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.