alpha m-1} \binom{m-1}{y} P^y (1-P)^{m-y-1} - \sum_{x=0}^{h-1} \binom{\lfloor h/\alpha_2 \rfloor - 1}{x} P^x (1-P)^{\lfloor h/\alpha_2 \rfloor - x - 1} \tag{5.1}$$

The power difference between $MCconv(m, \alpha)$ and $MCseq(\lfloor h/\alpha \rfloor + 1, h, \alpha)$ is given by

$$E\left[D(P)\right] = \int_0^1 D(P)\ f_P(p)\ dp \tag{5.2}$$

A crude bound for the difference in power is obtained by finding real numbers a and b such that $a \leq D(P) \leq b$. Let $b(m, \alpha; h, \alpha_2)$ be the upper bound for the power difference between $MCconv(m, \alpha)$ and $MCseq(\lfloor h/\alpha_2 \rfloor + 1, h, \alpha_2)$, respectively. Note that we can obtain crude bounds for $\alpha \neq \alpha_2$.

For example, The power difference between $MCconv(1000, 0.05)$ and $MCseq(801 = 40/0.05 + 1, 40, 0.05)$ is approximately 0.0296.

For small h, the crude bound is too large. However, this bound decreases quickly with h until reach zero in $h/\alpha + 1$.

6 DISCUSSION AND CONCLUSIONS

The sequential Monte Carlo test is a feasible and more economical way to reach decisions in a hypothesis testing under Monte Carlo sampling. We have shown that, for each conventional Monte Carlo test with m simulations, there is a sequential Monte Carlo procedure with the same significance level, power and execution time. The number of simulations is generally much smaller than m when the null hypothesis is true.

If execution time is crucial, the user can trade a small amount of power in the sequential test by a large decrease in number of simulations. To guide this trade-off choice, we develop bounds for the difference in power between the $MCconv$ e $MCseq$ tests. For $n \geq h/\alpha + 1$, an usual situation, the sequential MC test has a constant power and this leads to the suggestion of adopting $n = h/\alpha + 1$.

ACKNOWLEDGEMENTS

We are grateful to Martin Kulldorff for very useful comments and suggestions on an earlier draft of this paper. This research was partially funded by the National Cancer Institute, grant number RO1CA095979, Martin Kulldorff PI. The second author was partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). This research was partially carried out while the first author was at the Department of Ambulatory Care and Prevention, Harvard Medical School, whose support is gratefully acknowledged.

The first author received partial support from FAPEMIG, Fundação de Amparo à Pesquisa do Estado de Minas Gerais.

References

Assunção, R., Tavares, A. I., Correa, T., and Kulldorff M. (2007). Space-time cluster identification in point processes, *Canadian Journal of Statistics* 35: 9–25.

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

Booth, J. G. and Butler, R. W. (1999). An importance sampling algorithm for exact conditional tests in log-linear models, *Biometrika* 86: 321–332.

Caffo, B. S. and Booth, J. G. (2003). Monte Carlo conditional inference for log-linear and logistic models: a survey of current methodology, *Statistical Methods in Medical Research* 12: 109–123.

Diggle, P. J., Zheng, P., and Durr, P. A. (2005). Non-parametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK, *Journal of Royal Statistical Society*, Series C 54: 645–658.

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

Hope, A. C. A. (1968). A simplified Monte Carlo Significance Test Procedure, *Journal of Royal Statistical Society*, Series B 30: 582–598.

Jockel, K-H. (1986). Finite Sample Properties and Asymptotic Efficiency of Monte Carlo Tests, *Annals of Statistics* 14: 336–347.

Khalaf, L. and Kichian, M. (2005). Exact tests of the stability of the Phillips curve: the Canadian case, *Computational Statistics & Data Analysis* 49: 445–460.

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

Kulldorff M., Fang, Z., and Walsh, S. J. (2003). A tree-based scan statistic for database disease surveillance, *Biometrics* 59: 323–331.

Luger, R. (2006). Exact permutation tests for non-nested non-linear regression models, *Journal of Econometrics* 133: 513–529.

Manly, B. F. J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, third edition. Chapman & Hall/CRC, Boca Raton.

Marriott, F. H. C. (1979). Bernard's Monte Carlo Test: How many Simulations?, *Applied Statististics* 28: 75–77.