Parte X

Teste de geradores de números aleatórios

- Os usuários de uma simulação devem se certificar de que os números fornecidos pelo gerador de números aleatórios são suficientemente aleatórios.
- O primeiro passo é plotar o histograma e as distribuições acumuladas.
  - Este é um teste rápido que permite identificar erros grosseiros de implementação.
  - Para construir um histograma, particionamos a região de valores possíveis em k intervalos consecutivos. Em seguida verificamos a proporção de valores que ocorrem em cada intervalo, e verificamos se esta proporção se aproxima da esperada pela distribuição subjascente.
- Em seguida, deve-se realizar todos os testes estatísticos possíveis.
  - Descreveremos cada um destes testes nas próximas seções.
  - Passar em cada teste é uma condição necessária, mas não suficiente. Ou seja, se o teste falha podemos assumir que o gerador é ruim. Porém, passar no teste não é uma garantia que ele é bom, pois ele pode ele pode falhar
    * no teste seguinte,
    * ou no mesmo teste quando utilizamos outra semente ou outra parte do ciclo.
- Com o surgimento de novos testes, geradores que eram considerados bons deixaram de ser.
- Embora os testes apresentados aqui são para distribuições uniformes, muitos destes testes podem ser adaptados para outras distribuições.
  - A aleatoriedade dos geradores de números distribuídos uniformemente é mais importante, pois estes números são utilizados para produzir geradores de números que respeitam outras distribuições.
- Os testes podem ser classificados como empíricos e teóricos.
  - Os testes empíricos consideram o gerador como uma caixa preta. Os números que saem do gerador são avaliados com técnicas estatísticas.
  - Por outro lado, os testes teóricos avaliam propriedades matemáticas desejáveis nos geradores. Para isso, estes métodos levam em conta a forma como cada gerador produz os números.
- Os testes teóricos são sofisticados e envolvem ferramentas matemáticas mais complexas.
  - Eles indicam a qualidade de um gerador (de acordo com algum critério matemático) sem exigir a geração de números, mas avaliando sua definição. Ex.: nos geradores congruente lineares, eles avaliam as constantes a, c e m.
  - Estes testes também diferem dos empíricos por avaliar todo o ciclo do gerador, e não apenas um segmento dos números gerados.
  - Com excessão do teste espectral, todos os testes apresentados aqui são empíricos.

38 Testes de Ajuste à Distribuição

38.1 Teste Chi-Quadrado

- Este é o teste mais utilizado para determinar se os dados observados atendem a uma determinada distruição.
- Os passos são:
  1. Preparamos um histograma com k intervalos, obtendo assim o percentual de ocorrência o_i em cada intervalo $i = 1, \ldots, k$.
  2. Com base na distruibuição que estamos testando, determinamos o percentual de ocorrências esperado e_i para cada intervalo $i = 1, \ldots, k$.
     - No caso da distribuição uniforme, $e_i = 1/k$ para todo i.
  3. Calculamos a estatística
     $$D = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$
     que tem distruibuição chi-quadrada com $k-1$ graus de liberdade.
  4. Com significância α, se D é menor que $\chi^2_{1-\alpha,k-1}$ (fornecido em softwares ou tabelas), não podemos rejeitar a hipótese de que os números foram originados da distribuição.
- Ex.: determinar (com significância $\alpha = 0.10$) se o gerador $x_n = (125x_{n-1}+1) \bmod 2^{12}$, $x_0 = 1$, fornece números de acordo com uma distribuição uniforme entre 0 e 1.
  - 1000 números foram gerados, e um histograma com 10 intervalo foi criado (Tabela 27.1, Jain). Esperamos 100 observações em cada intervalo (10%).
  - Resultou em $D = 10.38$. Como $\chi^2_{0.9,9} = 14.68$, não podemos rejeitar a hipotése de que os números gerados são uniformes.

```
n=1000; m=2^12; a=125; b=1; x=zeros(1,n);
x(1)=1;
for i=2:n; x(i)=mod(a*x(i-1)+b,m); end;
x=x/m; hist(x); grid on; figure; autocorr(x)

[h pv]=chi2gof(x,'cdf',@(z)unifcdf(z),
   'nparams',0, 'alpha',0.05)
```

- Como o e_i aparece no denominador, erros em intervalos com e_i menores têm mais peso na estatística D.
  - Portanto, o teste funciona melhor se os intervalos do histograma são escolhidos de tal forma que os e_i's sejam iguais.
  - Uma forma aproximada de resolver é agrupar cada intervalo com e_i pequeno com algum intervalo vizinho.
  - Note que no caso da distribuição uniforme, basta utilizar um histograma com intervalos de mesmo tamanho.
- Se r parâmetros da distribuição são estimados utilizando a mesma amostra, o número de graus de liberdades cai de $k-1$ para $k-1-r$.
  - Ex.: se suspeitamos que os dados formam uma normal, e estimamos a médio e o desvio padrão com base na amostra, devemos subtrair 2 graus de liberdade.
  - No caso da distribuição uniforme entre 0 e 1 nenhum parâmetro precisa ser estimado.
- Este teste é mais indicado para distribuições discretas.
  - No caso de distribuições contínuas, o teste chi-quadrado é apenas uma aproximação. E portanto exige mais observações.
  - Pois o teste agrupa (discretiza) os dados em intervalos, unindo valores com probabilidades diferentes (isto pode ser evitado em distribuições discretas).
  - Assim, o nível de significância real apenas se aplica para um número infinito de observações (intervalos).
  - Na prática (amostras finitas), reduzimos este efeito evitando a ocorrência de intervalos com poucas observações: cada intervalo com menos de 5 observações é agrupado com algum intervalo vizinho.
    * Quando o intervalo tem mais observações temos mais confiança da proporção observada.

38.2 Teste Kolmogorov-Smirnov

- Passos:
  1. Determine a função de distribuição acumulada observada $F_o(x)$, e a função de distribuição acumulada esperada $F_e(x)$.
  2. Utilizando os n valores da amostra, calcule as estatísticas
  $$K^+ = \sqrt{n} \times \max_x \{F_o(x) - F_e(x)\}$$
  $$K^- = \sqrt{n} \times \max_x \{F_e(x) - F_o(x)\},$$
  que representam as maiores diferenças entre as distribuições para mais e para menos.
  3. Com significância α, não podemos rejeitar a hipótese de que os dados obedecem a distribuição se K^+ e K^- forem menores que $K_{1-\alpha,n}$ (obtido em tabela).
- A função de distribuição acumulada observada $F_o(x)$ é o percentual de observações com valor menor ou igual a x, ou seja:
  - Sejam $x_1 \leq x_2 \leq \ldots \leq x_n$ os elementos da amostra em ordem crescente. Então,
  $$F_o(x) = \begin{cases} 0, & \text{se } x < x_1 \\ i/n, & \text{se } x_i \leq x < x_{i+1}, \quad i = 1, \ldots, n-1 \\ 1, & \text{se } x \geq x_n \end{cases}$$
- Existe uma observação importante no cálculo de K^- (figura 27.1, Jain).
  - Toda função de distribuição acumulada é não decrescente.
  - Como $F_o(x)$ se baseia em um conjunto finito de observações, ela possui incrementos discretos (ou seja, $F_o(x)$ é constante entre observações consecutivas $x_i \leq x < x_{i+1}$).
  - Por outro lado, quando a distribuição é contínua, $F_e(x)$ será contínua.
  - Assim, o $\max_x\{F_e(x) - F_o(x)\}$ para $x_{i-1} \leq x < x_i$ ocorre imediatamente antes de x_i. Ou seja, $\max_{x_{i-1} \leq x < x_i}\{F_e(x) - F_o(x)\}$ converge para $F_e(x_i) - F_o(x_{i-1})$.
- Para a distribuição uniforme entre 0 e 1 temos que $F_e(x) = x$. Portanto,
  $$K^+ = \sqrt{n} \times \max_{i=1,\ldots,n} \left\{\frac{i}{n} - x_i\right\}$$
  $$K^- = \sqrt{n} \times \max_{i=1,\ldots,n} \left\{x_i - \frac{(i-1)}{n}\right\}$$
  - Ex.: tabela 27.2, Jain.
- Comparando com o teste Chi-quadrado, concluimos:

– Ao contrário do teste Chi-quadrado, que é mais apropriado para distribuições discretas e amostras grandes, o teste K-S foi projetado para distribuições contínuas e amostras pequenas.

– O teste K-S compara as distribuições acumuladas (teórica e observada), enquanto o teste Chi-quadrado compara as densidades de probabilidades.

– Ao contrário do teste Chi-quadrado, o teste K-S não faz agrupamento de observações. Neste sentido, o teste K-S faz melhor uso dos dados.

– A escolha dos tamanhos dos intervalos é um problema do teste Chi-quadrado (não existe regras bem definidas para isso). Esta escolha afeta o resultado.

– O teste Chi-quadrado é sempre aproximado, enquanto o teste K-S é exato sempre que os parâmetros da distribuição são conhecidos.

```
n=1000; m=2^12; a=125; b=1; x=zeros(1,n);
x(1)=1;
for i=2:n; x(i)=mod(a*x(i-1)+b,m); end;
x=x/m; hist(x); grid on; figure; autocorr(x)

[h p]=kstest(x',[x' unifcdf(x')],0.05)

MIDSQUARE
n=100; x=zeros(1,n); x(1)=1234;
for i=2:n;
   x(i)=mod(floor(x(i-1)*x(i-1)/100),10000);
end;
x=x/10000; hist(x)

[h pv]=chi2gof(x,'cdf',@(z)unifcdf(z),
   'nparams',0, 'alpha',0.05)
[h p]=kstest(x',[x' unifcdf(x')],0.05)
```

38.3 Teste de correlação serial

- A covariância é um método direto de testar dependência entre duas variáveis aleatórias.

  – Se a covariância é não nula, existe depedência entre elas (linear).

  – Porém, se a covariância é nula, pode haver dependência não linear entre elas.

- Dada uma sequência $X_1, \ldots, X_n$ de variáveis aleatórias, podemos testar a covariância entre variáveis separados por k posições na sequência, ou seja, entre X_i e X_{i+k} para $i = 1, \ldots, n-k$.

  – Esta medida é chamada de "autocovariância" de "lag" k, e é expressa por

$$R_k = \frac{1}{n-k} \sum_{i=1}^{n-k} \left(X_i - \mathrm{E}[X_i]\right) \left(X_{i+k} - \mathrm{E}[X_{i+k}]\right).$$

  – Se X é uniforme entre 0 e 1, então $\mathrm{E}[X] = 1/2$.

- Se n é grande, e as variáveis aleatórias são independentes com distribuição uniforme entre 0 e 1, então R_k (para $k \geq 1$) tem distribuição normal com média zero e variância $1/[144(n-k)]$.

  – A forma normal vem do "Teorema Central do Limite".

  – Quando duas variáveis X e Y são independentes (o que implica em correlação nula), temos $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$ (pois a correlação é igual a $\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$). Portanto, $\mathrm{E}[R_k] = 0$ e $\mathrm{E}[R_k^2]$

$$= \frac{1}{(n-k)^2} \sum_{i=1}^{n-k} \mathrm{E}\left[\left(X_i - \frac{1}{2}\right)^2 \left(X_{i+k} - \frac{1}{2}\right)^2\right]$$

$$= \frac{1}{144(n-k)},$$

  pois os produtos cruzados são nulos, e se X é uniforme entre $-1/2$ e $1/2$, então $\mathrm{E}[X^2] = \int_{-1/2}^{1/2} X^2 dx = 1/12$.

- Portanto, para $k \geq 1$ o intervalo de confiança $100(1-\alpha)\%$ para a autocovariância é

$$R_k \pm \frac{z_{1-\alpha/2}}{12\sqrt{n-k}},$$

  onde z_β (obtido em tabela) é tal que $\Pr[X \leq z_\beta] = \beta$ se X tem distribuição uniforme com média 0 e desvio padrão 1.

- Para cada valor de $k \geq 1$ desejado, o teste consiste em determinar se o valor zero pertence ao intervalo de confiança da autocovariância de lag k.

  – Se não pertencer, não podemos rejeitar a hipótese de que não existe autocovariância de lag k.

- Quando $k = 0$, R_0 é a variância da sequência, cujo valor esperado é $1/12$ quando os valores são independentes e uniformes entre 0 e 1.

```
n=1000; m=2^12; a=125; b=1; x=zeros(1,n);
x(1)=1;
for i=2:n; x(i)=mod(a*x(i-1)+b,m); end;
x=x/m;

%x=(1:n)/n;
for k=1:10,
  r=sum((x(1:end-k)-.5).*(x(k+1:end)-.5))/(n-k);
  s=norminv(.975)/(12*sqrt(n-k)); [r-s, r+s]
end
```

39 Testes de Uniformidade d-dimensional

- Uma observação feita por Marsaglia (1968) revelou a necessidade de se testar a uniformidade em dimensões maiores:
  - "Números aleatórios ocorrem principalmente em planos."
  - Isto significa que as tuplas sobrepostas $(x_1, x_2, \ldots, x_d), (x_2, x_3, \ldots, x_{d+1}), \ldots$ ocorrem em um número pequeno de hiperplanos (espaços vetoriais com dimensão $d-1$) no espaço vetorial de dimensão d, que cruzam o hipercubo $[0,1]^d$.
  - Isto também ocorre para tuplas não sobrepostas $(x_1, x_2, \ldots, x_d), (x_{d+1}, x_{d+2}, \ldots, x_{2d}), \ldots$.
  - Ex.: 2 dimensões: figura 7.2, Law. 3 dimensões: figura 7.4, Law.
  - Dependendo do espaçamento entre hiperplanos, o resultado pode ser mais aceitável. Ex.: figura 7.1, Law.
  - Além da inspeção visual, vamos considerar dois destes testes de uniformidade d-dimensional: o teste serial (empírico) e o teste espectral (teórico).

```
n=17; m=16; a=5; b=3; x=zeros(1,n); x(1)=7;
for i=2:n; x(i)=mod(a*x(i-1)+b,m); end;
x=x/m;

scatter(x(1:end-1), x(2:end))
%scatter3(x(1:end-2), x(2:end-1), x(3:end))
```

39.1 Teste Serial

- Passos:
  - Divida o espaço no hipercubo $[0,1]^d$ em K^d células de igual tamanho. Ex.: figura 27.5, Jain.
  - Utilizando uma amostra $x_1, \ldots, x_n$, produza n/d tuplas não sobrepostas.
  - Conte quantas tuplas ocorrem em cada célula. Esperamos encontrar $n/(dK^d)$ tuplas em cada célula.
  - Utilize o teste Chi-quadrado, com $K^d - 1$ graus de liberdade. Note que este teste exige que as observações sejam independentes, portanto as tuplas não podem ser sobrepostas.

```
n=18; m=16; a=5; b=3; x=zeros(1,n); x(1)=7;
for i=2:n; x(i)=mod(a*x(i-1)+b,m); end;
x=x/m;

d=2; k=3;
v=1:length(x);
x1=x(v(mod(v,2)==0)); x2=x(v(mod(v,2)==1));
c = zeros(1,d*k);
for i=1:k,
   for j=1:k,
    c(k*(i-1)+j) = sum((x1 >= (i-1)/k) & (x1 < i/k) ...
       & (x2 >= (j-1)/k) & (x2 < j/k));
   end
end

[h p]=chi2gof(c,'nbins', d*k, 'expected', ...
   ones(1,d*k)*n/(d*k^d), 'alpha',0.05)

---------------------

n=1000; m=2^12; a=125; b=1; x=zeros(1,n);
x(1)=1;
for i=2:n; x(i)=mod(a*x(i-1)+b,m); end;
x=x/m;

scatter(x(1:end-1), x(2:end))
figure; scatter3(x(1:end-2), x(2:end-1), x(3:end))

d=2; k=3;
v=1:length(x);
x1=x(v(mod(v,2)==0)); x2=x(v(mod(v,2)==1));
c = zeros(1,d*k);
for i=1:k,
   for j=1:k,
    c(k*(i-1)+j) = sum((x1 >= (i-1)/k) & (x1 < i/k) ...
       & (x2 >= (j-1)/k) & (x2 < j/k));
   end
end

[h p]=chi2gof(c,'nbins', d*k, 'expected', ...
   ones(1,d*k)*n/(d*k^d), 'alpha',0.05)
```

39.2 Teste Espectral

- Este teste determina a máxima distância entre hiperplanos adjacentes. Quanto maior esta distância, pior o gerador.
- Quando o período é pequeno, esta distância pode ser determinada por enumeração.
- Ex.: 27.7, Jain: comparar os geradores $x_n = 3x_{n-1} \bmod 31$ e $x_n = 13x_{n-1} \bmod 31$, ambos com $x_0 = 15$.
  - Como ambos produzem ciclos com os mesmo números (em ordem diferente), ambos apresentam o mesmo resultado no teste serial.
  - Porém, os pares sobrepostos mostram diferenças: figuras 27.6 e 27.7, Jain.
  - No primeiro gerador temos 3 retas com máxima distância: $x_n = 3x_{n-1} + 31k, k = 0, 1, 2$. No segundo temos: $x_n = -5x_{n-1}/2 + 31k/2, k = 0, \ldots, 5$.
  - A distância entre duas retas paralelas $y = ax + c_1$ e $y = ax + c_2$ vale $|c_2 - c_1|/\sqrt{1 + a^2}$.

- Assim, obtemos que a distância é menor no segundo gerador. Na inspeção visual também escolheríamos o segundo gerador.

blank

- Este método permite o uso de tuplas sobrepostas e não sobrepostas.
  - Portanto, as sobrepostas deveriam ser utilizadas, pois produzem mais pontos: $n-1$ contra n/d no caso sem sobreposição.
- Para m grande ou dimensões maiores, a enumeração dos hiperplanos é inviável.
  - Existem métodos para calcular as distâncias sem enumeração (ver Knuth 1981), mas devido à sua complexidade, não detalharemos aqui.

40 Conclusões (Law)

- Existe grande variedade de testes para números aleatórios, e existe muita controvérsia sobre quais testes são melhores, e se testes teóricos são mais definitivos que os empíricos.
  - Ex.: se a amostra utilizada em um teste empírico for muito pequena, o resultado pode corresponder apenas a um pequeno fragmento do ciclo. Por outro lado, um teste que leva em conta um ciclo inteiro pode não ser compatível com fragmentos deste mesmo ciclo.
    - Podemos ter segmentos não aleatórios em ciclos considerados aleatórios.
  - De fato, nenhuma quantidade de testes pode convercer alguém que certo gerador é o melhor em todos os casos.
- Portanto, a recomendação é que os testes sejam consistentes com o uso que será dado ao gerador.
  - Ex.: se os números serão utilizados aos pares, examinar o comportamento de pares (talvez com um teste serial) seja o mais apropriado.
- Assim, o projetista deve ser cauteloso na escolha e no teste do gerador se a simulação é cara, requer alta precisão, ou é um componente crítico de um estudo mais amplo.