

# MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística  
Universidade de São Paulo  
pam@ime.usp.br  
<http://www.ime.usp.br/~pam>

## Aula 20

2 de junho de 2025

# Sumário

1 Simulação estática

2 Métodos de reamostragem

# Preliminares

- Sabemos como construir alguns modelos probabilísticos para representar uma situação real, ou então para descrever um experimento aleatório. Notamos, também, que determinados um espaço amostral e probabilidades associadas aos pontos desse espaço, o modelo probabilístico ficará completamente determinado e poderemos, então, calcular a probabilidade de qualquer evento aleatório.
- Muitas vezes, mesmo construindo um modelo probabilístico, certas questões não podem ser resolvidas analiticamente e teremos que recorrer a **estudos de simulação** para obter aproximações de quantidades de interesse.
- Estudos de simulação tentam reproduzir num ambiente controlado o que se passa com um problema real. Para nossos propósitos, a solução de um problema real consistirá na simulação de **variáveis aleatórias** (**simulação estática**) ou de **processos estocásticos** (**simulação dinâmica**).
- A simulação de variáveis aleatórias deu origem aos chamados **métodos Monte Carlo** (MMC), que por sua vez supõem que o pesquisador disponha de um **gerador de números aleatórios**.

## Preliminares

- Um **número aleatório** (NA) representa o valor de uma variável aleatória uniformemente distribuída no intervalo  $(0, 1)$ . Originalmente, esses NA eram gerados manualmente ou mecanicamente, usando dados, roletas etc. Modernamente, usamos computadores para gerar NA.
- Os MMC apareceram durante a segunda guerra mundial, em pesquisas relacionadas à difusão aleatória de neutrons num material radioativo. Os trabalhos pioneiros devem-se a Metropolis e Ulam (1949), Metropolis et al. (1953) e von Neumann (1951). Veja Sóbol (1976), Hammersley e Handscomb (1964) e Ross (1997).
- Para ilustrar, suponha que se queira calcular a área da figura  $F$  contida no quadrado  $Q$  de lado unitário (Figura 1). Suponha que sejamos capazes de gerar pontos aleatórios em  $Q$ , de modo homogêneo, isto é, de modo a cobrir toda a área do quadrado, ou ainda, que estes pontos sejam *uniformemente distribuídos sobre  $Q$* . Se gerarmos  $N$  pontos, suponha que  $N'$  desses caiam em  $F$ . Então, poderemos aproximar a área de  $F$  por  $N'/N$ . Quanto mais pontos gerarmos, melhor será a aproximação.
- Note que o problema em si não tem nenhuma componente aleatória: queremos calcular a área de uma figura plana. Mas, para resolver o problema, uma possível maneira foi considerar um mecanismo aleatório. Veremos que esse procedimento pode ser utilizado em muitas situações.

# Preliminares

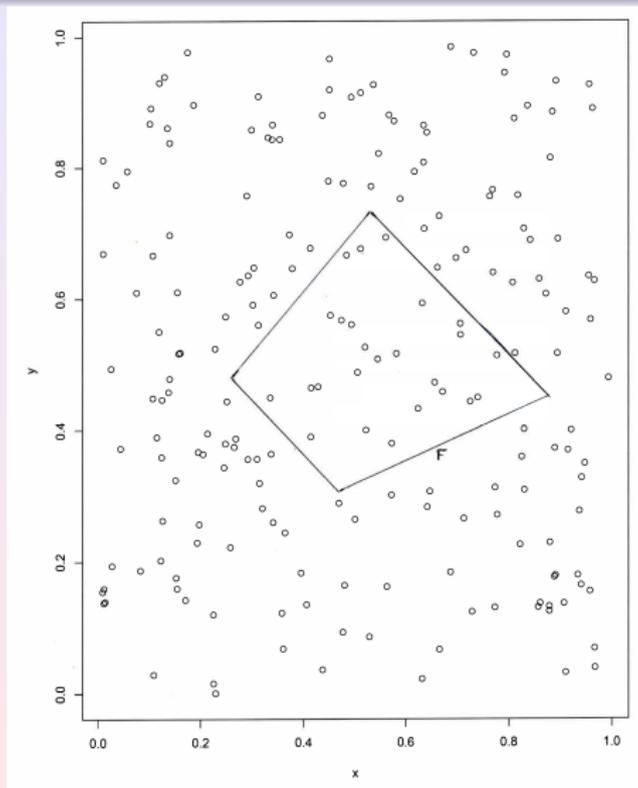


Figura: Área de uma figura por simulação.

## Preliminares

- Dissemos acima que um NA é um valor de uma variável aleatória uniformemente distribuída no intervalo  $(0, 1)$ . Vejamos algumas maneiras de obter um NA.
- **Exemplo 1.** Lance uma moeda 3 vezes e atribua o valor 1 se ocorrer cara e o valor 0 se ocorrer coroa. Os resultados possíveis são as *sequências* ou **números binários** abaixo:

000, 001, 010, 011, 100, 101, 110, 111.

- Cada um desses números binários corresponde a um número decimal. Por exemplo,  $(111)_2 = (7)_{10}$ , pois  $(111)_2 = 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$ . Considere a representação decimal de cada sequência acima e divida o resultado por  $2^3 - 1 = 7$ . Obteremos o números aleatórios  $0, 1/7, 2/7, \dots, 1$ . Qualquer uma das 8 sequências acima tem a mesma probabilidade, a saber,  $1/2^3 = 1/8$ .
- De modo geral, lançando-se a moeda  $n$  vezes, teremos  $2^n$  possibilidades e cada uma terá probabilidade  $1/2^n$  e os NA finais são obtidos por meio de divisão por  $2^n - 1$ .

## Preliminares

- O que se faz atualmente é fazer **simulação por meio de computadores**, que utiliza **números pseudo-aleatórios**, no lugar de NA.
- Os números pseudo-aleatórios (NPA) são obtidos por meio de técnicas que usam relações matemáticas recursivas **determinísticas**. Logo, um NPA gerado numa iteração dependerá do número gerado na iteração anterior, e portanto não será realmente aleatório, donde o nome pseudo-aleatório.
- Um método bastante utilizado em pacotes computacionais é o método congruencial
- O R usa o comando `runif(n,min,max)`, onde  $n$  é o número de valores a gerar e  $(\min, \max)$  é o intervalo no qual se quer gerar os NPA. No nosso caso,  $\min=0$  e  $\max=1$ .
- Outros pacotes têm seus próprios geradores de NPA, como o MatLab.
- **Exemplo 2.** O comando `u <- runif(10,0,1)` pede para gerar 10 NA e guardá-los no vetor `u`.

```
> u<-runif(10,0,1)
```

```
> u
```

```
[1] 0.80850094 0.56611676 0.75882010 0.89910843 0.48447125
```

```
[6] 0.02119849 0.06239355 0.30022882 0.12722598 0.49714446
```

# Preliminares

Existem três grandes grupos de métodos de simulação:

- 1) **Métodos de Simulação Estática**. Aqui, os procedimentos têm por objetivo gerar amostras independentes. Citamos os métodos Monte Carlo, aceitação/rejeição e reamostragem ponderada.
- 2) **Métodos de Simulação por Imputação**. A ideia desses métodos é aumentar os dados, introduzindo **dados latentes**, com o intuito de facilitar a simulação. Dentre esses métodos citamos o algoritmo EM ( de *expectation-maximization*) e o algoritmo de dados aumentados.
- 3) **Métodos de Simulação Dinâmica**. Esses métodos são denominados atualmente por MCMC (**Markov Chain Monte Carlo**) e têm por objetivo construir uma cadeia de Markov, cuja distribuição de equilíbrio seja a distribuição da qual queremos amostrar. Os métodos mais importantes aqui são o amostrador de Gibbs e os algoritmos de Metropolis e Metropolis-Hastings.

## Métodos Monte Carlo

- Suponha uma v.a. com distribuição  $F$  e desejamos calcular a média de uma função qualquer  $h(X)$ . Suponha, ainda, que exista um método para simular uma amostra  $X_1, \dots, X_n$  de  $F$ . Nas seções seguintes veremos alguns desses métodos. Então, o método Monte Carlo (MMC) consiste em aproximar  $\mu_F = E_F[h(X)]$  por

$$\hat{\mu}_F = \hat{E}_F[h(X)] = \frac{1}{n} \sum_{i=1}^n h(X_i). \quad (1)$$

- Observe que (14) aproxima a integral  $\int h(x)dF(x)$  ou  $\int h(x)f(x)dx$ , se existir a densidade de  $X$ . A lei (forte) dos grandes números garante que, quando  $n \rightarrow \infty$ ,  $\hat{\mu}_F$  converge para  $\mu_F$  com probabilidade um.
- O erro padrão da estimativa (14) é dado pela raiz quadrada da variância de  $\hat{\mu}_F$ , denotada por  $\text{Var}(\hat{\mu}_F)$ . Esta, por sua vez, pode ser estimada por  $\widehat{\text{Var}}(\hat{\mu}_F)$  e, portanto, uma estimativa do erro padrão de  $\hat{\mu}_F$  será

$$\widehat{EP}(\hat{\mu}_F) = \frac{1}{\sqrt{n}} \left[ \sum_{i=1}^n (h(X_i) - \hat{\mu}_F)^2 \right]^{1/2} = O(n^{-1/2}). \quad (2)$$

## MMC – Exemplo 3

- **Exemplo 3.** Suponha que se queira calcular o valor esperado de  $h(X)$ , onde  $h(x) = \sqrt{1 - x^2}$  e  $F \sim \mathcal{U}(0, 1)$ . Então, se  $X_1, \dots, X_n$  for uma amostra da distribuição uniforme padrão,

$$\hat{E}_F[h(X)] = \frac{1}{n} \sum_{i=1}^n \sqrt{1 - X_i^2}.$$

- Por exemplo, gerando-se 1.000 valores de uma  $\mathcal{U}(0, 1)$ , obtivemos o valor 0,7880834. Observe que essa, é também, uma estimativa de um quarto da área de um círculo unitário, ou seja,  $\pi/4 = 0,7853982$ . O erro padrão calculado por (2) é 0,0069437.
- Uma outra aplicação do MMC é na obtenção de amostras de distribuições marginais. Suponha, por exemplo, que as v.a.  $X$  e  $Y$  tenham densidade conjunta  $f(x, y)$  e marginais  $f_X(x)$  e  $f_Y(y)$ , respectivamente. Então,

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x) dx, \quad (3)$$

onde  $f_{Y|X}(y|x)$  é a densidade condicional de  $Y$  dado que  $X = x$ .

## MMC – Exemplo 3

Para obter uma amostra de  $f_Y(y)$  procedemos como segue (**método da composição ou mistura**):

- (a) obtemos um elemento  $x^*$  de  $f_X(x)$ ;
- (b) fixado  $x^*$ , obtemos um elemento  $y^*$  de  $f_{Y|X}(y|x^*)$ .

Repetimos os passos (a) e (b)  $n$  vezes, obtendo-se  $(x_1, y_1), \dots, (x_n, y_n)$  como uma amostra de  $f(x, y)$ , enquanto que  $y_1, \dots, y_n$  representa uma amostra de  $f_Y(y)$ .

É óbvio que devemos saber como amostrar das densidades  $f_X(x)$  e  $f_{Y|X}(y|x)$ ;  $x^*$  é chamado o **elemento misturador**.

## Simulação de variáveis discretas

- Vimos que a geração de NAs corresponde a gerar valores de uma distribuição uniforme no intervalo  $(0, 1)$
- Se  $U \sim \mathcal{U}(0, 1)$  e se  $0 < a < b < 1$ , então

$$P(a < U < b) = b - a. \quad (4)$$

- Considere, agora, uma v.a. qualquer  $X$ , com a distribuição de probabilidades dada abaixo:

$$\begin{array}{l} X : \quad x_1, \quad x_2, \quad \dots, \quad x_n \\ p_j : \quad p_1, \quad p_2, \quad \dots, \quad p_n \end{array}$$

## Simulação de variáveis discretas

- Geramos, agora, um NA  $u$ ; Coloque:

$$X = \begin{cases} x_1, & \text{se } u < p_1, \\ x_2, & \text{se } p_1 \leq u < p_1 + p_2, \\ \dots & \\ x_j, & \text{se } p_1 + \dots + p_{j-1} \leq u < p_1 + \dots + p_j. \end{cases} \quad (5)$$

- Como

$$P(X = x_j) = P(p_1 + \dots + p_{j-1} \leq U < p_1 + \dots + p_j) = p_j,$$

usando (4), vemos que  $X$  tem a distribuição que queremos.

- **Exemplo 4. Simulação de Uma Distribuição de Bernoulli.**
- Suponha que  $X$  tenha uma distribuição de Bernoulli, com  $P(X = 0) = 1 - p = 0,48$  e  $P(X = 1) = p = 0,52$ . Para gerar valores de tal distribuição basta gerar NA  $u$  e concluir:
- Se  $u < 0,48$ , coloque  $X = 0$ ;  
Se  $u \geq 0,48$ , coloque  $X = 1$ .

## Simulação de variáveis discretas

- Por exemplo, suponha que geramos dez NA : 0, 11; 0, 82; 0, 00; 0, 43; 0, 56; 0, 60; 0, 72; 0, 42; 0, 08; 0, 53. Então, os dez valores gerados da distribuição em questão são 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, respectivamente.
- **Exemplo 5. Simulação de Uma Distribuição Binomial.**
- Sabemos que se  $Y \sim b(n, p)$ , então  $Y$  é o número de sucessos num experimento de Bernoulli, com  $n$  repetições, e probabilidade de sucesso  $p$ .
- No Exemplo 4, obtivemos 5 sucessos, logo  $Y = 5$ . Portanto, se  $Y \sim b(10; 0, 52)$ , e quisermos, digamos, gerar 20 valores dessa distribuição, basta considerar 20 experimentos de Bernoulli, sendo que em cada um deles repetimos o experimento  $n = 10$  vezes, com probabilidade de sucesso  $p = 0, 52$ .
- Para cada experimento  $j$  consideramos o número de sucessos (número de 1),  $y_j$ ,  $j = 1, 2, \dots, 20$ . Obteremos, então, os 20 valores simulados  $y_1, \dots, y_{20}$  da v.a.  $Y$ . Observe que esses valores serão inteiros entre 0 e 20, inclusive esses dois valores.

## Simulação de variáveis discretas

● **Exemplo 6. Simulação de Uma Distribuição de Poisson.**

Se  $N \sim P(\lambda)$ , então  $P(N = j) = p_j$  é dada por

$$P(N = j) = \frac{e^{-\lambda} \lambda^j}{j!}, \quad j = 0, 1, \dots \quad (6)$$

- A geração de valores de uma distribuição de Poisson parte da seguinte relação recursiva, que pode ser facilmente verificada:

$$p_{j+1} = \frac{\lambda}{j+1} p_j, \quad j \geq 0. \quad (7)$$

- Seja  $F(j) = P(N \leq j)$  a função de distribuição acumulada ( f.d.a.) de  $N$ .
- Considere  $j$  o valor atual gerado e queremos gerar o valor seguinte. Chamemos simplesmente  $p = p_j$  e  $F = F(j)$ . Então o algoritmo para se gerar os sucessivos valores é o seguinte:

## Simulação de variáveis discretas

Passo 1: Gere o NA  $u$ ;

Passo 2: Faça  $j = 0$ ,  $p = e^{-\lambda}$  e  $F = p$ ;

Passo 3: Se  $u < F$ , coloque  $N = j$ ;

Passo 4: Faça  $p = \frac{\lambda}{j+1}p$ ,  $F = F + p$  e  $j = j + 1$ ;

Passo 5: Volte ao Passo 3.

Note que, no Passo 2, se  $j = 0$ ,  $P(N = 0) = p_0 = e^{-\lambda}$  e  $F(0) = P(N \leq 0) = p_0$ .

## Simulação de variáveis discretas

Suponha, por exemplo, que queiramos simular valores de uma distribuição de Poisson com parâmetro  $\lambda = 2$ . Então  $e^{-2} = 0,136$ . Obtemos:

Passo 1: Geramos  $u = 0,35$ ;

Passo 2:  $j = 0$ ,  $p = 0,136$ ,  $F = 0,136$ ;

Passo 3:  $u > F$ ;

Passo 4:  $p = 2(0,136) = 0,272$ ,  $F = 0,136 + 0,272 = 0,408$ ,  $j = 1$ ;

Passo 5: voltemos ao Passo 3; com  $u < F$ , colocamos  $N = 1$ . Temos, portanto o primeiro valor gerado da distribuição. Prosseguimos com o algoritmo para gerar outros valores.

## Simulação de variáveis discretas

O R e a planilha Excel possuem subrotinas próprias para simular valores de uma dada distribuição de probabilidades. A Tabela 1 traz as distribuições discretas contempladas por cada um e os comandos apropriados.

Tabela 1- Opções de Distribuições Discretas

Distribuição	Excel(Par.)	R(Par.)
Bernoulli	Bernoulli(p)	–
Binomial	Binomial(n,p)	binom(n,p)
Geométrica	–	geom(p)
Hipergeométrica	–	hyper(N,r,k)
Poisson	Poisson( $\lambda$ )	pois( $\lambda$ )

## Simulação de variáveis contínuas

- Considere uma v.a.  $X$ , com função de distribuição acumulada  $F$ , representada na Figura 2. Usando-se um gerador de NA, obtemos um valor  $u$ . Marca-se esse valor no eixo das ordenadas e por meio da função inversa de  $F$  obtém-se o valor  $x$  da v.a.  $X$  no eixo das abscissas. Ou seja, estamos resolvendo a equação

$$F(x) = u, \quad (8)$$

ou  $x = F^{-1}(u)$ . Formalmente, estamos usando o **método da transformação integral**, consubstanciada no seguinte resultado. Suponha  $F$  estritamente crescente.

- **Proposição.** Se  $X$  for uma v.a. com f.d.a  $F$ , então a v.a.  $U = F(X)$  tem distribuição uniforme no intervalo  $[0, 1]$ .
- O resultado pode ser estendido para o caso de  $F$  ser não decrescente, usando-se uma definição mais geral de inversa.

## Simulação de variáveis contínuas

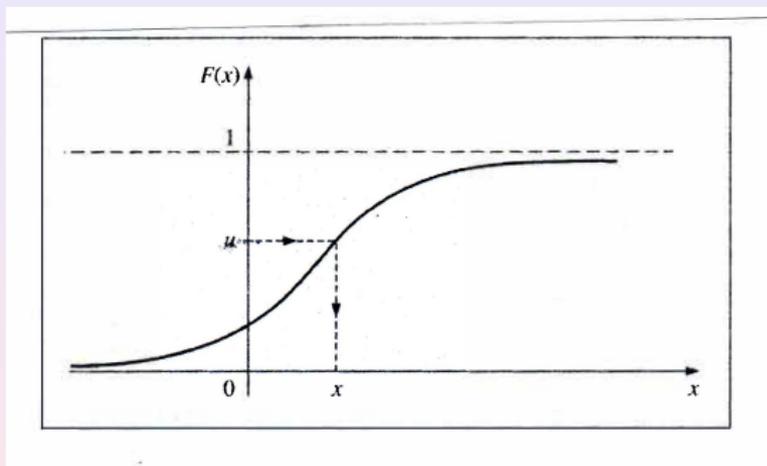


Figura: Função de distribuição acumulada de uma v.a.  $X$ .

## Simulação de variáveis contínuas

● **Exemplo 7.** Simulação de uma Distribuição Exponencial.

Se a v.a.  $T$  tiver densidade dada por

$$f(t) = \frac{1}{\beta} e^{-t/\beta}, \quad t > 0, \quad (9)$$

a sua f.d.a. é dada por

$$F(t) = 1 - e^{-t/\beta}, \quad (10)$$

logo temos que resolver esta equação para gerar  $t$ .

- Tomando logaritmo na base  $e$ , temos

$$1 - u = e^{-t/\beta} \Leftrightarrow \log(1 - u) = -\frac{t}{\beta} \Leftrightarrow t = -\beta \log(1 - u).$$

Logo, gerado um NA, um valor da distribuição  $\text{Exp}(\beta)$  é dado por  $-\beta \log(1 - u)$ .

## Simulação de variáveis contínuas

- Por exemplo, suponha  $\beta = 2$  e queremos gerar 5 valores de  $T \sim \text{Exp}(2)$ .

Gerados os valores

$u_1 = 0,57$ ,  $u_2 = 0,19$ ,  $u_3 = 0,38$ ,  $u_4 = 0,33$ ,  $u_5 = 0,31$  de uma distribuição uniforme em  $(0,1)$  (os números aleatórios), obteremos

$$t_1 = (-2)(\log(0,43)) = 1,68, \quad t_2 = (-2)(\log(81)) = 0,42,$$

$$t_3 = (-2)(\log(0,62)) = 0,96, \quad t_4 = (-2)(\log(0,67)) = 0,80,$$

$$t_5 = (-2)(\log(0,69)) = 0,74.$$

- Podemos reduzir um pouco os cálculos se usarmos o seguinte fato: se  $U \sim \mathcal{U}(0,1)$ , então  $1 - U \sim \mathcal{U}(0,1)$ . Resulta que poderemos gerar os valores de uma exponencial por meio de

$$t = -\beta \log(u).$$

- Usando esta fórmula para os valores de  $U$  acima, obteremos os seguintes valores de  $T$  : 1,12; 3,32; 1,93; 0,96; 2,34.

## Simulação de variáveis contínuas

● **Exemplo 8. Simulação de uma Distribuição Normal.**

Há vários métodos para gerar v.a. normais, mas uma observação importante é que basta gerar uma v.a. normal padrão, pois qualquer outra pode ser obtida desta. De fato, gerado um valor  $z_1$  da v.a.  $Z \sim N(0, 1)$ , para gerar um valor de uma v.a.  $X \sim N(\mu, \sigma^2)$  basta usar a transformação  $z = (x - \mu)/\sigma$  para obter

$$x_1 = \mu + \sigma z_1. \quad (11)$$

- Um método usa a transformação integral e uma tabela de probabilidades para a normal padrão.
- Vejamos um exemplo. Suponha que  $X \sim N(10; 0, 16)$ , ou seja,  $\mu = 10$  e  $\sigma = 0,4$ . Temos que resolver a equação (7), ou seja,

$$\Phi(z) = u,$$

onde estamos usando a notação  $\Phi(z)$  para a f.d.a. da  $N(0, 1)$ .

- Por exemplo, gerado um NA  $u = 0,230$ , temos que resolver

$$\Phi(z) = 0,230,$$

ou seja, temos que encontrar o valor  $z$  tal que a área à sua esquerda, sob a curva normal padrão, seja 0,230. Veja a Figura 3.

## Simulação de variáveis contínuas

Consultando uma tabela para a normal, encontramos que  $z = -0,74$ . Logo o valor gerado da normal em questão satisfaz

$$\frac{x - 10}{0,4} = -0,74,$$

ou seja,  $x = 10 + (0,4)(-0,74) = 9,704$ . Qualquer outro valor pode ser gerado da mesma forma.

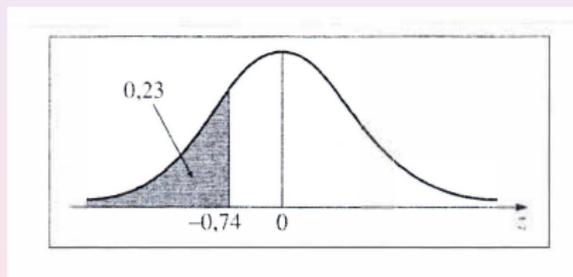


Figura: Geração de um valor  $z \sim N(0,1)$ .

## Simulação de variáveis contínuas

- Há outros métodos mais eficientes. Alguns são variantes do método de Box - Müller (1958).
- Nesse método, são geradas duas v.a.  $Z_1$  e  $Z_2$ , independentes e  $N(0, 1)$ , por meio das transformações

$$\begin{aligned}Z_1 &= \sqrt{-2 \log U_1} \cos(2\pi U_2), \\Z_2 &= \sqrt{-2 \log U_1} \sin(2\pi U_2),\end{aligned}\tag{12}$$

em que  $U_1$  e  $U_2$  são v.a. com distribuição uniforme em  $[0, 1]$ .

- Portanto, basta gerar dois NAs  $u_1$  e  $u_2$  e depois gerar  $z_1$  e  $z_2$  usando (12).
- O método de Box-Müller pode ser computacionalmente ineficiente, pois necessita calcular senos e cossenos. Uma variante, chamado de método polar, evita esses cálculos.

## Simulação de variáveis contínuas

Com o R podemos usar o comando `qnorm`, para obter um quantil de uma distribuição normal, a partir de sua f.d.a. Por exemplo, para gerar 1.000 valores de uma distribuição normal padrão, usamos:

```
u <- runif(1000,0,1) # gera 1000 observações de uma uniforme[0,1]
x <- qnorm(u,mean=0, sd = 1) # Calcula os quantis para o vetor
                               u simulado da uniforme
```

```
par(mfrow=c(1,2))
hist(u, freq=FALSE, main="Histograma da amostra da distribuição
                          Uniforme simulada")
hist(x, freq=FALSE, main="Histograma da variável X simulada
                          a partir do resultado do Teorema 15.1")
```

Os histogramas, da uniforme e da normal, estão na Figura 4.

# Simulação de variáveis contínuas

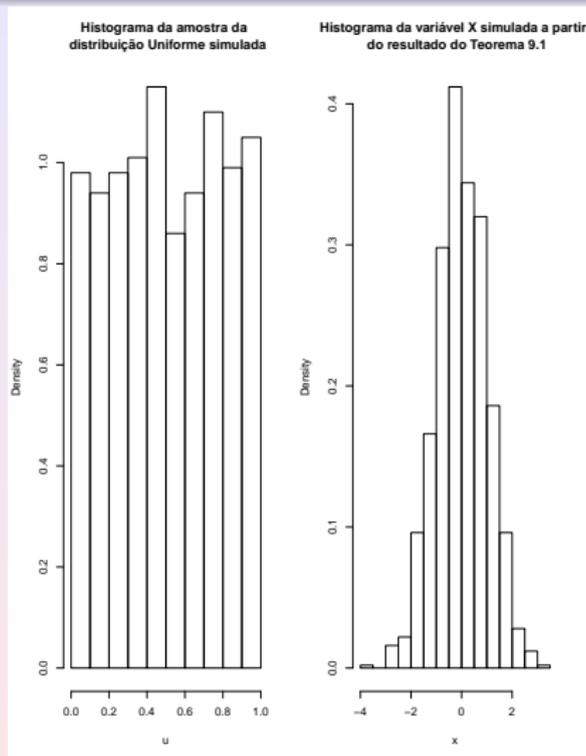


Figura: Simulação de uma normal padrão via R.

## Simulação de variáveis contínuas

O R e a planilha Excel têm subrotinas próprias para gerar muitas das distribuições estudadas nesta seção. A Tabela 2 mostra as opções disponíveis e os comandos apropriados. Além da  $N(0, 1)$  o Excel usa a função INV para gerar algumas outras distribuições contínuas.

Tabela 2: Opções de Distribuições Contínuas

Distribuição	Excel(Par.)	R(Par.)
Normal	Normal(0,1)	norm( $\mu, \sigma$ )
Exponencial	–	exp( $\beta$ )
t(Student)	–	t( $\nu$ )
F(Snedecor)	–	f( $\nu_1, \nu_2$ )
Gama	–	gamma( $\alpha, \beta$ )
Qui-Quadrado	–	chisq( $\nu$ )
beta	–	beta( $\alpha, \beta$ )

## Simulação de vetores aleatórios

- É mais complicado simular distribuições bidimensionais. No caso de  $X$  e  $Y$  serem **independentes**, então

$$f(x, y) = f_X(x)f_Y(y), \quad \forall x, y,$$

se elas forem contínuas, por exemplo. Logo, para gerar um valor  $(x, y)$  da densidade conjunta  $f(x, y)$ , basta gerar a componente  $x$  da distribuição marginal de  $X$  e a componente  $y$  da distribuição marginal de  $Y$ , **independentemente**.

- No caso de v.a. **dependentes**, temos que vale a relação:

$$f(x, y) = f_X(x)f_{Y|X}(y|x).$$

- Logo, por essa relação, primeiramente geramos um valor  $x$  da distribuição marginal de  $X$  e fixado esse valor,  $x_0$ , digamos, geramos um valor da distribuição condicional de  $Y$ , dado que  $X = x_0$ . Isso implica que devemos saber como gerar valores das distribuições  $f_X(x)$  e  $f_{Y|X}(y|x)$ .

## Simulação de vetores aleatórios

**Exemplo 8.** Distribuição Normal Bidimensional.

O método de Box-Müller gera valores de duas normais padrões independentes,  $Z_1$  e  $Z_2$ . Logo, se quisermos gerar valores da distribuição conjunta de  $X$  e  $Y$ , **independentes e normais**, com  $X \sim N(\mu_x, \sigma_x^2)$  e  $Y \sim N(\mu_y, \sigma_y^2)$ , basta considerar

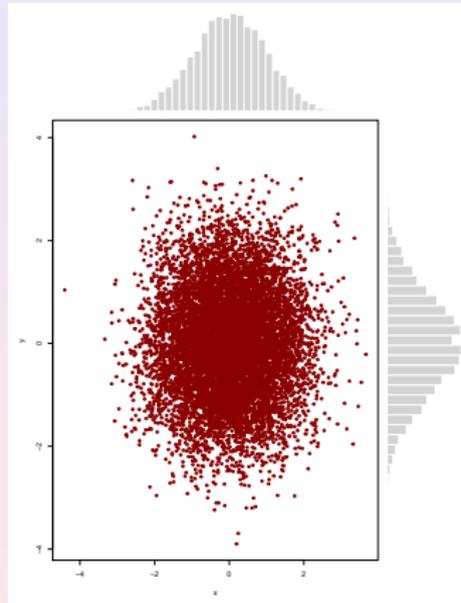
$$X = \mu_x + \sigma_x Z_1, \quad Y = \mu_y + \sigma_y Z_2.$$

O código em R, para o caso de duas normais padrões, seria:

```
u1<-runif(10000,0,1)
u2<-runif(10000,0,1)
P<-data.frame(u1,u2)
x<-qnorm(u1)
y<-qnorm(u2)
```

## Simulação de vetores aleatórios

Na Figura 5 temos os histogramas das duas curvas juntamente com o diagrama de dispersão bidimensional obtidos gerando-se 10.000 valores cada uma das normais padrões independentes.



**Figura:** Simulação de duas normais independentes (nas margens) e gráfico de dispersão.

## Simulação de vetores aleatórios

Na Figura 6 temos a densidade bidimensional normal padrão e as respectivas curvas de nível.

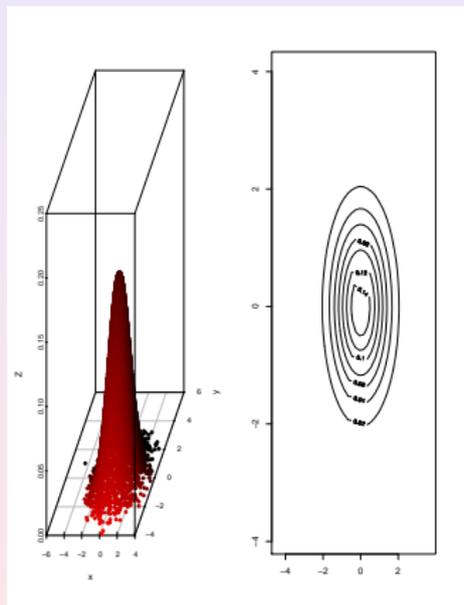


Figura: Distribuição normal padrão bidimensional gerada e curvas de nível.

## Métodos de reamostragem

- Suponha que se queira simular uma amostra da densidade  $\pi$ , mas que isso não seja fácil. O que se pode fazer é proceder em duas etapas.
- Suponha que haja uma densidade  $g$ , da qual seja fácil gerar valores e que esteja próxima de  $\pi$ . Então:
  - (a) na primeira etapa simulamos uma amostra de  $g$ ;
  - (b) na segunda etapa, exercemos um mecanismo de correção, de modo que a amostra de  $g$  seja “direcionada” para tornar-se uma amostra de  $\pi$ .
- Em geral o que se faz é, ao simular um valor de  $g$ , este é aceito com certa probabilidade  $p$  e escolhendo-se  $p$  adequadamente podemos assegurar que o valor aceito seja um valor de  $\pi$ .

## Aceitação – rejeição

- Nesta situação, supomos que exista uma constante finita conhecida  $A$ , tal que  $\pi(x) \leq Ag(x)$ , para todo  $x$ . Ou seja,  $Ag$  serve como um envelope para  $\pi$  (Figura 7).
- **Algoritmo:**
  - [1] Simule  $x^*$  de  $g(x)$ ;
  - [2] simule  $u$  de uma distribuição  $\mathcal{U}(0, 1)$ , independentemente de  $x^*$ ;
  - [3] se  $u \leq \pi(x^*)/Ag(x^*)$ , então aceite  $x^*$  como gerada de  $\pi(x)$ ; caso contrário, volte ao item 1.

## Aceitação – rejeição

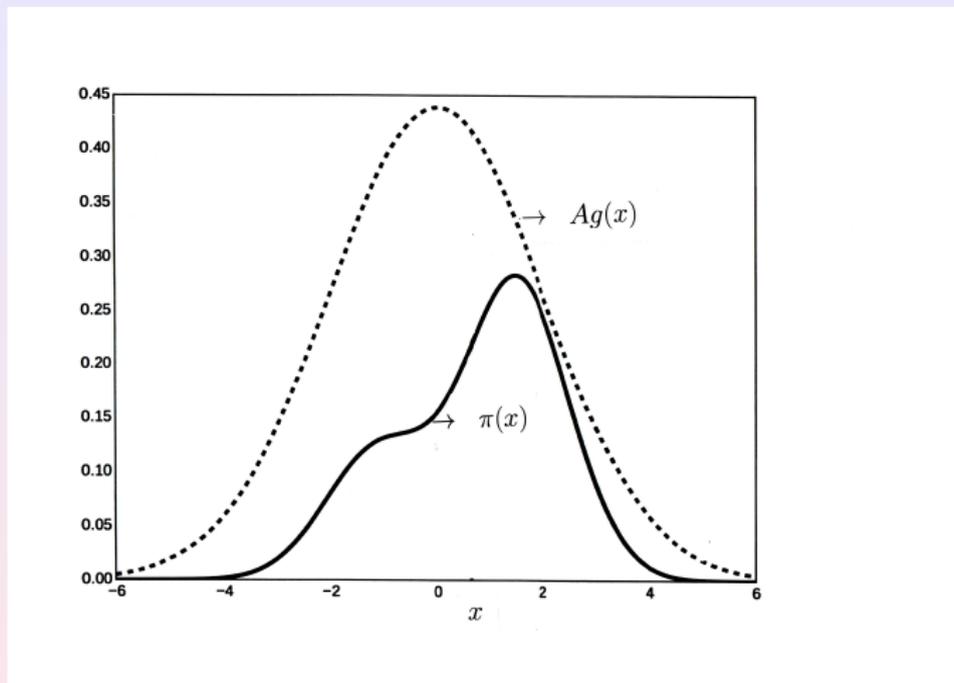


Figura: Densidades  $\pi$  (linha cheia) e  $Ag$  (linha pontilhada).

## Aceitação – rejeição

- Suponha que  $X \sim g(x)$ . Seja  $p(x)$  a probabilidade que  $x$  seja aceito; então,  $p(x) = \pi(x)/Ag(x)$ . Logo,

$$P(X \leq x \text{ e } X \text{ aceito}) = \int_{-\infty}^x p(y)g(y)dy$$

e

$$P(X \text{ aceito}) = \int_{-\infty}^{\infty} p(y)g(y)dy.$$

- Segue que

$$P(X \leq x \mid \text{aceito}) = \frac{\int_{-\infty}^x p(y)g(y)dy}{\int_{-\infty}^{\infty} p(y)g(y)dy} = \int_{-\infty}^x \pi(y)dy,$$

logo os valores aceitos têm realmente distribuição  $\pi(x)$ .

- Também,

$$P(\text{aceitação}) = \int_{-\infty}^{\infty} p(y)g(y)dy = \frac{1}{A} \int_{-\infty}^{\infty} \pi(y)dy = \frac{1}{A}. \quad (13)$$

## Aceitação – rejeição

- Observe que  $\pi$  deve ser conhecida a menos de uma constante de proporcionalidade, ou seja, basta conhecer o que se chama o **núcleo** de  $\pi(x)$ .
- Devemos escolher  $g(x)$  de modo que ela seja facilmente simulável e de sorte que  $\pi(x) \approx Ag(x)$ , pois a chance de rejeição será menor. Também de (13), devemos ter  $A \approx 1$ .
- **Observações:** (a)  $0 < \pi(x^*)/Ag(x^*) \leq 1$ ;  
(b) O número de iterações,  $N$ , necessárias para gerar  $\pi$  é uma v.a. com distribuição geométrica, com probabilidade de sucesso

$$p = P(U \leq \pi(x^*)/Ag(x^*))P(N = n) = (1 - p)^{n-1}p, \quad n \geq 1.$$

Potanto, em média, o número de iterações necessárias é  $E(N) = 1/p$ . Como vimos acima,  $p = 1/A$ , logo  $E(N) = A$ . Logo, é desejável escolher  $g$  de modo que  $A = \sup_x \{\pi(x)/g(x)\}$ .

(c) De (b), podemos dizer que o número esperado de iterações do algoritmo necessárias até que um valor de  $\pi$  seja gerado com sucesso é exatamente o valor da constante  $A$ .

- O método pode ser usado também para o caso de v.a.'s discretas.

## Aceitação – rejeição

- **Exemplo 1.** Considere a densidade de uma Beta(2,2), ou seja,

$$\pi(x) = 6x(1 - x), \quad 0 < x < 1.$$

- Suponha  $g(x) = 1$ ,  $0 < x < 1$ . O máximo de  $\pi(x)$  é 1,5, para  $x = 0,5$ .
- Logo, podemos tomar  $A = 1,5$  (o valor máximo de  $\pi(x)$ , para  $x = 0,5$ ), e teremos  $p = P(\text{aceitação}) = 1/A = 0,667$ .
- Portanto, para obter, por exemplo, uma amostra de tamanho 1.000 de  $\pi(x)$  deveremos simular em torno de 1.500 valores de uma uniforme padrão. Veja a Figura 8.



## Aceitação – rejeição

Um algoritmo equivalente é o seguinte:

- 1) Simule  $x^*$  de  $g(x)$ ;
- 2) Simule  $y^*$  de  $\mathcal{U}(0, Ag(x^*))$ ;
- 3) Aceite  $x^*$  se  $y^* \leq \pi(x^*)$ ; caso contrário, volte a 1.

Usando o código R do capítulo, podemos obter as Figuras 9 e 10; a primeira mostra o histograma dos valores gerados (com a verdadeira curva adicionada) e a segunda mostra as regiões de aceitação e rejeição.

## Aceitação – rejeição

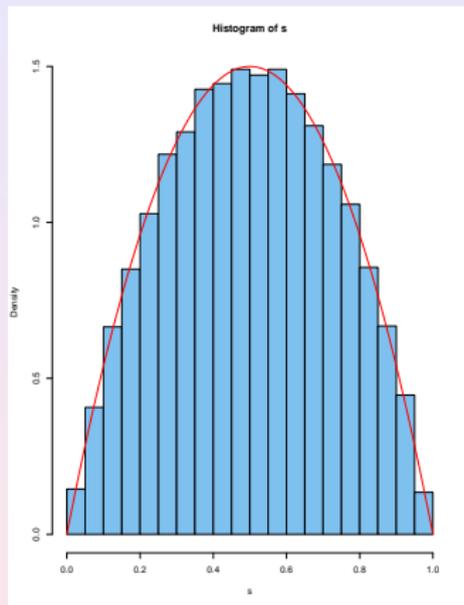


Figura: Histograma dos valores gerados e densidade (vermelho).

## Aceitação – rejeição

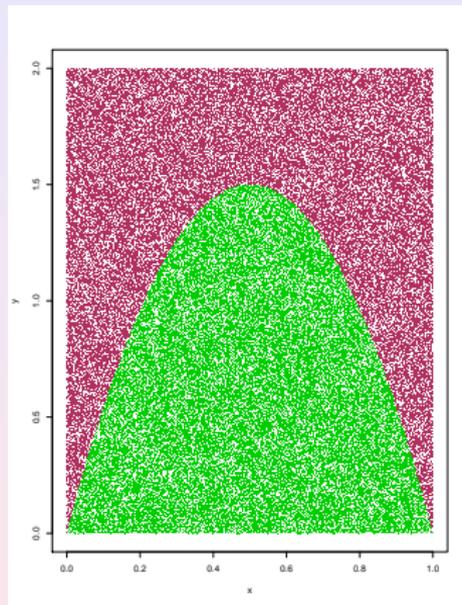


Figura: Região de aceitação (verde) e de rejeição (marrom).

# Reamostragem ponderada

Neste tipo de simulação temos essencialmente as duas etapas anteriores, mas  $Ag$  não precisa ser um envelope para  $\pi$ .

Algoritmo:

- 1) Retire uma amostra  $x_1, \dots, x_n$  de  $g(x)$ ;
- 2) construa os pesos  $w_i$  dados por

$$w_i = \frac{\pi(x_i)/g(x_i)}{\sum_{j=1}^n \pi(x_j)/g(x_j)}, \quad i = 1, 2, \dots, n;$$

- 3) reamostre da distribuição de probabilidades discreta  $(x_i, w_i)$ ,  $i = 1, \dots, n$ .

## Reamostragem ponderada

- Então, a amostra resultante tem distribuição  $\pi$ . De fato, se  $x$  for um valor simulado pelo método,

$$F_x(a) = P(X \leq a) = \sum_{i: x_i \leq a} w_i = \frac{\sum_{i=1}^n (\pi(x_i)/g(x_i)) I_{\{x_i \leq a\}}}{\sum_{j=1}^n (\pi(x_j)/g(x_j))},$$

e o último termo converge, quando  $n \rightarrow \infty$  (lei forte dos grandes números), para

$$\frac{\int (\pi(x)/g(x)) I_{\{x \leq a\}} g(x) dx}{\int (\pi(x)/g(x)) dx} = \frac{\int \pi(x) I_{\{x \leq a\}} dx}{\int \pi(x) dx} = F_\pi(x).$$

- O método de reamostragem ponderada (**importance sampling**) é também usado para reduzir a variância da estimativa MC. Lembremos que o MMC consiste em aproximar  $E_F[h(X)]$  por

$$\hat{E}_F[h(X)] = \frac{1}{n} \sum_{i=1}^n h(X_i). \quad (14)$$

Suponha que em (14)  $F$  tenha densidade  $\pi$ . Então

$$\theta_\pi = E_\pi[h(X)] = \int h(x)\pi(x)dx = \int h(x)\left[\frac{\pi(x)}{g(x)}\right]g(x)dx. \quad (15)$$

## Reamostragem ponderada

- Se chamarmos  $\varphi(x) = h(x)\pi(x)/g(x)$ , teremos

$$\theta_\pi = \int \varphi(x)g(x)dx.$$

- Segue-se que, obtendo-se uma amostra  $x_1, \dots, x_n$  de  $g(x)$ , poderemos estimar (15) por

$$\hat{\theta}_\pi = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) = \frac{1}{n} \sum_{i=1}^n w_i h(x_i), \quad (16)$$

onde  $w_i = \pi(x_i)/g(x_i)$ . Compare com o estimador MC de (14), que é não viesado, ao passo que (16) é viesado.

- Se quisermos um estimador não viesado, basta considerar

$$\theta_\pi^* = \frac{\sum_{i=1}^n w_i h(x_i)}{\sum_{i=1}^n w_i}. \quad (17)$$

- Note que o estimador dá mais peso a regiões onde  $g(x) < \pi(x)$ . Geweke (1989) provou que  $\theta_\pi^* \rightarrow \theta$ , com probabilidade um, se o suporte de  $g(x)$  inclui o suporte de  $\pi(x)$ ,  $X_i \sim \text{iid } g(x)$  e  $E[h(X)] < \infty$ . Mostrou, também, que o erro padrão da estimativa (17) é dado por

$$\frac{[\sum_{i=1}^n [h(x_i) - \theta_\pi^*]^2 w_i^2]^{1/2}}{\sum_{i=1}^n w_i}.$$

## Reamostragem ponderada

- **Exemplo 2.** Num modelo genético, 197 animais distribuem-se em quatro classes  $\mathbf{X} = (x_1, x_2, x_3, x_4)' = (125, 18, 20, 34)'$ , segundo as probabilidades  $(\theta + 2)/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4$ , e o objetivo é estimar  $\theta$ .
- A verossimilhança é

$$L(\theta|\mathbf{X}) \propto (2 + \theta)^{125} (1 - \theta)^{38} \theta^{34}.$$

- Supondo uma priori constante, a densidade a posteriori é

$$p(\theta|\mathbf{X}) \propto (2 + \theta)^{125} (1 - \theta)^{38} \theta^{34}.$$

- Um estimador de  $\theta$  é a média dessa densidade a posteriori, que é estimada por (17). Aqui, suponha que  $g(\theta) \propto \theta^{34} (1 - \theta)^{38}$ ,  $0 < \theta < 1$ , ou seja, temos uma Beta (35,39).

## Reamostragem ponderada

- Portanto,

$$\hat{\theta} = \frac{\sum_{i=1}^n w_i \theta_i}{\sum_{i=1}^n w_i},$$

onde  $\theta_1, \dots, \theta_n$  é uma amostra de  $g(x)$  e os pesos  $w_i$  são dados por

$$w_i = \frac{(2 + \theta_i)^{125}}{\sum_{j=1}^n (2 + \theta_j)^{125}}, \quad i = 1, \dots, n.$$

Gerando-se 10.000 valores de uma Beta(35,39) obtivemos  $\hat{\theta} = 0,6180$ .

## Referências

Box, G.E.P. and Müller, M.E.(1958). A note on the generation of random normal deviates. *The Annals of Statistics*, **29**, 610–611.

Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods*. New York: Wiley.

Kleijnen, J. and Groenendall, W. (1994). *Simulation: A Statistical Perspective*. Chichester: Wiley.

Ross, S.(1997). *Simulation, 2nd Ed.*, New York: Academic Press.

Sobol, I.M.(1976). *Método de Monte Carlo*. Moscow: Editorial MIR.

von Neumann, J.(1951). Various techniques used in connection with random digits, Monte Carlo Method. *U.S. National Bureau of Standards Applied Mathematica Series*, **12**, 36–38.