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

Introdução às Pilhas e Filas e Teste de Permutação

Luiz H.Duczmal, Lupércio F. Bessegato,
Marco A. da Cunha Santos e Sabino J. Ferreira Neto

Relatório Técnico
RTE-03/2003
Série Ensino

Introdução

Este trabalho tem o objetivo de introduzir as idéias básicas em pilhas, filas e testes de permutação. Há uma vasta literatura sobre os estes temas; o objetivo deste texto é o de introduzir, através de exemplos simples, as idéias básicas envolvidas. Assim, alguns exemplos de pilhas e filas são apresentados juantamente com os métodos computacionais básicos utilizados no estudo destes temas. Os testes de permutação são apresentados através de exemplos ilustrativos de série temporal e modelo de regressão simples.

O ponto em comum na discussão dos temas envolvidos neste trabalho (pilhas, filas e testes) é a ênfase na utilização de métodos computacionais, com a discussão dos algoritmos envolvidos e exemplos de implementação.

O trabalho foi concebido como um texto introdutório aos tópicos citados e dirigido principalmente aos estudantes de graduação na área de exatas. De um modo geral é suposto apenas do leitor alguma familiaridade com ambiente computacional e as noções básicas de probabilidade e variável aleatória.

Os Autores

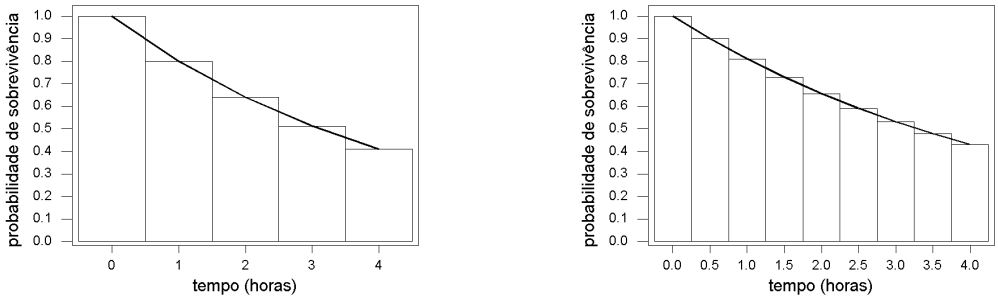
CAPÍTULO 1
Pilhas e Filas

INTRODUÇÃO

Muitas vezes não podemos prever exatamente quando um evento ocorre em um processo complexo. Por exemplo, é virtualmente impossível saber com antecedência exatamente quando uma central telefônica irá receber uma chamada, ou quando um avião irá pousar em um aeroporto muito movimentado. No entanto, pode ser possível descrever o *comportamento médio do sistema*, isto é, apesar de não sermos capazes de saber detalhadamente quando acontece cada evento individual, podemos prever com razoável precisão se uma central telefônica corre o risco de ficar congestionada ou se muitos aviões vão chegar atrasados. O *processo de Poisson* é a descrição matemática de um sistema em que eventos ocorrem ao longo do tempo, independentes uns dos outros, com um comportamento médio bem definido. Antes de definir formalmente esse tipo de sistema precisamos desenvolver alguns pré-requisitos.

1. A VARIÁVEL ALEATÓRIA EXPONENCIAL

Vamos tentar fazer um modelo da duração de uma lâmpada incandescente. Desde o momento em que a lâmpada é ligada, seu filamento corre um risco constante de se partir, devido a fissuras microscópicas que vão aparecendo em sua estrutura. Chega um certo momento que essas fissuras crescem a tal ponto que o filamento se rompe, e a lâmpada queima. Nosso modelo simplificado vai ser assim: Observamos a lâmpada uma vez por hora, verificando se ela ainda funciona. A cada hora, existe uma probabilidade p de que ela queime – senão ela continua funcionando por mais uma hora, e novamente ela corre um risco p de queimar, e assim por diante, até que por fim em alguma hora ela finalmente queima. O gráfico da figura 1A mostra a probabilidade de sobrevivência da lâmpada com o passar das horas, em que fizemos p=0.2. É claro que esse modelo pode ser melhorado; digamos, observando a lâmpada uma vez a cada 30 minutos, e a cada etapa ela corre um risco 0.1 de queimar. O gráfico de probabilidade de sobrevivência seria formado por colunas mais estreitas, mas o formato geral da curva que passa pelo ponto médio do topo das colunas seria bem similar (figura 1B).



Figuras 1A e 1B: Acompanhando a probabilidade de sobrevivência de uma lâmpada a cada hora, e a cada meia hora.

No caso limite, a lâmpada é observada continuamente, e vemos uma curva *f* decrescente, com a seguinte propriedade: a proporção da altura da curva entre quaisquer três instantes $t_1 < t_2 < t_3$ igualmente espaçados se mantém constante:

$\frac{f(t_1)}{f(t_2)} = \frac{f(t_2)}{f(t_3)}$, onde $\Delta t = t_3 - t_2 = t_2 - t_1$.

Sabemos do Cálculo Diferencial que funções f com essa propriedade são do tipo exponencial, $f(t) = Ce^{-kt}$, onde $C>0$ e $k>0$ são parâmetros. Vamos escolher cuidadosamente esses parâmetros de modo que a duração T da lâmpada seja uma variável aleatória cuja função de densidade de probabilidade (*fdp*) seja a função *f*. Estamos agora abandonando a interpretação inicial do gráfico de colunas como probabilidade de sobrevivência, em prol de uma interpretação mais útil (veja a figura 2).

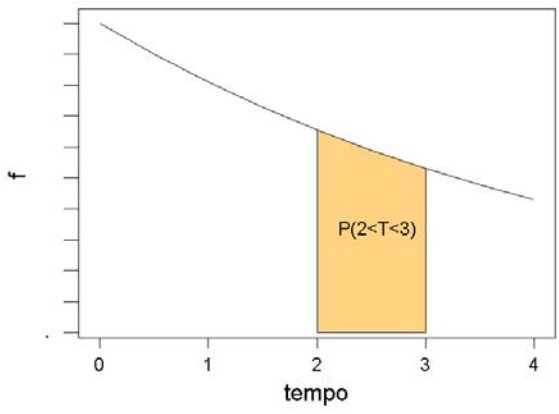


Figura 2: O parâmetro C é escolhido igual a k para que a função *f* seja interpretada como a f.d.p. da variável aleatória *T*. Por exemplo, a área sombreada é igual a $P(2<T<3)$.

A propriedade a ser satisfeita por f é

$$\int_0^\infty Ce^{-kt}\,dt = 1.$$

Assim

$$\frac{-C}{k}e^{-kt}\Big|_0^\infty = \frac{C}{k} = 1 \quad \Rightarrow \quad C = k.$$

Portanto $f(t) = ke^{-kt}, \quad t \geq 0.$

Valores altos de *k* fazem com que *f* decaia rapidamente, e a duração média é menor. O oposto ocorre para valores baixos de *k*.

Exercício 1A: Mostre que a variável aleatória T definida acima tem média $1/k$.
Exercício 1B: Calcule o tempo necessário para que a lâmpada tenha 50% de probabilidade de ainda estar funcionando.
Exercício 1C: Use o Método da Função Inversa (veja a referência 1) para mostrar que o seguinte algoritmo gera a variável aleatória exponencial T com fdp $f(t) = ke^{-kt}, \quad t \geq 0$:

```
U=RAND(0,1);   %gera um número aleatório uniformemente distribuído no intervalo (0,1)
T=-LN(U)/k;    % gera a variável aleatória exponencial T
```

2. A VARIÁVEL ALEATÓRIA DE POISSON

Definição: A *variável aleatória de Poisson* é uma variável discreta X tal que $P(X = k) = \frac{e^{-\alpha}\alpha^k}{k!}$,

para $k = 0,1,2,3,\ldots$, onde α é um número real positivo.

Por exemplo, se $\alpha = 2.5$, a distribuição da variável aleatória X obedece à seguinte tabela:

K	0	1	2	3	4	5	6	7	...
P(X=k)	.0821	.2052	.2565	.2138	.1336	.0668	.0278	.0099	...

Exercício 2A: Mostre que a variável aleatória de Poisson X está bem definida, isto é,

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{e^{-\alpha}\alpha^k}{k!} = 1.$$

Dica: Use o fato de que $e^{\alpha} = 1 + \alpha + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} + \cdots$.

Exercício 2B: Mostre que a variável aleatória de Poisson X tem média α, isto é,

$$E(X) = \sum_{k=0}^{\infty} k\,P(X = k) = \sum_{k=0}^{\infty} k\,\frac{e^{-\alpha}\alpha^k}{k!} = \alpha.$$

Dica: Use o exercício 2A, e também que $0.P(X = 0) = 0.$

3. O PROCESSO DE POISSON

Vamos supor agora que uma central telefônica esteja conectada a um grande número de telefones. As chamadas originadas de cada telefone atingem a central de forma completamente aleatória. Vamos supor neste exemplo que chegam *em média* $\alpha = 3$ chamadas por minuto na central. Um modelo muito simplificado do que acontece pode ser feito assim: A cada 10 segundos, jogue uma moeda honesta (com iguais probabilidades de sair cara ou coroa). Se sair cara, registre que uma chamada chegou na central. Caso contrário não registre nada. Vamos ter então uma seqüência de números binários, 1 indicando que houve uma chamada naquele intervalo de 10 segundos, e 0 indicando que não houve. Esse modelo tosco tem uma boa propriedade: em média chegam 3 chamadas por minuto. Isso ocorre porque o número X de chamadas que atingem a central, de acordo com esse modelo, é uma variável aleatória binomial:

$$P(X = k) = \binom{6}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{6-k}, \quad k = 0,1,...,6,$$

e sua média é $Np = 6.\frac{1}{2} = 3$.

No entanto esse modelo tem duas falhas óbvias:
1) Nunca ocorrem duas ou mais chamadas num mesmo sub-intervalo de 10 segundos;
2) Nunca ocorrem mais de 6 chamadas em um minuto.

Vamos tentar sanar essas falhas modificando nosso modelo. Ao invés de 6 sub-intervalos, construa N sub-intervalos de $\frac{1}{N}$ minutos cada. No lugar de uma moeda honesta, jogue em cada sub-intervalo uma moeda que tenha probabilidade $p = \frac{\alpha}{N}$ de sair cara. Assim, se N for grande, as falhas apontadas acima são minimizadas. A probabilidade de ocorrerem *k* chamadas em um minuto é

$$P(X = k) = \binom{N}{k}\left(\frac{\alpha}{N}\right)^k\left(1-\frac{\alpha}{N}\right)^{N-k}, \quad k = 0,1,...,N.$$

pois *X* é uma variável binomial com $p = \frac{\alpha}{N}$.

O modelo ideal pode então ser construído como o caso limite do modelo acima, fazendo $N \to \infty$.

Sendo *Y* a variável aleatória que mede o número de chamadas que ocorrem em um minuto nesse modelo ideal, podemos ver que

$$\begin{aligned}
&P(Y = k) = \\
&= \lim_{N\to\infty} P(X = k) \\
&= \lim_{N\to\infty}\binom{N}{k}\left(\frac{\alpha}{N}\right)^k\left(1-\frac{\alpha}{N}\right)^{N-k} \\
&= \lim_{N\to\infty}\frac{N(N-1)\cdots(N-k+1)}{k!}\frac{\alpha^k}{N^k}\left(1-\frac{\alpha}{N}\right)^N\left(1-\frac{\alpha}{N}\right)^{-k}
\end{aligned}$$

$$= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-k+1)}{N^k} \frac{\alpha^k}{k!} \left(1-\frac{\alpha}{N}\right)^N \left(1-\frac{\alpha}{N}\right)^{-k}$$
$$= \frac{\alpha^k}{k!} \lim_{N\to\infty} \frac{N(N-1)\cdots(N-k+1)}{N^k} \left(1-\frac{\alpha}{N}\right)^N \left(1-\frac{\alpha}{N}\right)^{-k}$$
$$= \frac{\alpha^k}{k!} \lim_{N\to\infty} \frac{N}{N}\frac{N-1}{N}\cdots\frac{(N-k+1)}{N} \left(1-\frac{\alpha}{N}\right)^N \left(1-\frac{\alpha}{N}\right)^{-k}$$

Como k é fixo,

$$\lim_{N\to\infty} \frac{N}{N}\frac{N-1}{N}\cdots\frac{(N-k+1)}{N} = 1 \quad \text{e}$$

$$\lim_{N\to\infty} \left(1-\frac{\alpha}{N}\right)^{-k} = 1 .$$

Do Cálculo Diferencial, sabemos que

$$\lim_{N\to\infty} \left(1-\frac{\alpha}{N}\right)^N = e^{-\alpha} .$$

Assim, concluímos que

$$P(Y=k) = \lim_{N\to\infty} P(X=k) = \frac{e^{-\alpha}\alpha^k}{k!},$$

mostrando que Y é de fato a variável aleatória de Poisson que vimos anteriormente, com $\alpha = Np$. Como ilustração, a tabela a seguir mostra alguns valores da distribuição de X (com N=6, 12, 60 e 600) e de Y (com α=3.0). Observe que à medida que N cresce, as distribuições binomiais se aproximam da distribuição de Poisson.

k	0	1	2	3	4	5	6
P(X=k)(N=6,p=0.5)	.015625	.093750	.234375	.312500	.234375	.093750	.015625
P(X=k)(N=12,p=0.25)	.031676	.126705	.232293	.258104	.193578	.103241	.040149
P(X=k)(N=60,p=0.05)	.046070	.145484	.225882	.229845	.172384	.101616	.049025
P(X=k)(N=600,p=0.005)	.049414	.148986	.224228	.224604	.168453	.100902	.050282
P(Y=k)($\alpha = 3.0$)	.049787	.149361	.224042	.224042	.168031	.100819	.050409

O intervalo de tempo T entre dois eventos consecutivos no modelo ideal apresentado anteriormente é uma variável aleatória exponencial, com média $1/\alpha$. Para se convencer de que realmente é isso o que acontece, imagine que a central telefônica corre um risco constante, ao longo do tempo, de receber uma chamada: o modelo é então idêntico ao modelo que vimos no início do capítulo para a duração de uma lâmpada. Como ocorrem em média α chamadas por minuto, então evidentemente a duração média de cada chamada deve ser $1/\alpha$. Associado a um processo de Poisson existem então duas variáveis aleatórias:

T : o intervalo de tempo entre duas chamadas consecutivas; T é v.a. contínua exponencial com fdp $f(t) = \alpha e^{-\alpha t}, t \geq 0$.

Y : o número de eventos que ocorrem no intervalo de tempo $(0, t_{MAX})$; Se escolhemos t_{MAX}=1, então Y é uma v.a. discreta de Poisson, $P(Y = k) = \frac{e^{-\alpha}\alpha^k}{k!}$, $k = 0,1,2,3,...$ (fig. 3).

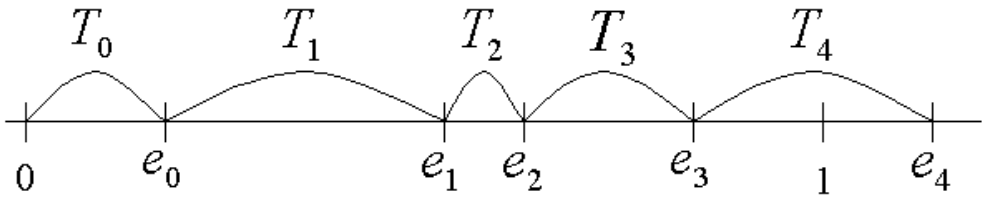


Figura 3: Nesse exemplo, os valores simulados T_0, T_1, T_2, T_3, T_4 da variável aleatória T formam os tempos acumulados dos eventos e_0, e_1, e_2, e_3, e_4, onde

$e_0 = T_0$

$e_1 = T_0 + T_1$

$e_2 = T_0 + T_1 + T_2$

$e_3 = T_0 + T_1 + T_2 + T_3$

$e_4 = T_0 + T_1 + T_2 + T_3 + T_4$.

O evento de tempo e_4 é maior que 1, e portanto $Y = 4$, pois foram gerados 4 eventos no intervalo (0,1).

Exercício 3A: Implemente o código abaixo, que gera eventos de um processo de Poisson. Para gerar a variável aleatória exponencial, foi usado o exercício1C.

```
% Algoritmo clássico para gerar eventos de um processo de Poisson
Y=-1          % inicializa o contador de eventos
SOMA=0  %  tempo acumulado
DO
    Y=Y+1              % incrementa o contador de eventos – Y=0 marca o primeiro evento
    U=RAND(0,1)        % gera número aleatório no intervalo (0,1)
    T=-LN(U)/ALFA   % gera variável aleatória exponencial com média 1/ALFA
                       % que é o intervalo de tempo médio entre eventos consecutivos
    SOMA=SOMA+T % acumula os tempos, formando o tempo em que ocorre
                       % o Y-ésimo evento
    TA[Y]=SOMA       % armazena o tempo acumulado no vetor TA[]
    PRINT(evento :Y  tempo :TA[Y])
WHILE(SOMA<TMAX)
% repita até sair do intervalo de tempo (0,TMAX)
% O último evento ocorre fora do intervalo (0,TMAX)
PRINT(ocorreram :Y eventos dentro do intervalo (0,TMAX) )
```

Exercício 3B: Um outro modo de gerar eventos de um processo de Poisson num intervalo de tempo *(a,b)* consiste em gerar pontos aleatórios uniformemente distribuídos em um intervalo *(c,d)* que contenha com bastante folga o intervalo *(a,b)*. Escreva um programa de computador que implemente essa idéia e compare com o algoritmo do exercício 3A. Porque não podemos usar *(c,d)=(a,b)*?

4. PILHAS E FILAS

Certos processos aleatórios podem ser descritos da seguinte forma: um servidor atende a pedidos que vão chegando à medida que o tempo passa. O servidor demora algum tempo, que é variável, para atender a cada pedido; por sua vez, os pedidos vão chegando de uma forma aleatória, ao longo do tempo. Vamos supor que a chegada de pedidos obedeça a um processo de Poisson: chegam em média α pedidos por minuto, e o intervalo de tempo entre pedidos consecutivos é uma variável aleatória exponencial com média $\frac{1}{\alpha}$. Se um pedido que chega for rapidamente atendido, o servidor fica esperando o próximo pedido chegar, que é por sua vez atendido rapidamente, e assim por diante. As coisas ficam mais interessantes quando o servidor recebe um novo pedido num momento em que ainda não acabou de processar o primeiro; nesse caso, o destino do novo pedido será um *buffer*, ou reservatório de espera, um local provisório em que o pedido fica aguardando até que o servidor possa atendê-lo. As coisas se complicam mais ainda se um terceiro pedido chega: o buffer já está ocupado com o segundo pedido, e o primeiro pedido ainda está sendo atendido. Existem duas saídas: o terceiro pedido entra no buffer, para ser atendido só depois que o segundo pedido for atendido (nesse caso, dizemos que o buffer é uma *fila*), ou então o terceiro pedido toma a frente do segundo pedido, para ser atendido logo que o primeiro pedido for finalizado pelo servidor (e dizemos então que o buffer é uma *pilha*). As idéias de fila e pilha serão formalmente descritas mais adiante.

Precisamos construir antes uma seqüência de eventos de dois tipos: *chegadas e atendimentos*. Uma seqüência de números reais em ordem crescente, gerada por um processo de Poisson com média μ_C (veja o exercício 3A) forma um vetor de tempos de chegadas TC[]. Analogamente, um processo de Poisson com média μ_A forma um vetor de tempos de atendimentos TA[]. A concatenação desses dois vetores (isto é, sua união e reordenação) forma um vetor de tempos de eventos TEMPO[], onde o tipo de evento (1 indica chegada e 0 indica atendimento) é registrado no vetor TIPOEV[].

```
% GERA OS TEMPOS DE CHEGADA
IC=-1
SOMAT=0
DO
    IC=IC+1
    U=RAND(0,1)
    T=-LN(U)/MC
    SOMA=SOMA+T
    TC[IC]=SOMA
WHILE(SOMA<TMAX)
```

```
% GERA OS TEMPOS DE ATENDIMENTO
IA=-1
SOMAT=0
DO
    IA=IA+1
    U=RAND(0,1)
    T=-LN(U)/MA
    SOMA=SOMA+T
    TA[IA]=SOMA
WHILE(SOMA<TMAX)

% CONCATENA OS VETORES DE CHEGADA E ATENDIMENTO
IAC=IA+IC
I1=0
I2=0
FOR(I=0;I<IAC;I++)
   IF(TA[I1]<TC[I2])
       TEMPO[I]=TA[I1]
       TIPOEV[I]=0 % ATENDIMENTO
       I1=I1+1
   ENDIF
   ELSE
       TEMPO[I]=TC[I2] % CHEGADA
       TIPOEV[I]=1
       I2=I2+1
   ENDELSE
ENDFOR
```

Por exemplo, seja $T_{\max} = 1.0$, $MC = \mu_C = 4.5$ e $MA = \mu_A = 3.2$, e suponha que sejam gerados os seguintes vetores $TC[]$ e $TA[]$:

TC
0.12
0.51
0.63
0.68
1.07

TA
0.15
0.42
0.74
0.88
1.06

Os vetores concatenados TEMPO[] e TIPOEV[] são (veja que IAC=8):

TEMPO	TIPOEV
0.12	1
0.15	0
0.42	0
0.51	1
0.63	1
0.68	1
0.74	0
0.88	0

Uma pilha *P* é um vetor munido de um número inteiro *i*, que chamamos de *topo* da pilha. Uma seqüência ordenada de valores reais $r_1, r_2, \ldots, r_M$, acompanhada de uma seqüência de instruções $s_1, s_2, \ldots, s_M$, onde cada s_k é *0* ou *1,* é *processada* pela pilha *P* de acordo com o seguinte algoritmo:

$i = 0$;
Para $k = 1$ até M faça {
  Se $s_k == 1$ faça { $P[i] = r_k$; $i = i + 1$; }
  Senão Se $i > 0$ faça { $i = i - 1$; }
}

É usual se chamar cada r_k de *item*, e cada s_k de *tipo de evento*. Quando $s_k = 1$ ocorre um *empilhamento*, e quando $s_k = 0$ ocorre um *desempilhamento*. O topo marca sempre a posição atualizada disponível da pilha para o próximo empilhamento. O topo também conta quantos itens estão na pilha. Se $i = 0$ quando tentamos fazer um desempilhamento, a pilha não se altera, e aparece um aviso de *underflow*.
Na prática podemos exigir ainda que a pilha tenha um tamanho máximo $TAM_{\max}$. Assim, um novo empilhamento será recusado quando o topo for igual a $TAM_{\max}$ (veja o algoritmo completo mais adiante).

```
% SIMULAÇÃO DE PILHA COM TAMANHO MÁXIMO TAMMAX
TOPO=0
FOR(I=0;I<IAC;I++)
   IF(TIPOEV[I]==1)  %CHEGADA
      IF(TOPO<TAMMAX)
         PILHA[TOPO]=TEMPO[I] %EMPILHA
         TOPO=TOPO+1 % AUMENTA O TAMANHO DA PILHA
         PRINT(CHEGADA:TEMPO[I]   TAMANHO:TOPO)
      ENDIF
      ELSE
         PRINT(OVERFLOW:TEMPO[I])
      ENDELSE
   ENDIF
   ELSE % ATENDIMENTO
     IF(TOPO>0)
         PRINT(ATENDIMENTO DE :PILHA[TOPO] NO :TEMPO[I])
         TOPO=TOPO-1 % DIMINUI O TAMANHO DA PILHA
      ENDIF
      ELSE
         PRINT(UNDERFLOW: TEMPO[I])
      ENDELSE
   ENDELSE
ENDFOR
```

Usando os vetores do exemplo anterior, e fazendo $TAM_{MAX} = 2$, a pilha seria processada assim:

posição 1						0.63	0.63		
posição 0		0.12			0.51	0.51	0.51	0.51	
comentários	Início: Pilha:vazia topo=0	cheg. t=0.12 topo=1	atend.0.12 t=0.15 topo=0	underflow t=0.42 topo=0	cheg. t=0.51 topo=1	cheg. t=0.63 topo=2	overflow t=0.68 topo=2	atend.0.63 t=0.74 topo=1	atend.0.51 t=0.88 topo=0

Exercício 4A: Acompanhe cada etapa do algoritmo de pilha para checar o processo acima. Verifique experimentalmente o que acontece a longo prazo quando $\mu_A = \mu_C$, $\mu_A > \mu_C$ e $\mu_A < \mu_C$.

Uma fila F é um vetor munido de dois números inteiros i e f, que chamamos respectivamente de *início* e *fim* da pilha. Uma seqüência ordenada de M valores reais $r_1, r_2, \ldots, r_M$, acompanhada de uma seqüência de instruções $s_1, s_2, \ldots, s_M$, onde cada s_k é *0* ou *1,* é *processada* pela fila F de acordo com o seguinte algoritmo:

i=0;
f=0;
Para $k = 1$ até M faça {
  Se $s_k == 1$ faça { $P[f] = r_k$; $f = f + 1$; }
  Senão Se $f > i$ faça { $i = i + 1$; }
}

É usual se chamar cada r_k de *item*, e cada s_k de *tipo de evento*. Quando $s_k = 1$ ocorre uma *chegada*, e quando $s_k = 0$ ocorre um *atendimento*. O fim f marca sempre a posição atualizada disponível da fila para se colocar o próximo item. O início i marca a posição do próximo item a ser atendido. O valor $f - i$ conta quantos itens estão na fila, e é chamado de *tamanho* da fila. Se $i = f$ quando tentamos fazer um atendimento, a fila não se altera, e aparece um aviso de *underflow*.

Na prática podemos exigir ainda que a fila tenha um tamanho máximo $TAM_{\max}$. Assim, uma nova chegada será recusada quando o fim f for igual a $TAM_{\max}$ (veja o algoritmo completo mais adiante).

Usando os vetores do exemplo anterior, e não colocando nenhum limite em TAM_{MAX}, a fila ficaria assim:

Posição4									
Posição3							0.68	0.68	0.68
Posição2						0.63	0.63	0.63	
Posição1					0.51	0.51	0.51		
Posição0		0.12							
Coment.	Início: fila:vazia início=0 fim=0 tam=0	cheg. t=0.12 início=0 fim=1 tam=1	atend.t=0.12 t=0.15 início=1 fim=1 tam=0	underflow t=0.42 início=1 fim=1 tam=0	cheg. t=0.51 início=1 fim=2 tam=1	cheg. t=0.63 início=1 fim=3 tam=2	cheg. t=0.68 início=1 fim=4 tam=3	atend.t=0.51 t=0.74 início=2 fim=4 tam=2	atend.t=0.63 t=0.88 início=3 fim=4 tam=1

Fica claro que as primeiras posições do vetor fila nunca mais serão utilizadas depois que o evento armazenado for atendido. Isso causa um desperdício de memória, uma vez que essas posições desocupadas nunca mais serão usadas. Uma outra estratégia mais eficiente pode ser feita utilizando-se uma fila em forma "circular", de tamanho máximo TAM_{MAX}: ao invés de incrementar uma unidade em *início* e *fim*, incremente uma unidade *módulo* TAM_{MAX}; isto é, o resto da divisão por TAM_{MAX}. Isso vai fazer com que apenas TAM_{MAX} posições de memória sejam ocupadas pela fila, que agora tem um formato de "anel". Vamos usar o seguinte algoritmo:

```
% SIMULAÇÃO DE FILA COM TAMANHO MÁXIMO TAMMAX
INICIO=0
FIM=0
TAM=0  % TAMANHO DA FILA
FOR(I=0;I<IAC;I++)
   IF(TIPOEV[I]==1)  %CHEGADA
      IF(TAM<TAMMAX)
         % COLOCA O TEMPO DE CHEGADA NO FIM DA FILA
         FILA[FIM]=TEMPO[I]
         FIM=(FIM+1) MOD TAMMAX  %DESLOCA A POSIÇÃO DO FIM DA FILA
         TAM=TAM+1  % AUMENTA O TAMANHO DA FILA
         PRINT(CHEGADA:TEMPO[I]   TAMANHO:TAM INICIO :INICIO FIM :FIM)
      ENDIF
      ELSE
         PRINT(OVERFLOW:TEMPO[I])
      ENDELSE
   ENDIF
   ELSE % ATENDIMENTO
     IF(TAM>0)
         % DESLOCA A POSIÇÃO DO INÍCIO DA FILA
         INICIO=(INICIO+1) MOD TAMMAX
         TAM=TAM-1 % DIMINUI O TAMANHO DA FILA
         PRINT(ATENDIMENTO DE :PILHA[FIM]  EM :TEMPO[I]  :INICIO  :FIM)
      ENDIF
      ELSE
         PRINT(UNDERFLOW: TEMPO[I])
      ENDELSE
   ENDELSE
ENDFOR
```

Usando os vetores do exemplo anterior, e fazendo neste exemplo $TAM_{MAX} = 2$, a fila ficaria assim:

Posição1					0.51	0.51	0.51		
Posição0		0.12				0.63	0.63	0.63	
Coment.	Início: fila:vazia início=0 fim=0 tam=0	cheg. t=0.12 início=0 fim=1 tam=1	atend.t=0.12 t=0.15 início=1 fim=1 tam=0	underflow t=0.42 início=1 fim=1 tam=0	cheg.t=0.51 início=1 fim=2mod2=0 tam=1	cheg. t=0.63 início=1 fim=1 tam=2	overflow t=0.68 início=1 fim=1 tam=2	atend.0.51 t=0.74 início=2mod2=0 fim=1 tam=1	atend.0.63 t=0.88 início=1 fim=1 tam=0

Exercício 4B: Acompanhe cada etapa do algoritmo de fila para checar o processo acima.

Exercício 4C: Implemente o algoritmo de fila. Verifique experimentalmente o que acontece a longo prazo quando $\mu_A = \mu_C$, $\mu_A > \mu_C$ e $\mu_A < \mu_C$. Calcule experimentalmente o valor médio do tamanho da fila e o número de underflows e overflows em cada uma dessas situações.

Exercício 4D: Repita o exercício 4C para pilhas.

Exercício 4E: Modifique os algoritmos de pilha e fila para:
a)Dois atendentes operando simultaneamente para a mesma pilha;
b)Duas pilhas sendo atendidas simultaneamente por um atendente;
c)Duas pilhas atendidas simultaneamente por dois atendentes.
Dica: Use índices diferentes no vetor TIPOEV[] para cada pilha e/ou atendente.

REFERÊNCIAS

1. Sheldon M. Ross – Simulation, 3rd Edition – Academic Press, 2002
2. Sheldon M. Ross- A First Course in Probability – Prentice Hall, 2002
3. Paul L. Meyer – Probabilidade – Aplicações à Estatística – LTC, 1987
4. Sabino J. Ferreira, Luiz Duczmal, Lupércio F. Bessegato, Marcos A. Santos – Introdução às Técnicas de Simulação em Estatística – RTE 02/2003 – Departamento de Estatística da UFMG.

CAPÍTULO 2

Variáveis aleatórias e testes de permutação

Com algoritmos que envolvem *permutações* e avanços recentes em técnicas computacionais é possível verificar, por exemplo, se há indícios de relação entre variáveis aleatórias baseando-se apenas em um conjunto de valores observados. Ou verificar, por exemplo, o grau de aleatoriedade de uma variável aleatória que evolui no tempo.

Esta é uma vasta e atual área de pesquisa. No meio estatístico este campo de estudo é conhecido como "*testes de hipóteses*". Para a realização de um teste é necessário definir o tipo de problema e as técnicas que serão utilizadas. Dentre os testes utilizados, há uma categoria denominada "*testes de permutação*". Estes testes são realizados com programas de computadores que utilizam permutações para simular o experimento em estudo. Apresentaremos alguns exemplos ilustrativos das principais idéias envolvidas neste tipo de teste.

A situação básica é a seguinte: dispomos de um conjunto de valores e há a suspeita de que estes valores se se relacionam de alguma maneira, distribuídos não completamente ao acaso. Com o uso de algumas técnicas e da idéia de permutação é possível "checar" esta possibilidade. A seguir apresentaremos alguns exemplos.

1 -Tendência em série temporal

A tabela 1 apresenta as temperaturas médias observadas para o mês de julho, em Celsius, período 1976-1994, na ilha Rei George, situada na Antártica (valores aproximados baseando em Ferron F.A., Simões J.C.; Aquino F.E (2001)). O gráfico 1 mostra a sequência de temperaturas observadas, em função do ano, de acordo com os dados da tabela.

.

TABELA 1

Ano	temp	Ano	temp
1976	-10,1	1986	- 6,2
1977	- 8,0	1987	-14,0
1978	- 11,0	1988	- 9,5
1979	- 4,1	1989	0,0
1980	-12,0	1990	- 6,0
1981	- 4,0	1991	-10,0
1982	- 7,2	1992	- 6,8
1983	- 5,0	1993	- 4,0
1984	- 4,1	1994	-14,0
1985	- 3,1		

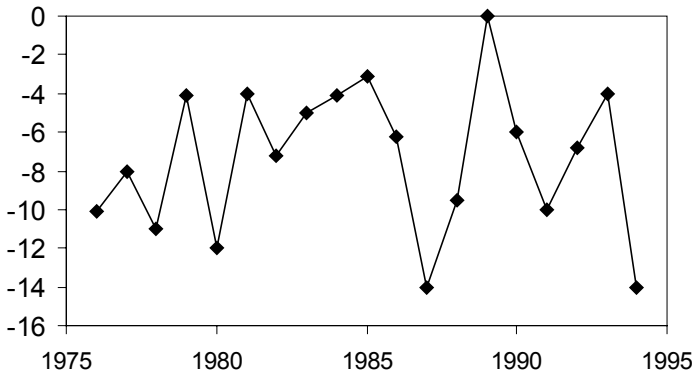


Gráfico 1 - Sequência de temperaturas da tabela I

Há a suspeita de que a temperatura desta ilha esteja em declínio ao longo das últimas décadas. Se de fato este fenômeno ocorre localmente, as temperaturas médias observadas devem apresentar uma tendência de queda ao longo do período, além dos efeitos aleatórios.

Testando hipóteses

Seja t o ano e X a temperatura. No exemplo considerado, há várias perguntas que poderiam ser feitas em relação ao comportamento da temperatura no período. Nos restringiremos a uma questão simples: *há evidências na tabela I de que a temperatura X é independente de t ou, apesar da aleatoriedade presente, há uma tendência de queda da temperatura ao longo dos anos*? Neste último caso é provável que exista uma relação entre X e t. Gostaríamos então de nos decidir sobre uma das seguintes hipóteses, que podemos formular para o comportamento da variável temperatura X e o tempo t:

- **Hipótese:** X é independente de t.
- **Hipótese alternativa**[1]**:** X e t estão relacionados da seguinte maneira:

$$X_i = g(t_i) + \varepsilon_i; \qquad (1)$$

onde $\{\varepsilon_i\}$ são variáveis aleatórias independentes com média zero e g é alguma função monótona decrescente, no sentido em que, se $t_1 < t_2$ então $g(t_1) > g(t_2)$.

Observe que, na hipótese alternativa, estamos propondo um *modelo de dependência* da variável aleatória X em relação ao tempo (não aleatório). É um teste portanto para verificar se X é independente de t *versus* X é estocasticamente decrescente em t. Observe ainda que, na hipótese alternativa, estamos assumindo que todo o efeito da aleatoriedade presente nas temperaturas é devido somente às variáveis aleatórias ε_i.

[1] Esta maneira de colocar o problema é típico da técnica conhecida como "teste de hipóteses". Esta técnica é útil para tomar decisões e a seguinte regra deve ser seguida: quando uma hipótese é formulada (denominada de "hipótese nula") é necessário especificar *qual hipótese se contrapõe à hipótese* nula (denominada "hipótese alternativa"). Uma decisão deve ser tomada comparando-se apenas estas duas hipóteses.

O próximo passo então é verificar se nos dados da Tabela I há indícios que favorecem a primeira hipótese ou a segunda. Caso a primeira hipótese não seja verificada, aceitaremos a segunda. Neste último caso diremos que há evidências de que a distribuição de probabilidades para a temperatura média no mês considerado apresenta uma tendência de queda ao longo dos anos, no período estudado.

Embaralhando os dados

Uma maneira de verificar se há uma tendência na sequência das temperaturas é "embaralhar" a sequência original, permutando os valores ($x_0, x_1, x_2 ... x_n$). A permutação é feita da seguinte maneira: para a sequência de pares *(t,x)* iguais a (1,3) (2,7) (3,9), por exemplo, uma seqüência permutada possível é (1,9) (2,3) (3,7).

Para a sequência de temperatura observada na tabela I, a idéia básica é gerar, em um programa de computador, *n sequências permutando-se os valores da sequência original*. Em seguida comparamos as sequências permutadas com sequência original observada (veja o esquema da figura 1). Se de fato há uma tendência de queda ao longo dos anos na sequência original, com pouca probabilidade isto se verificará nas sequências permutadas. Isto porque, para qualquer tipo de ordem presente na sequência original, é provável que esta ordem desapareça ou seja alterada devido às permutações.

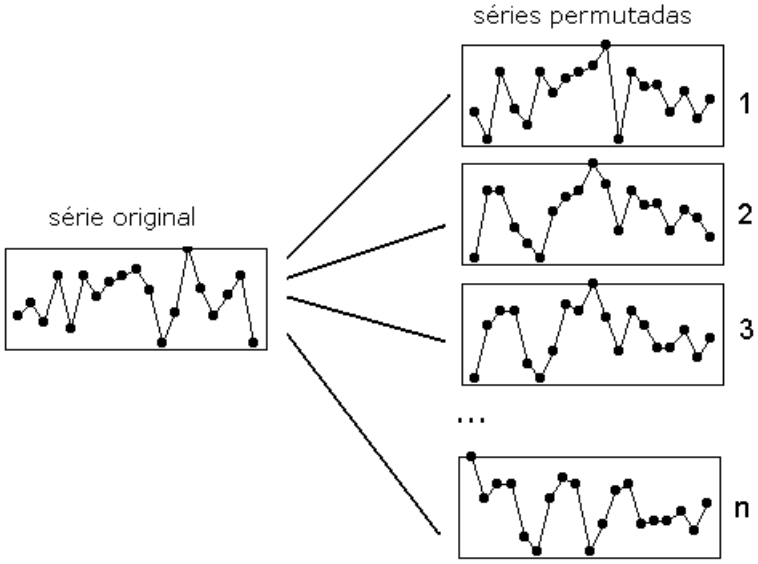


Figura 1

Como comparar sequências?

Uma comparação da sequência original com as sequências permutadas pode ser feita da seguinte maneira:

1. Renumere os valores de *t* para $t_1=0, t_2=1, t_3=2...$ ao invés de datas *(1974,1975..etc.)*.

2. Para cada sequência permutada $\{(t_1, x_1), (t_2, x_2)...(t_m, x_m)\}$ calcule a soma [Wald and Wolfowitz (1943)]:

$$w = \sum_{i=1}^{m} t_i x_i \qquad (2)$$

Esta soma é conhecida como *correlação de Pitman.* Compare os valores *w* obtidos para as séries permutadas com o obtido para sequência original, não permutada. Observe que, se de fato há uma tendência de queda dos valores de *X* ao longo dos anos, *o valor de w deve ser mais extremo do que os valores observado nas sequencias permutadas* (verifique isto tomando como exemplo uma sequencia pequena).

Portanto, para verificar se há evidência de tendência de queda nas temperaturas observadas podemos realizar um teste de permutação com o seguinte algoritmo:

AlGORITMO 1 – TESTE DE PERMUTAÇÃO PARA
TENDÊNCIA EM SÉRIE TEMPORAL

$\mathbf{X} = (x_0, x_1...x_n)$ vetor com *n* observações da série temporal observada;

m = número de permutações (usualmente > 100);

P1: INPUT $\mathbf{X} = (x_0, x_1...x_n)$

P2: INPUT *m*

P3 $w^* = \sum_{i=1}^{n} t_i x_i$

P4: $N = 0$

P5 FOR $k = 1$ TO m

P6: Obtenha a série permutada $\mathbf{X}_k = (x_{(0)}, x_{(1)}...x_{(n)})$ a partir de **X**

P7: Use $\mathbf{X}_k$ para calcular $w_k = \sum_{i=1}^{n} t_i x_i$

P8: **se** $w_k \leq w^*$ **então** $N = N + 1$

P9: $r = 100N/m$. Comentário: r é a porcentagem de valores w menores ou iguais ao valor observado w^*.

Valor muito baixo de *r* é evidência de que há tendência decrescente em **X.**

Observações:

1. Se o teste for para tendência *crescente*, altere

P8: **se** $w_k \geq w^*$ **então** $N = N + 1$.

2. Usualmente considera-se que há evidência se $r < 10\%$; evidência muito forte se $r < 1\%$.
3. É preciso garantir que as permutações geradas sejam independentes e geradas com a mesma probabilidade (isto é, a probabilidade do algoritmo de permutação gerar uma sequência permutada $\{x_i\}$ é a mesma de gerar qualquer outra sequência permutada).

Exercício 1: efetue um teste de permutação para as temperaturas da Ilha Rei Jorge (tabela 1). Há alguma evidência de tendência decrescente?

Exemplo de implementação

A seguir um código para o algoritmo 1, implementado na linguagem *S+ (softawres "Splus" ou "R"):*

```
# Teste de permutação para tendência - Splus
# No vetor X deve estar a série observada

# entrada dos dados da tabela 1 (de um modo abreviado):
X<-c(10.1,8,11,4.1,12,4,7.2,5,4.1,3.1,6.2,14,9.5,0,6,10,6.8,4,14)
X<- -X  # troca o sinal


#inicio do teste

m<-1000                              # numero de permutacões
 w<-0                                # valor inicial de w
 N<-0                                # contador
 t<-c(1:length(X))                   # vetor de indices 1,2...n
 w1<-sum(X*t)                        # valor observado na série original
 for(i in 1:m) {                     # inicio das m permutações
      y<-sample(X, replace=F)        # y é a série original permutada
      w[i]<-sum(y*t)                 # valor calculado na seq i permutada
      if (w[i]<=w1) N<-N+1;          # N é o número de valores no vetor w
                                     #  iguais ou abaixo ao valor observado
}
r<-N*100/m                           # porcentagem
```

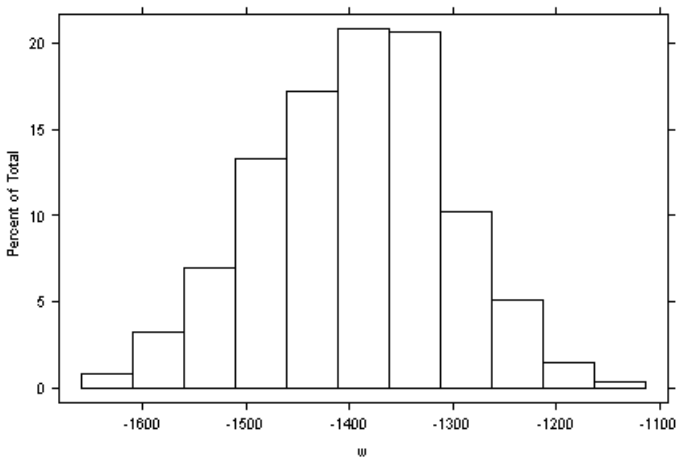


Fig 2 - Histograma para os valores de *w* (expressão (2)), com 1000 permutações da série original.

Na Fig 2 está o histograma obtido para 1000 permutações da série original. O valor de *w* obtido na série original foi de –1381,1 (compare este valor com os obtidos no histograma). O valor de r foi de 53,4%. Não há, portanto, evidência a favor da hipótese de tendência decrescente nos dados observados.

2 – Correlação em série temporal

No exemplo anterior o interesse é verificar se há evidências a favor ou contra a existência de tendência em uma série de valores no tempo. Há situações em que podemos suspeitar de uma relação entre X_t e t como em (1), porém considerando o seguinte modelo para a hipótese alternativa: X_t tende sempre a estar "próximo" de X_{t-1} quando $G(X_t)$ é "próximo" a $G(X_{t-1})$. Em outras palavras, gostaríamos de testar a hipótese de que a distribuição de X é independente de t contra a hipótese alternativa de que X depende de t através de alguma função desconhecida G. Como no exemplo anterior, estamos considerando aqui X(t) variável aleatória e t um índice (no caso, o tempo). Observe ainda que desta vez não estamos supondo a função G necessariamente monótona crescente ou decrescente.

Podemos efetuar o seguinte teste [Good, 1993]:

- **Hipótese:** *X* é independente de *t.*
- **Hipótese alternativa:** $X_i = G(t_i) + \varepsilon_i$, $i = 1...n$, onde $\{\varepsilon_i\}$ são variáveis aleatórias independentes com média zero e *g* uma função contínua com a propriedade de que, se t_1 "próximo" a t_2, então o valor de $G(t_1)$ deve ser "próximo" a $G(t_2)$.

Como verificar se há ou não evidências em favor da hipótese alternativa? A idéia neste caso é a mesma do exemplo anterior: permutamos a série observada para testar as hipóteses

formuladas. Neste caso necessitamos apenas de uma maneira diferente para comparar as séries, já que a expressão utilizada em (2) é útil para verificar a possível existência de tendências mas não propriamente a existência de uma relação de dependência como a descrita acima. Uma maneira mais apropriada para comparar as séries neste caso é ordenar os valores $t_1, t_2,...t_n$ e calcular o coeficiente

$$M = \sum_{i=1}^{n-1} (X_i - X_{i+1})^2 \qquad (3)$$

A variável *M* é denominada "coeficiente de correlação serial". Porque utilizar este coeficiente? Observe que, *se os valores de* X_i *e* X_{i+1} *tendem a estarem próximos na série observada, a soma das diferenças* $(X_i - X_{i+1})^2$ *na série observada deve ser menor do que a soma obtida para um grande número de séries permutadas.* Esta é a idéia básica do teste de permutação para este caso. Se de fato o valor de *M* para a série observada for muito diferente dos valores obtidos nas séries permutadas, rejeitamos a hipótese de independência. Neste caso assumimos então que há dependência, como descrito na hipótese alternativa.

AlGORITMO 2 – TESTE DE PERMUTAÇÃO PARA CORRELAÇÃO SERIAL

$\mathbf{X} = (x_0, x_1...x_n)$ vetor com *n* observações da série temporal observada;

m = número de permutações (usualmente > 100);

P1: INPUT $\mathbf{X} = (x_0, x_1...x_n)$

P2: INPUT *m*

P3 $M^* = \sum_{i=1}^{n-1} (X_i - X_{i+1})^2$

P4: $N = 0$

P5 FOR $k = 1$ TO m

P6: Obtenha a série permutada $\mathbf{X}_k = (x_{(0)}, x_{(1)}...x_{(n)})$ a partir de **X**

P7: Use $\mathbf{X}_k$ para calcular $M_k = \sum_{i=1}^{n-1} (X_i - X_{i+1})^2$

P8: **se** $M_k \le M^*$ **então** $N = N + 1$

P9: $r = 100N/m$

Valor muito baixo para *r* é evidência contra a hipótese nula e a favor da hipótese alternativa, ou seja, a variável X depende de t através de alguma função desconhecida G .

Exercício 2 : efetue um teste de permutação para correlação serial para as temperaturas da tabela 1.

3 - Comparação entre dois grupos

Tabela 2

cidade	antes	depois
Atlanta	95	123
Boston	151	160
Chicago	192	180
Denver	71	93
Los Angeles	86	99
Miami	215	193
New Orleans	254	311
New York	123	121
Philadelphia	97	131
St. Louis	153	169

Exemplo 2 - Na tabela 2 estão os resultados de uma pesquisa realizada para a Coca-Cola Company sobre o efeito de uma campanha publicitária realizada nos EUA, com o slogan *Twice the Cola, twice the fun*[2]. Para testar se a campanha foi eficaz, foram entrevistados 500 indivíduos escolhidos aleatoriamente em dez cidades americanas. A cada pessoa foi pedido para citar cinco marcas de refrigerante. A tabela mostra o número de pessoas que citaram Coca Cola, antes e depois da campanha.

Este é um exemplo aonde temos um conjunto de observações $(x_1, x_2...x_n)$ de um vetor aleatório $(X_1, X_2...X_n)$ no primeiro grupo e um conjunto de observações $(y_1, y_2...y_n)$ de um vetor aleatório $(Y_1, Y_2...Y_n)$ no segundo grupo. Suponha que todas as variáveis $(X_1, X_2...X_n)$ sejam independentes e com mesma função distribuição de probabilidade F_1. De modo similar, considere que os elementos $(Y_1, Y_{2...}Y_n)$ são independentes e com a mesma função distribuição F_2.

Um ponto de interesse no exemplo da tabela 2 é verificar se de fato os grupos possuem características diferentes ou se trata-se de observações de uma mesma variável aleatória, apenas divididas em dois grupos. Em outras palavras, gostaríamos de verificar se há evidências de que $F_1(x)=F_2(x)$.

Em termos de teste de permutação, o problema pode ser colocado da seguinte maneira: se não há de fato uma diferença de comportamento das pessoas nos dois grupos

[2] Computational Techiniques in Statistics - lecture notes, School of Mathematical Sciences, University of London. *http://www.maths.qmw.ac.uk/~bb/CTS_Chapter2_Students.pdf*

então não deve haver diferença significativa quando comparamos grupos formados, alocando-se ao acaso todas as vinte observações em dois grupos de dez.

Então podemos propor o seguinte teste de permutação para comparação de dois grupos:

ORITMO 3 – TESTE DE PERMUTAÇÃO PARA
DIFERENÇA DAS MÉDIAS ENTRE DOIS GRUPOS

Dado os grupos 1 e 2 com *n* observações cada,

a) Encontre a diferença das médias observadas: $D^* = m_x - m_{y;}$
b) Forme uma *tabela com elementos permutados:* dois grupos de *n* elementos cada, escolhendo sem reposição os elementos de cada grupo ao acaso, dentre todas as observações;
c) Repita (b) um grande número de vezes e para cada *tabela i permutada* calcule a diferença entre as médias D_i;
d) Verifique se os valores $\{D_i\}$ obtidos são tipicamente diferentes de D^*. Uma maneira de verificar isto é construir um histograma para os valores obtidos nas permutações com o valor observado na tabela 1. Outro modo é calcular diretamente a porcentagem de valores da diferença que são maiores ou iguais ao valor observado (ou menores e iguais, se o valor observado for negativo), como nos exemplos anteriores (cálculo de r).

Exercício 3: Efetue um teste de permutação para os dados da tabela 2. Há indícios de que a campanha publicitária foi eficaz?

Algumas observações:

1 – Neste exemplo estamos considerando que os dados não são *emparelhados*. Isto significa que em termos práticos não estamos considerando a cidade, na análise do efeito da campanha nem que os indivíduos entrevistados em cada grupo são os mesmos. Desconsiderar a cidade faz sentido neste caso visto que a campanha foi nacional. No entanto há situações em que devemos considerar que os pares (x_i, y_i) não podem ser descaracterizados. Neste caso o teste de permutação deve ser um *teste para dados emparelhados* e é realizado de modo um pouco diferente [ver Philips(1993)] .

2 - Ao invés de *média* é possível utilizar neste mesmo teste a *mediana* ou a *estatística-t.* Esta escolha não é relevante para o teste [cf Manly(1997)].

3 – Observe que não fizemos nenhuma suposição sobre as funções F_1 e F_2. Em linguagem mais técnica este tipo de teste é dito *não-paramétrico*.

Exercício 4 - A velocidade da luz e o experimento de Michelson[3]. Em 1879, A. A. Michelson realizou 100 determinações da velocidade da luz no ar. As medidas foram agrupadas em cinco tentativas de 20 observações cada. Os números estão em Km/s e a velocidade estimada é obtida somando-se 299.000 a cada um. O valor atualmente aceito para a "verdadeira" velocidade da luz no vácuo é 299.792,5 Km/s. A tabela 3 mostra os valores obtidos para a velocidade da luz na segunda, quarta e quinta tentativas.

Os dados de Michelson apresentam evidências de possíveis problemas experimentais: os valores observados parecem possuir diferentes distribuições de probabilidades em cada tentativa, com variâncias e médias diferentes.

a) Implemente o algoritmo 3.
b) Teste se as médias para os grupos "tentativa 4" e "tentativa 5" são significativamente diferentes.
c) Repita para a segunda e quinta tentativa.

Tabela 3 – Experimento de Michelson para determinação de v. Velocidade da luz = 299.000 + v

tentativa 1		tentativa 4		tentativa 5	
850	1000	890	910	870	890
740	980	810	920	870	840
900	930	810	890	810	780
1070	650	820	860	740	810
930	760	800	880	810	760
850	810	770	720	940	810
950	1000	760	840	950	790
980	1000	740	850	800	810
980	960	750	850	810	820
880	960	760	780	870	850

[3] Dados extraídos de "The Data and Story Library" - http://lib.stat.cmu.edu/DASL

4 – Regressão linear simples

O modelo de regressão linear simples é utilizado quando temos *n* pares $(x_1,y_1)(x_2,y_2)...(x_n,y_n)$ de observações de duas variáveis aleatórias (X,Y). A suposição é a de que existe uma relação linear da forma:

$$Y = \alpha + \beta x + \varepsilon \qquad (4)$$

onde α, β são constantes e somente os valores ε são variáveis aleatórias independentes, com média zero e mesma distribuição para cada observação. Este é um problema clássico em ciência e tecnologia. Os valores de α, β são *estimados* por a, b respectivamente, dado pelas expressões:

$$\begin{aligned} a &= \bar{y} - b\,\bar{x}, \\ b &= S_{xy} / S_{xx} \end{aligned} \qquad (5)$$

onde

$$\bar{x} = \sum x_i / n\,, \quad \bar{y} = \sum y_i / n\,, \quad S_{xx} = \sum (x_i - \bar{x})^2 \quad \text{e} \quad S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})\,.$$

Estas expressões são obtidas minimizando-se a soma dos quadrados $\sum (y_i - a - bx_i)^2$. Os símbolos $\bar{x}, \bar{y}$ nas expressões dadas são simplesmente as médias dos valores *x* e *y*, respectivamente.

Observe que no modelo proposto em (4), ocorre independência entre as variáveis aleatórias X e Y quando $\beta = 0$, pois neste caso os valores de Y são devido apenas a flutuações aleatórias em torno de valores independentes de X. Há situações práticas aonde o que se deseja saber é justamente se há evidências de que as variáveis sejam independentes. Isso pode ser realizado portanto a partir de testes para verificar se há evidências a favor da hipótese $\beta = 0$.

Há testes clássicos para verificar a partir dos dados se há evidências de que constante β seja zero. Um teste de permutação tambem pode ser utilizado; como nos exemplos anteriores, a idéia é considerar que *se de fato as variáveis* (X,Y) *são independentes* ($\beta = 0$)*, então todas as amostras formadas permutando-se os pares* (x_i, y_j)*, i,j=1,2...n da amostra original são igualmente prováveis de ocorrerem*[4].

Um teste de permutação para independência das variáveis no modelo de regressão pode ser efetuado permutando-se os pares dos dados originais ao acaso. Em seguida calculamos o valor de *b* para cada conjunto de dados permutados e comparamos com o valor obtido no conjunto original[5].

---

[4] Uma discussão mais detalhada da justificativa para um teste de permutação no modelo de regressão pode ser vista em Mainly (1997).

[5] Observe que o termo S_{xx} que aparece no cálculo de *b* será o mesmo para todas as permutações; portanto é suficiente estimar em cada caso apenas o valor S_{xy} ao invés de *b*.

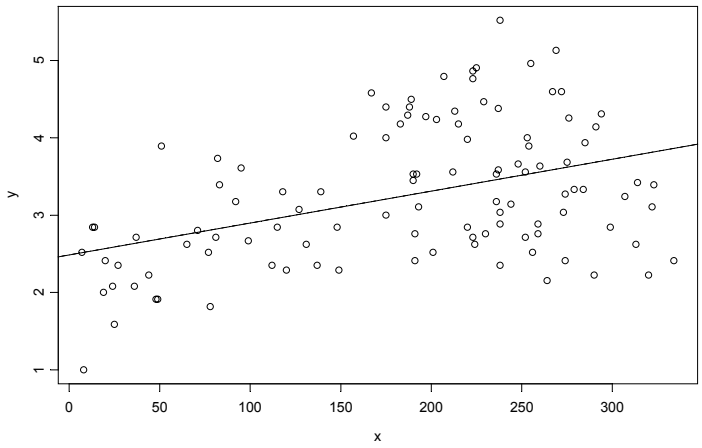


Fig 3

Exemplo 4 - Atmosfera em Nova York[6]. Os dados da figura 3 são observações durante 111 dias consecutivos para a concentração superficial de ozônio (ppm) e a radiação solar na superfície. No eixo-y está a concentração de ozônio. A reta indicada corresponde à reta ajustada pelo modelo de regressão simples.

A reta ajustada na figura 3 parece indicar uma influência da radiação solar no teor de ozônio na atmosfera da cidade. É necessário verificar até que ponto esta tendência é confirmada em um teste estatístico. A Fig 4 mostra 50 retas obtidas pelas mesmas expressões em (4) porém permutando-se os os pares (x,y) aleatoriamente. A reta tracejada é a reta obtida com os dados originais. Vemos que as retas obtidas com as permutações parecem indicar que a reta original é significativamente "diferente", com valores y sempre mais extremos. Isto é forte evidência a favor de real dependência nos dados.

A Fig 5 é o histograma para valores de *b* obtido em 1000 permutações. O valor que corresponde aos dados originais é b=0,00412. Este valor está acima de todos os valores obtidos com as permutações. Isto é evidência forte da dependência do teor de ozônio com a radiação solar[7].

[6] Estes dados estão no exemplo "exair" que acompanha o software "Splus 2000" da MathSoft Inc, extraídos de John M. Chambers and Trevor J. Hastie, (eds.) Statistical Models in S, Wadsworth and Brooks, Pacific Grove, CA 1992, pg. 348.

[7] Do ponto de vista do teste realizado a conclusão é correta, mas de um modo geral, em testes para independência, sempre é necessário verificar se há outras variáveis envolvidas que podem explicar a dependência encontrada. Por exemplo, para explicar o teor de ozônio talvez seja necessário considear outros fatores (clima, emissão de poluentes, etc).

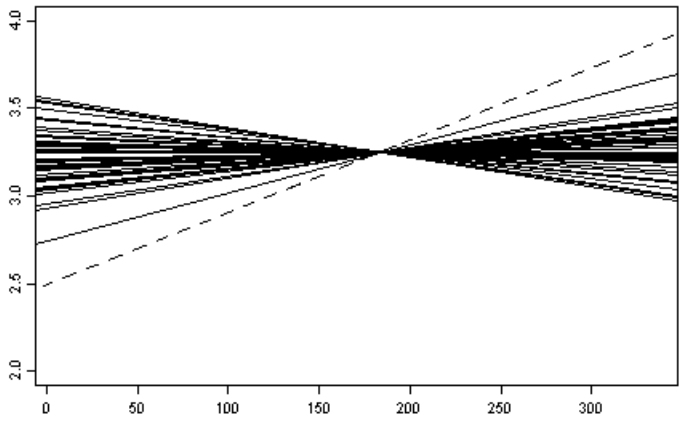


Fig 4

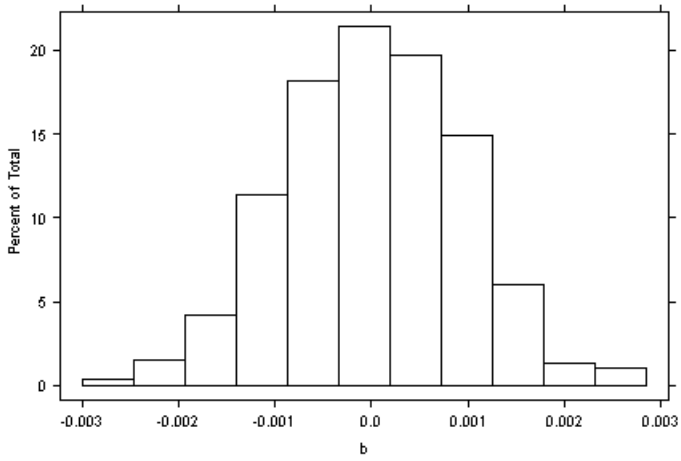


Fig 5

Para ler mais: os capítulos introdutórios do livro escrito por Bryan Manly (1997) formam uma ótima introdução ao assunto. Em Phillip (1993) há uma apresentação mais técnica de testes de permutação para várias situações.

5 – Notas sobre o uso de Testes de Permutação

Em estatística existem várias maneiras de testar hipóteses, dependendo do problema estudado. Em linhas gerais, podemos citar os testes tipo paramétricos, permutação, bootstrap e os tipos "rank". O objetivo deste texto foi o de introduzir as idéias básicas envolvidas em um teste de permutação e teste de hipótese em geral.

Em relação aos testes de permutação é interessante observar o seguinte:

1 – *Teste de permutação e outros testes* - Para os problemas analisados (séries temporais, regressão, comparação entre grupos) existem testes paramétricos clássicos já bem estudados para as questões como as apresentadas neste texto. Para o leitor familiarizado com os testes paramétricos e outros, pode surgir a questão de qual o melhor teste. A resposta à esta pergunta é que isto depende do problema em estudo. Em linhas gerais, depende das hipóteses de interesse e de alguns aspectos técnicos, como o nível de confiança e o poder desejado para o teste. Quando as amostras são grandes, decisões baseadas em testes paramétricos usualmente concordam com os correspondentes testes de permutação. Uma discussão sobre este ponto e uma comparação entre vários tipos de teste com os testes de permutação pode ser encontrada em Good (1993).

2 – *Tamanho da amostra* – Para amostras grandes um problema que ocorre é que o número de permutações possíveis torna-se muito grande. Gerar todas as permutações para determinar o nível de significância de uma estatística de teste pode vir a ser uma tarefa difícil e até mesmo inviável computacionalmente. Uma maneira de contornar este problema é utilizar o *método de Monte Carlo*. Neste método a estatística de teste é computada para *amostras* de permutações e os percentis são obtidos da distribuição resultante, ao invés da distribuição de todas as permutações. As amostras de permutações são escolhidas de maneira uniforme. O nível de significância p' obtido desta maneira pode diferir do nível de significância p de um teste baseado no conjunto de todas as permutações. No entanto o nível de significância p' é um estimador consistente de p, com variância $p(1-p)/N$, onde N é o número de permutações consideradas no Monte Carlo. Outras técnicas para contornar o problema de amostras grandes existem, como o *importance sampling*. Mais detalhes sobre o comportamento assintótico de p' e o importance sampling podem ser encontrados em [3].

Referências

1. Bogacka, B. Computational Techniques in Statistics - lecture notes, School of Mathematical Sciences, University of London. *http://www.maths.qmw.ac.uk/~bb/CTS_Chapter2_Students.pdf*

2. Ferron F.A., Simões J.C.; Aquino F.E (2001). Atmospheric temperature time series for King George Island. *Revista do Departamento de Geografia,* Universidade de São Paulo n 14, pp 25-32.

3. Good, Phillip *Permutation Tests* (1993) Springer Series in Statistics – Springer-Verlag, New York, Inc.

4. Manly, Bryan F. J.(1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology* Chapman & Hall – London, UK.

5. Wald A; J. Wolfowitz (1943) An Exact test for randomness in the nonparametric case based on serial correlation. *Annal Math Statist*; 14,378-388