SME0305 - 2016

Gustavo C. Buscaglia / Roberto F. Ausas

ICMC - Ramal 738176, gustavo.buscaglia@gmail.com
ICMC - Ramal 736628, rfausas@gmail.com

Integração Monte Carlo

- **Métodos de Monte Carlo** é uma classe de algoritmos computacionais baseados em amostragens aleatórias.
- É um método poderoso, flexível e direto. Muitas vezes é a maneira **mais simples** de resolver um problema. Frequentemente é a **única** maneira de resolver o problema.

Introdução

- Em geral se identificam vários passos:

  1. Formular o resultado procurado como sendo o valor esperado (média) de uma quantidade **X**, sobre um conjunto de **dados possíveis** $\Omega = \{\theta\}$. A probabilidade de cada dado específico θ (seu peso na média) é $P(\theta)$, conhecida.

  $$x = \mathbb{E}(\mathbf{X}) = \sum_{\theta \in \Omega} \mathbf{X}(\theta)\, P(\theta)$$

  2. Gerar uma sequência $\{\theta_i\}$ que amostre o conjunto Ω de acordo com a probabilidade P.
  3. Com cada dado θ_i realizar o **cálculo determinístico** $x_i = \mathbf{X}(\theta_i)$.

4. Os resultados são finalmente agregados, somados, por exemplo calculando médias como

$$\hat{x} = \frac{1}{N} \sum_{i=1}^{M} x_i = \langle \{ x_i \} \rangle$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{M} (x_i - \hat{x})^2 = \langle \{ (x_i - \hat{x})^2 \} \rangle$$

- Vejamos um exemplo. Seja x, o que queremos calcular, a razão entre a área do círculo e a do quadrado circunscrito ($= \pi/4$).

- **Passo 1:** x é o valor esperado da função que, escolhido um ponto $\theta = (\alpha, \beta)$ de maneira aleatória em

$$\Omega = \{(\alpha, \beta) \in [-1, 1] \times [-1, 1]\} ,$$

com probabilidade uniforme, vale

$$\mathbf{X}(\theta) = \begin{cases} 1 & \text{if } \alpha^2 + \beta^2 < 1 \\ 0 & \text{se não} \end{cases}$$

Sedimentos atmosféricos, em certas escalas, se distribuem com probabilidade uniforme.

- **Passo 2:** Simulação.

  1. Gerar $\theta_i = (\alpha_i, \beta_i)$, formado por dois números α_i e β_i entre -1 e 1, independentes e com probabilidade uniforme.
  2. Calcular $X_i = \mathbf{X}(\theta_i)$, definida como $= 1$ se $\alpha_i^2 + \beta_i^2 < 1$ e $= 0$ se não. Seja $X = \{X_i\}$ a sequência gerada.
  3. Tirar a média

$$\hat{x} = \langle X \rangle = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Pela **lei dos grandes números**, dado $\varepsilon > 0$, a probabilidade de que $|\hat{x} - x| > \varepsilon$ tende a zero quando N tende a infinito.

- **Octave**

```
som=0;
for i=1:N
x=-1+2*rand();
y=-1+2*rand();
if (x^2+y^2<1)
 som=som+1;
end
end
p=som/N
```

- Resultados para $4p$:

#	$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 10^6$	$N = 10^7$
1	3.0800	3.1492	3.1450	3.1418	3.1405
2	3.2320	3.1340	3.1396	3.1400	$--$
3	3.1240	3.1412	3.1469	3.1376	$--$
4	3.0800	3.1660	3.1488	3.1406	$--$

- Observar que os resultados são **mais precisos** quando N é maior.

- Observar que os resultados são **menos dispersos** quando N é maior.

- **Convergência**.

- **Que maneira cara** de calcular a integral

$$\pi = 4 \int_0^1 \sqrt{1 - x^2}\, dx \quad \textbf{!!}$$

- Vamos todavía fazer outro exemplo elementar: Seja $f(x) = e^x$ cuja integral indefinida ($= e^x$) infelizmente **esquecemos**. Desejamos calcular

$$J = \int_0^2 f(x)\ dx\ .$$

cujo valor exato é $J = e^2 - 1 = 6.389056\ldots$.

- J é o valor esperado da função $\mathbf{J}$, definida sobre $\Omega = [0,2]$,

$$\mathbf{J}(\theta) = 2f(\theta)$$

quando todos os $\theta \in \Omega$ são equiprováveis.

- Fazemos para isto uma simulação Monte Carlo (`som=0;`):

   1. Geramos números θ_i uniformemente em $[0, 2]$.
      `theta=2*rand();`
   2. Computamos, `f=e^theta;`
   3. Agregamos, `som=som+2*f;`

  Finalmente, o estimador $\hat{J}$ is `J=som/N`.

#	$N = 10^3$	$N = 10^4$	$N = 10^5$	$N = 10^6$
1	6.5378	6.4562	6.3955	6.3961
2	6.2635	6.4387	6.3960	6.3870
3	6.1529	6.3785	6.4007	
4	6.5339	6.4326	6.3932	
				6.389056

- Notar que o algoritmo devolve $\hat{J}$, um resultado aleatório. Toda vez que rodamos é diferente.

- À pergunta "quanto vale J?" deveremos responder realizando uma inferência a partir dos resultados, uma estimativa.

- A integral $\int_0^2 f(x)\ dx$ seria bem diferente (~ 16.4) se entre $x = 0$ e $x = 10^{-4}$ a função valesse 100000. Porém, se nenhuma das N amostras consideradas caiu nesse intervalo, sempre obteremos $J \simeq 6.4$. **Eventos raros podem demorar a convergência**.

Uma análise de risco

- Um bairro conta com uma rede hidráulica conhecida, alimentada desde um reservatório a pressão $P = 5$ (bar).
- Dois tipos de canos: **grossos**, com “compliância” $C = 10$, e os **finos**, com $C = 2$. De todos os nós é extraída uma vazão nominal $Q = 0.1$.
- Anualmente, os canos **finos** (apenas eles) ficam obstruidos com probabilidade $p = 0.1$. Quando obstruidos, seu valor de C se reduz a $C = 0.2$.
- Qual é a probabilidade de que, passado um ano, **algum dos nós esteja a pressão menor que** $P = 1.15$**?**

- Seja M o número de canos finos, todos distintos por ocuparem lugares distintos na rede.
  - Ω: Existem 2^M casos diferentes. $2^{100} = 1.27 \times 10^{30}$.
  - Cada caso $\theta \to$ um vetor `C` diferente.
  - Cada caso $\theta \to$ probabilidade $P(\theta) = p^m (1-p)^{(M-m)}$, sendo $m(\theta)$ o número de obstruções.
  - Para cada caso, podemos calcular deterministicamente todas as pressões, em particular a mínima:

    ```
    P=ResolveRede(nv,conec,C); Y=min(P);
    ```

  - A probabilidade desejada x é

$$x = \mathbb{E}(\mathbf{X}) = \sum_{\theta \in \Omega} \mathbf{X}(\theta)\, P(\theta) \ ,$$

o valor esperado da função

$$\mathbf{X}(\theta) = \begin{cases} 1 & \text{se } Y(\theta) < 1.15 \\ 0 & \text{se não} \end{cases}$$

- Dificuldade: Se o cálculo de $\mathtt{X}$ demora **1 nanosegundo** obteremos o resultado após $\mathbf{4 \times 10^{13}}$ **anos**.

- Tentemos um método Monte Carlo:

  1. Gerar um vetor `C` com a probabilidade adequada

```
function Cn = Cnew(nc,C)
Cn=C;
for i=1:nc
 if ((C(i)==2) && (rand() < 0.1))
  Cn(i)=0.2;
 end
end
```

2. Determinar as pressões $P(C)$, se $\min P(C) < 1.15$ considerar dentro dos casos falhos.

   ```
   som = 0;
   for i=1:N
    Cn = Cnew(nc,C); P = ResolveRede(nv,conec,Cn);
    Y=min(P);
    if (Y<1.15)
     som = som + 1;
    end
   end
   ```

3. A probabilidade pedida é estimada por $\hat{x}$,

   ```
   x = som/N
   ```

- Esse algoritmo percorre apenas uma minúscula fração do conjunto de possibilidades se N é pequeno.
- Quando $N \to +\infty$ podemos esperar um resultado correto.

Simulações

- Construimos os dados da rede sem obstruções. Começamos com apenas canos finos,

  ```
  [nv nc conec C coord] = RedeHidraBairro(10,10,2,2);
  ```

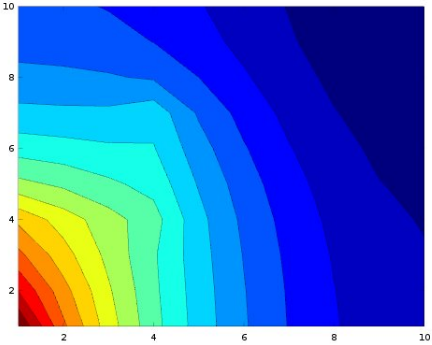
  e substituimos por canos grossos onde corresponda

  ```
  C(1,1) =20; C(2,1)  =20;  C(3,1) =20;
  C(94,1)=20; C(104,1)=20; C(114,1)=20;
  C(31,1)=20; C(32,1)=20; C(33,1)=20;
  ```

- A rede nominal se resolve como

```
P=ResolveRede(nv,conec,C);
```

dando

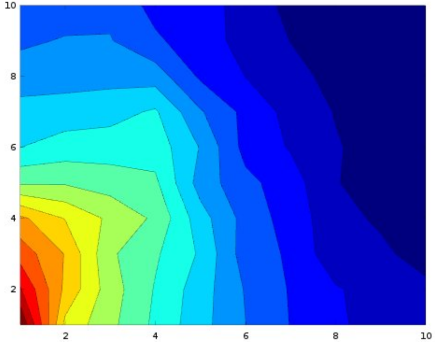


`max(P)=5`, `min(P)=1.6745`

- Outra realização

```
Cn = Cnew(nc,C); P = ResolveRede(nv,conec,Cn);
```

dando

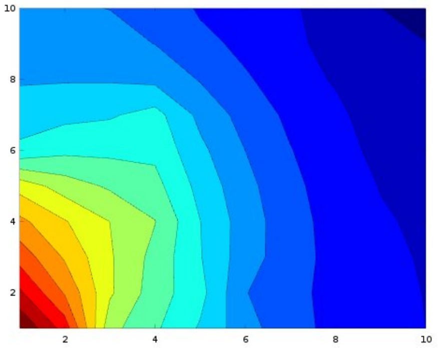


max(P)=5, min(P)=0.93263

- E outra

```
Cn = Cnew(nc,C); P = ResolveRede(nv,conec,Cn);
```

dando



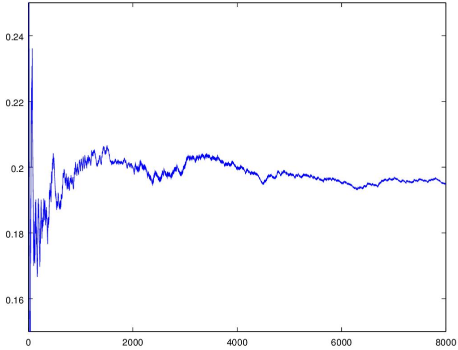
max(P)=5, min(P)=1.1423

- Podemos agora fazer N realizações:

#	$N = 1000$	$N = 4000$	$N = 16000$	$N = 64000$
1	0.176	0.179	0.193	0.190
2	0.185	0.195	0.197	
3	0.194	0.194		
3	0.236	0.186		
4	0.198	0.195		

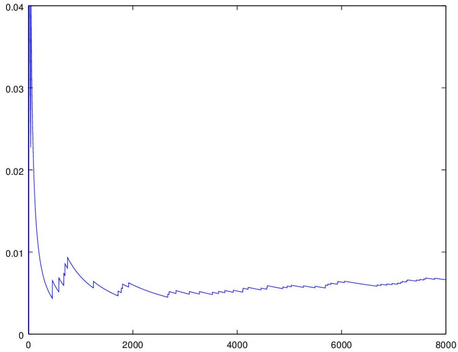
- Das quais poderia se inferir que a probabilidade de falha do sistema é aproximadamente 20%.

- A probabilidade calculada se estabiliza quando *N* cresce,



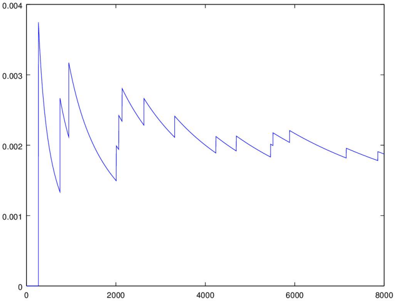
mas ainda é necessária uma **inferência** estatística para extrair alguma conclusão quantitativa ou qualitativa.

- Mudando a probabilidade p de obstrução de cada cano individual para $p = 0.05$, o resultado é bem diferente,



A estimativa de probabilidade de falha cai para ~0.7%.

- Quando p se reduz para 0.035 a probabilidade de falha cai para ~ 0.2%, e 8000 realizações são certamente insuficientes:



$p =$	20%	10%	5%	3.5%
$q \simeq$	94%	20%	0.7%	0.2%

Uma estimativa do erro

- Hipóteses: $x = \mathbb{E}(\mathbf{X}) \in \mathbb{R}$ (valor esperado finito) e

$$\sigma^2(\mathbf{X}) = \mathbb{E}((\mathbf{X} - x)^2) = \sum_{\theta} (\mathbf{X}(\theta) - x)^2 P(\theta) < +\infty \ .$$

- O resultado $\hat{x}_N$ de um cálculo Monte Carlo é uma variável aleatória, sobre o conjunto de sequências de N amostras possíveis.

- O método funciona porque

$$\mathbb{E}(\hat{x}_N) = x$$

e porque

$$\sigma^2(\hat{x}_N) = \mathbb{E}((\hat{x}_N - x)^2) = \frac{\sigma^2(\mathbf{X})}{N} \ .$$

- **Estimativa a priori do erro:** O desvio padrão de $\hat{x}_N - x$, que quantifica o erro, cumpre

$$\sigma(\hat{x}_N - x) = \sigma(\mathbf{X})/N^{1/2} \ .$$

- Lembremos que $\mathbb{E}(\cdot)$ é uma integração. O erro de uma integração determinística (Simpson) é de ordem $O(N^{-4/d})$, onde d é a dimensão de Ω. Para $d > 8$ Monte Carlo é mais preciso.

- A dimensão d não tem impacto sobre $\sigma^2(\hat{x}_N)$. A **suavidade** da função $\mathbf{X}(\theta)$ também não.

- **Estimativa a posteriori:** $\sigma^2(\mathbf{X})$ pode ser estimada por

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{X}(\theta_i) - \hat{x})^2$$

  porque $\mathbb{E}(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2(\mathbf{X})$.

- Assim, o erro de $\hat{x}$ é estimado por

$$\sigma(\hat{x}_N - x) \simeq \frac{\hat{\sigma}}{\sqrt{N}} \ .$$

- Exemplo: Estimativa do erro no cálculo da probabilidade de falha da rede hidráulica.

  – $x = \mathbb{E}(\mathbf{X})$ onde $\mathbf{X}(\theta)$ é 1 ou 0.

  – $\hat{x} = (1/N)\sum_i X_i$, e

$$\hat{\sigma}^2 = \frac{1}{N}\left(\sum_i X_i^2 - N\hat{x}^2\right) = \frac{1}{N}\left(\sum_i X_i - N\hat{x}^2\right) = \hat{x}\,(1-\hat{x}) \ .$$

  – Por tanto,

$$\sigma(\hat{x}_N - x) \simeq \sqrt{\frac{\hat{x}\,(1-\hat{x})}{N}} \ \leq \ \frac{1}{2\sqrt{N}} \ .$$

– O erro relativo de $\hat{x}$ é, então, $\epsilon \simeq \sqrt{\frac{1-\hat{x}}{\hat{x}}}\ \frac{1}{\sqrt{N}}$

	$N = 1000$	$N = 4000$	$N = 16000$	$N = 64000$
$\hat{x} = 0.20$	6.3%	3.2%	1.6%	0.8%
$\hat{x} = 0.07$	11.5%	5.7%	2.8%	1.4%
$\hat{x} = 0.01$	31.5%	15.7%	7.8%	3.9%
$\hat{x} = 0.001$	99.9%	50%	25%	12%

- Essas estimativas de erro dão uma ideia de quando surgem as dificuldades ($\hat{x}$ e N pequenos).
- Um resultado quantitativo sério requer **inferência**.

Sumário:

- Integramos o método matricial de resolução de redes (hidráulicas) com a simulação Monte Carlo para avaliação de risco, discutindo o básico.

- Introduzimos os estimadores

  $$\hat{x}_N = \langle \{X_i\}_{i=1:N} \rangle \simeq x = \mathbb{E}(\mathbf{X}) \quad \text{e} \quad \hat{\sigma}_N^2 = \langle \{X_i^2\}_{i=1:N} \rangle - \hat{x}^2 \simeq \sigma^2(\mathbf{X}).$$

  para expressar o resultado como $\hat{x} \pm \hat{\sigma}$.

- Visto como método iterativo ($N = 1, 2, \ldots$ a medida que são tomadas amostras), Monte Carlo gera uma sequência **convergente**.

- O conceito de **problemas de alta dimensão** foi exemplificado considerando dados aleatórios cuja dimensão correspondia ao número de canos finos de uma rede.