UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Otimalidade de Testes Monte Carlo

Ivair Ramos Silva

Tese de doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Doutor em Estatística.

Orientador: Prof. Dr. Renato Martins Assunção

Belo Horizonte/MG, 15 de Abril de 2011.

0.1 Resumo

A operacionalização de um teste de hipóteses é condicionada ao conhecimento da distribuição de probabilidade da estatística de teste sob a hipótese nula H_0. Caso não se conheça a distribuição da estatística de teste sob H_0, sua distribuição assintótica pode ser usada para que a decisão sobre a rejeição ou não de H_0 possa ser feita, o que exige o estudo dos tamanhos amostrais para os quais tal distribuição assintótica se verifica. Quando a distribuição assintótica também não pode ser deduzida analiticamente, métodos de reamostragem podem ser aplicados para a construção de um critério alternativo de decisão, tais como reamostragem Bootstrap, testes de permutação e simulação Monte Carlo (*MC*), sendo este último o objeto de estudo deste trabalho.

Os testes de hipóteses montados pela utilização de simulações Monte Carlo podem ser divididos em dois tipos: os que se baseiam em um número m fixo de simulações, que é a estratégia convencional; e os procedimentos sequenciais, com os quais o número de simulações a serem geradas não é previamente fxado. Apesar de atualmente contarmos com recursos computacionais que favorecem o processamento de grandes bases de dados de forma extremamente veloz, ainda existem situações em que o tempo de processamento de uma estatística de teste é longo, o que motiva o desenvolvimento e utilização dos procedimentos sequenciais.

Os aspectos que recebem forte atenção na literatura sobre testes *MC* são: a busca por procedimentos que apresentem reduzido tempo médio de simulação até que a decisão sobre H_0 seja efetuada; a estipulação de cotas para a probabilidade da decisão quanto à rejeição de H_0 ser diferente da que se tomaria com o teste exato (risco de reamostragem); a estimação do valor-p; e a estipulação de cotas para as possíveis perdas de poder do teste *MC* sequencial em relação ao teste *MC* convencional ou em relação ao respectivo teste exato.

Usando certas suposições sobre a distribuição de probabilidade e a função poder associadas à estatística de teste, a literatura mostra que o poder do teste *MC* convencional é praticamente igual ao poder do teste exato. Um dos objetivos desta tese é demonstrar que é possível obter cotas para a diferença de poder entre o teste *MC* convencional e o teste exato que valem para qualquer estatística de teste, ou seja, a validade de tais cotas não depende de suposições além das necessárias à existência de um teste exato.

Besag and Clifford (1991) propuseram um procedimento de teste *MC* sequencial que, sob H_0, apresenta baixo tempo de execução. Objetivamos mostrar aqui como deve ser feita a escolha dos parâmetros de operacionalização deste procedimento sequencial. Primeiramente, mostramos como otimizar a escolha do número máximo de simulações sem afetar seu poder e, em seguida, demonstramos a forma de aplicar a regra de interrupção das simulações de modo a garantir um mesmo poder que o teste convencional.

O procedimento sequencial de Besag and Clifford (1991) só apresenta redução no tempo de execução nos casos em que H_0 é verdadeira. Com a principal finalidade de tornar o teste *MC* sequencial mais veloz que o *MC* convencional também quando a hipótese nula é falsa, procedimentos sequenciais alternativos tem sido propostos na literatura. O cálculo analítico

exato do poder de tais procedimentos sequenciais, bem como do valor esperado do número de simulações, são intratáveis para o caso geral, pois envolve o conhecimento da distribuição de probabilidade do valor-p, que por sua vez depende de cada aplicação específica. Pelo uso de algumas restrições ao comportamento da distribuição de probabilidade do valor-p, alguns autores obtiveram cotas para o risco de reamostragem e para a esperança do número de simulações para procedimentos sequenciais para os quais, em cada tempo t de simulação, a regra de iinterrupção das simulações é dada por uma função linear em t. Nesta tese, construimos um procedimento sequencial que permite um formato geral para a regra de interrupção das simulações, o qual chamaremos de teste *MC* sequencial generalizado. Esta construção absorve as principais propostas apresentadas na literatura e permite um tratamento analítico do poder e do número esperado de simulações para uma estatística de teste qualquer. Isto é feito pela elaboração de cotas superiores para a perda de poder e para a esperança do número de simulações. Com base em conceitos desenvolvidos nesta tese, apresentamos a construção do procedimento ótimo em termos do número esperado de simulações. Nós também cotamos o risco de reamostragem dentro de uma extensa classe de distribuições de probabilidade para o valor-p.

0.2 Introdução

A avaliação da eficiência de um teste de hipóteses é fortemente embasada no comportamento das probabilidades dos erros tipo I e tipo II, que por sua vez dependem do critério de rejeição da hipótese nula. Submeter a decisão sobre a rejeição da hipótese nula à avaliação do valor-p é uma abordagem amplamente aceita, pois, além de ser um critério simples para o controle da probabilidade do erro tipo I, é um recurso que é comumente interpretado como medidor do quão típico é o valor observado da estatística de teste frente ao que se espera sob a hipótese nula. Não é nosso objetivo discutir os méritos desta interpretação do valor-p, mas destacamos que, independentemente da interpretação, o valor-p favorece a tomada de decisão quanto à rejeição ou não da hipótese nula de forma objetiva. Obviamente, vários poderiam ser os critérios, baseados na amostra, para se tomar a decisão sobre a hipótese nula. É neste aspecto que focamos nosso trabalho. Ou seja, a decisão sobre qual das hipóteses é verdadeira deve ser feita com base na redução e controle das probabilidades dos erros tipo I e tipo II. Sob essa ótica, se um critério de decisão não é baseado na utilização direta do valor-p, entendemos que devemos focar a avaliação das probabilidades de ambos os erros gerados por tal critério.

Nos contextos em que não se conhece a distribuição de probabilidade da estatística de teste sob a hipótese nula, sua distribuição assintótica pode ser usada para que uma aproximação do valor-p seja efetuada e usada na tomada de decisão sobre H_0. Quando os tamanhos amostrais não são satisfatórios, o artifício de se utilizar a distribuição assintótica para a decisão sobre H_0 gera probabilidades nominais de erro tipo I bem diferentes das reais. Ademais, não são raros os casos em que não é possível obter analiticamente a distribuição assintótica da estatística de teste, de onde surgem os testes baseados em reamostragem Bootstrap, testes de permutação e os testes *MC*. O teste *MC* pode ser interpretado como uma forma de estimar a decisão sobre H_0 quando da aplicação do teste exato. Uma interpretação alternativa também é cabível, a de que o teste *MC* define um novo critério de teste, o qual pode ter seu poder avaliado e sua probabilidade de erro tipo I controlada tal como se faz para qualquer critério concorrente. Esta interpretação é coerente, pois, apesar de se ter mesmo poder para os testes *MC* e exato se suas decisões forem sempre coincidentes, a não coincidência entre suas decisões não implica poderes menores para o teste *MC*. Segundo (Gleser, 1996), a primeira lei de estatística aplicada postula que dois pesquisadores, usando a mesma metodologia, não devem chegar a conclusões diferentes com um mesmo banco de dados. Apesar de reconhecermos a importância dessa lei, entendemos que o controle da probabilidade do erro tipo I, combinado a poderes satisfatórios, deve estar em primeiro plano na avaliação de um procedimento baseado em simulação *MC*.

Resultados teóricos envolvendo a eficiência dos testes *MC* o colocam em uma posição de destaque nos contextos onde sua aplicação é viável, sendo sugerido até mesmo como concorrente aos testes assintóticos, tal como colocado por Jockel (1984) e exemplificado em Christensen and Kreiner (2007). Uma argumentação mais geral que sustenta tal credibilidade

se baseia na igualdade da probabilidade do erro tipo I e o nível de significância pretendido pelo usuário. Nesta tese, apresentamos uma segunda característica que reforça o potencial de uso da abordagem *MC* em substituição aos testes assintóticos. Vamos mostrar que o poder do teste *MC* é comparável ao do respectivo teste exato para qualquer estatística de teste, independentemente do tamanho amostral. Nem sempre esta característica é assegurada nos testes assintóticos.

A viabilidade da aplicação do teste *MC* pode ser severamente comprometida quando a estatística de teste é computacionalmente intensa. Com o propósito de reduzir o tempo gasto com a simulação *MC*, procedimentos sequenciais têm sido propostos. Eles não exigem que o número de simulações seja pré-fixado. As simulações são interrompidas tão logo se perceba um comportamento favorável a uma das hipóteses, a nula H_0 ou a alternativa H_A. Isto ocasiona um número aleatório de simulações. Listamos a seguir os principais aspectos considerados pela literatura no que se refere ao desenvolvimento de procedimentos *MC* sequenciais.

A1 - Perda de poder em relação ao teste *MC* com m fixo: O ganho no tempo de execução proporcionado pela utilização de um procedimento sequencial deve ser avaliado frente às potenciais perdas de poder em relação ao teste que fixa o número de simulações em m. É desejável que o poder do teste sequencial com máximo de simulações em m seja da ordem daquele que se teria com o *MC* fixo em m simulações. Com isto, procura-se evitar que o tempo de execução seja reduzido em detrimento de um poder maior com o m fixo;

A2 - Perda de poder em relação ao teste exato: O teste *MC*, sequencial ou não, é uma opção ao teste exato quando este é inviável. Após atender A1, o próximo desafio a um procedimento é que ele possua poder comparável com o do teste exato;

A3 - Tempo esperado para o número de simulações sob as hipóteses nula e alternativa: Obviamente, este é o principal aspecto que motiva o desenvolvimento de testes *MC* sequenciais. O objetivo é reduzir o valor esperado para o número de simulações sob as hipóteses nula e alternativa;

A4 - Risco de Reamostragem:

Este é um conceito que surge pela necessidade de se antender à primeira lei de (Gleser, 1996). O teste *MC*, para um dado m finito de simulações, oferece uma probabilidade não nula de que a decisão obtida por uma aplicação seja diferente de uma segunda aplicação do teste ao mesmo banco de dados. Uma forma de estudar tal probabilidade é comparar a decisão que se obtém pelo uso do *MC* com a que se teria pelo uso do valor-p real (teste exato). O risco de reamostragem é facilmente confundido com a perda de poder do teste *MC* em relação ao teste exato. Entretanto, a ocorrência de risco de reamostragem consideravelmente maior que zero não implica redução de poder do teste *MC* em relação ao teste exato. De fato, mostramos que mesmo quando a perda de poder é nula, o risco de reamostragem pode ser consideravelmente maior que zero;

A5 - Estimação do valor-p: O estimador de máxima verossimilhança (EMV) é eventualmente usado para se estimar o valor-p. Esta abordagem não é criticada pela literatura em procedimentos para os quais este estimador possui distribuição uniforme discreta entre 0 e

1, condição suficiente, mas não necessária, para que um estimador do valor-p seja tido como "válido". Um estimador do valor-p é válido sua distribuição acumulada, sob a hipótese nula e avaliada no ponto α, vale no máximo α, onde α é o nível de significância do teste. Como alguns procedimentos *MC* geram EMV não uniformemente distribuídos no intervalo discreto (0,1], algoritmos auxiliares acompanham a descrição de alguns dos procedimentos apresentados na literatura a fim de oferecer estimadores "válidos". Por definição, um estimador válido implica probabilidade de erro tipo I menor ou igual a α. Porém, é fácil elaborar estimadores para o valor-p que geram probabilidade de erro tipo I igual a α e que não possuem distribuição uniforme. Portanto, pensando apenas no controle do erro tipo I, a validade do valor-p é uma propriedade importante de se garantir. No entanto, dispensável se um procedimento qualquer garante probabilidade de erro tipo I igual a α.

O aspecto A3 é o mais explorado na literatura. Como veremos nesta tese, propostas interessantes têm sido apresentadas no sentido de minimizar o valor esperado do número de simulações no teste *MC* sequencial. O risco de reamostragem tem sido o segundo ponto de maior importância nos procedimentos propostos até o momento. A estimação correta do valor-p vem em terceiro lugar.

0.3 Objetivos Gerais

O objetivo desta tese é, para qualquer estatística de teste e de forma analítica, demonstrar propriedades e desenvolver conceitos que favoreçam a maximização do poder e a minimização da esperança do tempo de execução do teste Monte Carlo.

0.4 Objetivos Específicos

O primeiro objetivo é mostrar como usar o teste *MC* convencional de modo a garantir que possua o mesmo poder que o teste exato, provando assim que, nas situações em que é possível simular sob a hipótese nula, é mais interessante usar o teste baseado em simulação *MC* do que aplicar testes baseados na distribuição assintótica da estatística de teste. Buscamos provar também que a escolha do número de simulações no *MC* com m fixo deve obedecer a um critério que leva em conta o nível de significância estipulado para a rejeição de H_0.

O segundo objetivo é provar que, do ponto de vista de preservação e controle das probabilidades dos erros tipo I e tipo II, o teste *MC* sequencial por Besag and Clifford (1991) sempre deve ser preferido no lugar do procedimento com número fixo de simulações. Mostramos também que o número máximo de simulações deste teste sequencial deve ser escolhido como uma função da regra de interrupção das simulações e da probabilidade de erro tipo I fixada pelo usuário.

O terceiro objetivo é propor um procedimento *MC* sequencial com duas barreiras de interrupção das simulações que permita a utilização de um formato geral para tais barreiras, com propriedades válidas para qualquer estatística de teste, que possa ser avaliado analiticamente

e que seja ótimo em termos do tempo médio de execução. Chamaremos tal procedimento de "teste Monte Carlo Sequencial Generalizado"(MC_G). Pretendemos mostrar que o MC_G apresenta tempo médio de simulação inferior aos apresentados por procedimentos concorrentes e que, ao mesmo tempo, possui o mesmo poder que o procedimento *MC* convencional. Pretendemos fazer isso sem o uso de suposições além das necessárias à existência do teste exato. Objetivamos também cotar o risco de reamostragem associado ao MC_G com auxílio de uma suposição razoável do ponto de vista prático.

0.5 Justificativa e Relevância da Pesquisa

O teste de hipóteses é um dos conceitos mais consagrados e utilizados da inferência estatística. Sua aplicação varre todas as áreas da ciência e, apesar da controvérsia acerca da significância estatística, ainda é amplamente usado em análise de dados, sendo ainda alvo de intensa pesquisa no ambiente acadêmico. Com isso, a simulação Monte Carlo é um método importante do ponto de vista prático, uma vez que oferece um tratamento viável quando ocorre a impossibilidade de realizar testes estatísticos exatos em situações mais complexas ou em que os tamanhos amostrais são insuficientes para o uso de resultados assintóticos. Pode-se encontrar uma vasta relação de aplicações do teste Monte Carlo, das quais as direcionadas à análise espacial de dados são tipicamente citadas como motivadoras do estudo teórico deste método. Como exemplos de aplicações na área de análise espacial podemos citar Ripley (1992), Kulldorff (2001), Assunção and Maia (2007) ou Peng et al. (2005). Exemplos de aplicações fora da estatística espacial podem ser vistos em Booth and Butler (1999), Caffo and Booth (2003) ou Wongravee et al. (2009). Os resultados já conhecidos, no que concerne ao poder do teste *MC*, dependem fortemente de suposições que, na prática, são de difícil verificação devido à própria situação em que a abordagem Monte Carlo é requisitada. Isto é, diante do total desconhecimento do comportamento da distribuição da estatística de teste, é difícil verificar as suposições usuais. Portanto, a valoração da aplicação dos testes *MC* carece da demonstração de resultados mais gerais sobre a magnitude do poder e sobre a escolha do número máximo de simulações m. Da mesma forma, os procedimentos sequenciais devem ser propostos sob aspectos gerais e de simples aplicação, de modo que possam garantir a sua utilização e confiabilidade, uma vez que pretendem ser priorizados em substituição ao *MC* com m fixo.

0.6 Organização

Este material é formado pela coleção de três artigos que tratam do teste Monte Carlo para testes de hipóteses. O primeiro deles considera a escolha dos parâmetros de operacionalização do teste Monte Carlo sequencial proposto em Besag and Clifford (1991). Este artigo, entitulado "Power of the Sequential Monte Carlo Test", foi publicado no volume 28, edição 2, do periódico "Sequential Analysis".

O segundo trabalho desta tese estuda as propriedades do poder e os critérios para escolha do número de simulações do teste Monte Carlo convencional para uma estatística de teste qualquer. Este trabalho está condensado no segundo artigo, entitulado "Monte Carlo Tests under General Conditions: Power and Number of Simulations", submetido em fevereiro de 2011 ao "Journal of Statistical Planning and Inference".

A generalização dos testes Monte Carlo sequenciais com duas barreiras, e a construção do teste sequencial ótimo em termos do tempo médio de execução, é o conteúdo do terceiro e último artigo desta coleção, entitulado "Optimal Generalized Sequential Monte Carlo Test". Este artigo será submetido ao "Journal of the American Statistical Association".

Referências Bibliográficas

Assunção, R. and Maia, A. (2007). A note on testing separability in spatial-temporal marked point processes. *Biometrics*, 63(1):290–294.

Besag, J. and Clifford, P. (1991). Sequential monte carlo p-value. *Biometrika*, 78:301–304.

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

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

Christensen, K. and Kreiner, S. (2007). A monte carlo approach to unidimensionality testing in polytomous rasch models. *Applied Psychological Measurement*, 31(1):20–30.

Gleser, L. (1996). *Comment on "Bootstrap Confidence Intervals"* . Number 11. Statistical Science, T. J. DiCiccio & B. Efron.

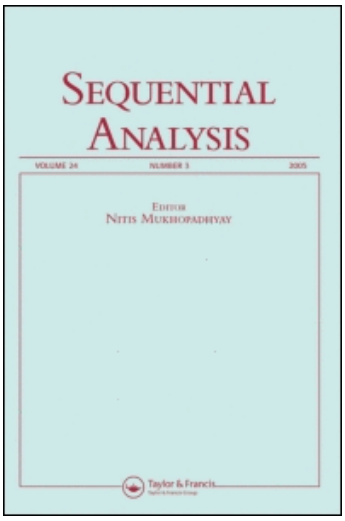
Jockel, K. (1984). Application of monte-carlo tests - some considerations. *Biometrics*, 40(1):263–263.

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

Peng, R., Schoenberg, F., and Woods, J. (2005). A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, 100(469):26–35.

Ripley, B. (1992). Applications of monte-carlo methods in spatial and image-analysis. *Lecture Notes in Economics and Mathematical Systems*, 376:47–53.

Wongravee, K., Lloyd, G., Hall, J., Holmboe, M., Schaefer, M., Reed, R., Trevejo, J., and Brereton, R. (2009). Monte-carlo methods for determining optimal number of significant variables. application to mouse urinary profiles. *Metabolomics*, 5(4):387–406.



Sequential Analysis
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597296

Power of the Sequential Monte Carlo Test

I. Silva [a]; R. Assunção [a]; M. Costa [a]
[a] Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Online Publication Date: 01 April 2009

To cite this Article Silva, I., Assunção, R. and Costa, M.(2009)'Power of the Sequential Monte Carlo Test',Sequential Analysis,28:2,163 — 174
To link to this Article: DOI: 10.1080/07474940902816601
URL: http://dx.doi.org/10.1080/07474940902816601

Sequential Analysis, 28: 163–174, 2009

ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474940902816601

Taylor & Francis
Taylor & Francis Group

Power of the Sequential Monte Carlo Test

I. Silva, R. Assunção, and M. Costa
Departamento de Estatística, Universidade Federal de Minas Gerais,
Belo Horizonte, Minas Gerais, Brazil

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.

Subject Classifications: 62L05; 62L15; 65C05.

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

Received April 15, 2008, Revised February 1, 2009, February 5, 2009, Accepted February 8, 2009

Recommended by Nitis Mukhopadhyay

Address correspondence to I. Silva, Departamento de Estatística, Universidade Federal de Minas Gerais, Avenida Presıdente Antonio Carlos, 6627, Prédio do ICEx, Sala 4054, Pampulha, Belo Horizonte, Minas Gerais 31270-901, Brazil; Fax: 55-31-3409-5924; E-mail: ivair@bdmg.mg.gov.br

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., 2005; Kulldorff, 2001) and data mining (Kulldorff et al., 2003; Rolka et al., 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 that 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 rth 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 $1/m, 2/m, \ldots, 1$. That is, the Monte Carlo p-value P_{mc} has a uniform distribution on the discrete set $\{1/m, 2/m, \ldots, 1\}$. Let W be the random variable $W = P_{mc} - X$ where $X \sim U(0, 1/m)$ and it is independent of P_{mc}. Then, $W \sim U(0, 1)$ under the null hypothesis. In this sense, Besag and Clifford (1991) say that the Monte Carlo p-value is exact, because P_{mc} has the same uniform distribution under the null hypothesis than the analytical p-value $P = 1 - F(U)$. In addition to that, irrespective of the validity of the null hypothesis, $P_{mc} \to P$ almost everywhere because, for any observed value u_0, we have

$$p_{mc} = \frac{1 + \#\{u_i \geq u_0\}}{m} = \frac{1}{m} + \left[\frac{\#\{u_i \geq u_0\}}{m - 1}\right]\frac{m - 1}{m} \to 1 - F(u_0)$$

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, ..., 10/1000, 9/1000, 8/1000, ..., 1/1000. Note that the support of the sequential Monte Carlo procedure is more concentrated on the lower end of the interval $(0, 1)$. This is a desirable characteristic because these are the p-value possible values that we want to know more precisely.

Let the support of p_s be denoted by

$$S = \{1/n, 2/n, \ldots, h/n, h/(n-1), \ldots, h/(h+2), h/(h+1), 1\}.$$

The values of the form $(h-q)/n$, with $0 \leq q \leq h-1$, occur when we need to run the procedure up to the maximum $n-1$ number of simulations without never getting h u_is larger than u. The other values, of the form $h/(h+q)$, with $h \leq h+q \leq n-1$, occur when we either stop the procedure earlier than the maximum $n-1$ or when the hth larger value occur exactly at the $(n-1)$th sequential observation and hence $l=n-1$. Therefore, we can also define the sequential p-value as

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

If $a \in S$, then $\mathbb{P}(P_s \leq a) = a$ under the null hypothesis. To see this, assume that $a = h/(h+q) \leq 1$ with $h \leq h+q < n$. Then

$$\begin{aligned} \mathbb{P}(P_s \leq a) &= \mathbb{P}(P_s \leq h/(h+q)) = \mathbb{P}(L \geq h+q) \\ &= \mathbb{P}(L > h+q-1) = \frac{h}{h+q} \end{aligned}$$

because $L > h+q-1$ if, and only if, after $l-1$ Monte Carlo withdrawals, the observed u is among the largest h of the sample of equally probable $(l-1)+1$ elements. Consider now that $a=(h-q)/n$ with $0 \leq q \leq h-1$. Then

$$\begin{aligned} \mathbb{P}(P_s \leq a) &= \mathbb{P}(P_s \leq (h-q)/n) \\ &= \mathbb{P}(L = n-1 \text{ and only } h-q-1 \text{ exceeding } u) \\ &= \mathbb{P}(u \text{ is the } (h-q)\text{th largest among } n) \\ &= (h-q)/n \end{aligned}$$

We can transform P_s by subtracting a random variable X such that the sequential p-value also has a continuous uniform distribution in $(0, 1)$. For that, define X conditionally on the observed value of the discrete p-value P_s. Suppose that $p_s = b \in S$. Let a be the largest element of S that is smaller than b. Define $a=0$ if $b=1/n$. Then $X \sim U(a, b)$ and $P_s - X$ has a uniform distribution in $(0, 1)$ under the null hypothesis, exactly as the p-values P and P_{mc}. Because this is less common to be carried out in practice, in the remaining of this paper we will not transform P_s in this way, keeping its definition as in (2.1).

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

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

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, n, h, and 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. Whether 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. Because 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 given by:

$$\mathbb{P}(L = l \mid P = p) = \begin{cases} \binom{l-1}{h-1} p^h (1-p)^{l-h} & \text{if } l = h, h+1, \ldots, n-2 \\ \sum_{x=n-1}^{\infty} \binom{x-1}{h-1} p^h (1-p)^{x-h} & \text{if } l = n-1 \end{cases}$$

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

$$\begin{aligned}\mathbb{P}(\text{Reject } H_0 \mid P = p) &= \mathbb{P}(L \geq (h/\alpha) \mid P = p) \\ &= \mathbb{P}(L > (h/\alpha) - 1 \mid P = p) \\ &= \sum_{x=0}^{h-1} \binom{h/\alpha - 1}{x} p^x (1-p)^{h/\alpha - x - 1} \end{aligned} \tag{3.1}$$

Because the last expression does not involve n, the power of the sequential Monte Carlo test is constant as long as $n \geq (h/\alpha) + 1$. Because 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, α) with $n < h/\alpha + 1$

The power of the procedure *MCseq*(n, h, α) 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 $h/(n-1)$. Because $n < h/\alpha + 1$, 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

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

Figure 1(a) shows the shape of $\mathbb{P}(G \leq n\alpha - 1 \mid P = p)$ for $\alpha = 0.05$, $h = 5$, and $n = 21, 61, 101$. The power for $n < h/\alpha + 1$ is given by integrating out (3.2) with respect

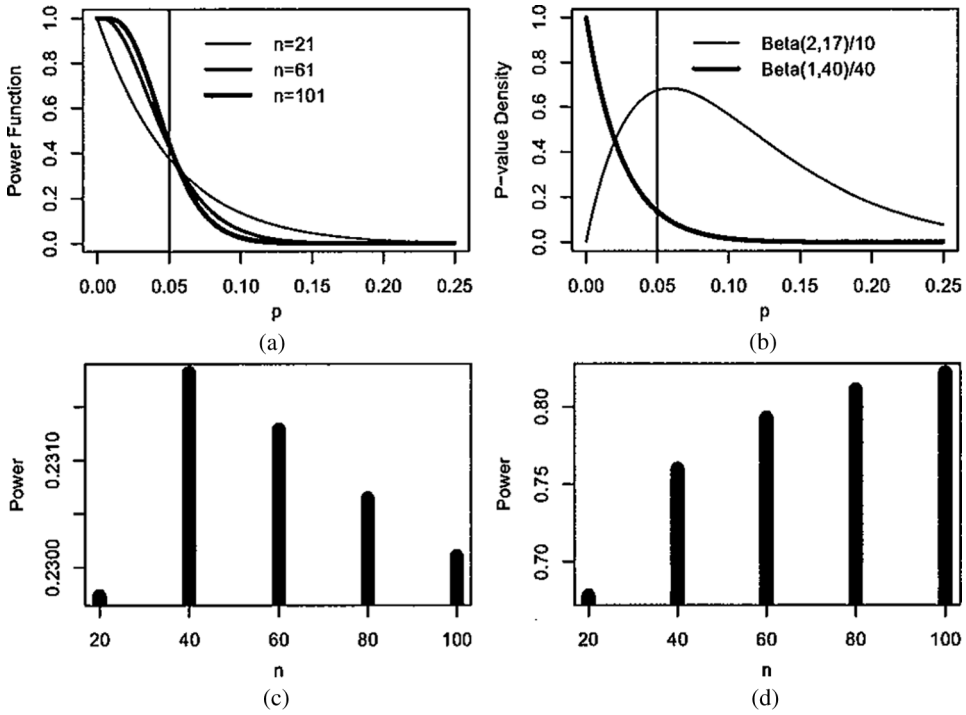


Figure 1. Power behavior of the sequential Monte Carlo test and the p-value density.

to the p-value probability distribution F_p:

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

Denote by $\pi(n, h, \alpha, F_P)$ the power function of the sequential procedure. Depending on F_P, the power curve can be non-monotone. To illustrate this result, consider two different F_P distributions. One of them assumes that P is distributed according to a Beta distribution with parameters 2 and 17. The other assumes a Beta distribution with parameters 1 and 40 (see Figure 1(b)). The graph in Figure 1(c) shows the power (3.3) using F_P as a Beta$(2, 17)$ distribution with $h = 5$, $\alpha = 0.05$, and $n = 20, 40, 60, 80, 100$. We can see that the power does not increase with n in the range $61 \leq n \leq 101$.

In contrast, Figure 1(d) shows the power using $F_P \sim \text{Beta}(1, 40)$ and the same tuning parameters as before. In this case, the power is increasing with n. Indeed, Hope (1968) and Jockel (1986) showed that we always have the power increasing with n if the p-value distribution F_P belongs to certain distribution classes, which include the Beta$(1, 40)$ distribution.

This illustrative example shows that, 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

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

The $MCconv(m, \alpha)$ power is

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

whereas 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. Table 1 shows the quartiles of the null distribution of L for the sequential MC test $MCseq(m+1, \alpha m, \alpha)$ equivalent to the conventional MC test $MCconv(m, \alpha)$ with different significance level α between 0.01 and 0.05, and with $m = 1000$ and $n = 1001$. 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.

Table 1. Quartiles of the null distribution of L for some sequential MC test equivalent to the conventional MC test with $m = h/\alpha$

α	h	1st quartile	Median	3rd quartile	$\mathbb{E}(L)$
0.01	10	12.33	19.00	39.00	55.56
0.02	20	25.67	39.00	79.00	97.74
0.03	30	39.00	59.00	119.00	134.69
0.04	40	52.33	79.00	159.00	168.24
0.05	50	65.67	99.00	199.00	199.26

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 $P = p$. Because 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

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

A crude bound for the difference in power is obtained by finding real numbers a and b such that $a \le D(P) \le b$. 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 $\alpha \neq \alpha_2$.

Figure 2 shows a graph of the power difference $D(p)$ between $MCconv(1000, 0.05)$ and $MCseq(\lfloor h/0.05 \rfloor + 1, h, 0.05)$. Each curve $(p, D(p))$ represents $D(p)$ for a given value of h, with $h = 5, 10, \ldots, 45$. The curve showing both the highest peak and deepest valley corresponds to $h = 5$. As h increases, the curves dampen and have less pronounced extreme values. Hence, the larger the value of h, the smaller $D(p)$.

For $h = 25$, $D(p)$ assumes its maximum value when $p \approx 0.0423$, and the minimum when $p \approx 0.0586$. The maximum value of $|D(p)|$ is equal 0.0921, and it is a crude bound for the power difference between $MCconv(1000, 0.05)$ and $MCseq(501, 25, 0.05)$. This is so because $D(p)$ varies from its maximum (and positive) value to its minimum (and negative) value within a short interval in p. Because the power $E[D(p)]$ simply integrates $D(p)$ with respect to $f_P(p)$, the extreme

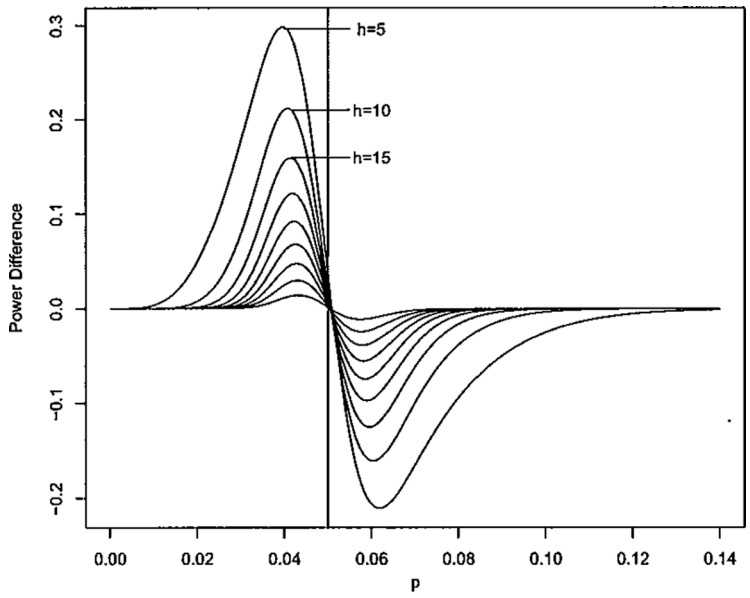


Figure 2. Graph of the power difference function between the fixed-size and the sequential Monte Carlo tests.

value of $|D(p)|$ will be approximately the value of 0.0921 only if the density $f_P(p)$ is tightly concentrated around the point of maximum. This is not likely to happen in practice and hence we can expect a much lower value for the power $E(D(p))$.

Table 2 shows the values of the crude bound $\max_p D(p) = b(m, \alpha; h)$ for the power difference between $MCconv(m, \alpha)$ and $MCseq(n, h, \alpha)$ with $m = 1000, 10000$, $\alpha = 0.05$ and $h = 5, 15, \ldots, 500$. We set $n = \lfloor h/\alpha \rfloor + 1$ in all cases. The fourth column shows these bounds for $m = 1000$ and $m = 10000$ separately. The numbers in this column are also the power difference bounds between two conventional Monte Carlo tests, namely between $MCconv(m, 0.05)$ and $MCconv(m_2, 0.05)$, where m_2 is shown in the third column.

For small h, the crude bound is too large. However, this bound decreases quickly with h. For example, when $h = 25$, the $MCseq(501, 25, 0.05)$ has power smaller than $MCconv(1000, 0.05)$ by, at most, 0.09. For $m = 10000$, the bound $\max_p D(p)$ decreases even faster.

Table 2. Crude bound $\max_p D(p) = b(m, \alpha; h)$

$m = 1000$				$m = 10000$			
h	n	m_2	$\max_p D(p)$	h	n	m_2	$\max_p D(p)$
5	101	100	0.2980	100	2001	2000	0.1931
15	301	300	0.1595	150	3001	3000	0.1469
25	501	500	0.0921	250	5001	5000	0.0858
40	801	800	0.0296	400	8001	5000	0.0278
50	1001	1000	0.0000	500	7001	7000	0.0000

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 and power. More important, this sequential Monte Carlo test requires at most one additional simulation and its expected number of simulations is generally much smaller than m, specially 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* and *MCseq* tests. This is more relevant if we consider that the true differences are likely to be much smaller than the bounds suggest. These bounds can also be used to compare two conventional Monte Carlo 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 $n = h/\alpha + 1$.

ACKNOWLEDGMENTS

The authors 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 R01CA095979, 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.

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 Perrmutation 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*, 3rd ed., Boca Raton: Chapman & Hall/CRC.

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

Rolka, R., Burkom, H., Cooper, G. F., Kulldorff, M., Madigan, D., and Wong, W. (2007). Issues in Applied Statistics for Public Health Bioterrorism Surveillance Using Multiple Data Streams: Research Needs, *Statistics in Medicine* 26: 1834–1856.

Monte Carlo test under general conditions: Power and number of simulations

I.R.Silva*, R.M.Assunção

Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract

The statistical tests application demands the probability distribution of the test statistic U under the null hypothesis. When that distribution can not be obtained analytically, it is necessary to establish alternative methods to calculate the p-value. If U can be simulated under the null hypothesis, Monte Carlo simulation is one of the ways to estimate the p-value. Under some assumptions about the probability distribution and the power function of U, the literature has obtained thin upper bounds for the power difference between exact test and Monte Carlo test. The motivation of this paper is to dispense any assumptions and to demonstrate that the Monte Carlo test and exact test have same magnitude in power, for any test statistic, even for moderated m. This demonstration is possible by exploring the trade-off between the power and the significance level.

Keywords:
MC Test, Exact Test, Power, p-value, Significance Level

1. Introduction

A current obstacle in the construction of a test statistic U is the fact that in several cases is impossible to obtain analytically its probability distribution function. Regularly, even its asymptotic distribution can not be found. An example of that situation is the Scan statistic, Kulldorff (2001), which is applied to detect spatial clusters. Even when it is possible to deduce the asymptotic distribution F_d of U the problem is not properly solved. An undesirable consequence of using F_d to obtain the p-value is that the real type one error and the power loss compared to the exact test[1] are not controlled for finite samples.

If it is possible to generate samples from the test statistic under the null hypothesis H_0, the maximum-likelihood estimator for the p-value is given by the ratio between the number of simulated statistics u_i's greater than or equal to the observed value u_0, $i = 1, 2, ..., m-1$. The null hypothesis is rejected if the estimated p-value is smaller than or equal to the desired significance level α_{mc}. That approach is known as Monte Carlo test (MC test), proposed by Dwass (1957), introduced by Barnard (1963), and extended by Hope (1968) and Birnbaum (1974). A first property which must be mentioned is that estimating the p-value in that way and rejecting H_0 if the estimative is smaller than or equal to α_{mc} induces a probability of type one error smaller than or equal to α_{mc}. An additional positive property of MC test is that the number m of simulations can be random, Besag and Clifford (1991), without power loss, Silva et al. (2009). Another intensifier aspect of MC test application, showed by Dufour (2005), is the fact that MC test is viable even when, under H_0, the test statistic involves nuisance parameters.

*Corresponding author
Email addresses: irs@ufmg.br (I.R.Silva), assuncao@est.ufmg.br (R.M.Assunção)

[1] Exact test here is based on the p-value obtained directly from the real distribution of U under the null hypothesis

Methods directed to data spatial analysis are strongly cited in the literature to justify the theoretical development of MC test. Applications of MC test on spatial analysis are seen in Ripley (1992), Kulldorff (2001), Assunção and Maia (2007) or Peng et al. (2005). Applications are also sorted in a variety of areas, as we can see in Booth and Butler (1999), Caffo and Booth (2003) or Wongravee et al. (2009).

An important researching aspect involving MC tests is its power compared to that from the exact test. Hope (1968) has considered MC tests on situations when the likelihood ratio $f_1(u)/f_0(u)$ is a monotone increasing function of a fixed test statistic value u, with $f_1(u)$ and $f_0(u)$ the densities of the test statistic under the alternative and null hypothesys, respectively . Using $\alpha_{mc} = j/m$ and $j \in N$, he has showed that the uniformly most powerfull test (UMP) based on U exists and, consequently, the power of MC test converges, with m, to the power of the UMP. With a less restrictive condition of concavity for the exact test power function, Jockel (1986) has proved that the MC test power converges uniformly to the one of the exact test, and establishes upper bounds to the ratio between them for finite m. Marriott (1979) has used a normal distribution to approximate the power of MC test, what made possible to deduce that its convergence to the exact test power is fast. Fay and Follmann (2002) treated the probability of the sequential Monte Carlo test to conduce to different decisions about rejecting or not H_0 faced to the decisions by using the exact test, called resampling risk (RR). They proposed a sophisticated sequential implementation of MC tests and bound the RR by considering a certain class of distributions to the random variable p-value. Sequential implementation of MC tests are important for saving time when test statistic is computationally intense. Other efficient designs to implement MC tests sequentially has been proposed, as well the study of the expected number of simulations under the alternative hypothesis and the control of the resampling risk, and concerning these aspects, we can cite Fay and Follmann (2002), Fay et al. (2007), Gandy (2009) and Kim (2010). However, our focus here is to treat MC tests with fixed m, which is the simpler design when the test statistic is computationally light. We believe that extensions of our results to the sequential implementation is a natural path for future explorations.

We shall show here an approach for establishing upper bounds to the power difference between the exact test and MC test alternative to those described previouslly. Our proposal produces an upper bound sufficiently slim, such that, in practical terms, the possibility of losing power by using MC test should be unconsidered. This result was possible by considering α_{mc} slightly greater than the significance level α of the exact test. Assumptions involving the distribution of the test statistic, the exact power function or over the likelihood in U, are not required here.

Choosing a significance level greater than the initialy planned is not something enjoyable, neither it is a new idea use this artifice for obtaining a better power. But, if we consider that the required differences between α_{mc} and α, sufficient to guarantee slim upper bounds, are negligible even with short m, and such differences are controlled by the user, we conclude that this proposal offers an appeal to treat satisfactorily an inconvenient arbitration demanded in MC tests, that is the choice of m combined to the necessity of having a power equal to that from the respective exact test.

When the control of type one and type two errors is imperative, a possible interpretation of the results here is that MC test is as useful as the own exact test. Therefore, under that aspect, the asymptotic approximation of the distribution of U could be replaced by Monte Carlo procedure, once the test based in asymptotic distributions has no analitical control of the probability of the type one error and upper bounds for the power loss comparatively to the exact test for each sample size must be studied, analitically or empirically, for each specific application.

It is easy to see that, for having a probability of type one error equal to α_{mc} in MC test, it is necessary to choose m as a multiple of $1/\alpha_{mc}$, for rational α_{mc}. However, non-compliance with the rule $m = j/\alpha_{mc}$, j integer, results in power reduction of MC test as m increases. Let $\pi(m, \alpha_{mc}, F_P)$ be the MC test power for fixed significance level α_{mc} and F_P be the probability distribution of the p-value. The existence of F_P depends solely on the existence of the distributions of U under H_0 and the alternative hypothesis H_1, what

can be well understood in Sackrowitz and Samuel-Cahn (1999). It is intuitive that, as larger is m, larger is the power. But that is not true at the generic case. Silva et al. (2009) offers an example about how MC test power can decrease with m, even when $m = j/\alpha_{mc}$. We shall show here that $\pi(m, \alpha_{mc}, F_P)$ has potential for increasing with m only if $m = j/\alpha_{mc}$. Hope (1968) showed that, when the likelihood ratio is a monotone increasing function of u, for $m = j/\alpha_{mc}$, $\pi(m, \alpha_{mc}, F_P)$ is a monotone increasing function of m. Concerning this topic, our focus is to show what happens with the MC power if m is not a multiple of $1/\alpha$ for any likelihood ratio shape, and with a very simple reasoning, it can be showed that $\pi(m, \alpha_{mc}, F_P)$ is always non-increasing for m on the range $[\lfloor j/\alpha_{mc} \rfloor, \lfloor (j+1)/\alpha_{mc} \rfloor)$.

The next section develops a proposal to obtain upper bounds for the power difference between the exact test and MC test. In practice, those upper bounds allow the construction of MC tests that present same power that the exact tests. Section 3 shows that m must be a multiple of $1/\alpha_{mc}$ and Section 4 finishes this paper with a discussion.

2. Upper Bound for the Power Difference Between Exact Test and Monte Carlo Test

The calculation of the p-value depends on how the alternative hypothesis is defined. Without loss of generality, this paper works only the cases where the alternative hypothesis is formulated so that large values of U lead to the rejection of H_0.

As adopted by Silva et al. (2009), for a given test statistic U, with observed value u_0, let consider the event $[U_i \geq u_0]$ as a success, where $U_1, U_2, ..., U_{m-1}$ are independent copies from U under the null hypothesis. Let F be the probability distribution function of U_i under the null hypothesis. Then, the probability $\mathbb{P}(U_i \geq u_0) = 1 - F(u_0)$ is the p-value.

Let Y be the number of successes for a fixed m. The random variable Y has a binomial distribution with $m-1$ essays and success probability equal to the observed p-value. We shall use P for denoting the p-value as a random variable. Thus, given $P = p$, the MC procedure leads to the rejection of H_0 with probability

$$\mathbb{P}(Y \leq m\alpha_{mc} - 1 \mid P = p) = \pi(m, \alpha_{mc}, p) = \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \left(\begin{array}{c} m-1 \\ x \end{array} \right) p^x (1-p)^{m-1-x}. \tag{1}$$

The power of MC test is obtained by integrating, in continuous case, or by summing, in discrete case, (1) with respect to the distribution F_P. Without loss of generality, we shall work in this paper only with the continuous case. The discret case can be treated by applying the randomized test. Therefore, the power of MC test is:

$$\pi(m, \alpha_{mc}, F_P) = \int_0^1 \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \left(\begin{array}{c} m-1 \\ x \end{array} \right) p^x (1-p)^{m-1-x} F_P(dp). \tag{2}$$

Let α be the significance level of the exact test. The probability of rejecting H_0 by using the exact test is 1, if $p \leq \alpha$, and it is 0, otherwise. The power of the exact test can be expressed as follows:

$$\pi(\alpha, F_P) = \int_0^\alpha F_P(dp). \tag{3}$$

It is easy to prove the convergence of the MC test power to the exact power. Take $\alpha_{mc} = \alpha$. When $m \to \infty$, $Y/m \overset{ae}{\to} p$, then, for $p < \alpha$, $\mathbb{P}(Y \leq m\alpha - 1 \mid P = p) \overset{ae}{\to} 1$, and if $p \geq \alpha$, $\mathbb{P}(Y \leq m\alpha - 1 \mid P = p) \overset{ae}{\to} 0$.[2] Thus, according to the dominated convergence theorem, $\lim_{m\to\infty} \pi(m, \alpha, F_P) = \pi(\alpha, F_P)$. However, under the pratical point of view, it is more important to understand the relation between the exact and MC powers

[2]The abbreviation ae means almost everywhere convergence

for finite m.

Bounding the power loss of MC test, compared to the exact test, demands a cost to pay, and in the literature the usual payment is to restrict the p-value behaviour. That is a very expensive price, because, in practice, and under the main context where MC tests are required, assumptions about the unknown p-value density behaviour are rarely feasible to verify. We propose here a more practical exchange currency. We suggest to change the restrictions over the p-value density by using significance levels for the MC test slightly larger than the exact test ones.

From expressions (1) and (2), the trade-off between the significance level and the power can be manipulated. With that, we propose to sacrifice the significance level α_{mc} in an irrelevant way comparatively to α, $\alpha_{mc} = \alpha + \delta$, $\delta > 0$ and, at the same time, producing satisfatory upper bounds for the power loss of MC procedure.

Formally, the comparation in power between two or more tests presupposes identity between the associated significance levels. But, we perform here the comparation in power by using different significance levels α_{mc} and α. Nevertheless, the abuse of terminology is justified, because sufficiently thin upper bounds for the power loss of MC test are obtained by using despicable magnitudes for δ under the practical aspect.

The power difference between the exact test and the MC test is:

$$E[D(P)] = \int_0^1 D(p) F_P(dp) \tag{4}$$

with $D(p)$ given by

$$D(p) = 1_{(0,\alpha]}(p) - \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \left(\begin{array}{c} m-1 \\ x \end{array} \right) p^x (1-p)^{m-x-1} \tag{5}$$

where $1_{(0,\alpha]}(p)$ is the step function of $p \in (0, \alpha]$. As $\pi(m, \alpha_{mc}, p)$ decreases monotonously with p, the maximum value for $D(p)$ is:

$$b(m, \alpha, \alpha_{mc}) = \max_p \{D(p)\} = 1 - \sum_{x=0}^{\lfloor m\alpha_{mc} \rfloor - 1} \left(\begin{array}{c} m-1 \\ x \end{array} \right) \alpha^x (1-\alpha)^{m-x-1} \tag{6}$$

that is the maximum distance between the step function $1_{(0,\alpha]}(p)$ and the power function of the MC test $\pi(m, \alpha_{mc}, p)$.

Table 1 offers upper bounds $b(m, \alpha, \alpha_{mc})$ for some values of m, α and α_{mc} by applying directly Equation (6), and it highlights that just an unexpressive increase of α_{mc} is sufficient to lead MC test to a power pratically equal to that from exact test. It is important to note that these upper bounds are thin even with small m values, for example, observe $\alpha_{mc} = 0.015$ and $\alpha = 0.01$ or $\alpha_{mc} = 0.06$ and $\alpha = 0.05$, with $m = 3000$, where the upper bounds are smaller than 0.006 and 0.008, respectively.

The expression (6) can be used to find a value for m that, except by a known constant $\epsilon > 0$ chosen by the user, it induces the same power for MC test and the exact test, pointing out that the probability of type one error $\alpha_{mc} = \alpha + \delta(\epsilon)$, with desirable $\delta(\epsilon)$. With that, is possible to calibrate the upper bounds for the

m	$\alpha = 0.01$ $\alpha_{mc} = 0.013$	$\alpha_{mc} = 0.015$	$\alpha = 0.05$ $\alpha_{mc} = 0.055$	$\alpha_{mc} = 0.060$
1000	0.2065387	0.0818942	0.2505276	0.0855520
2000	0.1106199	0.0211332	0.1633004	0.0247740
3000	0.0636350	0.0059943	0.1124737	0.0078300
4000	0.0378388	0.0017746	0.0795444	0.0025768
5000	0.0229464	0.0005386	0.0571672	0.0008684
6000	0.0140999	0.0001661	0.0415376	0.0002973
7000	0.0087467	0.0000518	0.0304238	0.0001029
8000	0.0054652	0.0000163	0.0224209	0.0000359
9000	0.0034342	0.0000052	0.0166038	0.0000126
10000	0.0021679	0.0000016	0.0123450	0.0000045

Table 1: Upper bounds for the power differences between the exact test, with significance levels α= 0.01, 0.05, and MC test, with significance levels α_{mc}.

power loss according to the convenience of the user, who has the prior expertise to define subjectively what is a despicable value for $\delta(\epsilon)$. That fact confirms the assertion from Jockel (1984), which is, MC test has potential to compete against the tests based on asymptotic theory, mainly for small sample sizes.

As m increases, smaller is the $\delta(\epsilon)$ sufficient for ensuring an acceptable value for $b(m, \alpha, \alpha_{mc}) = \epsilon$. To see that, consider a random variable $X \sim Bin(n, p)$. Okamoto (1958) has established that, for $p < 1/2$, $\mathbb{P}(\sqrt{X/n} - \sqrt{p} \geq c) < \exp(-2nc)$, where c is a constant. Manipulating the last inequality, we have the following:

$$\mathbb{P}(Y \leq \alpha_{mc}m - 1 \mid P = p) \geq 1 - \exp\left\{-2(m-1)\left[\left(\frac{\alpha_{mc}m}{m-1}\right)^{1/2} - \sqrt{p}\right]^2\right\}. \tag{7}$$

Fixing $b(m, \alpha, \alpha_{mc}) = \epsilon$, from 6 and 7,

$$\epsilon \leq \exp\left\{-2\left[\sqrt{\alpha_{mc}m} - \sqrt{\alpha(m-1)}\right]^2\right\}. \tag{8}$$

The right hand side of the inequality (8) is decreasing with m for $m > (1 - \alpha/\alpha_{mc})^{-1}$, then, as m soars, smaller is $\delta(\epsilon)$ to keep an arbitrary ϵ. However, we must emphasize the importance of increasing m by multiples of $1/\alpha_{mc}$, which is the subject treated in the next section.

3. m as a Multiple of $1/\alpha_{mc}$

Hope (1968), by considering a monotone increasing function for the likelihood ratio in U and $m = j/\alpha_{mc}$, with $j \in N$, has studied the power behavior of MC test when m soars. Under such conditions, he has proved both, the existence of the uniformily more powerful test (UMP) and the convergence of MC test power to the UMP one. We offer here an additional argument for the using of m in the form j/α_{mc} that is the fact of $\pi(m, \alpha_{mc}, F_P)$ decreasing with m for $\lfloor (j-1)/\alpha_{mc} \rfloor < m < \lfloor j/\alpha_{mc} \rfloor$, $j > 1$.

We reject H_0 if, among the $(m-1)$ simulated U_is, the number of values greater than or egual to u_0 is not greater than $\alpha_{mc}m - 1$. Obviously, that requires $m \geq 1/\alpha_{mc}$, because, otherwise, H_0 is never rejected.

Consider two MC tests which differ from the number of simulations, m and $m + k$, $k > 0$. For the first MC test, which is based on m, we reject H_0 only if the number of simulated values U_is exceeding u_0 is at most $\lfloor \alpha_{mc}m \rfloor - 1$, whereas, for the second test, with $(m + k)$ simulations, H_0 is rejected if such number is

at most $\lfloor \alpha_{mc}(m+k) \rfloor - 1$. According to (1), for any observed $P = p$, the power of the second test is greater than that from the first test if

$$\sum_{y=0}^{\lfloor \alpha_{mc}(m+k) \rfloor - 1} \left(\begin{array}{c} m+k-1 \\ y \end{array} \right) p^y (1-p)^{m-y-1} > \sum_{y=0}^{\lfloor \alpha_{mc} m \rfloor - 1} \left(\begin{array}{c} m-1 \\ y \end{array} \right) p^y (1-p)^{m-y-1} \tag{9}$$

Observe that this inequality is equivalent to

$$\mathbb{P}(X \leq \lfloor \alpha_{mc}(m+k) \rfloor - 1) > \mathbb{P}(Y \leq \lfloor \alpha_{mc} m \rfloor - 1)$$

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

$$\mathbb{P}(X \leq \lfloor \alpha_{mc}(m+k) \rfloor - 1) = \mathbb{P}(X \leq \lfloor \alpha_{mc} m \rfloor - 1)$$

and the inequality (9) becomes

$$\mathbb{P}(X \leq \lfloor \alpha_{mc} m \rfloor - 1) > \mathbb{P}(Y \leq \lfloor \alpha_{mc} m \rfloor - 1)\,,$$

which is not valid. Thus, if $0 < k < 1/\alpha$, $\pi(m+k, \alpha_{mc}, F_P) \leq \pi(m, \alpha_{mc}, F_P)$. For example, for $\alpha_{mc} = 0.01$, a MC test with $m_1 = 1050$ is less powerful than a test with $m_2 = 1000$.

4. Discussion

A considerable parcel of the researchers, who are devoted to the development of statistical tests, are restricted to study the limit distribution of the test statistic to guide the criterium decision about acceptance/rejection the null hypothesis. Such works should consider Monte Carlo test application with the same entusiasm that is currently devoted to asymptotic study. What support this assertion is that Monte Carlo approach preserves power comparatively to the exact test and controls the real probability of the type one error, what is not always true in the asymptotic treatment, particularly for moderated sample sizes.

References

Assunção, R., Maia, A., 2007. A note on testing separability in spatial-temporal marked point processes. Biometrics 63 (1), 290–294.

Barnard, G., 1963. Discussion of professor bartlett´s paper. J. R. Statist. Soc. 25B (294).

Besag, J., Clifford, P., 1991. Sequential monte carlo p-value. Biometrika 78, 301–304.

Birnbaum, Z., 1974. Computers and unconventional test-statistics. in: F. proschan and r.j. serfling. Reliability and Biometry, 441–458.

Booth, J., Butler, R., 1999. An importance sampling algorithm for exact conditional tests in log-linear models. Biometrika 86, 321–332.

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

Dufour, J., 2005. Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. Jornal of Econometrics 133 (2), 443–477.

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

Fay, M., Follmann, D., 2002. Designing monte carlo implementations of permutation or bootstrap hypothesis tests. The American Statistician 56 (1), 63–70.

Fay, M., Kim, H.-J., 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.

Gandy, A., 2009. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. Journal of the American Statistical Association 104 (488), 1504–1511.

Hope, A., 1968. A simplified monte carlo significance test procedure. Journal of the Royal Statistical Society 30B, 582–598.

Jockel, K., 1984. Application of monte-carlo tests - some considerations. Biometrics 40 (1), 263–263.

Jockel, K., 1986. Finite sample properties and asymptotic efficiency of monte carlo tests. The Annals of Statistics 14, 336–347.

Kim, H.-J., 2010. Bounding the resampling risk for sequential monte carlo implementation of hypothesis tests. Journal of Statistical Planning and Inference (140), 1834–1843.

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

Marriott, F., 1979. Bernard's monte carlo test: How many simulations? Applied Statististics 28, 75–77.

Okamoto, M., 1958. Some inequalities relating to the partial sum of binomial probabilities. Annals of the Institute of Statistical Mathematics 10, 29–35.

Peng, R., Schoenberg, F., Woods, J., 2005. A space-time conditional intensity model for evaluating a wildfire hazard index. Journal of the American Statistical Association 100 (469), 26–35.

Ripley, B., 1992. Applications of monte-carlo methods in spatial and image-analysis. Lecture Notes in Economics and Mathematical Systems 376, 47–53.

Sackrowitz, H., Samuel-Cahn, E., 1999. P values as random variables - expeted p values. The American Statistician 53, 326–331.

Silva, I., Assunção, R., Costa, M., 2009. Power of the sequential monte carlo test. Sequential Analysis 28 (2), 163–174.

Wongravee, K., Lloyd, G., Hall, J., Holmboe, M., Schaefer, M., Reed, R., Trevejo, J., Brereton, R., 2009. Monte-carlo methods for determining optimal number of significant variables. application to mouse urinary profiles. Metabolomics 5 (4), 387–406.

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.

Optimal Generalized Sequential Monte Carlo Test

I. R. Silva*, R. M. Assunção

Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract

Conventional Monte Carlo tests require the simulation of m independent copies from the test statistic U under the null hypothesis H_0. The execution time of these procedures can be substantially reduced by a sequential monitoring of the simulations. The sequential Monte Carlo test power and its expected time are analytically intractable. The literature has evaluated the properties of sequential Monte Carlo tests implementations by using some restrictions on the probability distribution of the p-value statistic. Such restrictions are used to bound the resampling risk, the probability that the accept/reject decision is different from the decision from the exact test. This paper develops a generalized sequential Monte Carlo test that includes the main previous proposals and that allows an analytical treatment of the power and the expected execution time. This results are valid for any test statistic. We also bound the resampling risk and obtain optimal schemes minimizing the expected execution time within a large class of sequential design.

Keywords: Sequential Monte Carlo test, Power loss, p-value density, Resampling risk, Sequential design, Sequential probability ratio test.

1. Introduction

In the conventional Monte Carlo (MC) tests, the user selects the number m of simulations of the test statistic U under H_0. A Monte Carlo p-value is calculated based on the proportion of the simulated values that are larger or equal than the observed value of U, assuming that large values of U lead to the null hypothesis rejection. This procedure can take a long time to run if the test statistic requires a complicated calculation as, for example, those involved in complex models. These situations are exactly those where the MC tests are likely to be most useful, as analytical exact or asymptotic results concerning the test statistic U is hard to obtain. The adoption of sequential procedures to carry out MC tests is a way to reach a faster decision. In contrast with the fixed size MC procedure, in the sequential MC test the number of simulated statistics is a random variable. The basic idea is to stop simulating as soon as there is enough evidence either to reject or to accept the null hypothesis. For example, it is intuitively clear that, if the observed value of

*Corresponding author
Email addresses: irs@ufmg.br (I. R. Silva), assuncao@est.ufmg.br (R. M. Assunção)

U is close to the median of the first 100 simulated values, the null hypothesis is not likely to be rejected even if we perform another 950 simulations. If a valid p-value could be provided, most researchers would be confident to stop at this point. Sequential Monte Carlo tests are procedures that provide valid p-values in these situations.

Let X_t be the number of simulated statistics under H_0 exceeding the observed value u_0 at t-th simulation. In general, sequential MC procedures track the X_t evolution by checking if it crosses an upper or a lower boundary. When it does, the test is halted and a decision is reached. Typically, crossing the lower boundary leads to the rejection of the null hypothesis while the upper boundary crossing leads to the acceptance of the null hypothesis.

There are different proposals for a sequential Monte Carlo test in the statistical literature. Besag and Clifford (1991) proposed a very simple scheme that provides valid p-value for a sequential test with an upper bound $n-1$ in the number of simulations of U. It depends on a single tuning parameter h, making it extremely simple to use. We stop the simulations when $X_t = h$ for the first time and $t < n$. If $X_{n-1} < h$, the simulations are halted. If $h \leq l \leq n-1$ is the number of simulations carried out and if we stop at time t, the sequential p-value is given by

$$p_{BC} = \begin{cases} X_t/t, & \text{if } X_t = h, \\ (X_t+1)/n, & \text{if } X_t < h. \end{cases} \tag{1}$$

The support set of p_{BC} is

$$S = \{1/n, 2/n, \ldots, h/n, h/(n-1), \ldots, h/(h+2), h/(h+1), 1\} \ .$$

and we have $\mathbb{P}(P_s \leq a) = a$ under the null hypothesis if $a \in S$. This is a valid p-value estimator, because, a p-value estimator P_e is valid if $\mathbb{P}(P_e \leq b) \leq b$, where b is an element from the support set of P_e. Additional randomization can provide a continuous p-value with uniform distribution in the interval $(0, 1)$, rather than distributed on the discrete set S.

Therefore, the boundaries of Besag and Clifford (1991) are given by the horizontal line $X_t = h$ and the vertical line $t = n-1$. There is no lower boundary but only a predetermined maximum number of simulations, typically called a truncated sequential Monte Carlo test. The Besag and Clifford sequential MC test brings a reduction in execution time only when the null hypothesis is true. When it is false, one will often run the Monte Carlo simulation up to its upper bound $n-1$. Therefore, additional gains could be obtained by adopting a stopping criterium based on large values of X_t. For any fixed type I error probability α, Silva et al. (2009) showed that one can design a Besag and Clifford sequential MC test with the same power as a conventional Monte Carlo test and with shorter running time. Silva et al. (2009) showed also the puzzling result that this sequential Monte Carlo should have a maximum sample size equal to $h/\alpha + 1$, because, for $n \geq h/\alpha + 1$, the power is constant.

In addition to Besag and Clifford (1991), alternative sequential Monte Carlo tests have been suggested recently. These other procedures are mainly concerned with the resampling risk, defined by Fay and Follmann (2002) as the probability that the test decision of a realized MC test will be different from a theoretical MC test with an infinite number of replications. Fay and Follmann (2002) proposed the curtailed sampling design, where, if $X_t \geq \lfloor \alpha(n+1) \rfloor$, the procedure is interrupted and H_0 is not rejected, and, if $t - X_t \geq \lceil (1-\alpha)(n+1) \rceil$ or the number of simulations reaches n, the procedure is interrupted and H_0 is rejected, where n is the maximum number of simulations. They also introduced the interactive push out (IPO) procedure that requires a sequential algorithm to define the boundaries of the sequential procedure. This procedure is not proven to be optimal but simply to decrease the sample size with respect to a curtailed sampling design. For all their results, Fay and Follmann (2002) assumed a specific class of distribution for the p-value statistic, that distribution implied by a test statistic U that follows the standard normal distribution under the null hypothesis and follows a $N(\mu, 1)$ under the alternative hypothesis. Conditional to this class of distributions, they found numerically the worst distribution to bound the resampling risk. IPO has a smaller expected execution time than the curtailed sampling design but its implementation is not practical for bounding the resampling risk in arbitrarily low values such as 0.01, for example. Also, we think that the assumption on the p-value distribution is too restrictive and, in fact, we show that it is not necessary to obtain optimal procedures.

Fay et al. (2007) proposed an algorithm (and an R package) to implement a truncated Sequential Probability Ratio Test (tSPRT) to bound the resampling risk and studied its behaviour as a function of the p-value. The algorithm, denoted here as the FKH algorithm, calculates a valid p-value, which depends on the calculation of the number of ways to reach each point on the stopping boundary of the MC test.

Gandy (2009) proposed an algorithm to build a sequential MC test that uniformly bounds the resampling risk in arbitrarily small values and provides lower bounds to the expected number of simulations. His algorithm is not truncated and the expected number of simulations can be infinite for p-values close to α. Therefore, the simulations may go on indefinitely. One missing issue in his paper is the lack of results concerning the type I error probability when the number of simulations is truncated.

Kim (2010) explored the approach from Fay and Follmann (2002) to bound the resampling risk using their same restrictive class of p-value distributions. She used the B-value boundaries proposed by Lan and Wittes (1988) and applied the algorithm of Fay et al. (2007) to obtain valid p-values estimates. She was able to obtain arbitrarily low bounds to the resampling risk and showed empirically that the B-value boundaries produces a smaller expected number of simulations than the IPO designs. In this paper, she also defined an approximated B-value procedure, which is easy to calculate and has analytical formulas that give insights on the choice of parameter values of the exact B-value design.

These B-value boundaries have the main advantages from the other procedures cited and, in our opinion,

is the best alternative for a sequential MC test at the moment. However, its main results, concerning the resampling risk and the expected number of simulations, depend on the same restrictive class of p-value distributions of Fay and Follmann (2002). Moreover, important topics were not explored for the B-value boundaries such as, for example, its power with respect to the conventional MC test or the establishment of lower bounds for the expected number of simulations for any test statistic.

In this paper, we introduce a generalized sequential Monte Carlo allowing any monotonic shapes for the boundaries. For example, it is possible to construct boundaries which are close to each other in the beginning of the simulations, departing from each other as the simulations proceed and approaching each other again in the end of the simulations. We have been able to obtain bounds for the power loss of the sequential MC test. In fact, we establish boundaries shapes such that the sequential MC test has the same power as the conventional MC test for any α level. These boundaries are simple to calculate and they are valid in the general case of any p-value distribution. Moreover, we are able to provide an algorithm to find the truncated boundaries that lead to a design with minimum expected sampling size. Concerning the resampling risk, we consider a larger class of distributions for the p-value than Fay and Follmann (2002) and we show that it is suitable to explicit algebraic manipulation allowing simple bounding of the resampling risk for any sequential MC test design.

This paper is organized in the following way. In the next section, we describe the B-value boundaries. Section 3 defines our sequential MC test and develops its properties. In Section 4 we discuss a general class for the p-value distribution and provide some analytical results for the sequential tests. Section 5 presents a numeric routine for the preliminary choice of our boundaries and some specific suggestions for practical use. Section 6 offers a comparison between the B-value procedure and our procedure. Section 7 closes the paper with some discussion.

2. The B-value Procedure

Consider a hypothesis test of a null hypothesis H_0 against an alternative hypothesis H_a by means of a test statistic U. The MC test can be seen as an estimation procedure to the unknown decision from the exact test based on the null hypothesis distribution of U. Kim (2010) has adopted this point of view by seeing the MC test as a decision procedure concerning in which $(0, 1)$ interval, either $(0, \alpha]$ or $(\alpha, 1)$, does belong the exact p-value associated with the test statistic U. The parameter α is the significance level of the exact test. This interpretation leads to the following pair of hypotheses:

$$\begin{aligned} H_0^* &: p \leq \alpha \\ H_A^* &: p > \alpha \end{aligned} \tag{2}$$

where p is the observed and unknown p-value generated from the random variable p-value. Viewed as a random variable, we denote the p-value by P. Clearly, the decision in favor of any hypotheses above leads to a decision concerning the original hypotheses H_0 and H_A.

Let U be the test statistic, u_0 be its observed value for a fixed sample and $u_i, i = 1, ...,$ be the independently simulated values from U under H_0. Let

$$X_t = \sum_{i=1}^{t} 1_{\{[u_0,\infty)\}}(u_i)\,,$$

where $1_{\{[u_0,\infty)\}}(u_i)$ is the indicator function that $u_i \geq u_0$.

Kim (2010) used the B-value introduced by Lan and Wittes (1988) to propose a sequential procedure to test H_0^* versus H_A^*. Define:

$$V(t) = min\left\{s \geq 0 : x - tx \geq c_1\sqrt{n\alpha(1-\alpha)}\right\}$$

and

$$L(t) = max\left\{s \geq 0 : x - t\alpha \leq c_2\sqrt{n\alpha(1-\alpha)}\right\}\,.$$

Define also:

$$B_{\text{Sup}} = \left\{(t,x) = (t, \min\{V(t), r_1\}) : t = t_0^+, t_0^+ + 1, ..., n)\right\}\,,$$

the upper boundary, and

$$B_{\text{Inf}} = \left\{(t,x) = (t, \max\{L(t), t - r_0\}) : t = t_0^-, t_0^- + 1, ..., n)\right\}\,,$$

the lower boundary of a sequential Monte Carlo test, where t_0^+ is the smaller value of t such that $V(t) \leq t$ and t_0^- is the smaller value of t such that $L(t) \geq 0$. Similarly, let t_1^+ be the smaller value of t such that $V(t) \geq r_1$ and t_1^- the smaller value of t such that $L(t) \leq t - r_0$. The stopping boundaries from Kim (2010) are given by $B = B_{\text{Inf}} \cup B_{\text{Sup}}$. The B boundaries are formed by the union of linear functions in t. Figure 1 illustrate the B-boundaries B_{Sup} and B_{Inf} using $c_1 = -c_2 = 1.282$, $n = 600$ and $\alpha = 0.05$.

The upper boundary B_{Sup} is formed by the union of the line $V(t) = c_1\sqrt{n\alpha(1-\alpha)} + t\alpha$ until $t = t_1^+$, when the upper boundary becomes the horizontal line with height $r_1 = \lfloor\alpha(n+1)\rfloor$. The lower boundary B_{Inf} is formed by the line $L(t) = c_2\sqrt{n\alpha(1-\alpha)} + t\alpha$ up to $t = t_1^-$ when it becomes the vertical line $r_0 = t - \lceil(1-\alpha)(n+1)\rceil$.

Kim (2010) uses ϕ_{FKH}, the test criterium based on the valid p-value presented in Fay et al. (2007). The valid p-value is defined as $\hat{p}_v(X_t, t) = F_{\hat{p}_{MLE}}(X_t/t)$, where $\hat{p}_{MLE}$ is the maximum likelihood estimator of p and $F_{\hat{p}}$ is defined in (5.2) from Fay et al. (2007). The estimate $\hat{p}_v(X_t, t)$ of the p-value can be computed using the FKH algorithm. The test adopted by Kim (2010) for the B boundaries is given by:

$$\phi_{FKH}(t,x) = \begin{cases} 1, & \text{if } \hat{p}_v(x,t) \leq \alpha \\ 0, & \text{if } \hat{p}_v(x,t) > \alpha. \end{cases}$$

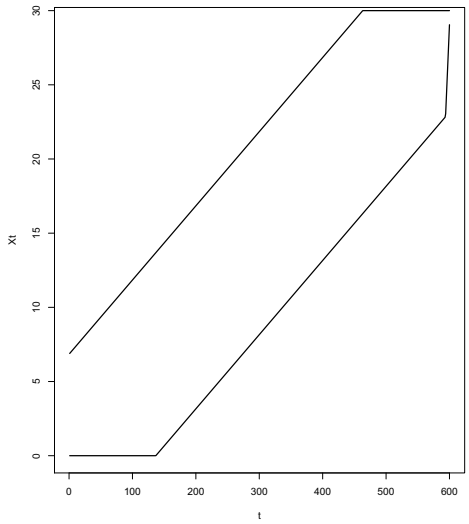


Figure 1: Example of the B boundaries with $\alpha = 0.05$ and maximum number of simulations equal to $n = 600$.

When $\phi_{FKH}(t, x) = 0$, H_0 is not rejected (because $H_0^* : p \leq \alpha$ is rejected). For $\phi_{FKH}(t, x) = 1$, H_0 is rejected (that is, $H_0^* : p \leq \alpha$ is not rejected). Henceforth, this procedure is called MC_B.

It is very important to remark that there is no need to check the value of X_t at every moment t. To see this, noticed that the boundaries B_{Sup} and B_{Inf} are composed by non-integer numbers while X_t is a count. As a consequence, there will be times t for which the simulations can not be interrupted by B_{Inf} and therefore there is no need to check against the lower boundary at these times. To illustrate this, consider the example from Kim (2010) illustrated in Figure 1. Table 1 shows the values of B_{Sup} and B_{Inf} between the times 134 and 179. The lower boundary is equal to zero until $t = 136$ and it is formed by numbers smaller than 1 until $t = 156$. Therefore, X_t reach the lower boundary during this period if $X_{137} = 0$ and there is no need to check against it for $t \leq 136$. Likewise, if X_t is not interrupted by B_{Inf} at $t = 156$ (that is, $X_{156} \geq 2$), it will not reach it at least until $t = 176$. Therefore, in practice, there is no need to check against the lower boundary for every simulated value. One needs to check only on those times t such that

$$B_{\text{Inf}}(t-1) < B_{\text{Inf}}(t)$$

for $t = 2, \ldots, m$ where $B_{\text{Inf}}(t)$ is the value of the lower boundary at time t. This will be explored by our generalized sequential Monte Carlo method described in Section 3.

Since B_{Sup} will typically be non-integer, it is always possible to define step functions equivalent to the upper boundary. To see this, consider again the Table 1. From $t = 134$ to $t = 143$, it is clear that the values $B_{\text{Sup}}(t)$ could be all substituted by 14 and the procedure would remain the same.

2.1. Bounding the resampling risk of MC_B

Fay and Follmann (2002) considered the IPO procedure that, interactively with the current simulations, adjusts the initial boundaries. This method allows the bounding of the resampling risk. The IPO procedure is not described in details here, but it should be noted that it is a computationally intensive procedure, and its implementation is intractable for bounding the resampling risk in arbitrarily small values (see (Kim, 2010)). Fay and Follmann (2002) considered a rather restrictive class of p-value distributions, with cumulative distribution function given by:

$$H_{\alpha,1-\beta}(p) = 1 - \Phi\left\{\Phi^{-1}(1-p) - \Phi^{-1}(1-\alpha) + \Phi^{-1}(\beta)\right\} \tag{3}$$

where $\Phi(.)$ is the cumulative distribution function of a standard Normal distribution, α is the desired significance level and β is the type II error probability. When $\alpha = 1 - \beta$, the cumulative distribution $H_{\alpha,1-\beta}(p)$ has a uniform distribution on $(0, 1)$, as is expected when H_0 is true.

The p-value distribution defined in (3) assumes a variety of shapes, but the analytical manipulation of the resampling risk or of the expected number of simulations is intractable. To circumvent this problem, Fay and Follmann (2002) used a Beta(a, b) distribution to approximate $H_{\alpha,1-\beta}(p)$, and this approximation is denoted by $\tilde{H}_{\alpha,1-\beta}(p)$. This approximation is chosen such that the expected value of P coincides with that from $H_{\alpha,1-\beta}(p)$ and such that $\tilde{H}_{\alpha,1-\beta}(\alpha) = H_{\alpha,1-\beta}(\alpha) = 1 - \beta$. Numerical studies were performed by Fay and Follmann (2002) to obtain the worst case $\tilde{F}$ within the class (3) in the sense of having the largest resampling risk. Let $\tilde{F}^*$ be the correspondent Beta distribution approximation to $\tilde{F}$.

Although MC_B is simpler and present a smaller expected time execution than the IPO procedure, it depends on the FKH algorithm which requires rather complex modifications for each type of sequential design. Kim (2010) proposes an approximation for the MC_B procedure. With this approximation, if B_{Sup} is reached before B_{Inf}, H_0^* is rejected, while H_A^* is accepted if B_{Inf} is reached first. The approximation may be used to gain analytic insights on the properties of the MC_B procedure or to help on choosing the parameters c_1, c_2, and n, as well as providing an approximation for the expected number of simulations. An undesirable characteristic of the approximated MC_B is that it is not truncated and the expected number of simulations must be calculated letting the maximum number of simulations go to infinity. Moreover, the approximation to MC_B does not offer guarantee that the type I error probability is under control for any choice of c_1 and c_2. For this reason, the approximation MC_B will not be explored here.

3. Our proposed generalized sequential Monte Carlo test

The analytical treatment of the MC test power function, when it is based on two interruption boundaries, is a cumbersome task. The reason is that it involves the calculation of the large number of possible trajectories of the random variable X_t responsible for H_0 rejection. Fay et al. (2007) present an algorithm to calculate

the terms associated with such number, and they used this algorithm to obtain both, the expected number of simulations and the resampling risk, for each fixed p-value. Fay et al. (2007) emphasize that such algorithm is valid only for the specific sequential procedure treated in that article, and adjustments are needed to use it with other sequential designs. Kim (2010) also used that algorithm for her calculations, and the approximate MC_B is an attempt to escape from the dependence on special algorithms.

Aiming to overcome this limitation, we propose a truncated sequential procedure with two boundaries that have the shape of step functions. The values of $X(t)$ are checked against the upper boundary for every t while they are checked against the lower boundary in an arbitrary set of predetermined discrete moments, possibly a smaller set than all integers between 1 and m. As we showed in Section 2, the B-boundaries can also be expressed by step functions with jumps equal to positive integer numbers. Therefore, the boundaries of MC_B and of our sequential procedure can be expressed in the same way. To express the boundaries by means of step functions is more cumbersome in terms of notation. The motivation for this design, where the lower boundary monitoring is not carried out for every time t, is mainly to allow for the analytical treatment of the power function, the expected number of simulations of the sequential MC test for any test statistic. We also bound the resampling risk of our sequential MC test.

Let $\eta^I = \{n_1^I, n_2^I, ..., n_{k_1}^I\}$, with $n_j^I < n_{j+1}^I$, be a set containing the moments when X_t must be checked against the lower boundary given by the values $I = \{I_1, I_2, ..., I_{k_1}\}$. If $X_{n_j^I} < I_j$, the simulations are interrupted and H_0 is rejected.

The monitoring of X_t with respect to the upper boundary crossing is carried out at all moments $t = 1, \ldots, m$ and this upper boundary is a step function. Let $\eta^S = \{n_1^S, n_2^S, ..., n_{k_2}^S\}$, with $n_j^S < n_{j+1}^S$ be the jump moments for the upper boundary. For $n_{j-1}^S \leq t < n_j^S$, the upper boundary is given by S_j where $n_0^S = 0$ and $S_1 < S_2 \ldots < S_{k_2}$. Let $S = \{S_1, S_2, ..., S_{k_2}\}$. Therefore, the simulations are interrupted if $D_t = 1$, where:

$$D_t = \begin{cases} 1, & \text{if } (t \in \eta^I \text{ and } X_t < I_j, \text{ for } t = n_j^I) \text{ or } (X_t = S_j, \text{ for } n_{j-1}^S < t \leq n_j^S) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

or if the number of simulations reach a predetermined maximum equal to m.

Let x_t be the observed value of the random variable X_t. The p-value can be estimated by:

$$p_I = \begin{cases} x_t/t, & \text{if } x_t = S_j, n_{j-1}^S < t \leq n_j^S \\ (x_t + 1)/(t + 1), & \text{if } x_t < I_j, t = n_j^I. \end{cases}$$

We define the test decision function for this sequential test:

$$\phi_I(t, x) = \begin{cases} 1, & \text{if the lower boundary } I \text{ is reached before the upper } S \text{ or the simulations reach } m \\ 0, & \text{if the upper boundary } S \text{ is reached before the lower } I. \end{cases}$$

The hypothesis H_0 is rejected if $\phi_I = 1$ and it is not rejected if $\phi_I = 0$. This sequential MC test will be denoted by MC_G.

As an example, take $k_1 = k_2 = 10$, $m = 600$, and consider $I = \{0, 1, 2, 3, 9, 15, 20, 24, 27, 29\}$ for the lower boundary values, $S = \{5, 7, 9, 13, 17, 23, 26, 29, 29, 30\}$ for the upper boundary values, and $\eta^I = \eta^S = \{20, 50, 79, 119, 239, 359, 459, 539, 569, 600\}$. Figure 2 shows these boundaries as dashed lines.

The choice of the boundaries is closely linked to the desired α_{mc}, which is equal to 0.05 in this example. In Section 5, we present an algorithm to obtain the appropriate boundaries for any α_{mc} and m in an easy and fast way. The solid lines are the B boundaries calculated by Kim (2010) using $c_1 = -c_2 = 1.282$, $n = 600$, and $\alpha = 0.05$.

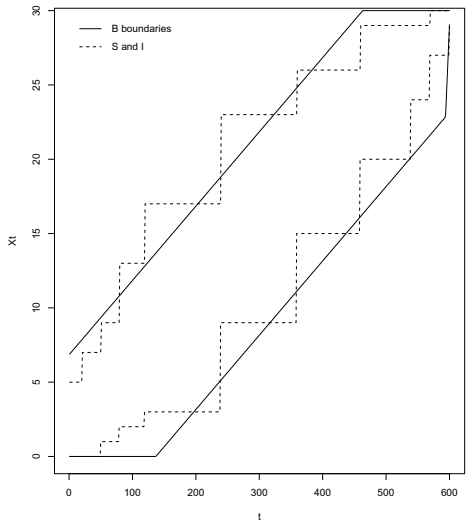


Figure 2: Example of the MC_G (S and I) and B boundaries with $\alpha = 0.05$ and a maximum number of simulations equal to $m = 600$.

3.1. Power and Size of the MC_G

In the MC_G procedure, the rejection of H_0 occurs in the first moment $t = n_j^I$ such that $x_t < I_j$. The power calculation is simpler if we merge the two sets η^I and η^S. Define $\eta = \eta^I \cup \eta^S = \{n_1, n_2, ..., n_k\}$ with $k = \#\eta$. Let $S^{'} = \left\{S_1^{'}, S_2^{'}, ..., S_k^{'}\right\}$ be the upper boundary adjusted for each $n_i \in \eta$ in the following way. If $n_i = n_j^S \in \eta^S$ for some j, then $S_i^{'} = S_j$. If $n_i \in \eta^I \cap \left(\eta^S\right)^c$, then $S_i^{'} = S_j$ where j is such that $n_j^S = \max\left\{n_r^S < n_i\right\}$. Thus, if n_i matches with some jump time in the set η^S, then $S_i^{'}$ is equal to the value in S for the time n_i. If n_i is not an element in η^S, then $S_i^{'}$ is the jump value of the time immediately preceding n_i.

Similarly, let $I^{'} = \left\{I_1^{'}, I_2^{'}, ..., I_k^{'}\right\}$ be the adjusted lower boundary. That is, when $n_i = n_j^I \in \eta^I$ or some j, then $I_i^{'} = I_j$. If $n_i \in \eta^S$ but $n_i \notin \eta^I$, then $I_i^{'} = I_j$ where j is such that $n_j^I = \max\left\{n_j^I < n_i\right\}$.

Thus, for a given value of $p \in (0,1)$, the power function of the MC_G procedure, is given by:

$$\begin{aligned}
\pi_G(p) &= \sum_{x_1=0}^{I_1'-1} \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} + \\
&+ \sum_{x_1=I_1'}^{\min\left\{S_1'-1, I_2'-1\right\}} \sum_{y=0}^{\min\left\{I_2'-x_1-1, n_2-n_1\right\}} \binom{n_2-n_1}{y} p^{y}(1-p)^{n_2-n_1-y} \times \\
&\times \left[\binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} \right] + \\
&+ \sum_{j=2}^{k-1} \sum_{x_j=I_j'}^{\min\left\{S_j'-1, I_{j+1}'-1\right\}} \sum_{y=0}^{\min\left\{I_{j+1}'-x_j-1, n_{j+1}-n_j\right\}} \sum_{x_{j-1}=I_{j-1}'}^{\min\left\{S_{j-1}'-1, x_j\right\}} \cdots \\
&\cdots \sum_{x_1=I_1'}^{\min\left\{S_1'-1, x_2\right\}} \binom{n_{j+1}-n_j}{y} \times \\
&\times \binom{n_1}{x_1} p^{y+x_j}(1-p)^{n_{j+1}-y-x_j} \prod_{i=2}^{j} \binom{n_i-n_{i-1}}{x_i-x_{i-1}}. \qquad (5)
\end{aligned}$$

This expression is composed by k summands. If k is not too large, the direct application of this expression produces results quickly and easily. The calculation would be computationally hard if we used a similar expression for sequential procedure where $k = m$. Note that, in the MC_B procedure, the number of summands in (5) can reach up to $2\alpha_{mc}m$.

Under the null hypothesis, p follows the $U(0,1)$ distribution. Hence, by integrating out (5) with respect to p with a $U(0,1)$ density, we obtain the type I error probability for MC_G:

$$\begin{aligned}
\mathbb{P}\left(\text{type I error}\right) &= \int_0^1 \pi_G(p)dp = \frac{I_1'}{n_1+1} + \\
&+ \sum_{x_1=I_1'}^{\min\left\{S_1'-1, I_2'-1\right\}} \sum_{y=0}^{\min\left\{I_2'-x_1-1, n_2-n_1\right\}} \frac{\binom{n_2-n_1}{y}\binom{n_1}{x_1}}{(n_2+1)\binom{n_2}{y+x_1}} + \\
&+ \sum_{j=2}^{k-1} \sum_{x_j=I_j'}^{\min\left\{S_j'-1, I_{j+1}'-1\right\}} \sum_{y=0}^{\min\left\{I_{j+1}'-x_j-1, n_{j+1}-n_j\right\}} \sum_{x_{j-1}=I_{j-1}'}^{\min\left\{S_{j-1}'-1, x_j\right\}} \cdots \\
&\cdots \sum_{x_1=I_1'}^{\min\left\{S_1'-1, x_2\right\}} \frac{\binom{n_{j+1}-n_j}{y}\binom{n_1}{x_1}\prod_{i=2}^{j}\binom{n_i-n_{i-1}}{x_i-x_{i-1}}}{(n_{j+1}+1)\binom{n_{j+1}}{y+x_j}}.
\end{aligned}$$

Similarly to Silva and Assunção (2011), an upper bound for the power difference between MC_G and the exact test can be obtained by:

$$b_G = \max_{p\in(0,1)} \left\{1_{(0,\alpha]} - \pi_G(p)\right\} \tag{6}$$

where α is the significance level of the exact test.

The power function of $\pi_G(p)$ evaluated for a fixed p is equal to the probability of X_t reaching I before reaching S, and this probability is decreasing with p. In this way, the largest power loss of MC_G as compared to the exact test is given by:

$$b_G = \max_{p\in(0,\alpha]} \{1 - \pi_G(\alpha)\} = 1 - \pi_G(\alpha). \tag{7}$$

Let MC_m be the conventional MC test performed with a fixed number m of simulations. An upper bound for the power difference between MC_m and MC_G is given by:

$$b_{m,G} = \max_{p\in(0,1)} \{\pi_m(p) - \pi_G(p)\} \tag{8}$$

where $\pi_m(p) = \mathbb{P}(G \leq \lfloor m\alpha_{mc} \rfloor - 1)$ is the power function of MC_m for a given p, and G is distributed according to a binomial distribution with parameters $m-1$ and p.

3.2. Expected Number of Simulations for MC_G

Let L be the random variable that represents the number of simulations carried out until the halting moment. To perform the computation of the expectation of L, obtained by $\mathbb{E}(L|P=p) = \sum_{l=1}^{n_k} l\, \mathbb{P}(L = l|P=p)$, for each fixed p. The probability $\mathbb{P}(L=l|P=p)$ is given by:

$$\mathbb{P}(L=l|P=p) = \begin{cases} \begin{pmatrix} l-1 \\ l-S_1' \end{pmatrix} p^{l-S_1'}(1-p)^{S_1'} & \text{if } l < n_1 \\ \begin{pmatrix} l-1 \\ l-S_1' \end{pmatrix} p^{l-S_1'}(1-p)^{S_1'} + \sum_{x=0}^{I_1'-1} \begin{pmatrix} n_1 \\ x \end{pmatrix} p^x(1-p)^{n_1-x} & \text{if } l = n_1 \\ \sum_{x=0}^{I_1'-1} \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \begin{pmatrix} l-n_1-1 \\ l-n_1-(S_2'-x) \end{pmatrix} p^{S_2'}(1-p)^{l-S_2'} & \text{if } n_1 < l < n_2 \\ \sum_{x=0}^{I_1'-1} \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \begin{pmatrix} l-n_1-1 \\ l-n_1-(S_2'-x) \end{pmatrix} p^{S_2'}(1-p)^{l-S_2'} + & \\ \sum_{x=I_1'}^{min\left\{S_1'-1, I_2'-1\right\}} \sum_{y=0}^{min\left\{I_2'-x-1, n_2-n_1\right\}} \begin{pmatrix} n_2-n_1 \\ y \end{pmatrix} p^y(1-p)^{n_2-n_1-y} \times & \\ \times \left[\begin{pmatrix} n_1 \\ x \end{pmatrix} p^x(1-p)^{n_1-x} \right] & \text{if } l = n_2. \end{cases}$$

We need to consider this calculation depending on l being equal to one of the n_j or not. For $l = n_j, j =$

$3, ..., k-1$, we have:

$$\begin{aligned}
\mathbb{P}(L=l|P=p) &= \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{y=0}^{min\left\{I'_j-x_{j-1}-1, n_j-n_{j-1}\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
&\cdots \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \begin{pmatrix} n_j-n_{j-1} \\ y \end{pmatrix} \times \\
&\times \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} p^{y+x_{j-1}}(1-p)^{n_j-y-x_{j-1}} \prod_{i=2}^{j-1} \begin{pmatrix} n_i-n_{i-1} \\ x_i-x_{i-1} \end{pmatrix} + \\
&+ \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
&\cdots \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \begin{pmatrix} n_j-n_{j-1}-1 \\ n_j-n_{j-1}-(S'_j-x_{j-1}) \end{pmatrix} \times \\
&\times \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} p^{l-S'_j}(1-p)^{S'_j} \prod_{i=2}^{j-1} \begin{pmatrix} n_i-n_{i-1} \\ x_i-x_{i-1} \end{pmatrix}.
\end{aligned}$$

For $n_{j-1} < l < n_j, j = 3, ..., k$:

$$\begin{aligned}
\mathbb{P}(L=l|P=p) &= \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
&\cdots \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \begin{pmatrix} n_j-n_{j-1}-1 \\ n_j-n_{j-1}-(S'_j-x_{j-1}) \end{pmatrix} \times \\
&\times \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} p^{l-S'_j}(1-p)^{S'_j} \prod_{i=2}^{j-1} \begin{pmatrix} n_i-n_{i-1} \\ x_i-x_{i-1} \end{pmatrix}.
\end{aligned}$$

Finally, for $l = n_k$:

$$\begin{aligned}
\mathbb{P}(L=l|P=p) &= \sum_{x_{j-1}=I'_{j-1}}^{min\left\{S'_{j-1}-1, I'_j-1\right\}} \sum_{y=0}^{min\left\{I'_j-x_{j-1}-1, n_j-n_{j-1}\right\}} \sum_{x_{j-2}=I'_{j-2}}^{min\left\{S'_{j-2}-1, x_{j-1}\right\}} \cdots \\
&\cdots \sum_{x_1=I'_1}^{min\left\{S'_1-1, x_2\right\}} \begin{pmatrix} n_j-n_{j-1} \\ y \end{pmatrix} \times \\
&\times \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} p^{y+x_{j-1}}(1-p)^{n_j-y-x_{j-1}} \prod_{i=2}^{j-1} \begin{pmatrix} n_i-n_{i-1} \\ x_i-x_{i-1} \end{pmatrix}. \qquad (9)
\end{aligned}$$

Using that p has a $U(0,1)$ distribution under the null hypothesis, we have

$$\mathbb{E}\left(L|H_0 \text{ is true}\right) = \int_0^1 \mathbb{E}(L|P=p)dp\,. \tag{10}$$

To calculate $\mathbb{E}(L)$ under H_A it is necessary to know the p-value distribution. However, a bound is easier to calculate as

$$b_{E(L)} = max_{p\in(0,1)}\left\{\mathbb{E}(L|P=p)\right\}. \tag{11}$$

$b_{E(L)}$ is a very conservative upper bound for $E(L)$. However, as we will illustrate in Section 5, this bound is useful to bound the expectation of L in values less than 65% of m.

4. A class of distributions for the p-value

Kim (2010) showed that, for $p = \alpha$, the resampling risk is at least 0.5. Hence, it is not possible to bound the resampling risk in relevant values if we allow all distributions of p-values. This is the reason to define a class for the p-value distribution, taken as the set $\Im$ of all continuous probability distributions in $(0,1)$ with differentiable densities that are non-increasing (that is, $f'_P(p) \leq 0$, for all $p \in (0,1)$, with f'_P representing the first derivative with respect to p of the p-value density function f.

From the p-value definition, its probability distribution function can be written in the following way:

$$\mathbb{P}(P \leq p) = 1 - F_A\left\{F_0^{-1}(1-p)\right\} \tag{12}$$

where F_A denotes the probability distribution function of the test statistic U under H_A and F_0 is the distribution of U under H_0.

Assuming the existence of densities functions $f_A(u)$ and $f_0(u)$ of U under H_A and H_0, respectively, the p-value density can be written as:

$$f_P(p) = \frac{f_A\left\{F_0^{-1}(1-p)\right\}}{f_0\left\{F_0^{-1}(1-p)\right\}}. \tag{13}$$

Hence, we can study the behavior of the p-value distribution by studying the behavior of the ratio between $f_A(u)$ and $f_0(u)$.

In the majority of the real applications, the ratio (13) is non-increasing with p and this is the motivation to restrict the analysis of the resampling risk to the set $\Im$. Let $\Im_B$ be the class of p-value distributions defined in Fay and Follmann (2002) with cumulative distribution $H_{\alpha,1-\beta}(p)$, as described in Section 2. Let π be the power of the exact test. We will show now that, for $\pi \geq \alpha$, $\Im$ is more general than $\Im_B$.

From the expression (3), the densities $h(p) \in \Im_B$ can be indexed by α and β and they are given by:

$$h_{\alpha,1-\beta}(p) = \exp\left\{-\frac{1}{2}\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha)\right]\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha) + 2\Phi^{-1}(1-p)\right]\right\} \tag{14}$$

where Φ^{-1} is the inverse function of the standard normal cumulative distribution function $\Phi(\cdot)$. The first derivative of $h_{\alpha,1-\beta}(p)$ with respect to p is equal to:

$$h'_{\alpha,1-\beta}(p) = \frac{\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha)\right]}{\phi(\Phi^{-1}(1-p))} h_{\alpha,1-\beta}(p) \tag{15}$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. For $1-\beta \geq \alpha$, we have $h'_{\alpha,1-\beta}(p) \leq 0$ for all $p \in (0,1)$.

Consider the subset of densities $\Im_B^* = \{f_P(p) \in \Im_B : 1-\beta \geq \alpha\}$. That is, $\Im_B^*$ is a subset from $\Im_B$ formed only by densities that implies an exact test power greater or equal to α. Therefore, $\Im_B^* \subset \Im$. Thus, at least for useful test statistic ($\mathbb{P}(P \leq p) \geq \alpha$), the class $\Im_B$ is a particular case from $\Im$.

The formulation of the class $\Im_B$ in Fay and Follmann (2002) was inspired on the behavior of the p-value distribution for the cases where $U_0 \sim N(0,1)$ and $U_A \sim N(\mu,1)$, with $\mu = \Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)$, which results in a distribution with shape $H_{\alpha,1-\beta}(p)$. Fay and Follmann (2002) have explained that this same distribution can be derived from the cases where $U_0 \sim \chi^2_1(0)$ and $U_A \sim \chi^2_1(\mu^2)$, where $\chi^2_1(\mu^2)$ is the random variable with non-central Chi-square distribution with 1 degree of freedom and non-centrality parameter equal to μ^2. They argued that, for the cases in which $U \sim F_{1,d}(\mu^2)$ the p-value distribution converges in distribution to $H_{\alpha,1-\beta}(p)$ when $d \to \infty$, where $F_{1,d}(\mu^2)$ is the random variable with F distribution with 1 and d degrees of freedom and non-centrality parameter equal to μ^2.

The class $\Im_B$ is smaller than $\Im$ and does not cover all cases of interest. For example, the spatial scan statistic developed by Kulldorff (2001) to detect spatial clusters follows very closely a Gumbel distribution under the null hypothesis and a chi-square distribution under H_A (see Abrams et al. (2010)). Therefore, even in interesting applied situations, there is not guarantee that $f_P(p) \in \Im_B$ and a larger class such as our $\Im$ may be useful.

It is worth mentioning that $h_{\alpha,1-\beta}(p)$ is a convex function when $1-\beta \geq \alpha$ and $p \leq 0.5$. Indeed, the second derivative of $h_{\alpha,1-\beta}(p)$ with respect to p is given by:

$$h''_{\alpha,1-\beta}(p) = \frac{\left[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha) - \phi'(\Phi^{-1}(1-p))\right]}{\phi(\Phi^{-1}(1-p))} h'_{\alpha,1-\beta}(p) \tag{16}$$

and we have that

$$\phi'(\Phi^{-1}(1-p)) = \frac{\Phi^{-1}(1-p)}{\sqrt{2\pi}\phi(\Phi^{-1}(1-p))} \exp\left\{-1/2\left[\Phi^{-1}(1-p)\right]^2\right\} \geq 0$$

if $p \leq 0.5$.

Cases where the real situation of the data presents a small distance from H_0 are examples of applications in that the density of the p-value could escape from $\Im_B$. When the p-values tend to small values, in direction to α, that is the situation where the p-value density is deforming, from an uniform density, to an asymmetric

curve to the left hand, the convexity could not be verified for $p \leq 0.5$. For example, suppose $U_0 \sim \chi^2_1(0)$ e $U_A \sim \chi^2_{1,01}(0)$. The corresponding p-value density from this conjecture is not concave for $p > 0.32$.

The family $\Im$ for bounding the resampling risk is not restricted to families such as the normal, chi-square or F distributions. It also contains p-value densities with mixed shapes, with concave and convex parts. As an additional benefit, $\Im$ allows the bounding of the resampling risk in a very simple way.

In the next subsections, we analyze the power, the expected number of simulations and the resampling risk of our generalized Monte Carlo test procedure when the p-value distribution belongs to the class $\Im$. It is important to remember that, when using the MC_G, the class $\Im$ is not needed neither to calculate a bound for the power loss with respect to the MC_m or to the exact test nor to establish the bound for the expected number of simulations under H_A. Indeed, the results in the Sub-sections 3.1 and 3.2 are valid for any test statistic. However, when the additional assumption that the p-value density $f_P(p)$ belongs to $\Im$ holds, stronger results can be obtained.

4.1. Upper bound for the power difference between the exact test and MC_G

The power of the generalized Monte Carlo test is given by integrating out the probability $\pi_G(p)$ of rejecting the null hypothesis conditioned on the p-value p with respect to the p-value density:

$$\pi_G = \int_0^1 \pi_G(p) f_P(p) dp \, .$$

The power difference between the exact test and MC_G is given by:

$$\delta^*_G = \int_0^1 \left(1_{(0,\alpha_{mc}]}(p) - \pi_G(p)\right) f_P(p) dp. \tag{17}$$

An upper bound for δ^*_I can be obtained if we use $f_{P,w}(p) = 1/\alpha_{mc}$ if $p \in (0, \alpha_{mc}]$, and $f_{P,w}(p) = 0$, otherwise:

$$\begin{aligned} \delta^*_G \leq b^*_I &= \int_0^1 \left(1_{(0,\alpha_{mc}]}(p) - \pi_G(p)\right) f_{P,w}(p) dp = \int_0^{\alpha_{mc}} \frac{1}{\alpha_{mc}} dp - \int_0^{\alpha_{mc}} \pi_G(p) \frac{1}{\alpha_{mc}} dp \\ &= 1 - \frac{1}{\alpha_{mc}} \int_0^{\alpha_{mc}} \pi_G(p) dp. \end{aligned} \tag{18}$$

Because the function (5) is a sum of Beta(a, b) density kernels, the integral (18) can be rewritten as a function of incomplete Beta(a, b) functions, all of them evaluated at $p = \alpha_{mc}$, with a and b depending only of the parameters I, S e η. In the same way, an upper bound for the power difference between MC_m and MC_G is given by:

$$b^*_{m,G} = \int_0^{\alpha_{mc}} \left(\pi_m(p) - \pi_G(p)\right) \frac{1}{\alpha_{mc}} dp. \tag{19}$$

As before, (19) can also be expressed using incomplete beta functions.

4.2. An upper bound for the expected number of simulations

For values of p near 0, the simulation time is around n_1, the first checking point of the lower boundary. For values of p near 1, the simulation time is around $S_1^{'}$, the smallest height of the upper boundary. Numerically, we find that $\mathbb{E}(L|P=p)$ is maximized for p around α_{mc}. Let

$$p_{\max} = \arg\max_p \mathbb{E}(L|P=p)$$

and define $f_{P,max}(p) = 1/p_{max}$, for $p \in (0, p_{max}]$, and $\bar{f}_{P,max}(p) = 0$, otherwise. Thus, it follows that

$$\mathbb{E}(L) = \int_0^1 \mathbb{E}(L|P=p)\bar{f}_P(p)dp \leq \int_0^1 \mathbb{E}(L|P=p)\bar{f}_{P,max}(p)dp = \int_0^{1/p_{max}} \mathbb{E}(L|P=p)\frac{1}{p_{max}}dp. \tag{20}$$

The right hand side of the inequality (20) defines an upper bound $b^*_{E(L)}$ for $\mathbb{E}(L)$.

4.3. An upper bound for the resampling risk

Let RR be the resampling risk in a MC test defined as:

$$RR = \mathbb{P}_{mc}(H_0 \text{ is not rejected}|P \leq \alpha)\mathbb{P}(P \leq \alpha) + \mathbb{P}_{mc}(H_0 \text{ is rejected}|P \geq \alpha)\mathbb{P}(P \geq \alpha) \tag{21}$$

where $\mathbb{P}_{mc}$ is the probability measure associated with the events generated by MC simulations. For the MC_G test, denote its resampling risk by RR_G, which is computed as:

$$RR_G = \int_0^\alpha [1 - \pi_G(p)]\, f_P(p)dp + \int_\alpha^1 \pi_G(p) f_P(p)dp. \tag{22}$$

As $\pi_G(p)$ is a decreasing function, the function $\left[1_{p\in(0,\alpha]}(p) - \pi_G(p)\right]$ is maximum at $p = \alpha$. Thus, RR_G is maximum when $f_P(p)$ puts the largest possible mass at α, which is the worst case $f_{P,w}(p)$. Substituting $f_P(p)$ in (22) by $f_{P,w}(p)$ and setting $\alpha = \alpha_{mc}$, we have :

$$RR_G \leq 1 - \frac{1}{\alpha_{mc}} \int_0^{\alpha_{mc}} \pi_G(p)dp. \tag{23}$$

Therefore, an upper bound for RR_G is equal to the upper bound (18) for the power loss with respect to the exact test. That is, $b^*_{RR_G} = b^*_G$.

The expression (22) can be rewritten in a way that emphasizes another property. The situation where $\pi \geq \pi_G$ is that where the control of RR_G is important. If $\pi \geq \pi_G$, then $RR_G \geq \delta_I$, where δ_I is the power difference between the exact test and MC_G. Therefore, equal power of the exact test and the MC_G test does not imply a null resampling risk. To see this:

$$\begin{aligned} RR_G &= \int_0^{\alpha_{mc}} f_P(p)dp - \int_0^{\alpha_{mc}} \pi_G(p) f_P(p)dp + \int_{\alpha_{mc}}^1 \pi_G(p) f_P(p)dp \\ &= \pi - \pi_G + 2\int_{\alpha_{mc}}^1 \pi_G(p) f_P(p)dp = \delta_I + 2\int_{\alpha_{mc}}^1 \pi_G(p) f_P(p)dp. \end{aligned} \tag{24}$$

5. Choosing Parameters to Operate MC_G

This section aims to provide the reader with a useful set of choices for the parameters I, S and η to run the MC_G test. The choices we suggest produce a MC_G test with power equal to a MC_m test for any test statistic with small expected number of simulations.

Optimizing $\mathbb{E}(L)$ analytically is undoubtedly a complex task. In contrast, a numeric approach is feasible and simple to operate, and this is the approach adopted here. Define the class M, the set of MC_G procedures that, under H_0, leads to the same decision about rejecting H_0 than the MC_m. Conditioned on this class M, the three next steps were developed to estimate the parameters of the MC_G with minimum $\mathbb{E}(L)$. Let $MC_{I_{op}}$ be such scheme with minimum $\mathbb{E}(L)$.

1. This step is intended to emulate the X_t path under H_0. Generate N observations from an $U(0,1)$ distribution, and label them as p_i, $i = 1, ..., N$. For each p_i, generate m values x_{ij}, with $j = 1, \ldots, m$ following a Bernoulli distribution with success probability p_i. Define the partial sum processes

$$S_i = \left\{ S_{it}, \text{ such that } S_{it} = \sum_{l=1}^{t} x_{il}, t = 1, ..., m \right\} .$$

2. We build envelopes for the path X_t based on the simulated ones. For that, select those S_i sequences leading to the rejection of H_0 by MC_m. That is, to be selected the sequence S_i must satisfy $\max_t \{S_{it}\} < m\alpha_{mc}$. Suppose there are s of those sequences and they form the set $\mathcal{R}$. If N is large, we expect $s/N \approx \alpha_{mc}$. Define the sequence $\hat{S}_t = \{\max_i \{S_{it}\} + 1, i \in \mathcal{R}\}$. The curve $\hat{S}_t$ is an estimator for the upper boundary of $MC_{G_{op}}$.

   Next, take the r sequences S_i such that $\max_t \{S_{it}\} \geq m\alpha_{mc}$ and collect them in the set $\mathcal{A}$. These are the sequences $S_i's$ that do not reject H_0. Define the sequence $\hat{I}_t = \{\min_i \{S_{it}\}, i \in \mathcal{A}\}$. The curve $\hat{I}_t$ is the estimator for the lower boundary of $MC_{G_{op}}$.

3. Take the set $\hat{\eta^S}$ containing the jumping moments of $\hat{S}$. $\hat{\eta^S}$ is an estimator for η^S associated to $MC_{G_{op}}$. Take also the set $\hat{\eta^I}$ formed by the jumping moments of $\hat{I}$. $\hat{\eta^I}$ is an estimator of η^I associated to $MC_{G_{op}}$. Formally:

$$\hat{\eta^I} = \left\{ \hat{n}_t^I = \hat{n}_{t-1}^I \; if \; \left\lceil \hat{I}_t \right\rceil = \left\lceil \hat{I}_{t-1} \right\rceil, \text{ or } \hat{n}_t^I = t \text{ if } \left\lceil \hat{I}_t \right\rceil > \left\lceil \hat{I}_{t-1} \right\rceil \right\} \tag{25}$$

   with, $t = 2, ..., m$ and $\hat{n}_1^I = \min \left\{ l : \hat{I}_l > 0 \right\}, l = 1, ..., m$. Also,

$$\hat{\eta^S} = \left\{ \hat{n}_t^S = \hat{n}_{t-1}^S \; if \; \left\lfloor \hat{S}_t \right\rfloor = \left\lfloor \hat{S}_{t-1} \right\rfloor, \; or \; \hat{n}_t^S = t \; if \; \left\lfloor \hat{S}_t \right\rfloor > \left\lfloor \hat{S}_{t-1} \right\rfloor \right. \tag{26}$$

   with, $t = 2, ..., m$ and $\hat{n}_1^S = min \left\{ \hat{S}_t \right\}$. The estimation procedure ends here.

As an heuristic argument to show that these boundaries estimated using this algorithm are indeed estimates of the $MC_{G_{op}}$ boundaries, consider that, for N sufficiently large, $\hat{S}_t$ and $\hat{I}_t$ are constructed to ensure

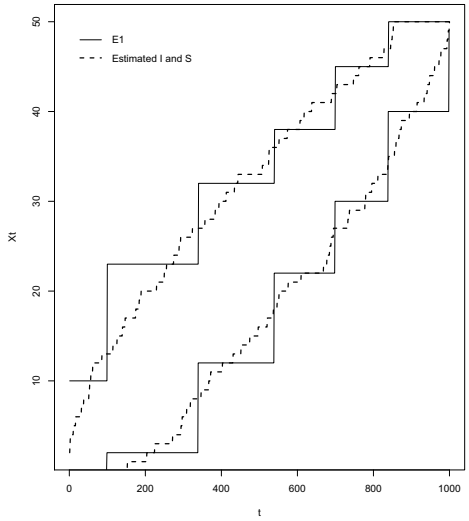


Figure 3: Scheme E_1 from Table 5 versus estimates for I and S considering $m = 1000$, $N = 100000$ and $\alpha_{mc} = 0.05$

that, under H_0, the decision of the MC_G will be always the same as that reached with MC_m. In addition, if $f_P(p) \in \Im$, we have

$$\mathbb{P}(X_t \text{ reach } I \text{ before reach } S \mid H_0) \leq \mathbb{P}(X_t \text{ reach } I \text{ before reach } S \mid H_A),$$

and then the power of these estimated boundaries is, at least, equal to that from MC_m. Concerning the expected number of simulations, $\mathbb{E}(L)$ decreases by increasing the I elements and by decreasing the S elements. The construction of $\hat{S}_t$ and $\hat{I}_t$ follows this logic. The reasoning is to scan each t taking the maximum value for I which does not restrict the X_t trajectories which would not reject H_0 by using MC_m. Simultaneously, it takes the minimum value for S which does not restrict the X_t trajectories which would reject H_0 by using MC_m.

The resulting estimators of η^I and η^S are quite sparse using this algorithm, while they could be computationally costly if calculated by means of the expressions developed in Section 3. An alternative and satisfactory way to construct η^I and η^S is based on the identification of the moments with high incidence of impact of X_t with the estimated boundaries. For that, let $n_q^* = \min\left\{t \in [1, ..., m], t : S_{it} \leq \hat{I}, i \in \mathcal{R}\right\}$ be an element of the sequence formed by the impact moments of each sequence S_{it} with $\hat{I}$ (considering only those sequences $i \in \mathcal{R}$). The most frequent impact moments of these sequences S_{it} with $\hat{I}$ are appropriate candidates for composing η^I. Apply the same reasoning for constructing η^S, and denote the correspondent sequence by n_r^*. Thus, as an alternative way to construct η^I and η^S, by arbitrary and conveniently low values k_1 and k_2, choose the most frequent elements in n_q^* and n_r^* to compose η^I and η^S, respectively. Extensive simulated examples indicate that, for $k_1 = k_2 \geq 5$, the exact computation of power loss and expected number

of simulations have a small computational cost, and the results are close to that using $\hat{\eta^I}$ and $\hat{\eta^S}$.

Figure 3 shows the estimates $\hat{I}$ and $\hat{S}$ obtained according to the steps 1, 2, and 3, of our algorithm, using $N = 100000$, $m = 1000$ and $\alpha_{mc} = 0.05$. The estimated boundaries are not parallel, but they are characterized by a funnel in the extremities. This behavior was verified in all of the simulations performed by us. For these specific estimates, if we take $\eta = \eta^S = \eta^I$, we obtain the times $\hat{\eta} = \{99, 339, 539, 699, 839, 999\}$. From the estimates plotted in Figure 3, we obtain $\hat{I} = \{2, 12, 22, 30, 40, 49\}$, and $\hat{S} = \{10, 23, 32, 38, 45, 50\}$. This specific scheme for our generalized sequential MC test is available in Table 5 and it is labeled as E_1. As we can see in the Table 2, column $b_{m,I}$, this scheme is efficient, presenting practically the same power than MC_m for $m = 1000$, with size equal to 0.049864. From Table 3, columns $\mathbb{E}(L|H_0)$ and $b_{E(L)}$, we see that this scheme have a small expected number of simulations, equal to 58.606 under H_0, and with an upper bound under H_A for any statistic, that is approximately 65% of the maximum 999. By using the class $\Im_B$ and the larger class $\Im$ for the p-value distribution, the bounds are expressively low, equal to 172.612 and 246.354, respectively. We consider that this scheme is a good option to replace MC_m. We must enphasize that, although the boundaries presented in Table 5 were guided by the algorithm above, all results in tables 2 and 3 are exact, because they were obtained by applying the expressions from Sections 3 and 4. Such algorithm is useful to construct preliminary choices of boundaries. The validation of an arbitrary design to practical use must be based on such exact calculations.

We provide other interesting schemes in Table 5. For each scheme, Table 2 offers the type I error probability, the upper bound for the power loss comparatively to MC_m and to the exact test, and the upper bound for the resampling risk. Table 3 gives the expected number of simulations under H_0 and the upper bounds under H_A. We adopt $b_{m,G}$ to denote the general upper bound for the power loss comparatively to MC_m, b^*_G and $b^*_{RR_G}$, the upper bounds for the power loss, with respect to the exact test, and for the resampling risk, respectively, where the super index $*$ indicates that the calculations are restricted to the p-value distribution on the class $\Im$. The same symbol was used to indicate the use of this class for the bounds in Table 3. Upper bounds using the class $\Im_B$ are also available in Tables 2 and 3, and they are indicated by a tilde accent. Concerning the use of $\Im_B$, the numerical explorations of Fay and Follmann (2002) were not used here to define the worst case of a p-value distribution with shape $H_{\alpha;1-\beta}(p)$. As discussed in Section 4.3, $h_{\alpha,1-\beta}(p)$ (for $1 - \beta \geq \alpha$) and π_G are decreasing with p. Therefore, the worst case within the class $\Im_B$, in the sense of bounding RR_I, occurs at the point of maximum of the function $H_{\alpha;1-\beta}(\alpha)$ with respect to β. For $1 - \beta \geq \alpha$, the point of maximum in β for the function $H_{\alpha;1-\beta}(\alpha)$ is 0.5. Then, the analytical worst case is given by $H_{\alpha;0.5}(\alpha)$. We used this result to compute $\tilde{b}_I$, $\tilde{b}_{RR_G}$ and $\tilde{b}_{E(L)}$ here.

6. MC_G versus MC_B

In this Section we offer a comparison between the MC_B and MC_G sequential test procedures. We use an example of the MC_B test given by Kim (2010). In this comparison, we focus on resampling risk bound and on the expected number of simulations. We assume that the p-value distribution $f_P(p)$ belongs to the class $\Im_B$. We did not consider other important characteristics of a test, such as the power loss with respect to the exact test and expected number of simulations for an arbitrary $f_P(p)$, because they were not treated by Kim (2010). We built our MC_G boundaries using the algorithm from Section 5. After securing an upper bound for the resampling risk to MC_G equal to that presented by the MC_B scheme developed in Kim (2010), we compared the average simulation time of the two procedures.

An obvious fact is that the B boundaries are particular cases of I and S, because MC_G was designed to be a generalized sequential with two stopping boundaries. We can rewrite the B boundaries using the MC_G notation, based on the sets I, S, η^I and η^S. In this way, for an MC_B test, the user can apply the general expressions for the power and the expected number of simulations developed in Section 3.

Define:

$$T_1^* = \left\{t > 2 : \left\lceil B_{\text{Inf}}(t-1)\right\rceil < \left\lceil B_{\text{Inf}}(t)\right\rceil\right\}$$

and

$$T_2^* = \left\{t > 2 : \left\lfloor B_{\text{Sup}}(t-1)\right\rfloor < \left\lfloor B_{\text{Sup}}(t)\right\rfloor\right\} .$$

Let $t_{11}^* < t_{12}^* < ... < t_{1k_1}^*$ be the ordered elements of T_1^*, and $t_{21}^* < t_{22}^* < ... < t_{2k_2}^*$ be the ordered elements of T_2^*. Rewritten in terms of I and S, the B boundaries are denoted by I^*, S^*, η^{I*} and η^{S*}, and they are built as follows:

$$I^* = \left\{\left\lceil B_{Inf}(t_{11}^*)\right\rceil, \left\lceil B_{Inf}(t_{12}^*)\right\rceil, ..., \left\lceil B_{Inf}(t_{1k_1}^*)\right\rceil\right\}$$

$$S^* = \left\{\left\lfloor B_{Inf}(t_{21}^*)\right\rfloor, \left\lfloor B_{Inf}(t_{22}^*)\right\rfloor, ..., \left\lfloor B_{Inf}(t_{2k_2}^*)\right\rfloor\right\}$$

$$\eta^{I*} = \left\{t_{11}^*, t_{12}^*, ..., t_{1k_1}^*\right\}$$

$$\eta^{S*} = \left\{t_{21}^*, t_{22}^*, ..., t_{2k_2}^*\right\}.$$

It should be noted that some important shapes for I and S, as the funnel behavior estimated in Section 5, can not be represented by the MC_B boundaries.

Table 4 shows the upper bounds for the resampling risk and for the expected number of simulations presented in Kim (2010) for $n = 600$, $\alpha = 0.05$ and $c_1 = -c_2 = 1.282$, as well the bounds associated to MC_G

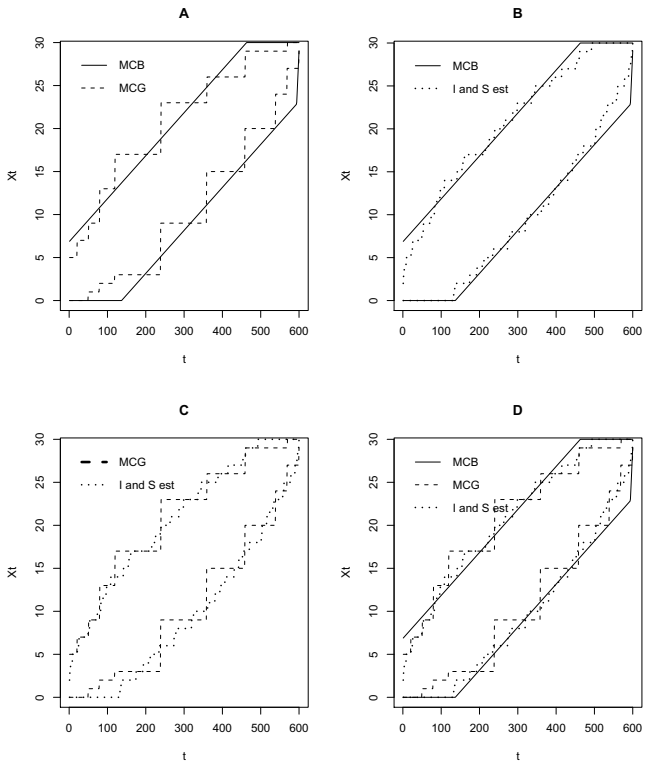


Figure 4: Boundaries for MC_G and MC_B sequential test procedures.

using the scheme E_{12} detailed in Table 5. Concerning the worst case of distribution within the class $\Im_B$, Kim (2010) adopted the numerical studies from Fay and Follmann (2002) and she found $\tilde{F}^* = H_{0.05;0.47}(p)$ for $\alpha = 0.05$, with approximation $\tilde{H}_{0.05;0.47}(p) := Beta(0.389; 2.523)$.

These bounds are also computed here for the sequential procedure proposed by Besag and Clifford (1991), which will be denoted by MC_h. This MC_h procedure is a very simple way to perform sequential tests, because it is based just in an upper boundary fixed in a value denoted by h and truncated in a maximum number of simulations n. Silva et al. (2009) had showed that MC_h has the same power than MC_m if $h = \alpha_{mc}m$ and its power is constant for $n \geq h/\alpha_{mc} + 1$, noting that the combination of this two last rules implies that MC_m must be replaced by MC_h, because they have the same power for a maximum number of simulations practically equal (that is, for $n = m + 1$).

Under either hypothesis, MC_G is substantially faster than the MC_B procedure with their expected time ratio being around 60%. This illustrates the gain provided by our MC_G algorithm. The estimated boundaries for $m = 600$ and $\alpha_{mc} = 0.05$ for the MC_G procedure are available in Figure 4, where we can also see the B boundaries and the boundaries I and S (built according to the scheme E_{12}). It is clear the greater flexibility

given by the MC_G boundaries. While the B boundaries are parallel almost up to the end of the experiment, the I and S boundaries are tapered when t gets close to the maximum number of simulations. This can be intuitively thought as if the boundaries were using the information that X_t not touching the boundaries after a long time was inducing a narrower vigilance.

7. Discussion

The generalized sequential Monte Carlo test has properties that recommend it in substitution to the conventional Monte Carlo test for any test statistic. In this paper, we gave simple expressions for the calculation of the test size, the expected number of simulations under the null hypothesis, and upper bounds for the expected number of simulations under the alternative hypothesis and for the power loss with respect to the fixed length conventional Monte Carlo test.

Exact calculations for some specific design indicates that our generalized sequential test has a substantially smaller execution time than other sequential methods proposed in the literature. Under a wide class of distributions for the p-value statistic, the bounds for the execution time under the alternative hypothesis is even more substantial. Under this class, the generalized sequential test has power virtually identical to the exact test, even for such intermediate maximum number of simulations as 4999. The use of this class allows the construction of optimal boundaries from a simple algorithm and they have a surprising funnel-type shape. These optimal boundaries, or any other generalized design, for any test statistic, can be evaluated by calculating the size, expected number of simulations under H_0, upper bound for the power loss and for the expected number of simulations under H_A, by using the expressions developed in section 3.

t	B_{Sup}	B_{Inf}	t	B_{Sup}	B_{Inf}
⋮	⋮	⋮	⋮	⋮	⋮
134	13.54	0.00	157	14.69	1.01
135	13.59	0.00	158	14.74	1.06
136	13.64	0.00	159	14.79	1.11
137	13.69	0.01	160	14.84	1.16
138	13.74	0.06	161	14.89	1.21
139	13.79	0.11	162	14.94	1.26
140	13.84	0.16	163	14.99	1.31
141	13.89	0.21	164	15.04	1.36
142	13.94	0.26	165	15.09	1.41
143	13.99	0.31	166	15.14	1.46
144	14.04	0.36	167	15.19	1.51
145	14.09	0.41	168	15.24	1.56
146	14.14	0.46	169	15.29	1.61
147	14.19	0.51	170	15.34	1.66
148	14.24	0.56	171	15.39	1.71
149	14.29	0.61	172	15.44	1.76
150	14.34	0.66	173	15.49	1.81
151	14.39	0.71	174	15.54	1.86
152	14.44	0.76	175	15.59	1.91
153	14.49	0.81	176	15.64	1.96
154	14.54	0.86	177	15.69	2.01
155	14.59	0.91	178	15.74	2.06
156	14.64	0.96	179	15.79	2.11
⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Boundary values $B_Sup(t)$ and $B_{Inf}(t)$ for $134 \leq t \leq 179$.

α	n_k	Scheme	$\mathbb{P}$(erro tipo I)	$b_{m,I}$	$\tilde{b}_I$	$b_I^* = b_{RR_G}^*$	$\tilde{b}_{RR_G}$
		E_1	0.049864	0.031032	0.000000	0.060000	0.023000
		E_2	0.049681	0.031748	0.000000	0.060895	0.023337
0.05	999	E_3	0.049920	0.027828	0.000000	0.060118	0.022348
		E_4	0.049982	0.024751	0.000000	0.053977	0.022826
		E_5	0.049998	0.020890	0.000218	0.058518	0.021326
0.01	999	E_6	0.009999	0.030136	0.002921	0.136997	0.028048
		E_7	0.009993	0.000000	0.003344	0.140063	0.028443
0.05	4999	E_8	0.050401	0.000000	0.014908	0.021841	0.016746
0,01		E_9	0.010003	0.027643	0.000577	0.061952	0.012354
0.05	9999	E_{10}	0.050036	0.008435	0.000000	0.017947	0.001838
0.01		E_{11}	0.009992	0.023415	0.000106	0.043559	0.008821

Table 2: Effective test size, upper bound for the power loss, and for the resampling risk associated to the schemes in the Table 5.

α	n_k	Scheme	$\mathbb{E}(L\|H_0)$	$b_{E(L)}$	$\tilde{b}_{E(L)}$	$b_{E(L)}^*$
		E_1	58.606	644.654	172.612	246.354
		E_2	96.739	807.572	235.119	402.704
0.05	999	E_3	113.597	792.104	264.587	407.943
		E_4	108.472	670.761	276.496	396.895
		E_5	128.276	698.242	335.969	474.473
0.01	999	E_6	42.717	677.409	225.362	550.574
		E_7	41.728	625.841	225.334	523.601
0.05	4999	E_8	384.591	4283.575	1016.884	1716.704
0.01		E_9	121.992	3580.891	1816.592	687.744
0.05	9999	E_{10}	731.131	9477.250	1951.502	3502.344
0.01		E_{11}	190.355	8706.908	1182.462	3968.661

Table 3: Expected Number of Simulations for Schemes from Table 5.

	MC_G E_{12}	MC_B with $n = 600$ $c_1 = -c_2 = 1.282$	IPO with $n = 576$	MC_h with $h = 30$ and $m = 600$
P(erro tipo I)	0.050000	0.050000	0.050000	0.050000
$\tilde{b}$	0.001888	≤ 0.025000	≤ 0.025000	0.001888
$\tilde{b}_{RR}$	0.024562	0.025000	0.025000	0.024562
$\mathbb{E}(L\|H_0)$	33.720	51.169	62.850	119.351
$\tilde{b}_{E(L)}$	91.877	163.118	213.508	390.507

Table 4: Upper bounds for the resampling risk and expected number of simulations for comparison among MC_G using E_{12}, MC_B, with $n = 600$, $\alpha = 0.05$ and $c_1 = -c_2 = 1.282$, and IPO.

E_1	$I = \{2, 12, 22, 30, 40, 49\}$ $S = \{10, 23, 32, 38, 45, 50\}$ $\eta = \{99, 339, 539, 699, 839, 999\}$
E_2	$I = \{2, 12, 20, 35, 49\}$ $S = \{19, 30, 37, 45, 49\}$ $\eta = \{99, 379, 539, 779, 999\}$
E_3	$I = \{7, 16, 24, 35, 49\}$ $S = \{28, 35, 41, 49, 49\}$ $\eta = \{199, 499, 699, 899, 999\}$
E_4	$I = \{13, 22, 30, 35, 49\}$ $S = \{28, 35, 42, 45, 49\}$ $\eta = \{299, 559, 719, 799, 999\}$
E_5	$I = \{17, 26, 35, 42, 49\}$ $S = \{34, 39, 43, 46, 49\}$ $\eta = \{399, 639, 799, 899, 999\}$
E_6	$I = \{1, 5, 7, 8, 9\}$ $S = \{8, 9, 9, 9, 9\}$ $\eta = \{299, 599, 799, 899, 999\}$
E_7	$I = \{2, 5, 7, 8, 9\}$ $S = \{8, 8, 8, 9, 9\}$ $\eta = \{399, 599, 769, 899, 999\}$
E_8	$I = \{27, 100, 199, 249\}$ $S = \{80, 150, 219, 249\}$ $\eta = \{799, 2499, 3999, 4999\}$
E_9	$I = \{5, 20, 35, 49\}$ $S = \{20, 31, 43, 49\}$ $\eta = \{799, 2499, 3999, 4999\}$
E_{10}	$I = \{17, 79, 499\}$ $S = \{79, 249, 499\}$ $\eta = \{499, 2999, 9999\}$
E_{11}	$I = \{79, 199, 499\}$ $S = \{199, 499, 499\}$ $\eta = \{2399, 4999, 9999\}$
E_{12}	$I = \{0, 1, 2, 3, 9, 15, 20, 24, 27, 29\}$ $S = \{5, 7, 9, 13, 17, 23, 26, 29, 29, 30\}$ $\eta = \{20, 50, 79, 119, 239, 359, 459, 539, 569, 600\}$

Table 5: Appropriate Schemes for Replacing MC_m by MC_G.

References

Abrams, A., Kleinman, K., Kulldorff, M., 2010. Gumbel based p-value approximations for spatial scan statistics. International Journal of Health Geographics 9 (61).

Besag, J., Clifford, P., 1991. Sequential monte carlo p-value. Biometrika 78, 301–304.

Fay, M., Follmann, D., 2002. Designing monte carlo implementations of permutation or bootstrap hypothesis tests. The American Statistician 56 (1), 63–70.

Fay, M., Kim, H.-J., 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.

Gandy, A., 2009. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. Journal of the American Statistical Association 104 (488), 1504–1511.

Kim, H.-J., 2010. Bounding the resampling risk for sequential monte carlo implementation of hypothesis tests. Journal of Statistical Planning and Inference 140, 1834–1843.

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

Lan, K., Wittes, J., 1988. The b-value: a tool for monitoring data. Biometrics 44, 579–585.

Silva, I., Assunção, R., 2011. Monte carlo test under general conditions: Power and number of simulations. Paper submitted to Journal of Statistical Planning and Inference.

Silva, I., Assunção, R., Costa, M., 2009. Power of the sequential monte carlo test. Sequential Analysis 28 (2), 163–174.

Acknowledgements

We are grateful to Martin Kulldorff for very useful comments and suggestions in the manuscript.