Testes Monte Carlo Seqüenciais

Renato Assunção
Universidade Federal de Minas Gerais
Departamento de Estatística

Roteiro

1. Testes de Hipótese via Monte Carlo
2. Teste Monte Carlo Seqüencial
3. Escolha dos parâmetros de sintonia (tuning)
4. Comparação dos poderes de testes seqüenciais e de tamanho fixo
5. Análise empírica

Método Monte Carlo

- São algoritmos para simular o comportamento de sistemas estocásticos.
- Os sistemas estocásticos podem ser:
  - complexos (tais como simular os efeitos de terremotos para efeito de cálculo de risco atuarial)
  - ou simples (tais como gerar valores de uma variável aleatória X)
- Métodos de simulação Monte Carlo: existem desde que ... ????
- Parece que existem desde sempre...

Inventando Monte Carlo

- Os inventores do método Monte Carlo foram
  - Stanislaw Ulam, um matemático polonês.
  - John von Neumann
- Ambos trabalhavam no Projeto Manhattan dos EUA durante a Segunda Grande Guerra.
- Ulam é famoso for planejar a bomba de hidrogênio com Edward Teller em 1951.
- John von Neumann dispensa apresentações.
- Eles inventaram o método Monte Carlo em 1946-1947
- Em 2007: 60 anos da criação de Monte Carlo

THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
PRINCETON, NEW JERSEY

May 21, 1947

Mr. Stan Ulam
Post Office Box 1663
Santa Fe
New Mexico

Dear Stan,

Thanks for your letter of the 19th. I need not tell you that Klari and I are looking forward to the trip and visit at Los Alamos this Summer. I have already received the necessary papers from Carson Mark. I filled out and returned mine yesterday; Klari's will follow today.

I am very glad that preparations for the random numbers work are to begin soon. In this connection, I would like to mention this: Assume that you have several random number distributions, each equidistributed in $0, 1$: (x^i), (y^i), (z^i), ... Assume that you want one with the distribution function (density) $f(\xi)\,d\xi$: (ξ^i) . One way to form it is to form the cumulative distribution function: $g(\xi) = \int^{\xi} f(\xi)\,d\xi$ to invert it $h(x) = \xi \rightleftarrows x = g(\xi)$, and to form $\xi^i = h(x^i)$ with this $h(x)$, or some approximant polynomial. This is, as I see, the method that you have in mind.

An alternative, which works if ξ and all values of $f(\xi)$ lie in $0, 1$, is this: Scan pairs x^i, y^i and use or reject x^i, y^i according to whether $y^i \leqq f(x^i)$ or not. In the first case, put $\xi^j = x^i$ in the second case form no ξ^j at that step.

The second method may occasionally be better than the first one. In some cases combinations of both may be best; e.g., form random pairs

$$\xi = \sin x, \quad \eta = \cos x$$

with x equidistributed between 0° and 360°. The obvious way consists of using the sin - cos - tables (with interpolation). This is clearly closely related to the first method. This is an alternative procedure:

Put $\xi = \frac{2t}{1+t^2}$, $\eta = \frac{1-t^2}{1+t^2}$, $t = \operatorname{tg} y$,

with y (which is $\frac{x}{2}$) equidistributed between 0° and 180°. Restrict y to 0° to 45°. Then the ξ, η will have to be replaced randomly by η, ξ and again by $\pm\xi, \pm\eta$. This can be done by using random digits $0, \ldots, 7$. It is also feasible with

Data

Destinatário

Aqui, von Neumann descreve o método de Ulam: gerar $X \sim F^{-1}(U)$ onde F é a acumulada de X e $U \sim \text{Unif}(0,1)$

Aqui, von Neumann descreve o seu método de aceitação-rejeição pela primeira vez como uma alternativa

Random digits 0, . . . , 9:

0	Replace ξ, η	by ξ, η
1	"	$-\xi, \eta$
2	"	$\xi, -\eta$
3	"	$-\xi, -\eta$
4	"	η, ξ
5	"	$\eta, -\xi$
6	"	$-\eta, \xi$
7	"	$-\eta, -\xi$
8	Reject this digit	
9	" " "	

Now $t = tg\, y$, $0° \leqq y \leqq 45°$, lies between 0 and 1, and its distribution function is $\frac{dt}{1+t^2}$. Hence one may pick pairs of numbers t, s both (independently) equidistributed between 0 and 1, and then

use t } for $(1+t^2)\, s \leqq 1$

reject t, s and form no t at this step } for $(1+t^2)\, s > 1$

Of course, the first part requires a divider, but the method may still be worth keeping in mind, especially when the ENIAC is available.

* * *

With best regards from house to house.

Yours, as ever,

John

~~John von Neumann~~

JvN:MW

É uma carta de apenas duas páginas.

Stan Ulam e Metropolis

- **Stanislaw Ulam sofreu uma cirurgia no cérebro em 1946 e inventou o método Monte Carlo enquanto convalescia.**
- **O primeiro artigo sobre o método apareceu em 1949 no Journal of the American Statistical Association.**
- **Ele tinha Metropolis como co-autor (mas não von Neumann).**

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 247 SEPTEMBER 1949 *Volume 44*

THE MONTE CARLO METHOD

NICHOLAS METROPOLIS AND S. ULAM
Los Alamos Laboratory

We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences.

ALREADY in the nineteenth century a sharp distinction began to appear between two different mathematical methods of treating physical phenomena. Problems involving only a few particles were studied in classical mechanics, through the study of systems of ordinary differential equations. For the description of systems with very many particles, an entirely different technique was used, namely, the method of statistical mechanics. In this latter approach, one does not concentrate on the individual particles but studies the properties of *sets of particles*. In pure mathematics an intensive study of the properties of sets of points was the subject of a new field. This is the so-called theory of sets, the basic theory of integration, and the twentieth century development of the theory of probabilities prepared the formal apparatus for the use of such models in theoretical physics, i.e., description of properties of aggregates of points rather than of individual points and their coordinates.

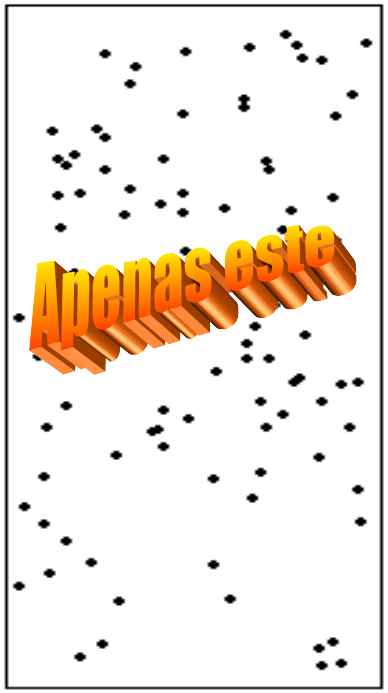
Soon after the development of the calculus, the mathematical apparatus of partial differential equations was used for dealing with the problems of the physics of the continuum. Hydrodynamics is the most widely known field formulated in this fashion. A little later came the treatment of the problems of heat conduction and still later the field theories, like the electromagnetic theory of Maxwell. All this is very well known. It is of course important to remember that the study of the

- O primeiro artigo...
- **E o resto é história...**

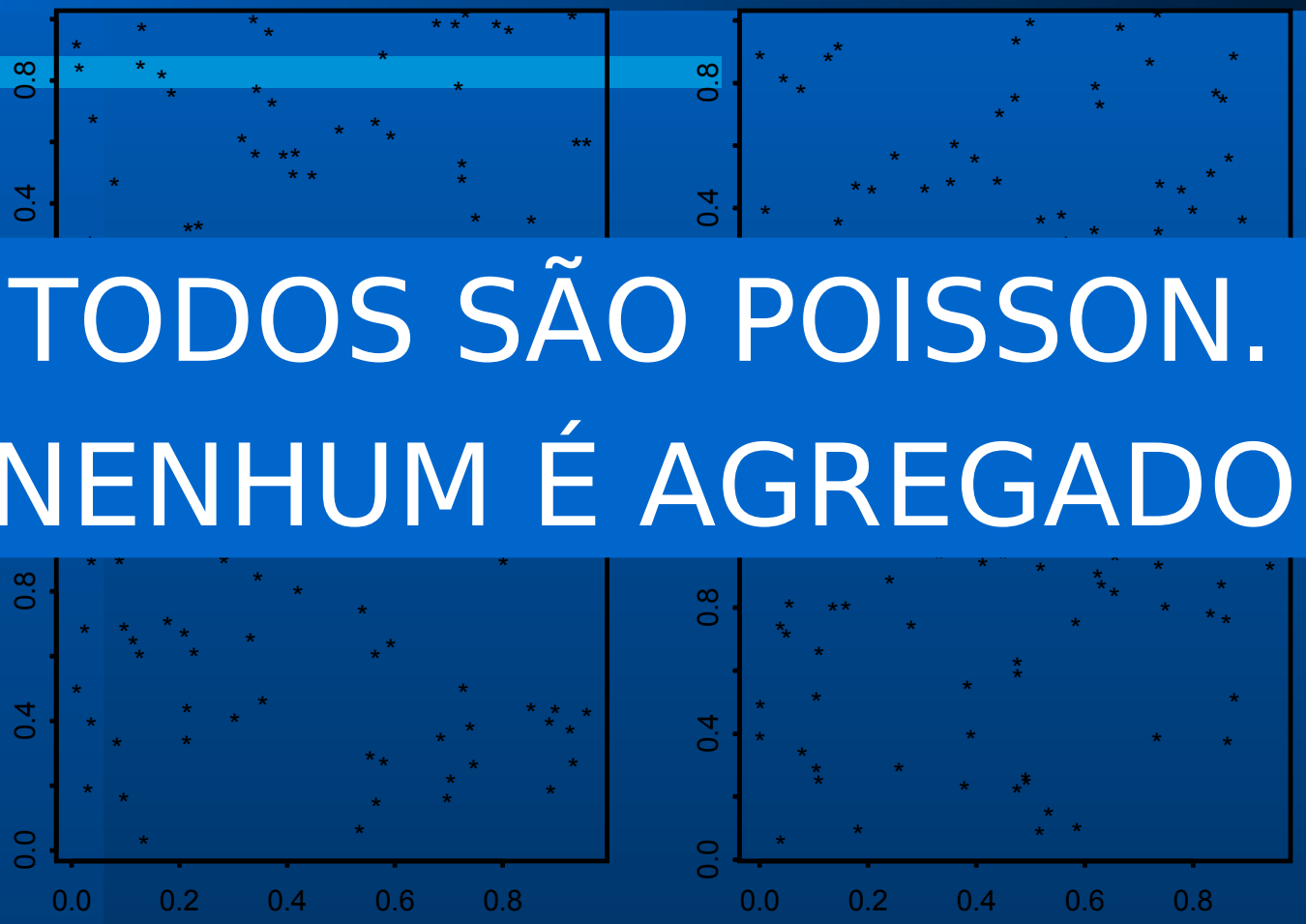
Testes estatísticos de hipóteses

- Considere uma seqüência de variáveis aleatórias $X_1,...,X_n$ (ou um processo estocástico em R^2 ou R^3 ....)
- Deseja-se testar uma hipótese sobre a distribuição conjunta dessas variáveis.
- Por exemplo, deseja-se testar se um processo pontual em R^2, observado apenas num polígono A, é um processo de Poisson homogêneo

Algum (ou alguns) desses padrões foi gerado por um processo de Poisson homogêneo ?



Quais são agregados e quais são gerados por Poisson homogêneo ?



TODOS SÃO POISSON.
NENHUM É AGREGADO.

Função K de Ripley

Função K de Ripley

- **Definição da função K:**

$$K(h) = \frac{1}{\lambda} \; E\left(n^o \; eventos \; à \; distância \leq h\right)$$

- **Como estimar a partir do mapa ?**
- **A partir do número médio de vizinhos à distância h:**

$$\hat{K}(h) = \frac{1}{n/R} \left(n^o \; médio \; de \; vizinhos \; à \; distância \; h\right)$$

Função K de Ripley

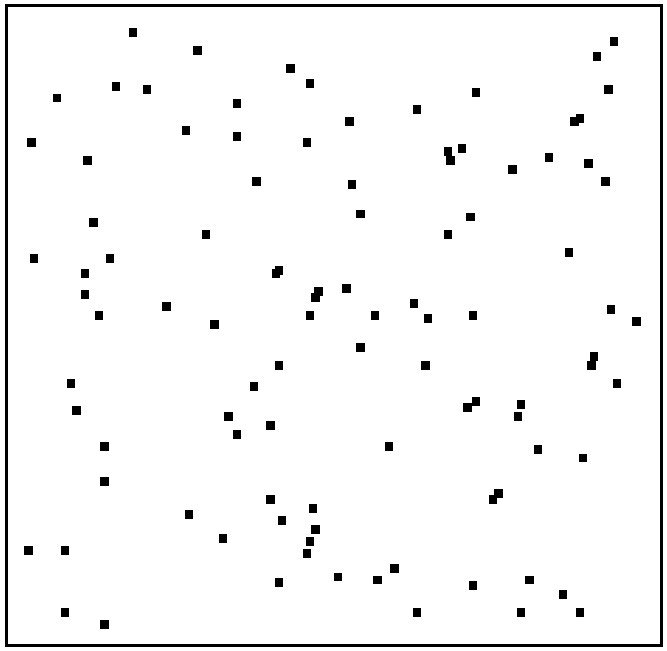
- **Estimativa:**

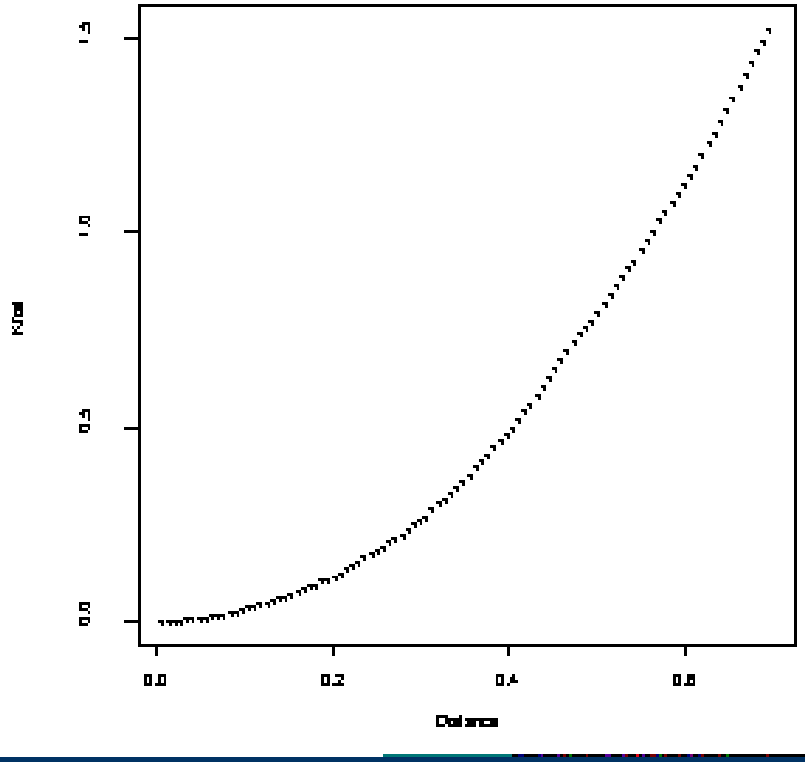
$$\hat{K}(h) = \frac{1}{n/R}\left(\frac{1}{n}\sum_{i \neq j}\frac{I_h(d_{ij})}{w_{ij}}\right)$$

onde d_{ij} é a distância entre os eventos *i* e *j* em R que possui R e $I_h(d_{ij})$ é 1 se $d_{ij} \leq h$ e 0, caso contrário. w_{ij} é correção de fronteira: proporção da circunferência centrada em i e passando em j que está na região de estudo.

Função K de Ripley

- **Gráfico da função K(h) versus h:**





Endireitando a função K

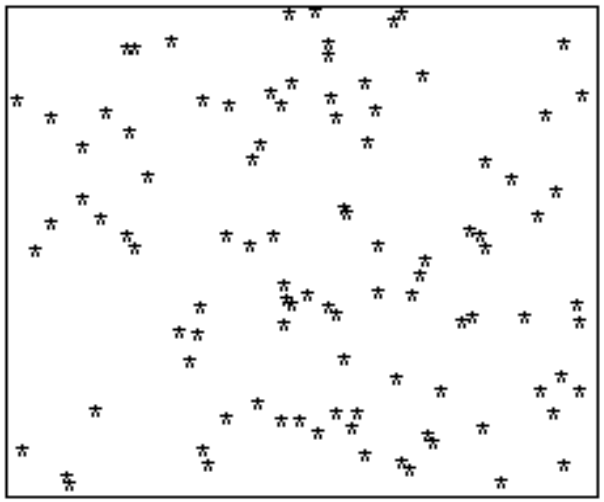
- **Sob regularidade,** $K(h) < \pi h^2$
- **Com aglomerados,** $K(h) > \pi h^2$
- **Assim, plot L(h) versus h onde**

$$L(h) = \sqrt{\frac{K(h)}{\pi}} - h$$

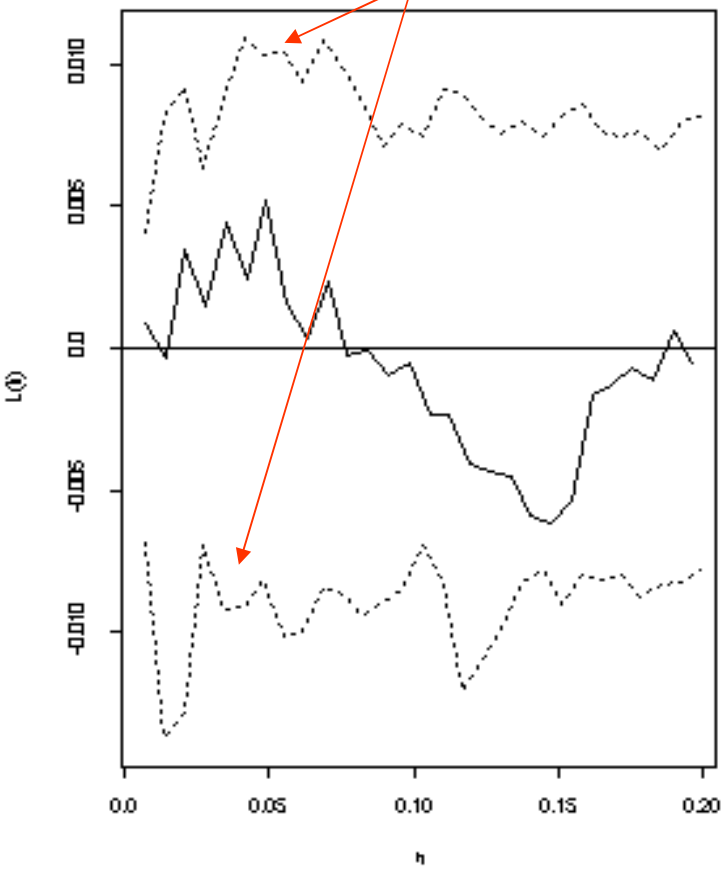
e procure ver se existem picos positivos e vales negativos. Se Poisson Homogêneo, L(h) = 0 para todo h.

Exemplo de função L(h)

Processo de Poisson

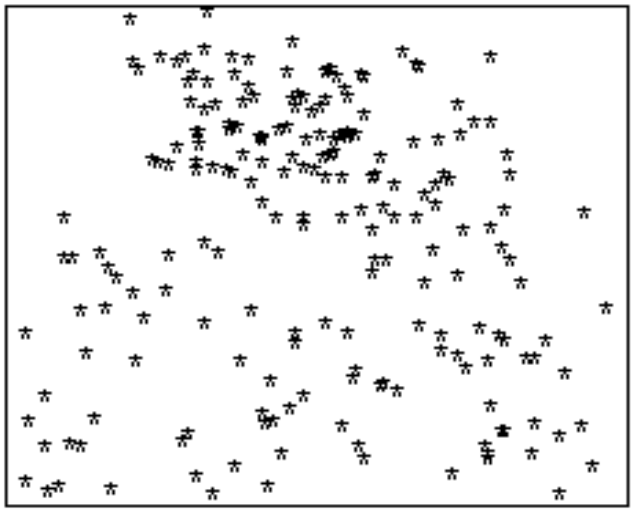


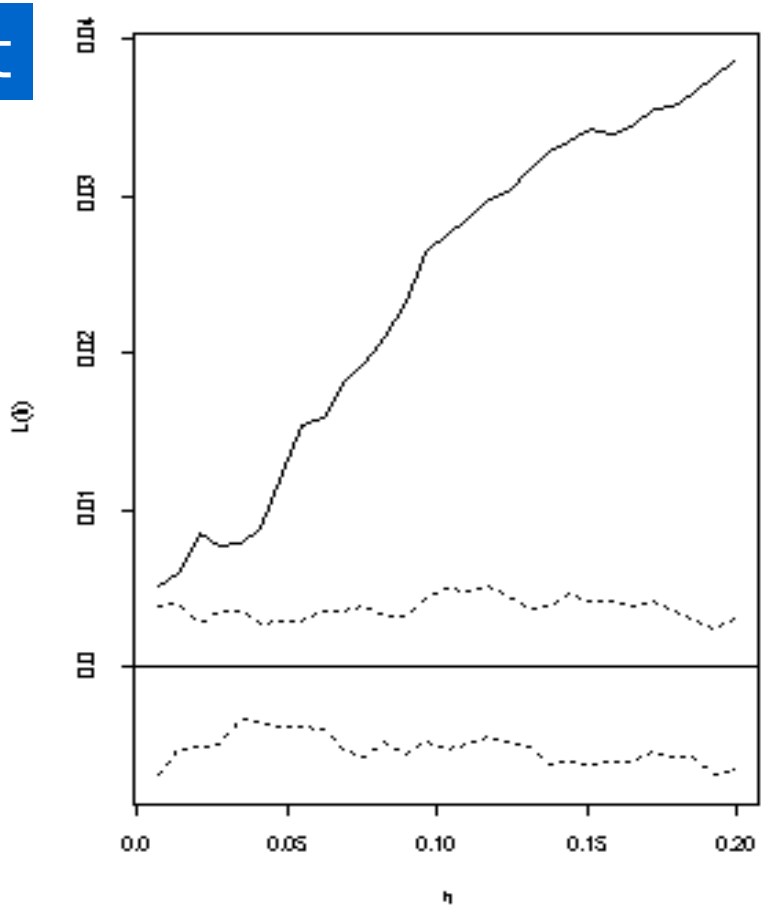
Envelopes para teste



Exemplo de função L(h)

Processo Neyman-Scott



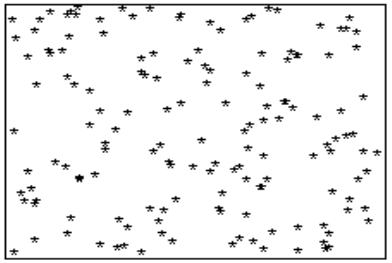


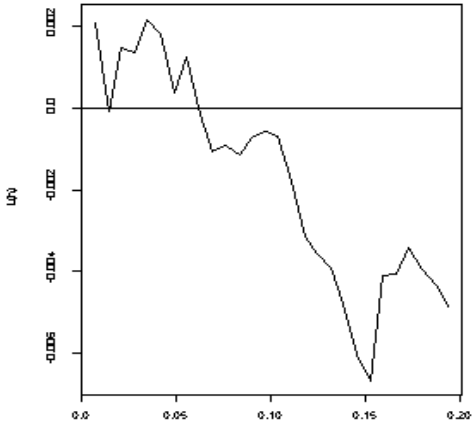
Teste de hipótese usando a função L

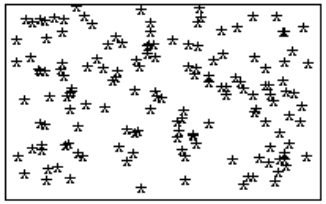
- **Suponha um padrão com 87 pontos**
- **Usar 1000 simulações de 87 eventos aleatórios de acordo com PPH na região**
- **Para cada padrão simulado, calcular a função L(h) associada**
- **Construir envelopes superiores e inferiores com os percentis 2.5% e 97.5% das 1000 L(h)'s em cada h**

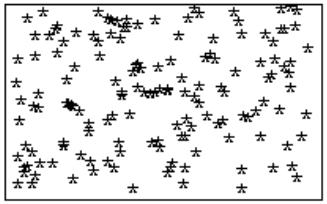
Construindo os envelopes

Primeiro padrão aleatório

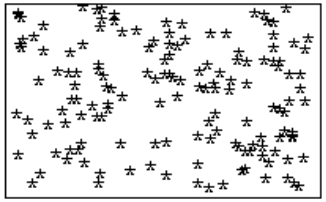


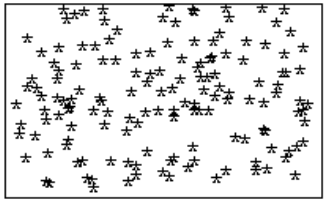


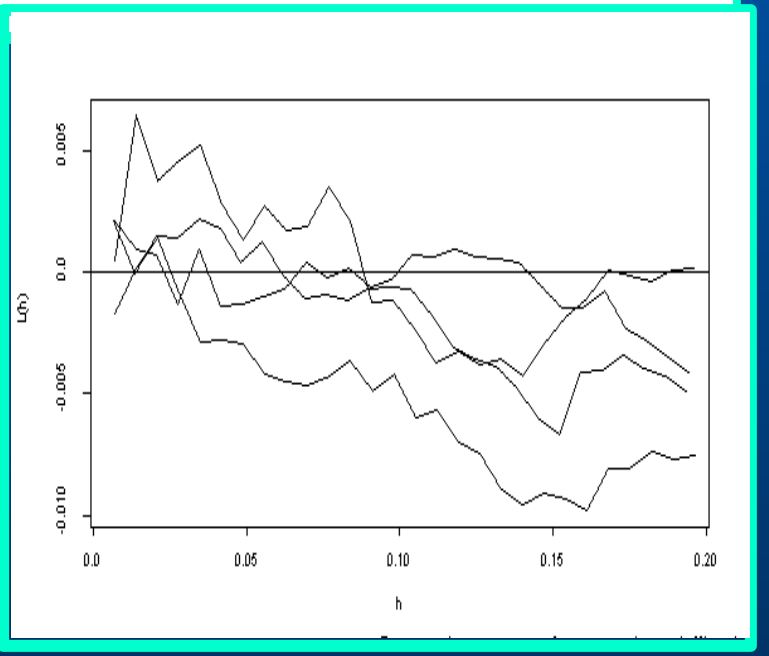




4 padrões sucessivos

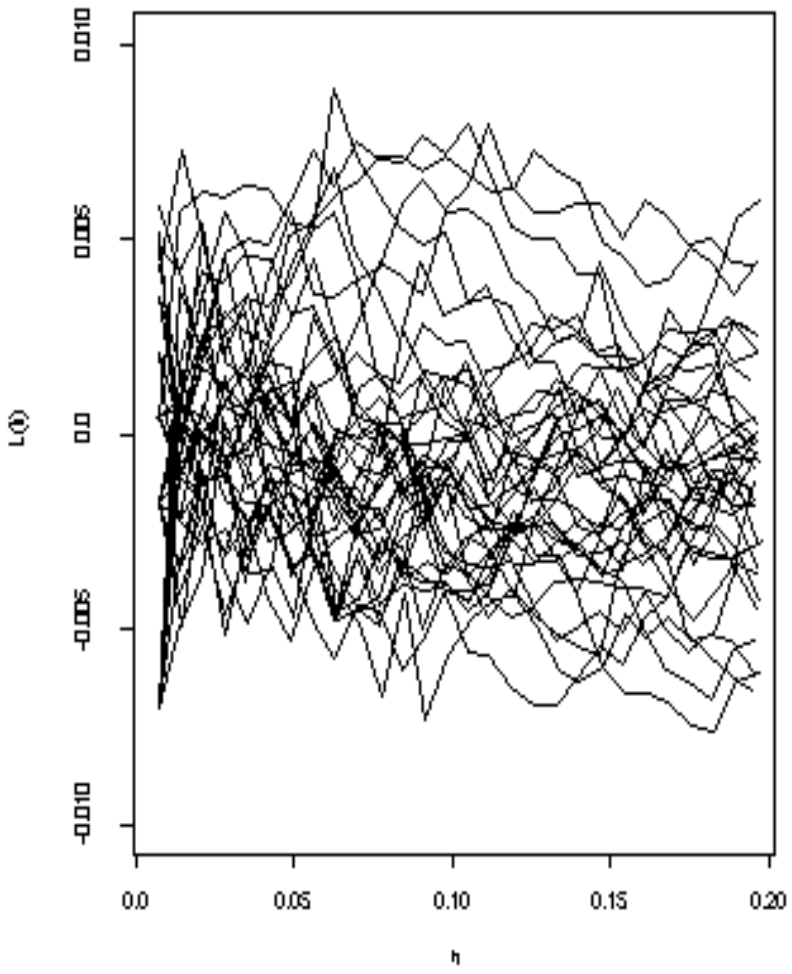




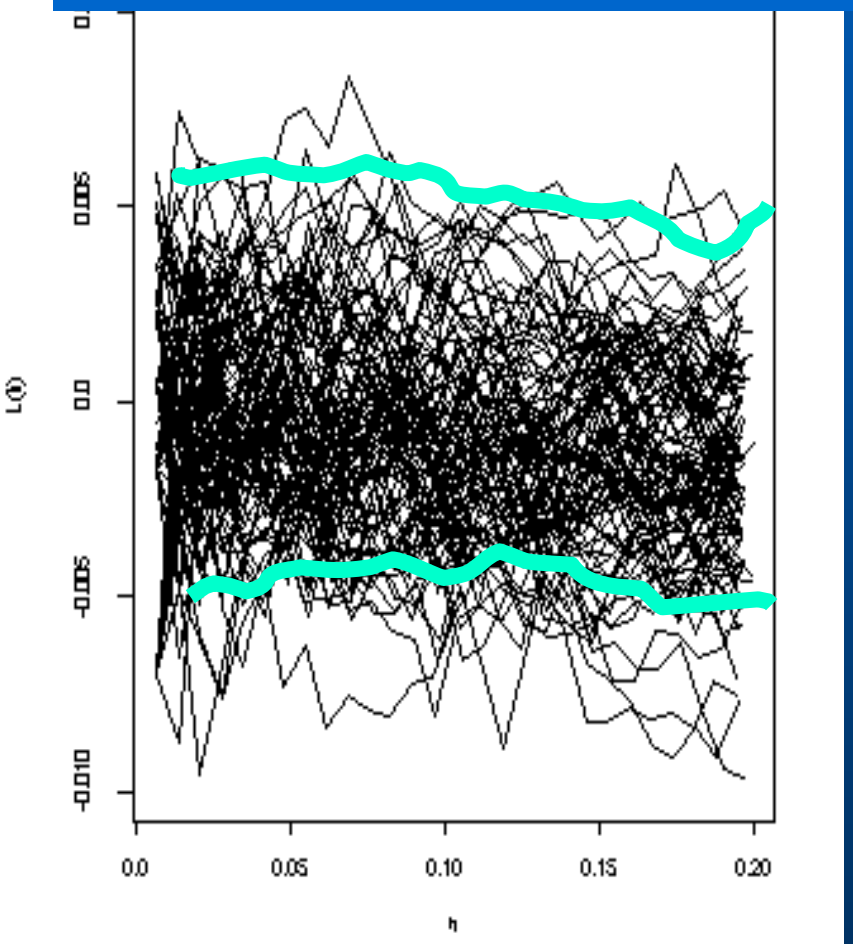


1000 Curvas L(h) simuladas

Primeiras 35 curvas simuladas



Primeiras 100 curvas simuladas



Testes Estatísticos de Hipótese

Observa-se seqüência aleatória $X_1, ..., X_n$

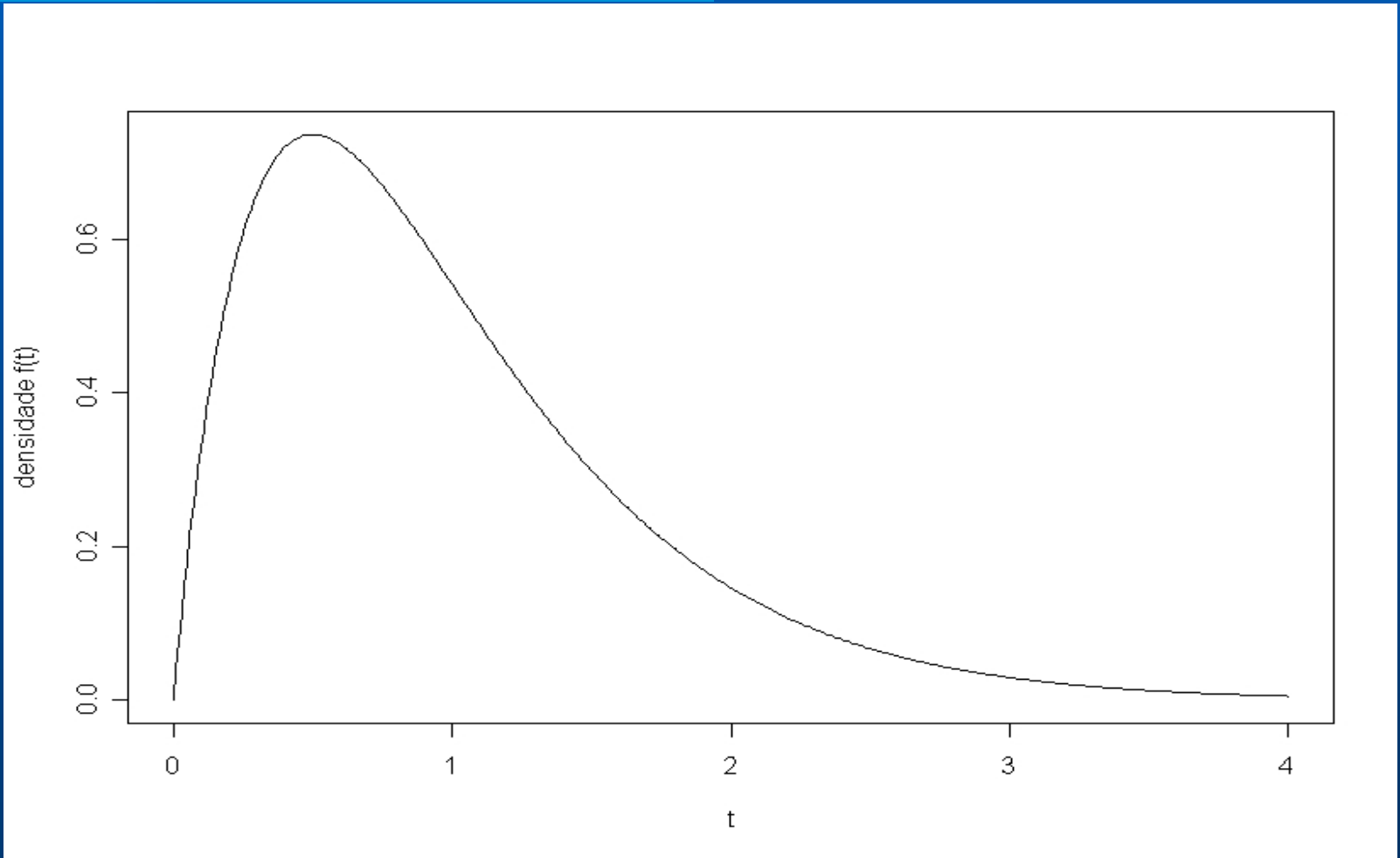
Calcula-se uma estatística $T = T(X_1, ..., X_n)$

Valores grandes de T levam a rejeição de certa hipótese H_0 (hipótese nula)

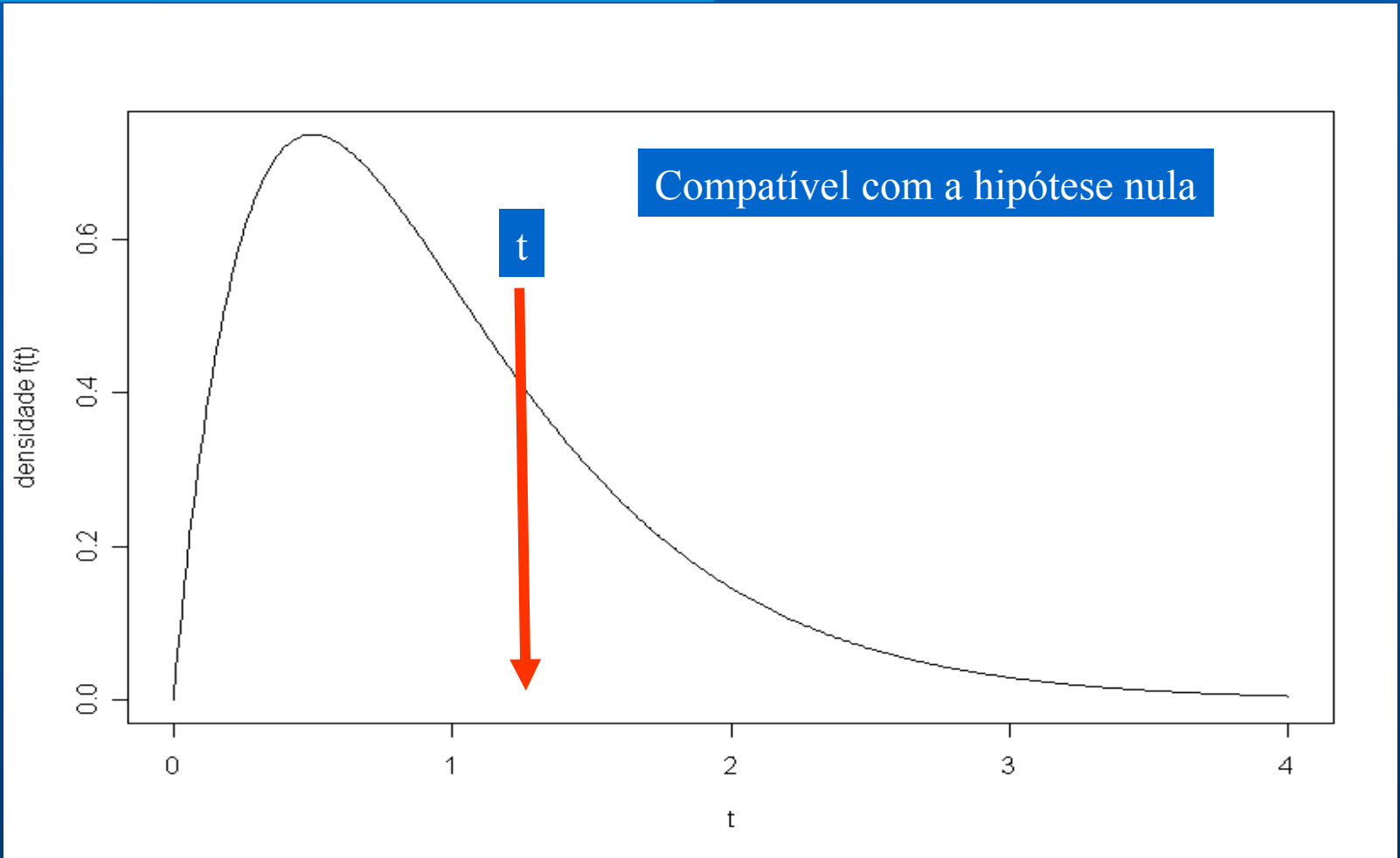
Calcula-se a distribuição de T sob a hipótese nula.

Calcula-se o realmente valor observado t da variável T na seqüência $X_1, ..., X_n$

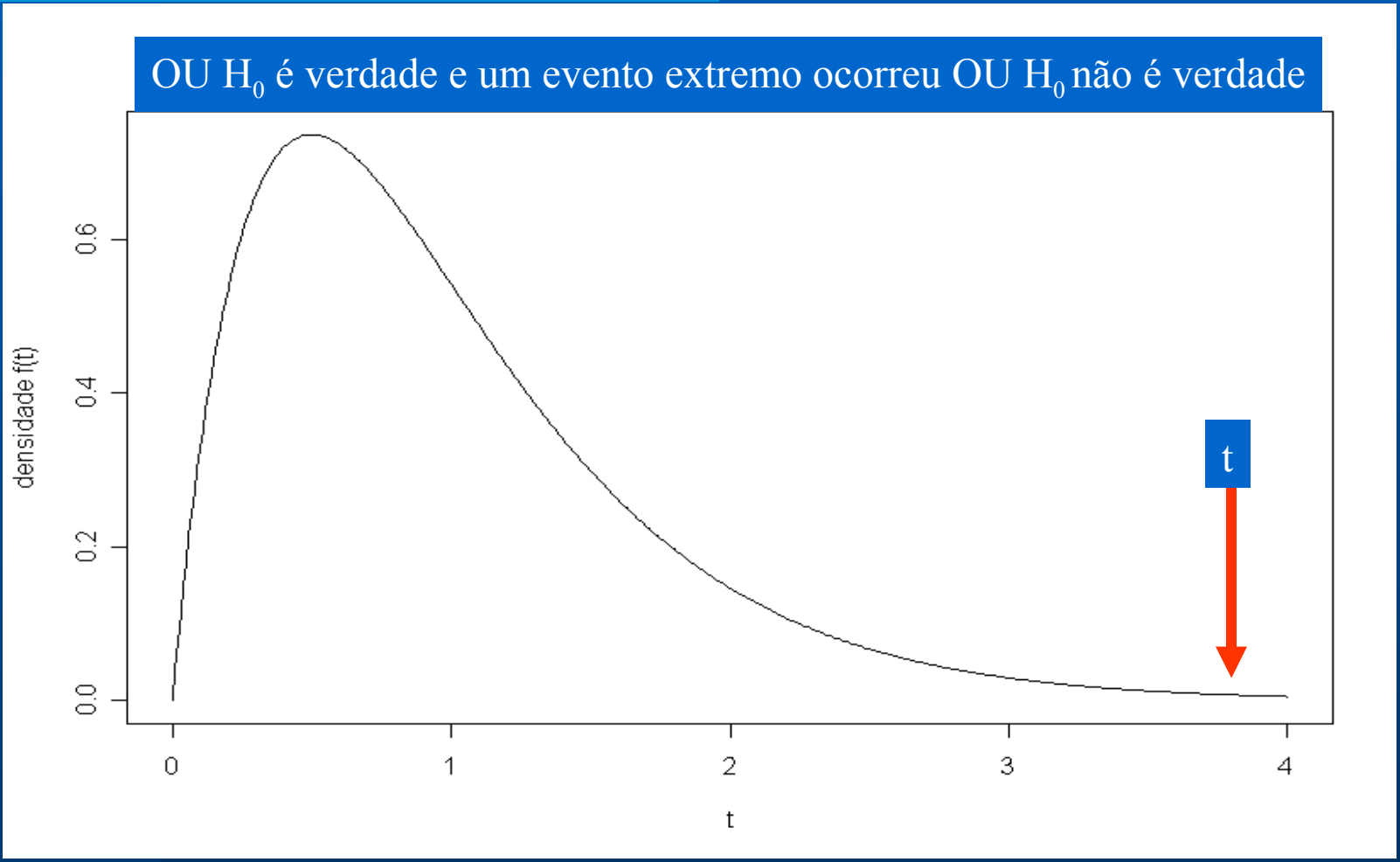
Distribuição de T sob H_0



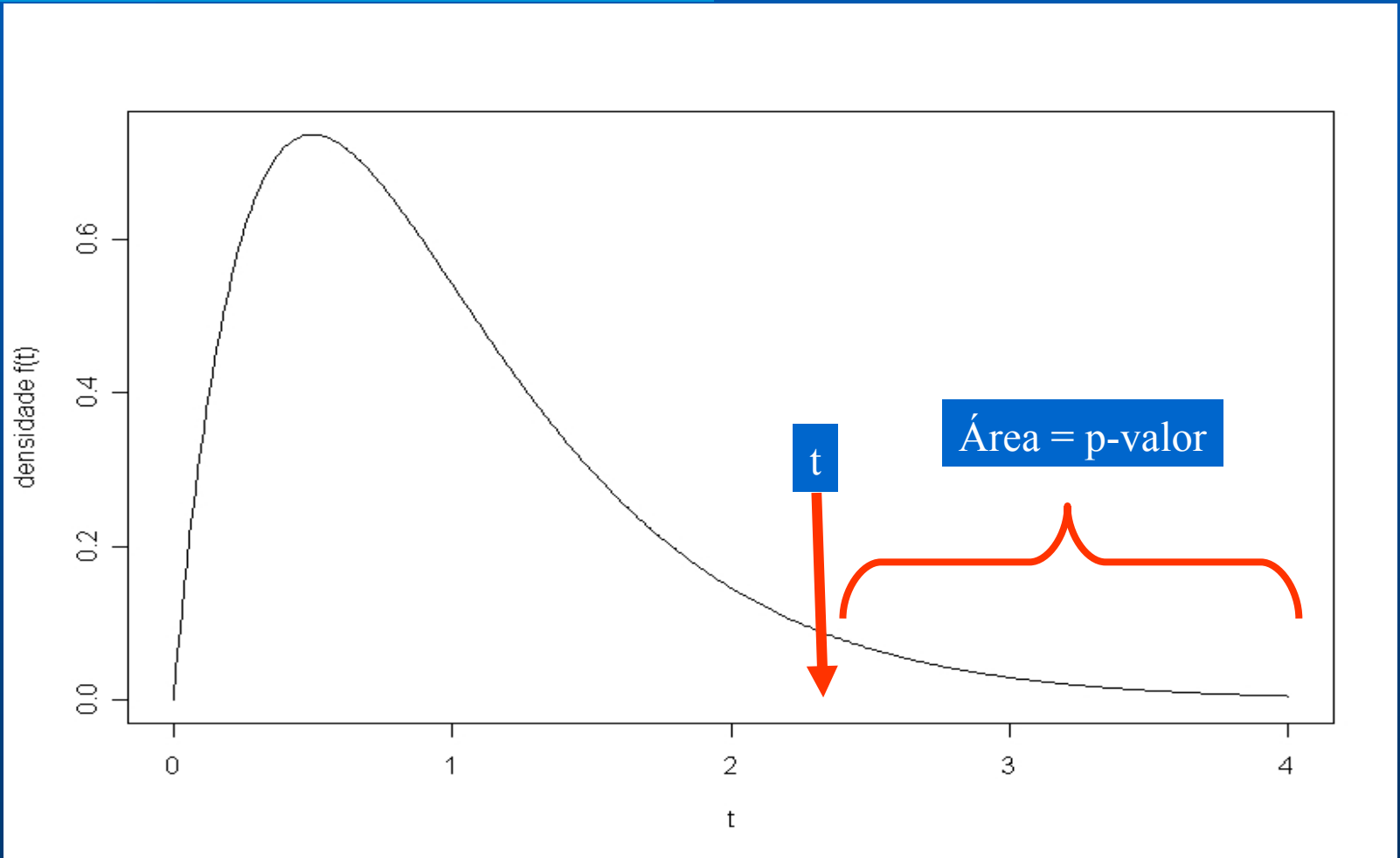
Valor t realmente observado de T



Outro cenário ...



P-valor



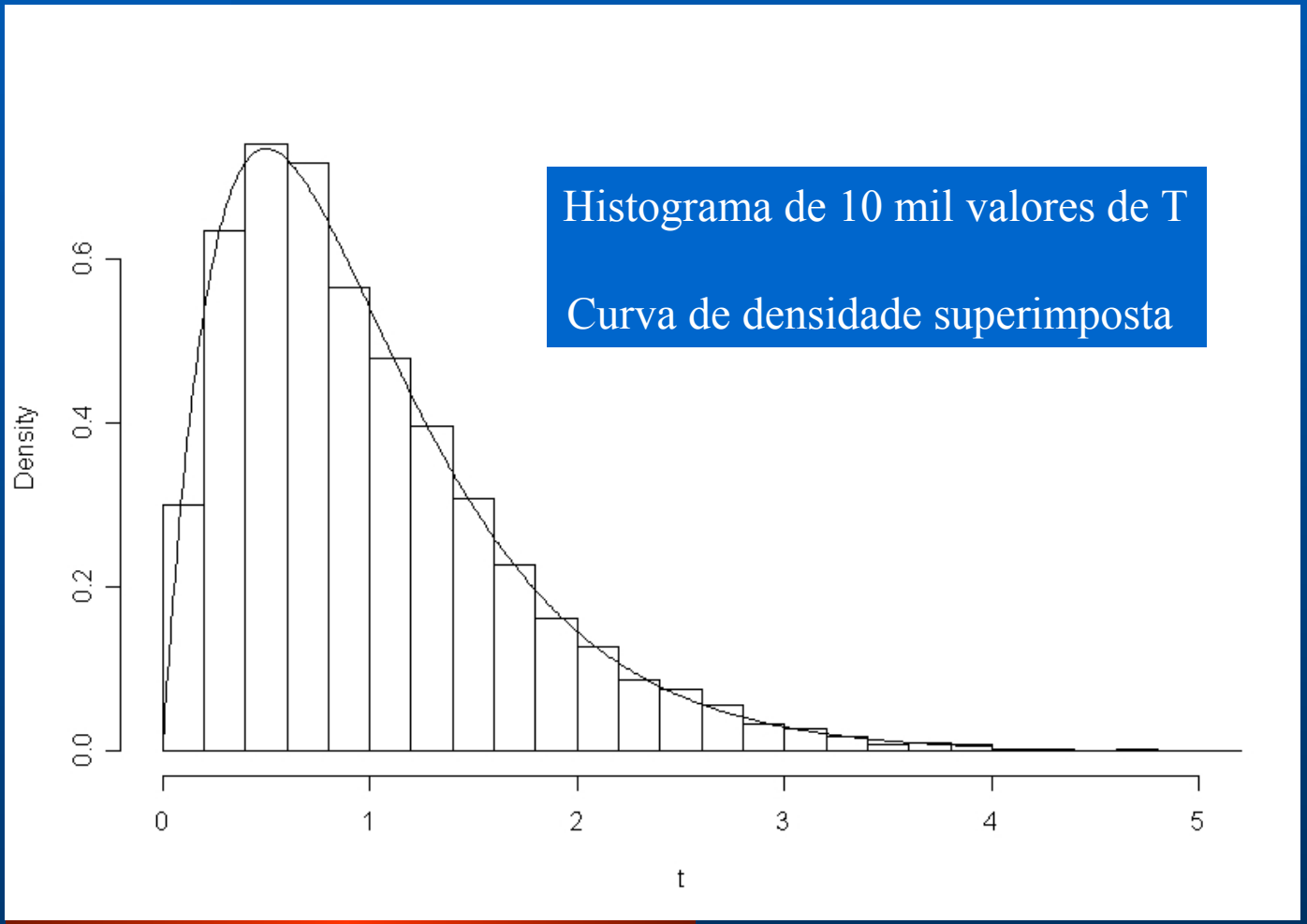
P-valor como resumo do teste

- $T = T(X_1, ..., X_n)$ é a estatística de teste
- Seja $F_T(t \mid H_0) = P(T \leq t \mid H_0)$ a distribuição acumulada de T SOB H_0
- t_{obs} é o valor realmente observado de T
- P-valor é $P(T > t_{obs} \mid H_0) = 1 - F_T(t_{obs} \mid H_0)$
- Se P-valor é pequeno, rejeitamos H_0
- Pequeno significa menor que α , tipicamente menor que 0.05 ou 0.01

Para quê um teste Monte Carlo?

- P-valor é $P(T > t_{obs} \mid H_0) = 1 - F_T(t_{obs} \mid H_0)$
- Mas nem sempre conseguimos obter a distribuição $F_T(t)$ de T analiticamente ou mesmo de forma aproximada.
- Mas às vezes conseguimos SIMULAR por Monte Carlo a distribuição de T sob H_0
- Isto é, geramos tantos valores de T quanto quisermos sob H_0
- Faça um histograma dos valores obtidos e isto dá uma excelente idéia do que é a densidade de T

Simulando T e fazendo histograma



3- Teste Monte Carlo Convencional

- **Genericamente**, atende aos casos em que a distribuição da estatística de teste, sob H_0, não é conhecida. Se baseia em **gerar *m-1* Estatísticas t_i's sob H_0** usando-se a própria amostra. A estimativa do valor-p por este procedimento é:

$$\hat{p} = \hat{P}(U_0 \geq U) = g/m$$

g: n° de estatísticas t_i's excedendo u.

A inconveniência deste procedimento está relacionada à escolha de *m* (geralmente 1000). **Pode tornar-se custoso computacionalmente. Seria necessário gerar *m*-1 t_i's?**

4-Teste Monte Carlo Seqüencial
BESAG AND CLIFFORD (1991)

Interpretação intuitiva da proposta seqüencial

4- Teste Monte Carlo Seqüencial

BESAG AND CLIFFORD (1991)

-Se baseia em **gerar Estatísticas t_i's sob H_0** até que sejam obtidos ***h*** t_i´s maiores ou iguais a u ou até que o n° de simulações atinja *n-1*. A estimativa do valor-p por este procedimento é:

$$p_r = \begin{cases} \frac{g}{l} & (g = h) \\ \frac{g+1}{n} & (g < h) \end{cases}$$

l: N° de simulações no momento em que observou-se *h* t_i*'s*$\geq u$.

g : N° de ti´s excedendo u.

-Sob a hipótese nula, o valor-p obtido desta forma é exato, e tem uma distribuição discreta uniforme em:

$$\{h/h, h/h+1, \ldots, h/m, h-1/m, h-2/m, \ldots, 1/m\} \subset (0,1)$$

5- Escolha de *n* no Procedimento Seqüencial

No procedimento descrito na seção 4 a escolha de *n* é feita de maneira arbitrária, sem que sejam observadas as probabilidades dos erros tipo I e tipo II. É possível mostrar que se fizermos:

$$n \geq (h/\alpha) + 1$$

- **O nível de significância não é afetado;**
- **O poder se mantém constante.**

5- Escolha de *n* no Procedimento Seqüencial MC

Então, seria conveniente usarmos $n=(h/\alpha)+1$ **, pois para os casos em que H_0 é falsa, o tempo de execução do procedimento é reduzido significativamente, sem perda de poder e com o mesmo nível de significância.**

Por exemplo, para *h* = 5:

n	E(*L*) sob H_0
1001	31
101	19

6- Escolhas de *m* no MC convencional

- Para que o procedimento MC convencional tenha poder maior que zero, **é necessário que *n* seja no mínimo 1/α**. Por exemplo, para α=0.01, *n* deve ser no mínimo 100.

- Devemos sempre escolher ***n* como múltiplo de 1/α**. Para α=0.01, *n* deve ser sempre múltiplo de 100, pois do contrário, o procedimento possivelmente terá um poder inferior a um outro que apresenta tempo de execução inferior.

7- Comparação dos poderes dos testes Monte Carlo convencional e seqüencial

7.1- Equivalência em poder

Qual é o esquema seqüencial que corresponde, em termos de poder, ao teste de Monte Carlo convencional?

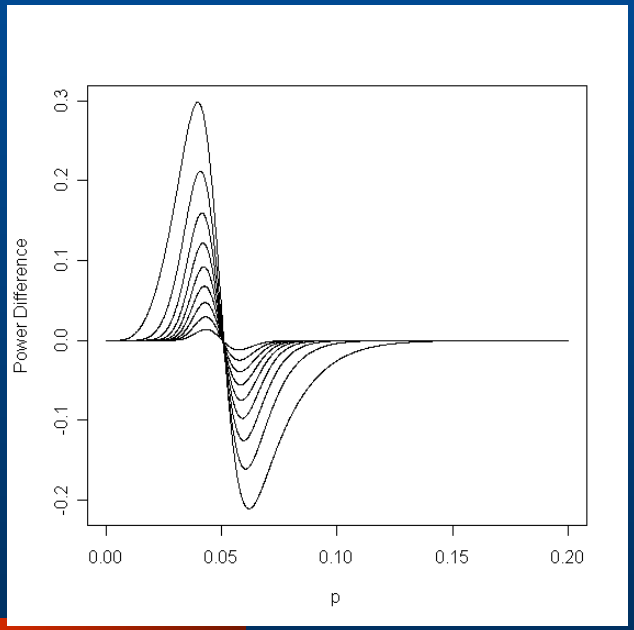
Para α e *m* fixos, os poderes dos testes seqüencial e convencional são exatamente iguais se $h = \alpha m$.

7- Comparação dos poderes dos testes Monte Carlo convencional e seqüencial

7.2- Cotas para a diferença de poder

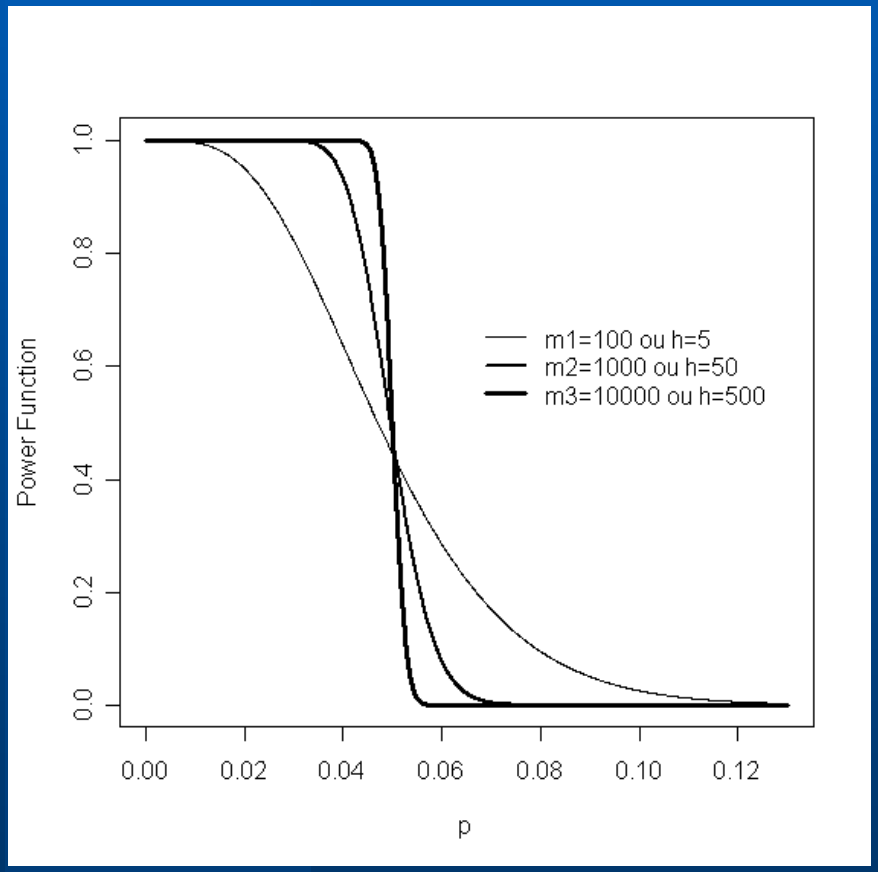
Mostrou-se que, os limites de variação da expressão abaixo, representam cotas para a diferença de poder entre MC convencional e seqüencial.

$$D(P)=\left[\sum_{y=0}^{\alpha m-1}\binom{m-1}{y}P^{y}(1-P)^{m-y-1}-\sum_{x=0}^{h-1}\binom{(h/\alpha)-1}{x}P^{x}(1-P)^{(h/\alpha)-x-1}\right]$$



7- Comparação dos poderes dos testes Monte Carlo convencional e seqüencial

7.2- Cotas para a diferença de poder



h	*m* = 1000	TC/TS máximo	TC/TS médio	TC/TS mínimo
10	0.2117599	100	25	5,52
20	0.1217025	50	13	2,49
30	0.0678584	33,3	8	1,66
40	0.0295589	25	6	1,25

8- Análise Empírica do Poder

Abordagem 1: Gerar dados sob condições controladas e aplicar a estatística Varredura espacial com MC convencional e sequencial.

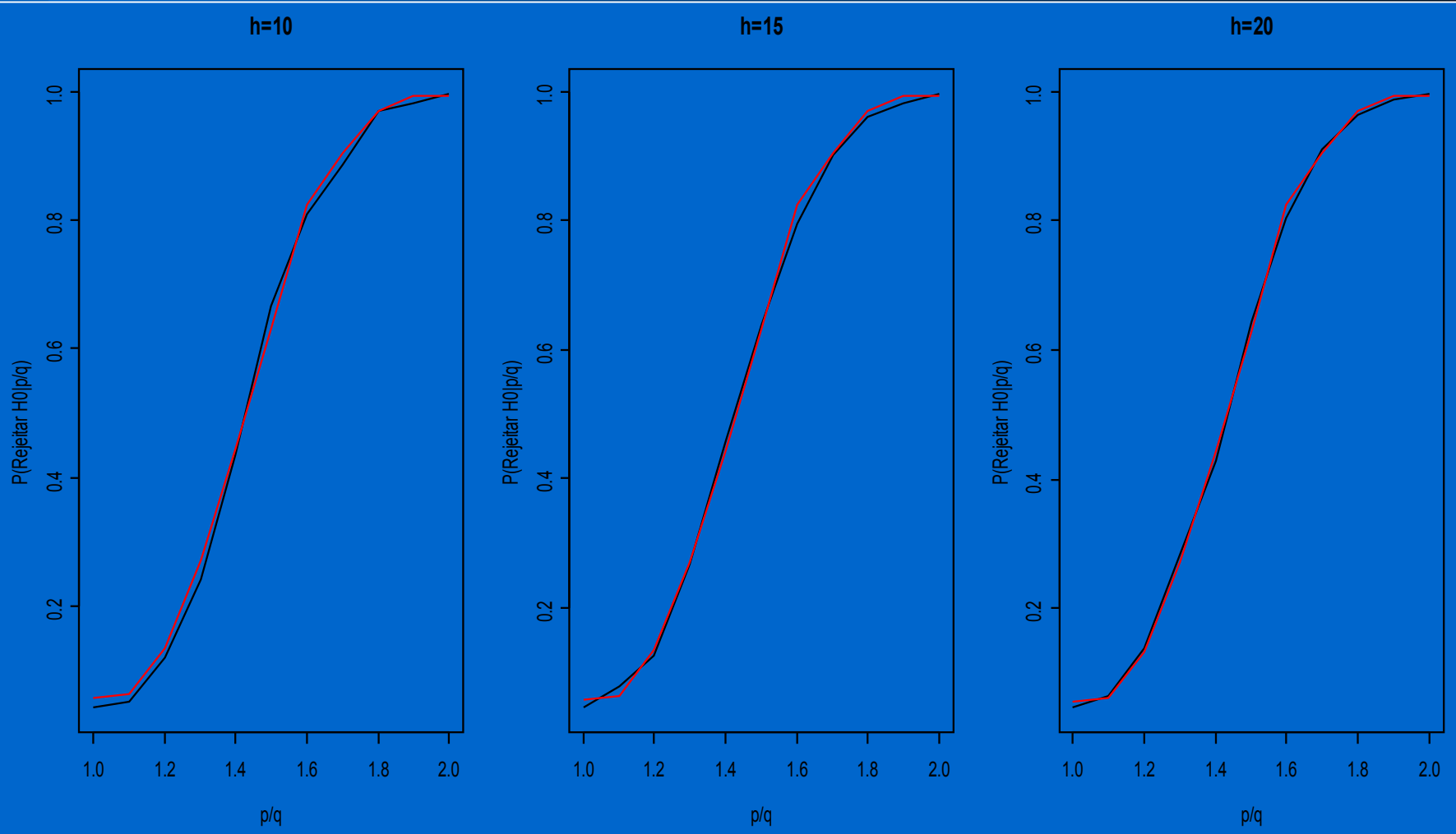
- Mapa do acervo Satscan com 32 áreas;
- Estratégia de simulação

  - Fixar cenário a partir da razão p/r ;

  - Simular 1000 amostras de casos para cada cenário, induzindo um conglomerado por amostra;

  - Verificar o percentual de vezes em que se rejeitava a hipótese nula ao se utilizar o teste seqüencial, com um nível de significância de 5%.

8- Análise Empírica do Poder – Abordagem 1



8- Análise Empírica do Poder

Abordagem 2: Utilização dos dados do acervo Benchmark, que pode ser obtido no endereço *www.satscan.com.*

Aplicando-se o procedimento seqüencial, para as mil primeiras configurações, a hipótese nula foi rejeitada **87.4%** das vezes, e aplicando-se o teste de varredura espacial convencional, rejeitou-se a hipótese nula para **88.4%** das configurações. Uma diferença de **0,01** entre os poderes.

9- Análise Empírica do Tempo de execução

O interesse é verificar a vantagem que constitui, em termos do tempo de execução do teste Monte Carlo, utilizar o teste seqüencial em substituição ao convencional.

Aplicou-se o teste Varredura espacial no software Satscan utilizando-se dados gerados aleatoriamente.

Para $h=5$, $\alpha = 0.05$, $n = 101$ e $m = 1000$, o procedimento convencional foi executado em **82 minutos,** enquanto que para o mesmo banco de dados, o seqüencial seria operacionalizado em no **máximo 12 minutos**.

10- Discussões

O uso do teste Monte Carlo seqüencial, em substituição ao convencional, é viável e conveniente pois é sempre possível realizá-los com o **mesmo poder**, mas com o seqüencial apresentando tempos de execução **esporadicamente inferiores**.

Também é possível formular testes seqüenciais com tempos **sempre inferiores** e que acarretam em **pouca diminuição de poder**, apresentando tempos de execução significativamente inferiores.

Poderíamos citar como exemplo de substituição o teste **Varredura espacial**, que poderia ser facilmente adaptado no software Satscan para obter o valor-p via metodologia seqüencial de MC.

Referências

BAHLO, M; STANKOVICH, J; SPEED, TP; RUBIO, JP; BURFOOT, RK; FOOOTE, SJ. Detecting genome wide haplotype sharing using SNP or microsatellite haplotype data. HUMAN GENETICS 119 (1-2): 38-50 MAR 2006.

BESAG, Julian; CLIFFORD, Peter. Sequential Monte Carlo p-value. **Department of Statistical, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K.** 1991.

BESAG, J; GREEN, P; HIGDON, D; MENGERSEN, K. BAYESIAN COMPUTATION AND STOCHASTIC-SYSTEMS. STATISTICAL SCIENCE 10 (1): 3-41 FEB 1995.

BROWNING, BL. FLOSS: flexible ordered subset analysis for linkage mapping of complex traits. BIOINFORMATICS 22 (4): 512-513 FEB 15 2006.

DIGLLE, Peter; GRATTON, Richard. Monte Carlo Methods of Inference for Implicit Statistical Models. Royal Statistical Society. (46), 2, pp: 193-227 Jan 1984.

Fay, MP; Follmann, DA. Designing Monte Carlo Implementations of Permutation or Bootstrap Hypothesis Tests. AMERICAN STATISTICIAN 56 (1): 63-70 FEB 2002.

FILL, JA. An interruptible algorithm for perfect sampling via Markov chains. ANNALS OF APPLIED PROBABILITY 8 (1): 131-162 FEB 1998.

JONHNSON, Norman L.; KOTZ, Samuel; KEMP, Adrienne W. **Univariate Discrete Distributions.** Second Edition, 1997.

KULLDORFF, Martin. A SPATIAL SCAN STATISTIC. **Biometry Branch, DCPC, National Cancer Institute.** EPN 344, 6130 Executive Blvd, Bethesda MD 20892, USA and Department of Statistics, Uppsala University, 751 20 Uppsala, Sweden. 1997.

Referências

KULLDORFF, Martin; HEFFERNAN, Richard; HARTMAN, Jessica; ASSUNÇÃO, Renato; MOSTASHARI, Farzad. A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. **PLOS Medicine, Vol. 2.** March 2005.

LIMA, Max Souza. **"Avaliação do Poder do Teste da Estatística Scan para Múltiplos Clusters".** 2004. L732 a. Dissertação (Mestrado em Estatística) – Universidade Federal de Minas Gerais.

LINDSTROM, S; WIKLUND, F; JONSSON, BA; ADAMI, HO; BALTER, K; BROOKES, AJ; XU, JF; ZHENG, SL; ISAACS, WB; ADOLFSSON, J; GRONBERG, H. Comprehensive genetic evaluation of common E-cadherin sequence variants and prostate cancer risk: strong confirmation of functional promoter SNP. HUMAN GENETICS 118 (3-4): 339-347 DEC 2005.

SATSCAN: Software for the Spatial, temporal, and space-time scan statistics. Disponível em: <http://www.satscan.org/references.html> Acesso em 12 nov. 2005.

SACKROWITZ, Harold; CAHN, Ester Samuel. P Values as Random Variables-Expected P Values. The American Statistician 53,4: 326-331 Nov 1999.

SONG, KK; WEEKS, DE; SOBEL, E; FEINGOLD, E. Efficient simulation of P values for linkage analysis. GENETIC EPIDEMIOLOGY 26 (2): 88-96 FEB 2004.

WIGGINTON, JE; ABECASIS, GR. An evaluation of the replicate pool method: Quick estimation of genome-wide linkage peak p-values. GENETIC EPIDEMIOLOGY 30 (4): 320-332 MAY 2006.