Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Noções de Estatística Computacional

M. A. C. Santos

Relatório Técnico RTE-03/98

Relatório Técnico
Série Ensino

Meus agradecimentos aos Professores
Frederico Rodrigues Borges da Cruz e
Renato Martins Assunção,
ambos do Departamento de Estatística da UFMG, pelas sugestões.

Noções de Estatística Computacional

Marcos Antônio da Cunha Santos
msantos@est.ufmg.br
Departamento de Estatística
Instituto de Ciências Exatas da UFMG
1998

Objetivos

Esta é a primeira versão de ume texto que tem por finalidade introduzir conceitos básicos e noções do uso de métodos computacionais em testes estatísticos e probabilidade, em uma rápida introdução ao tema.

O texto foi concebido como material adicional para um primeiro curso de probabilidade a nível de graduação. Deve ser utilizado após a introdução dos conceitos básicos de variável aleatória, funções distribuição e densidade de probabilidade.

Os exercícios podem ser utilizados como sugestões para temas de trabalhos práticos.

O texto pressupõe ainda o domínio, em nível básico, de uma linguagem de programação.

Conteúdo

notação

A notação adotada neste texto, com poucas exceções, segue a da maioria dos textos de estatística e probabilidade:

P(..)	é a probabilidade da ocorrência do evento entre parênteses;
X, Y...	maiúsculas simbolizam variável aleatória;
x,y.....	minúsculas são valores *pré-fixados* da variável aleatória;
F(x)	função distribuição de x, definida como $P(X \leq x)$;
f(x)	função densidade de probabilidade (caso contínuo);
p(x)	função probabilidade (variável discreta).
x*, y*..	(letra com asterisco) são valores aleatórios gerados em computador
U	variável aleatória com função densidade uniforme no intervalo [0,1].

1 - Geração de variáveis aleatórias

1.1 Introdução

Atualmente é possível gerar de maneira imediata números aleatórios (rigorosamente, pseudoaleatórios), com distribuição uniforme no intervalo [0,1], em qualquer computador. Em geral as linguagens de programação incorporam esta função para a geração de números aleatórios.

Normalmente o que se deseja é simular o comportamento de variáveis aleatórias com funções densidade ou probabilidade mais complexas do que simplesmente uma distribuição uniforme [0,1]. O problema básico é: gerar uma sequência de valores x_1, x_2, x_3...x_n de uma variável aleatória X, dado a função probabilidade p(x) (caso discreto) ou função densidade f(x) (caso contínuo). Evidentemente pode-se estender a idéia para o caso de variáveis aleatórias multidimensionais. Neste caso, o objetivo é gerar vetores aleatórios (X_1, X_2...X_n) dada a função densidade conjunta f(x_1, x_2...x_n). Usualmente estes valores devem ser gerados a partir da variável aleatória U, cuja função densidade é uniforme no intervalo [0,1].

Entre as diversas técnicas existentes para a geração de vetores aleatórios, algumas são específicas para certos tipos de variáveis; outras de aplicabilidade bastante geral. Os métodos específicos geralmente são bem mais eficientes. Nesta texto, no entanto, abordaremos métodos suficientemente gerais para uma rápida aplicabilidade em simulações simples.

1.2 Variáveis Discretas

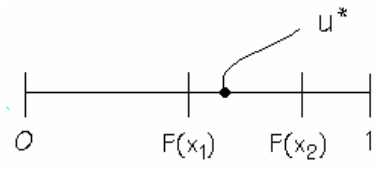
Seja X variável aleatória discreta com função probabilidade p(x). Para gerar uma sequência aleatória de inteiros x^*_1, x^*_2, x^*_3,...x^*_n com função probabilidade p(x):

- Calcule os valores da função distribuição acumulada $F(x_i)= \Sigma\ p(x_i)$;
- Divida o intervalo [0,1] em segmentos proporcionais a p(x)

(figura 1);

- Gere u* (variável aleatória com função densidade uniforme [0,1]).
- Para cada valor de u* gerado associe o valor inteiro que correspondente ao Segmento.

Figura 1



Algoritmo:

Repita

Gerar u* de uma uniforme [0,1]

Determinar x_n tal que $u^* \in (F(x_{n-1}), F(x_n)]$;

Até fim

Exemplo 1

Gerar valores de X, sendo X variável aleatória com valores possíveis 0, 1, 2, 3 ou 4, com função probabilidade

p(0) = 0.2 p(1) = 0.1 p(2) = 0.1 p(3) = 0.4 p(4) = 0.2

Solução: A função distribuição acumulada é

F(0) = 0.2 F(1) = 0.3 F(2) = 0.4 F(3) = 0.8 F(4) = 1.0

Abaixo os 12 primeiros valores gerados u* e os correspondentes valores para X:

u*	**x***	**u***	**x***
0.384610	2	0.653466	3
0.547862	3	0.022103	0
0.648772	3	0.151686	0
0.924035	4	0.725421	3
0.794113	3	0.645511	3
0.312040	2	0.740516	3

Sequência gerada*: 2 3 3 4 3 2 3 0 0 3 3 3 ...*

Exercícios

1 - Dois dados são lançados n vezes. Para cada lançamento, uma variável aleatória X é definida da seguinte maneira:

X = 0, se os dois números são pares;
= 1, se os dois são ímpares
= 2, se um dos dados é par e o outro ímpar.

Escreva um programa para gerar uma sequência de 100 valores para a variável X.

2 - (Ross, 1990) – Escreva um programa que, dado uma função probabilidade $p(x_j)$ $(j = 1,2...n)$ retorna como saída *k* valores de uma variável aleatória com esta função de probabilidade.

1.3 Variável Contínua: método da inversão

Este método é um método direto e de fácil aplicação; o inconveniente é que é necessário a expressão analítica da inversa da função distribuição acumulada F(x). Ou seja, dado y = F(x) é necessário explicitar $x = F^{-1}(y)$.

O problema pode ser abordado da seguinte maneira: Seja y = F(x) a função distribuição acumulada de X. Qual a *função densidade* de y? Para determinar a função distribuição F(y), temos

$$F(y) = P(Y \leq y) = P[F(x) \leq y] = P[x \leq F^{-1}(y)] = F[F^{-1}(y)] = y.$$

Derivando F(y), vemos que a função densidade de y será f(y) = 1, no intervalo $0 < y < 1$. Portanto,

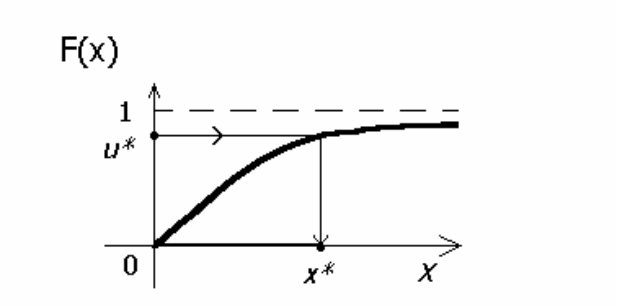
Se y = F(x) é a função distribuição de probabilidade acumulada da variável aleatória X, a função densidade de Y é uniforme no intervalo [0,1].

Isto pode ser visualizado da seguinte maneira (figura 2): em um gráfico de F(x), se pontos no eixo x são gerados aleatoriamente de acordo com a função densidade f(x*), os valores correspondentes no*

eixo y terão distribuição uniforme.

O método da inversão faz exatamente uma inversão no gráfico de F(x): simplesmente geramos pontos no eixo y com distribuição uniforme e verificamos os valores correspondentes no eixo x (figura 2)

Figura 2



Para a gerar valores de X, com função probabilidade acumulada F(x):

- ***gerar um valor de u* (distribuição uniforme [0,1]);***
- ***determinar $x^* = F^{-1}(u^*)$;***
- ***repetir n vezes para gerar $x^*_1, x^*_2, x^*_3 ... x^*_n$.***

1.4 Variável contínua: método do envelopamento

O método do envelopamento é suficientemente geral para ser usado em quase todas as situações práticas, quando não é possível obter a inversa da função distribuição de x. O inconveniente é que pode ser pouco eficiente, dependendo do caso.

Seja a função densidade f(x), da qual se deseja gerar valores $x^*_1, x^*_2, x^*_3 ... x^*_n$. A idéia é escolher uma função densidade auxiliar g(x), da qual é possível gerar valores x* pelo método da inversão. Esta função deve ser tal que Ag(x) "envelopa" f(x), onde A = constante arbitrária.

A expressão "envelopar" f(x) significa escolher uma função auxiliar g(x) e uma constante A tal que **Ag(x) ≥ f(x),** para todo x. A constante deve ser escolhida de tal forma a permitir o envelopamento completo. Uma vez escolhido A e g(x), proceda da seguinte forma:

enquanto não é fim faça:
    gerar x*
    gerar u*
    se $Ag(x^*)u^* \leq f(x^*)$ então

$x^*_i = x^*$

$i = i + 1$

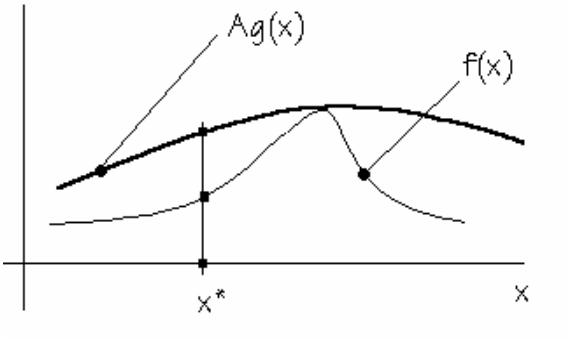
fim se

fim enquanto

Pode ser demonstrado que os valores *x* aceitos* são distribuídos de acordo com a função densidade f(x) (Ross, 1990).

Para compreender melhor este processo, seja x* o valor gerado a partir de g(x) (figura 3). O produto Au*g(x) terá valor mínimo igual a zero e máximo igual a Ag(x); se somente são aceitos valores menores do que f(x), a probabilidade de um valor *x* ser aceito* é portanto f(x)/Ag(x).

Figura 3 método do envelopamento



Observe que, nas regiões aonde Ag(x) é muito maior do que f(x), o processo pode ficar excessivamente lento, já que grande parte dos valores gerados de g(x) serão rejeitados. O ideal é que a função que envelopa tenha valores sempre próximos f(x).

1.5 O Orbital Atômico do Hidrogênio

A concepção na física moderna de orbitais atômicos é probabilística: o que se pode determinar é a *probabilidade* de encontrar um elétron a uma certa distância do núcleo. Podemos falar, portanto, de funções densidade para as coordenadas espaciais do elétron. Estas funções são obtidas, na física quântica, a partir das funções de onda complexa ψ que são, por sua vez, soluções da equação de Schrodinger. Esta equação é fundamental na física quântica e relaciona os estados quânticos possíveis com a energia do sistema. Para o átomo de hidrogênio, como o número atômico igual a 1, as soluções analíticas desta equação são simples e tratáveis

analiticamente.

Os orbitais atômicos são caracterizados pelos números quânticos *n, l, m*. Para o caso do hidrogênio, as funções de onda complexa ψ são conhecidas e podem ser facilmente encontradas em textos básicos de física moderna (ver, por exemplo, Tipler, 1981).. A função densidade de probabilidade para a posição de um elétron, em coordenadas esféricas, é obtida multiplicando-se $\psi(r, \theta, \varphi)$ pela função conjugada complexa $\psi^*(r, \theta, \varphi)$. Para visualizar o orbital no plano devemos ainda multiplicar por $4\pi r^2$, para obtermos a densidade eletrônica na casca esférica $r < R < r + dr$.

O orbital mais simples possui números quânticos n = 1, l = 0 e m = 0 (abreviadamente, orbital "*um-zero-zero"* ou "100"). A função densidade é

$$f(r) = C_{100}(4\pi r^2)\exp(-2r) \qquad (1)$$

onde C_{100} é a constante de normalização e r é a distância ao núcleo, em números de raios de Bohr (1 raio de Bohr = raio do primeiro orbital do átomo de hidrogênio previsto pela teoria clássica = 5,29 x 10^{-11} m). Para o orbital 200 (n = 2, l = 0, m = 0),

$$f(r) = C_{200}(4\pi r^2)(2 - r)^2 \exp(-r) \qquad (2)$$

Para os orbitais 210 e 211, há uma dependência angular:

$$f(r) = C_{210}(4\pi r^2)\, r^2 \exp(-r)\cos^2\theta \qquad (3)$$

$$f(r) = C_{211}(4\pi r^2)\, r^2 \exp(-r)\mathrm{sen}^2\theta \qquad (4)$$

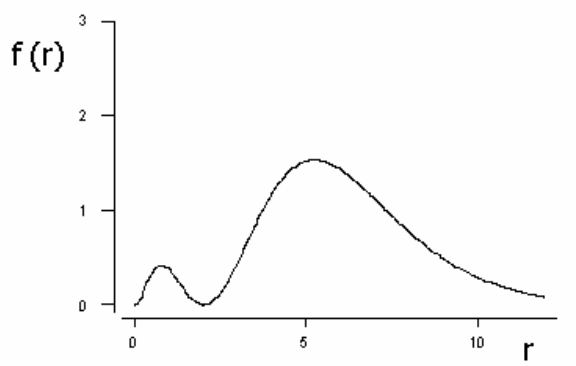


Figura 4
Gráfico para
$f(r)=r^2(2 - r)^2 \exp(r)$

Exemplo 2

Gerar pontos aleatórios, no plano, para a visualização do orbital 200

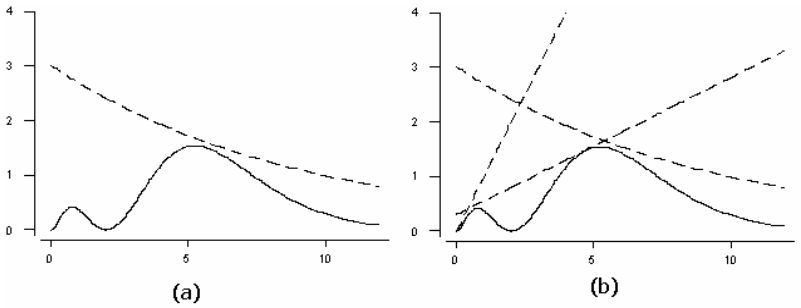
do átomo de hidrogênio (figura 4).

Solução: Podemos gerar pontos no plano em coordenadas polares (r, θ) com coordenada θ uniformemente distribuída no intervalo $[0,2\pi]$ e, de modo independente, gerar valores para a coordenada r seguindo a função densidade da figura 4. A dificuldade aqui é que não podemos usar o método da inversão; usaremos, portanto, o envelopamento.

Como o objetivo é apenas visualizar o orbital, não importando a escala, não é necessário o conhecimento da constante de normalização. Assim, podemos trabalhar com qualquer função proporcional a f(r) para a geração dos pontos, por exemplo $h(r) = r^2(2 - r)^2 \exp(-r)$ e escolher uma função g(r) apropriada.

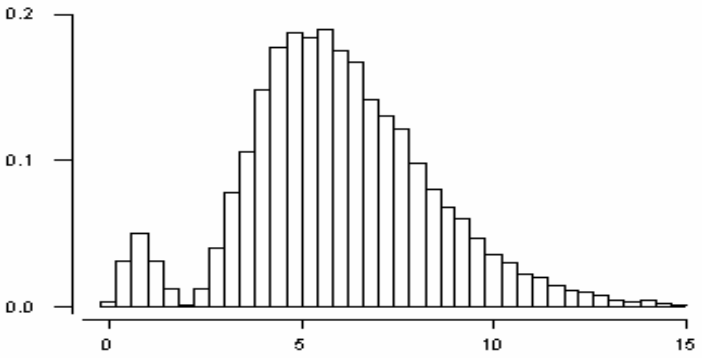
Há inúmeras escolhas possíveis para g(r). Na figura 5a a função h(r) está "envelopada" (grosseiramente) por Ag(r), onde $g(r) = (1/10)\exp(-x/10)$ e A = 30. É possível obter um envelopamento mais eficiente adotando-se diferentes funções para diferentes intervalos (figura 5b).

Figura 5 "envelopamento" para a função da figura 4



A figura 6 mostra um histograma para 12.000 valores gerados de r, utilizando-se o envelopamento da figura 5a. A figura 7 mostra a visualização do orbital em 2D.

Figura 6 Histograma para 12.000 valores gerados para r, com o envelopamento da figura 4a



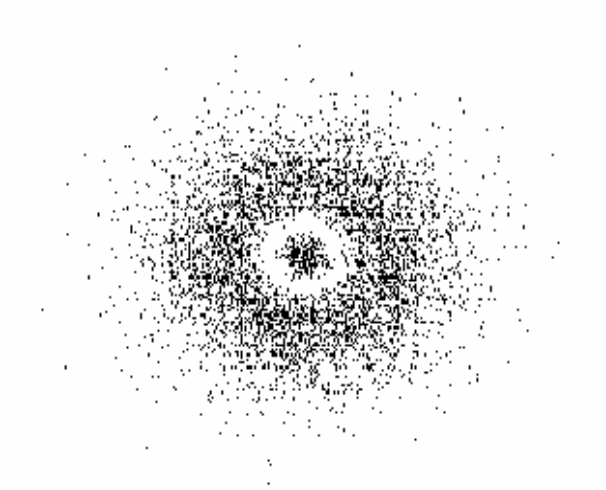


Figura 7
Visualização no plano do orbital 200 do átomo de hidrogênio

Exercícios

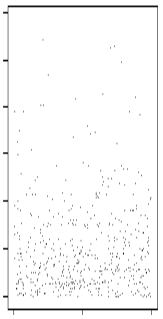


figura 8

3 - *Lei das atmosferas* - A figura 8 é uma simulação da distribuição de alturas de equilíbrio de partículas em um campo gravitacional (Tipler, l978). De acordo com a teoria cinética da matéria a função densidade de probabilidade para a altura de equilíbrio y de uma partícula é

$$f(y) = Ce^{-mgy/kT}$$

onde *m* em é a massa da partícula, *g* a constante gravitacional, *k* a constante de Boltzmann, T temperatura absoluta e *C* a constante de normalização.

Para elaboração da figura 8 foram gerados pontos de coordenadas aleatórias (x,y), geradas de maneira independente. A coordenada X foi gerada com distribuição uniforme no intervalo [0,1] e variável aleatória y com função densidade $f(y) = ae^{-ay}$. Determinar a expressão que permite gerar *y* a partir dos valores gerados *x*, dado a constante *a*.

4 – Gerar valores x^*_1 , x^*_2 , x^*_3, x^*_4 ...x^*_n de uma variável aleatória X tendo função densidade

$$f(x) = a/[\pi(x^2 + a^2)]$$ (distribuição de Cauchy).

(figuras 9 e 10). Escreva a expressão que permite gerar os valores x

a partir dos valores u^* gerado de uma uniforme [0,1]. Use *o método da inversão*.

Figura 9
Função densidade de Cauchy

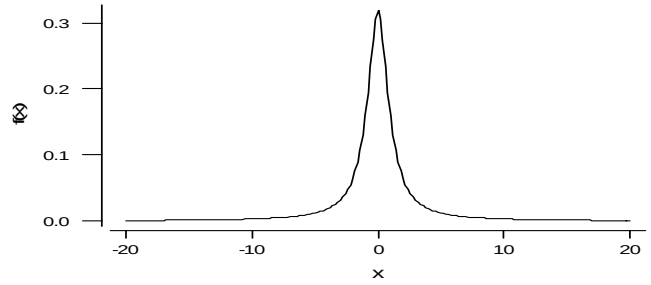


Figura 10
Histograma para 500 valores gerados com distribuição de Cauchy (intervalo -20 < x < 20**)**

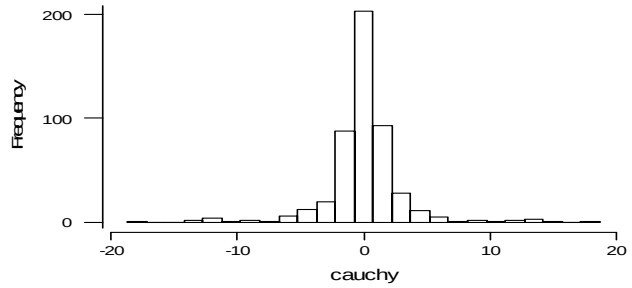
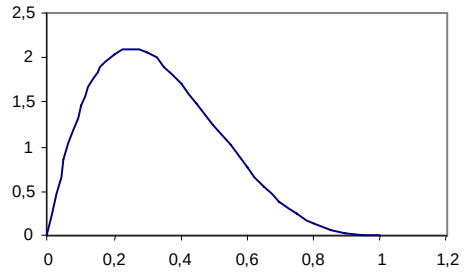


Figura 11



5 - Deseja-se simular uma variável aleatória X com função densidade

$$f(x) = 20x(1-x)^3, \qquad 0 < x < 1$$

(figura 11). O método escolhido foi o do envelopamento, com a função g(x) = 1, 0 < x < 1. Para a constante c tal que $f(x)/g(x) \leq c$ foi escolhido o valor 2,11, que é o menor valor possível para que ocorra o envelopamento no intervalo. Escreva um algoritmo para gerar valores de X com função densidade f(x), a partir dos números aleatórios U_1 e U_2, ambos com função densidade uniforme no

intervalo [0,1].

6 - Faça um programa para gerar 1000 valores de uma variável aleatória com a função distribuição do problema 2, com a = 1. A saída do programa deve ser um vetor disposto em coluna única em arquivo no formato texto. Faça um histograma para os valores obtidos (use aplicativos para este fim). Compare com a figura 6, que é um histograma para 500 valores de *x* gerados com esta distribuição.

7 - Reproduzir o exemplo 2, adotando outras funções para "envelopar" h(r). Obtenha um histograma como o da figura 6.

8 - Elabore um programa para visualizar, como na figura 7, o orbital 100 do hidrogênio, de acordo com a expressão (1).

9 - Elabore um programa para visualizar os orbitais 200 e 211 do hidrogênio, utilizando as expressões (3) e (4).

2 -Testes Estatísticos

2.1 Teste para diferença entre médias

Considere um experimento em que os resultados sejam aleatórios. Uma função dos dados observados (por exemplo, a média, mediana ou funções mais complexas) é denominada *estatística.* Uma questão relevante ao se trabalhar com estatísticas é: até que ponto os dados observados podem ser explicados como resultado do acaso ou se, pelo contrário, são *significativos,* do ponto de vista de uma tendência ou de uma característica observada?.

Este problema é analisado efetuando-se *testes estatísticos*. Normalmente existem testes específicos para cada tipo de situação e para alguns existem tabelas que permitem avaliar o nível de significância da estatística observada. Recentemente, os *testes de permutação* tem sido utilizados, principalmente devido às atuais facilidades computacionais.

O que é um teste de permutação

Um teste de permutação é basicamente a comparação do valor da estatística observada com as que são obtidas por uma simples permutação nos dados originais. O pressuposto básico é que, se uma estatística observada é devida ao mero acaso, uma permutação dos dados originais, mantendo-se a estrutura, deverá produzir estatísticas com valores semelhantes. Caso isto não ocorra, há indícios de não casualidade nos valores observados.

Os testes de permutação apresentam a vantagem sobre os testes convencionais de não necessitarem da suposição de aleatoriedade na coleta dos dados nem de uma suposição inicial para a função densidade da estatística observada (Edgington, 1995)

Bryan (1991) apresenta este exemplo em bioestatística de um teste de permutação:

Exemplo 3

Comprimento da mandíbula (em mm) de golden jackals (Canis aureus) para cada sexo, coleção de História Natural do Museu Britânico:

Machos:

120 107 110 116 114 111 113 117 114 112

Fêmeas:

110 111 107 108 110 105 107 106 111 111

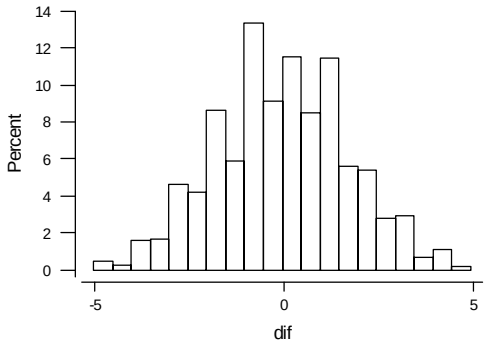
Há uma diferença entre *as médias* dos dois grupos de 4,8 mm. Mas são realmente "dois grupos"? Isto é, há evidências de que esta característica (comprimento da mandíbula) são diferentes para os dois grupos ou é um valor observado devido apenas a uma casualidade na disposição destes dados? Mesmo que nesta espécie os machos e as fêmeas tenham a mesma média, qualquer amostra com a estrutura acima poderá apresentar uma diferença aleatória entre as médias.

Em termos da teoria da probabilidade: os dois conjuntos de dados acima estão associados a funções densidade distintas, com médias diferentes, ou podem ser associados com maior probabilidade a uma única função densidade?

A idéia básica do teste de permutação neste caso é que *se não existe diferença entre as médias dos machos e fêmeas, então a*

diferença observada na amostra será tipicamente o valor observado para qualquer conjunto de dados, obtidos alocando-se ao acaso os vinte valores em dois grupos de 10.

Figura 12 - histograma para a diferença entre as médias dos dois grupos (machos-fêmeas), permutando-se os valores (1000 permutações).



Neste exemplo, o valor observado nos dados originais é 4,8 mm para a diferença entre as médias do grupo macho e do grupo fêmea. A figura 12 mostra um histograma obtido após 1000 permutações dos dados. Observe que o valor 4,8 mm é pouco provável de ocorrer, através de permutações. Podemos concluir que há evidências de que a diferença observada nos dados coletados é significativa, ou seja, existe uma diferença de fato entre as duas populações (machos e fêmeas).

Exercícios

10 – Escreva um algoritmo para permutar 20 elementos, dispostos em dois grupos de 10, permutando-se os dados entre os dois grupos. *Sugestão: a permutação pode ser feita sorteando-se, com igual probabilidade, os integrantes de cada grupo entre os 20 elementos, sem reposição.*

11 - Reproduza o teste do exemplo 3. Etapas:

a) Gerar 1000 amostras com 20 elementos cada, dispostos em dois grupos de 10, permutando-se os dados do exemplo 3 entre os dois grupos

b) Para cada amostra gerada calcule X = diferença entre as médias entre os dois grupos. Observe que X é uma variável aleatória.

c) Após as 1000 simulações calcular a porcentagem de dados que ficaram abaixo do valor observado (se X original negativo) ou

acima (se X original positivo). Isto é uma estimativa para uma grandeza denominada “p valor” do teste. Observe que p-valor abaixo de 10% é evidência a favor da hipótese de uma real desigualdade na média dos dois grupos.

d) Faça um histograma, com os valores de X obtido das 1000 simulações, como o da figura 12 (*sugestão: elabore uma programa cuja saída seja um arquivo texto e use aplicativos tipo Excel para visualização do histograma*).

12 - (Devore, 1987). As observações abaixo são o nível de pH de dezoito amostras de camadas superficiais de solo para duas localidades diferentes (*Central Soil Salinity Research Institute).* A questão de interesse é verificar se há diferença no nível médio de pH entre estas duas localidades

Localidade A

8.53 8.52 8.01 7.99 7.93 7.89 7.85 7.82 7.80

Localidade B:

7.85 7.73 7.58 7.40 7.35 7.30 7.27 7.27 7.23

Faça o teste de permutação para a diferença entre as médias, como o do exercício anterior. Verifique o p-valor do teste e conclua se a diferença observada entre as médias dos dois grupos, nos dados originais, é realmente significativa. Faça um histograma, como o da figura 12.

1.2 Séries Temporais

Série temporal é um conjunto de valores observados ordenados em função do tempo. Se as observações no tempo são igualmente espaçadas, a série é dita *regular.* Inúmeros fenômenos físicos, econômicos, biológicos, etc. podem ser visualizados como séries temporais e várias perguntas sobre o comportamento da série podem ser de interesse, dependendo do contexto. O estudo de tais séries é uma área ampla na estatística e há uma grande variedade de testes e técnicas disponíveis para análise das mesmas.

Como exemplo, pode ser de interesse verificar a hipótese de que uma série dada seja o resultado do seguinte processo:

$$X_i = \beta X_{i-1} + \varepsilon$$

onde i é o índice associado ao tempo da observação, β é uma constante e ε uma perturbação aleatória. Este processo é um processo de Markov e o modelo denominado *auto-regressivo de primeira ordem (*ou simplesmente AR(1)). Neste modelo é suposto ainda que os ε_i são independentes, com média zero e variância constante ao longo do processo (usualmente supõe-se ainda que os ε_i sejam *normalmente distribuídos*, mas para um teste de permutação esta suposição não é necessária).

Dada uma série observada, β pode ser estimado pelo método dos mínimos quadrados. A estatística de teste indicada neste caso é (Mainly, 1993, p.277):

$$v = \sum_{i=2}^{n} (x_i - x_{i-1})^2 \; / \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1)$$

Um teste de permutação para este problema consiste em comparar o valor da estatística v observada na série original e comparar com a distribuição obtida permutando-se os elementos x_i da série original.

Observe que, dependendo do tamanho da série, os cálculos podem se tornar muito intensivos, com as permutações excessivamente lentas. Uma alternativa é permutar as séries por blocos de 3 elementos (Efron,1991 p.99). A justificativa para que as permutações sejam por blocos é evitar destruir nas séries permutadas a eventual correlação existente entre instantes sucessivos. De um modo geral, a decisão de permutar por blocos deve ser analisada com cuidado, verificando-se as consequencias deste procedimento em cada caso.

Exercícios

13 - *luteinnizing hormone data (Efron, p.92).* - Os dados abaixo constituem uma série temporal para o nível de hormônio, monitorado a cada 10 minutos em um ciclo 8 horas O modelo AR(1) associado é

$$z_t = 0{,}586 z_{t-1} + \varepsilon_i \qquad (2)$$

Efetue um teste gerando 200 séries simuladas, permutando-se as posições de blocos de 3 elementos da série original (construa as séries sorteando os blocos componentes, com reposição). Para cada

série calcule a estimativa para β (equação 1) e construa um histograma dos valores obtidos. Estime o nível de significância do valor adotado em (2).

período	**nível**	período	**nível**	período	**nível**	período	**nível**
1	**2.4**	13	**2.2**	25	**2.3**	37	**1.5**
2	**2.4**	14	**1.8**	26	**2.0**	38	**1.4**
3	**2.4**	15	**3.2**	27	**2.0**	39	**2.1**
4	**2.2**	16	**3.2**	28	**2.9**	40	**3.3**
5	**2.1**	17	**2.7**	29	**2.9**	41	**3.5**
6	**1.5**	18	**2.2**	30	**2.7**	42	**3.5**
7	**2.3**	19	**2.2**	31	**2.7**	43	**3.1**
8	**2.3**	20	**1.9**	32	**2.3**	44	**2.6**
9	**2.5**	21	**1.9**	33	**2.6**	45	**2.1**
10	**2.0**	22	**1.8**	34	**2.4**	46	**3.4**
11	**1.9**	23	**2.7**	35	**1.8**	47	**3.0**
12	**1.7**	24	**3.0**	36	**1.7**	48	**2.9**

14 - Repita o teste do problema 13, utilizando exemplo de outras séries reais (procure em *sites* de banco de dados na Internet).

2.3 Teste de independência para duas séries binárias

Séries representando comportamento animal - Solow (1995) aplica um teste de permutação neste experimento. Trata-se de um experimento efetuado para verificação da sociabilidade de duas baleias, que se encontravam confinadas em um tanque. O experimento foi realizado no Hole Oceanographic Institut e consistiu em monitorar a posição de duas baleias A e B no tanque, que foi subdividido em dois (tanque "0" e tanque "1"). O ambiente dos dois subtanques foram controlados para oferecerem as mesmas condições ambientais e a posição das baleias registrada a cada minuto, durante 1 hora. A saída dos dados foi no seguinte formato:

Experimento 1

A::

0110101001101101100001000011101001100100000000000000000000001

B:

1110101111111101110111101111111001100110001010100000010000000

Experimento 2

A:

1011101000000011110010001110110001111110111010100000111001000

B:
0000000001111101011100111000000100000000101001001000000100100

Assim, no experimento 1, após o primeiro minuto A ocupava o tanque 1, o indivíduo B o tanque 0 e assim por diante. Há indícios de dependência entre as duas séries? Ou seja, há indícios de que a presença de um indivíduo afeta o comportamento do outro, para o caso dos experimentos 1 e 2 acima?

Uma maneira de verificar a independência ou não é calcular o coeficiente de correlação ρ. Valores de ρ próximos a 1 (ou –1) indicam forte correlação positiva (ou negativa); valores próximos a zero indicam ausência de correlação. Novamente nos deparamos com o problema de verificar se o coeficiente de correlação observado é realmente significativo.

Solow verificou a existência de independência calculando o coeficiente de correlação e permutando a série, porém mantendo para cada série permutada o mesmo número de transições 1 - 0 e 0 - 1. No entanto, uma permutação apenas das transições dos elementos individuais podem destruir a informação contida na seqüência, além de necessitar da suposição de que a cadeia é Markoviana.

Um método alternativo é deslocar uma cadeia em relação à outra, de 1 unidade, 2 unidades...n-1 unidades e para cada "shift" calcular o novo coeficiente de correlação. É necessário supor uma estrutura circular dos dados. O cálculo de ρ^* para cada deslocamento permite a geração de n -1 valores; a comparação com o valor observado e o cálculo do nível de significância permite rejeitar ou não a idéia de independência.

Um teste alternativo

Ao invés de utilizar, em cadeias binárias, o coeficiente de correlação podemos utilizar outras estatísticas. Para uma medida de associação entre as cadeias podemos utilizar a seguinte estatística (Assunção e Santos, 1997): para cada dígito 1 da cadeia A, computar a média observada em uma janela com h elementos na cadeia B. Calcular K = média observada em todas as as janelas. Efetuando o

deslocamento das cadeias é possível determinar o nível de significância para K, em função de h. Observe que o que está sendo testado é se a ocorrência de "uns" em uma cadeia induz o aparecimento de "uns" na outra, em uma vizinhança de até h passos (figura abaixo).

i : 1.... 2 3 4.... 5..... . 6 7 8...
A: 1 0 1 1 1 0 1 0 0 0 0 0 0 1 1 1 1 0 0 1 0 0 0 1 0 1 1 ...
___•___
B: 0 0 1 0 1 1 1 0 1 0 0 0 0 0 0 0 1 1 1 1 0 0 1 0 1 1 1 0 ...

deslocamento de A em relação B:

A: ⇒ 1 0 1 1 1 0 1 0 0 0 0 0 0 0 1 1 1 1 0 0 1 0 0 0 1 0 1 1 ⇒
___|___
B: 0 0 1 0 1 1 1 0 1 0 0 0 0 0 0 0 1 1 1 1 0 0 1 0 1 1 1 0 ...

Exercícios

15 a) - Faça um programa que, a partir de um arquivo texto com duas sequências dedados binários cada, efetue um teste para verificar independência entre as séries. Ao invés de efetuar o teste permutando as cadeias, faça o teste deslocando uma cadeia em relação à outra (será necessário assumir uma estrutura circular para os dados, isto é, o valor $X_{n+1} = X_1$). Passos:

1. deslocar a cadeia Y de 1 unidade em relação à cadeia X calcular o coeficiente de correlação ρ^* ;
2. repetir, deslocando até n -1 unidades;
3. calcular o nível de significância do ρ observado na sequência original.

A saída do programa deverá ser o valor observado do coeficiente de correlação e o nível de significância observado.

b) - Faça um programa que gere duas cadeias binárias, X e Y, com 200 elementos cada e que sejam *independentes*. Uma saída do programa deve ser um arquivo texto com as colunas X, Y e que possa ser lido pelo programa elaborado no item a.

Sugestão para gerar as cadeias: Escolha p_1 , p_2 e gere os valores de X e Y de modo independente com
$P(X_i = 1) = p_1$; $P(X_i = 0) = 1 - p_1$
$P(Y_i = 1) = p_2$ $P(Y_i = 0) = 1 - p_2$, para todo i.

Utilize o programa do item a para efetuar o teste de independência. Como as cadeias são de fato independentes, o nível de significância para o coeficiente de correlação deverá ser alto.

c) Faça um programa que gere duas cadeias binárias, X e Y, como no exercício anterior porem que sejam *dependentes*.

Sugestão: Escolha um valor de p e gere primeiro a cadeia X com uma estrutura binomial: $P(X_i = 0) = p$; $P(X_i = 1) = 1- p$, para todo i. Gere os valores Y_i com

$P(Y_i = 1) = 2/3$, se $X_i = 1$;
$= 1/3$ se $X_i = 0$.

(veja o item sobre geração de variáveis aleatórias discretas). Observe que desta maneira a cadeia Y é aleatória mas com certo grau de dependência da cadeia X.

d) Faça o teste de independência para os dados reais dos experimentos 1 e 2 item 2.2.

16 – Gere sequências binárias com uma outra estrutura de dependência, diferente da proposta no item 13c. Faça o teste para estas sequências e verifique o resultado.

17 - Elabore um programa para efetuar o teste alternativo proposto no final do item 2.2 para as cadeias geradas no problema 15, com h máximo igual a 10.

2.4 Teste Espacial

Considere o padrão de pontos da figura 13. Padrões como este podem representar árvores de uma determinada espécie em uma floresta, focos de incêndio em uma região geográfica ou um outro problema cujos dados podem ser associados a uma distribuição espacial de pontos no plano. O interesse é verificar se a distribuição espacial pode ser explicada como o resultado de uma completa aleatoriedade ou se está associada, com maior probalidade, a padrões espaciais com características específicas.

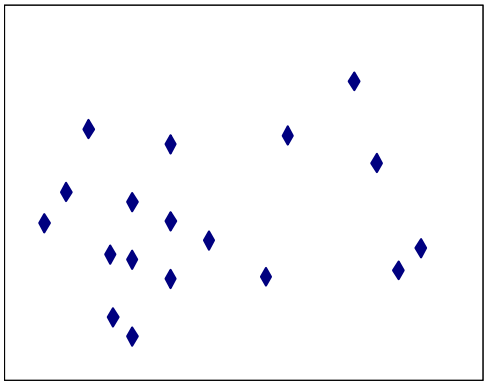


Figura 13
Distribuição espacial de pontos

Para abordar este tipo de problema existem os testes estatísticos para padrões espaciais. É uma área importante da estatística e com vários avanços teóricos recentes. Neste texto apresentaremos um teste relativamente simples, baseando em simulações de padrões aleatórios com distribuição uniforme.

Este teste é apresentado por Bryan (1991). Dado um padrão espacial de pontos, do qual se deseja verificar se a distribuição dos pontos pode ser atribuída completamente ao acaso ou não:

1 - defina as variáveis

g_1 = média das distâncias entre os pontos e seus vizinhos mais próximos (para N pontos, haverá N distâncias a serem consideradas);
g_2 = média das distâncias entre os pontos e o segundo vizinho mais próximo;...
g_i = média das distâncias entre os pontos e o i-ésimo vizinho mais próximo.

Calcule g_1, g_2 ,...g_{10}. Estes valores são as *médias observadas.*

2 – Gere 999 padrões em área equivalente com N pontos, com coordenadas dos pontos geradas a partir de uma distribuição uniforme de probabilidade. Para cada simulação calcular o vetor [g_1,

g_2 ,...g_{10}].

3 – Verificar, para cada *i*, o nível de significância das médias observadas. Nível de significância é a probabilidade estimada de se obter um valor igual ou mais extremo do que o observado. Pode ser calculado como a proporção de valores iguais ou mais extremo ao valor observado.

Analisando-se o nível de significância estimado para os g_i, é possível concluir se o padrão pode ser atribuído ou não, com maior probabilidade, ao acaso.

A tabela abaixo é um exemplo de resultado de teste efetuado para um conjunto de 24 plantas, dispostas em um quadrado com 2m de lado (Bryan). Neste exemplo há clara que evidência de que as distâncias médias entre os vizinhos, exceto para o primeiro caso, são valores pouco prováveis de serem encontrados se as posições das plantas fossem realmente completamente ao acaso.

Tabela 1 – Resultado de um teste para verificação de completa aleatoriedade de uma distribuição de 24 plantas em uma área quadrada com 2m de lado. (Bryan)

	g1	*g2*	*g3*	*g4*	*g5*	*g6*	*g7*	*g8*	*g9*	*g10*
Observado	0,217	0,293	0,353	0,419	0,500	0,559	0,606	0,646	0,698	0,739
Signifcância (%)	72,9	8,0	0,9	0,3	0,9	0,5	0,8	0,4	0,4	0,2

Exercícios

18 – Efetuar o teste para completo aleatoriedade para os dados do apêndice.

O teste deverá ser realizado nas seguintes etapas:

a) desenvolver um programa que, dado as coordenadas X,Y de N pontos (tipicamente 20 a 30 pontos), efetua as simulações e calcula o nível de significância para os g_i. Uma saída do programa deve ser uma tabela como a tabela 1.

b) Gere um padrão de pontos uniforme e efetue o teste;

c) Gere um padrão não uniforme e efetue o teste. *Sugestão para um padrão não uniforme: gere coordenadas X e Y independentes com distribuições não uniforme - veja a figura 14.*

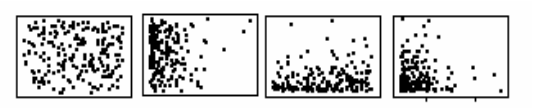


Figura 14
Padrões gerados com a) X e Y uniforme; b) X exponencial, Y uniforme; c) X uniforme, Y exponencial e d) X e Y exponencial

d) Teste dados reais, escolhendo uma das espécies de árvores do apêndice.

Apêndice
Coordenadas espaciais de 174 árvores de floresta tropical,de quatro diferentes espécies, Região de Linhares, ES

Dados gentilmente cedidos pelo Prof. João L.F.Batista Depto de Ciências Florestais, ESALQ/USP.

x,y são as coordenadas retangulares em uma área 100 x 50 m

árvore	espécie	X	y
1	1	7,9	2,8
2	1	90,5	31,6
3	1	92,9	33,6
4	1	90,3	34,0
5	1	83,7	47,9
6	1	86,2	28,5
7	1	84,5	28,5
8	1	83,6	20,2
9	1	73,1	20,4
10	1	78,0	24,0
11	1	76,3	26,1
12	1	72,5	32,4
13	1	70,7	41,5
14	1	60,2	45,1
15	1	67,4	40,7
16	1	60,3	37,2
17	1	60,9	32,3
18	1	60,8	30,8
19	1	64,9	28,1
20	1	60,6	26,4
21	1	61,9	18,9
22	1	67,0	17,5
23	1	69,4	18,2
24	1	69,5	14,3
25	1	68,0	15,7
26	1	55,2	21,9
27	1	55,2	21,9
28	1	57,5	32,4
29	1	59,2	32,7
30	1	60,0	33,0
31	1	57,3	35,4
32	1	59,6	38,8
33	1	50,5	45,3
34	1	45,6	42,7
35	1	47,8	37,2
36	1	42,5	37,2
37	1	40,4	34,6
38	1	41,3	34,1
39	1	45,7	31,3
40	1	46,9	28,8
41	1	43,4	21,8
42	1	45,4	18,4
43	1	49,1	17,0
44	1	45,8	15,1
45	1	44,1	15,9
46	1	40,5	14,8
47	1	40,5	14,8
48	1	35,9	10,9
49	1	40,5	14,8
50	1	31,7	16,8
51	1	36,4	21,1
52	1	35,1	26,9
53	1	31,8	27,2
54	1	35,6	30,6
55	1	35,2	31,4
56	1	39,0	34,3
57	1	35,8	34,4
58	1	38,1	36,9
59	1	35,9	38,0
60	1	35,5	39,3
61	1	32,3	35,6
62	1	29,4	37,4
63	1	29,3	30,4
64	1	27,9	28,1
65	1	20,5	21,5
66	1	26,0	24,1
67	1	29,8	20,6
68	1	20,1	18,2
69	1	19,5	14,3
70	1	18,9	18,1
71	1	17,5	21,3
72	1	14,6	23,8
73	1	13,0	26,9
74	1	14,2	31,8
75	1	17,2	34,2
76	1	15,2	37,3
77	1	8,6	40,2
78	1	6,5	39,5
79	1	7,0	34,8
80	1	7,2	34,1
81	1	1,8	34,2
82	1	1,8	31,0
83	1	2,9	24,4
84	1	6,0	9,4
85	2	82,3	17,8
86	2	73,2	3,1
87	2	76,5	4,0
88	2	50,4	10,9
89	2	50,3	30,8
90	2	47,4	45,8
91	2	34,1	15,2
92	2	32,4	48,0
93	2	27,2	38,4
94	2	26,1	13,8

95	2	9,9	19,1
96	2	17,3	47,9
97	2	10,6	49,0
98	2	2,0	38,3
99	2	0,8	37,6
100	2	9,5	25,0
101	2	3,9	20,2
102	2	2,4	13,9
103	3	8,1	42,6
104	3	13,6	44,9
105	3	12,4	43,8
106	3	19,1	43,8
107	3	14,2	32,3
108	3	25,3	34,7
109	3	30,8	47,0
110	3	34,3	32,6
111	3	36,7	1,6
112	3	40,6	14,1
113	3	51,4	27,4
114	3	67,6	36,6
115	3	68,4	40,9
116	3	70,3	24,2
117	3	87,3	23,3
118	3	89,7	46,2
119	3	95,2	44,8
120	3	98,2	9,6
121	4	5,9	6,4
122	4	8,3	7,0
123	4	3,9	13,5
124	4	9,8	19,4
125	4	6,1	25,2
126	4	15,6	40,4
127	4	18,8	37,3
128	4	15,5	31,2
129	4	11,0	10,8
130	4	14,0	8,4
131	4	29,2	5,5
132	4	20,7	10,7
133	4	29,2	24,6
134	4	29,5	28,1
135	4	25,6	43,0
136	4	32,7	39,2
137	4	32,3	35,2
138	4	35,9	33,9
139	4	38,1	4,0
140	4	45,4	4,3
141	4	49,4	8,5
142	4	44,2	18,2
143	4	45,5	21,2
144	4	41,2	31,4
145	4	45,0	33,8
146	4	48,0	37,9
147	4	54,9	33,5
148	4	55,9	25,3
149	4	52,4	16,7
150	4	57,2	8,0
151	4	5,8	6,5
152	4	53,3	6,8
153	4	60,3	0,9
154	4	61,1	5,7
155	4	68,8	25,2
156	4	66,0	31,6
157	4	60,6	42,6
158	4	78,9	41,8
159	4	75,6	42,2
160	4	72,2	30,9
161	4	71,2	18,9
162	4	77,6	12,4
163	4	80,9	1,3
164	4	86,5	14,2
165	4	90,0	27,6
166	4	85,9	32,9
167	4	9,7	48,2
168	4	96,3	42,7
169	4	94,8	38,4
170	4	96,7	27,9
171	4	93,7	20,6
172	4	9,3	5,4
173	4	91,3	3,6
174	4	97,4	1,1

Bibliografia

Devore, Jay L. (1987), *Probability and Statistics for Engineering and the Sciences..* Brooks/Cole Publishing Company, USA

Edgington, Eugene S. (1995) *Randomization Tests* (3rd ed.). Marcel Dekker, Inc. New York

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap.* Chapman & Hall, London

Maily, Bryan F. J. (1991) *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, London

Renato M. Assunção, Marcos A C Santos (1997) *Analysing Correlation Between Two Discrete Sthocastic Processes With Applications to Animal Behavior.* Relatório Técnico - Departamento de Estatística, UFMG

Ross, Sheldon M. (1990) *Simulation*, 2nd edition. Academic Press, New York

Solow, A. R., Smith, W. K., and Recchia, C. (1995) *A conditional test of independence between two Markov chains*. Biometrical Journal, 8, 973-977

Tipler, Paul A. (1981) *Física Moderna.* Guanabara Dois, Rio de Janeiro