V.A.s continuas

O paradoxo do contínuo

- Seja X uma v.a. cujos valores possíveis formam um intervalo da reta [a,b]
- Temos uma situação paradoxal:
  - Seja x <u>qualquer</u> valor específico em [a,b]. Por exemplo, x=0.2367123
  - Então P(X=0.2367123) = 0
  - Isto vale para qualquer valor especifico x em [a,b]
  - No entanto, $P(X \in [a,b]) = 1$
  - Assim, todo valor especifico em [a,b] tem probabilidade ZERO de acontecer mas algum numero em [a,b] acontece com certeza
- Isto e’ similar ao paradoxo de uma barra de densidade uniforme ser representada pelo segmento [0,1] ter uma massa 1 kg mas nenhum ponto no segmento [0,1] poder ter massa > 0

Função densidade

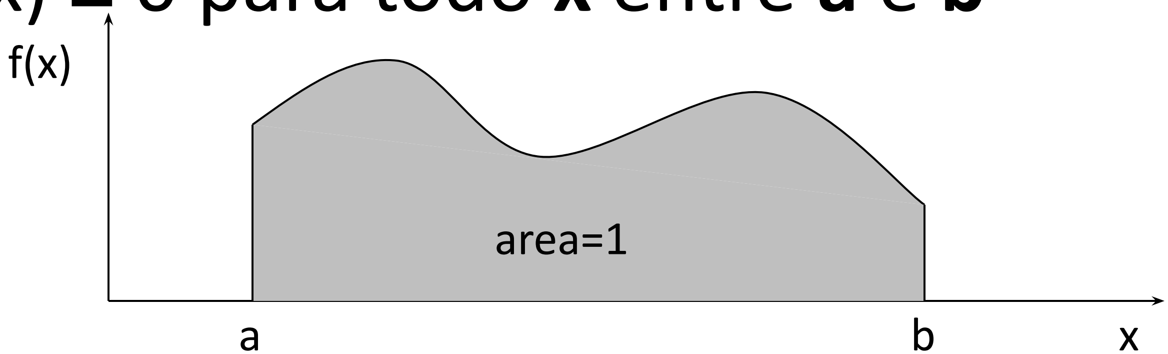
- Como sair dessa situação e trabalhar com v.a.'s continuas?
- A resposta esta na função densidade de probabilidade.
- Uma função f(x) > 0 definida na reta tal que

$$1 = \int_{-\infty}^{+\infty} f(x)dx$$

- QUALQUER função f(x) com esta propriedade representa a densidade de uma v.a.

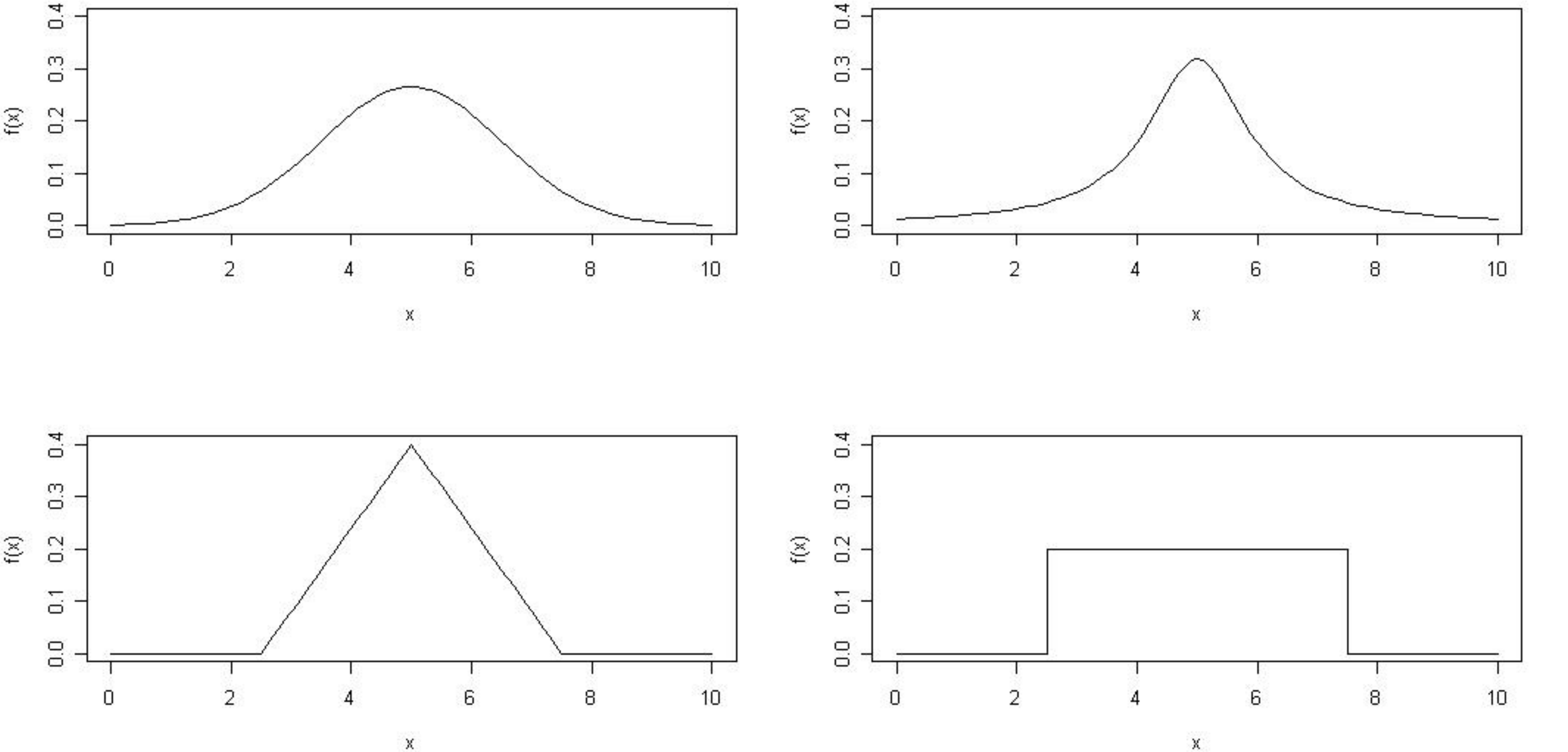
Funcao densidade de probabilidade

- f(x) e’ chamada uma ***funcao densidade de probabilidade*** (na regiao $-\infty \leq a \leq x \leq b \leq \infty$) se ela atende aos requisitos:
- f(x) $\geq$ 0 para todo **x** entre **a** e **b**

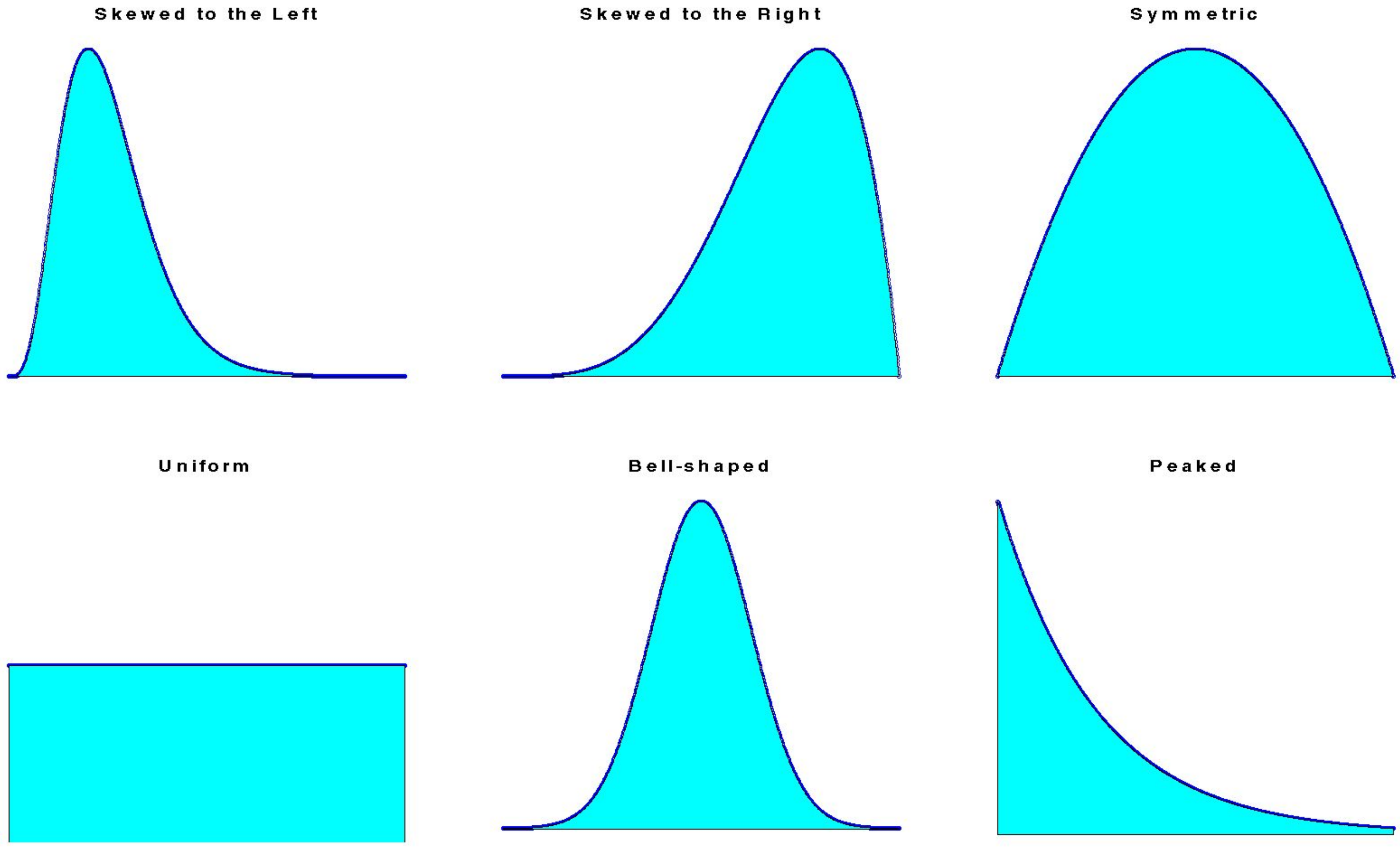


1) A area total sob a curva f(x) entre **a** e **b** e’ 1.0

Exemplos de possíveis shapes para f(x)



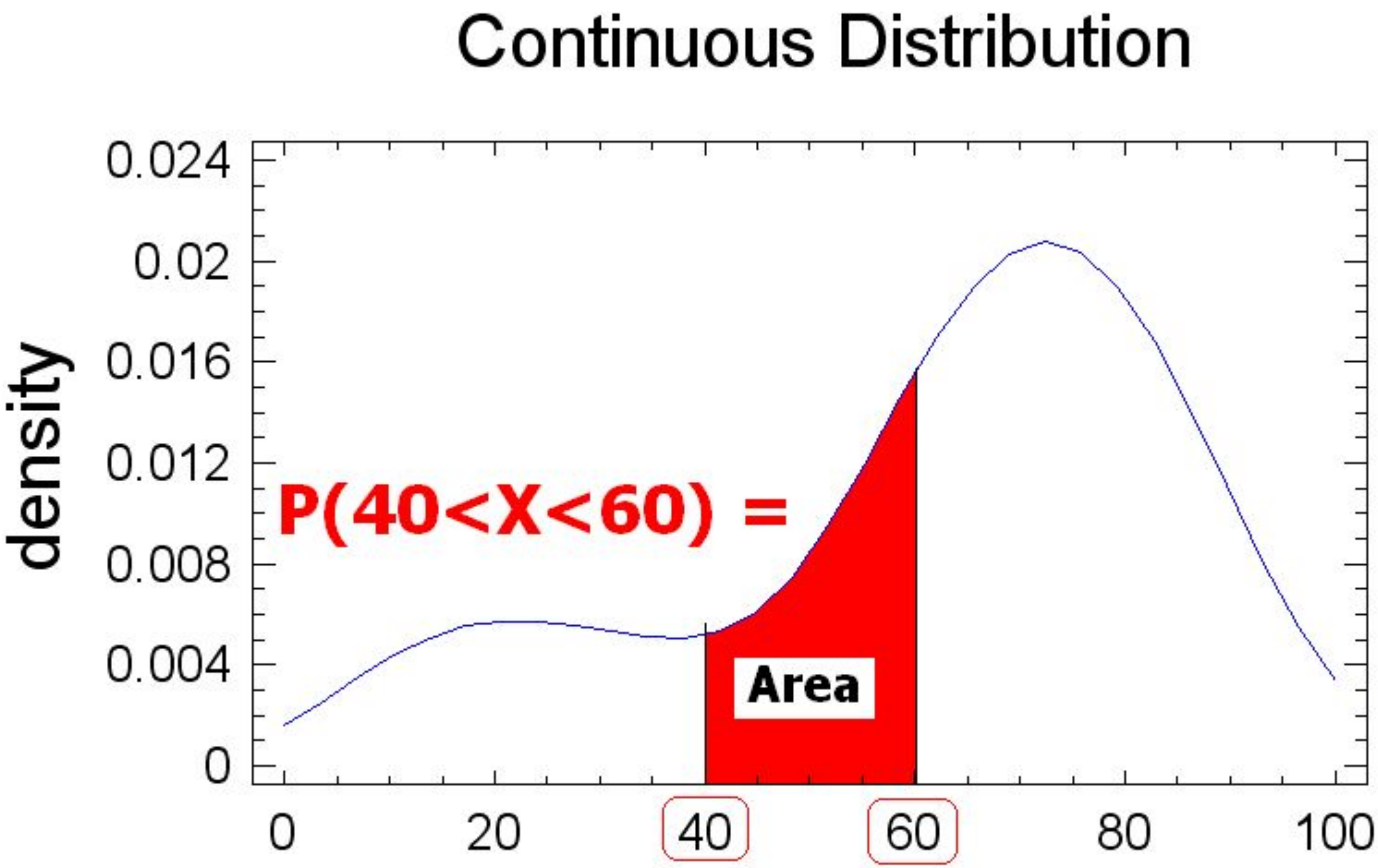
Mais exemplos de shapes para f(x)



Probab = area sob a curva

- Densidade fornece uma maneira visual de calcular probabilidades

$$P(X \in [c, d]) = \int_c^d f(x)dx = \text{area sob a curva}$$

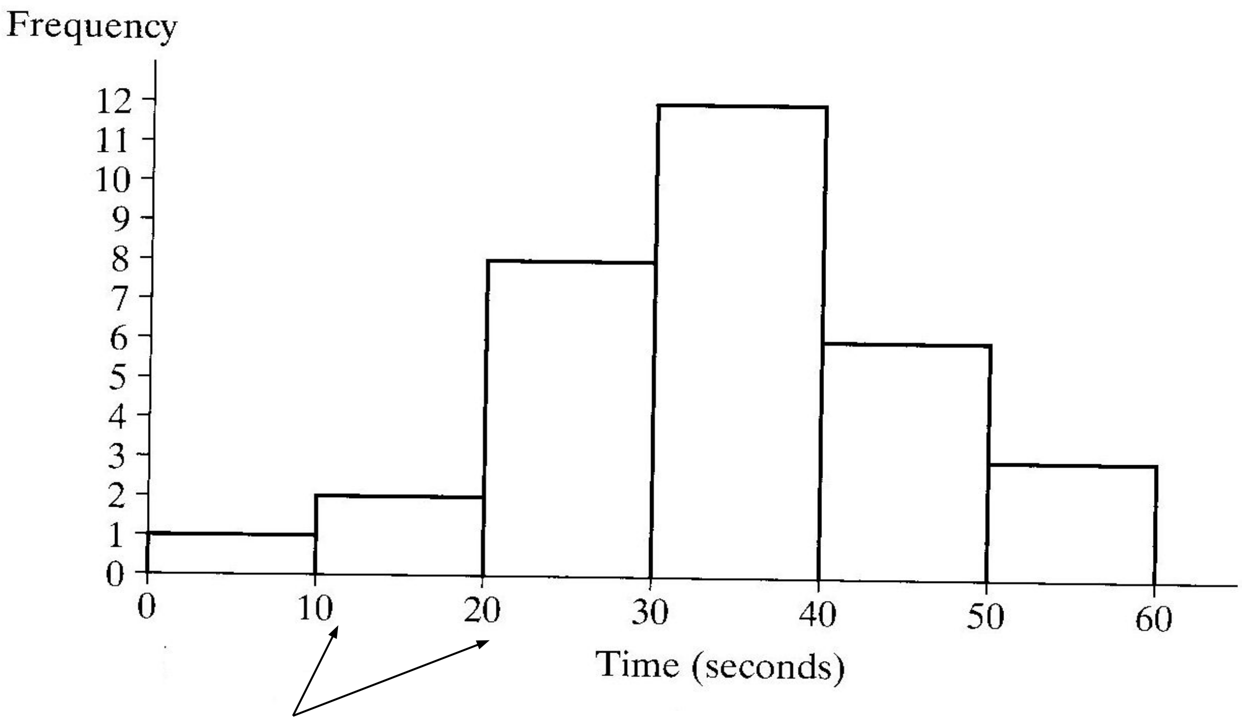

Continuous Distribution
density
0.024
0.02
0.016
0.012
0.008
0.004
0
P(40<X<60) =
Area
0
20
40
60
80
100

x

Identificando uma densidade

- Uma amostra de n instancias
- Medimos v.a.'s continuas nas n instancias
- $X_1, X_2, \ldots, X_n$: copias independentes de uma v.a. X
- Qual a distribuição da v.a. X?
- Podemos ter uma ideia da densidade de X olhando para o histograma dos dados.
- O que e' o histograma?

Histogram



A histogram is made up of a series of bars or rectangles

For continuous data, the class boundaries are written as part of a continuous scale

The area of each rectangle represents the frequency of a class interval.

Histograma normalizado

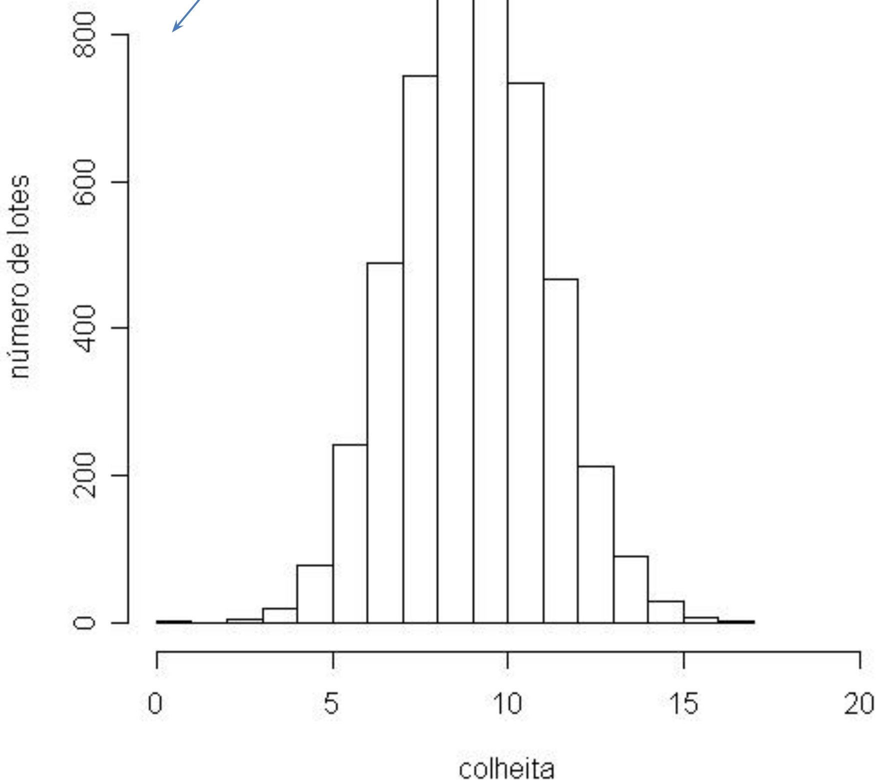
- Normalize a histogram: make the total area of all rectangles equal to 1.
- Se altura do retangulo = frequencia na classe
- Solucao: divida a altura dos retangulos por n*Δ onde Δ e' o comprimento da classe.
- Verifique que a area total agora e' 1:

$$\text{area total} = \sum_i \frac{\#\{X_i \in \text{classe i}\}}{n\Delta} * \Delta = \frac{1}{n} \sum_i \#\{X_i \in \text{classe i}\} = 1$$

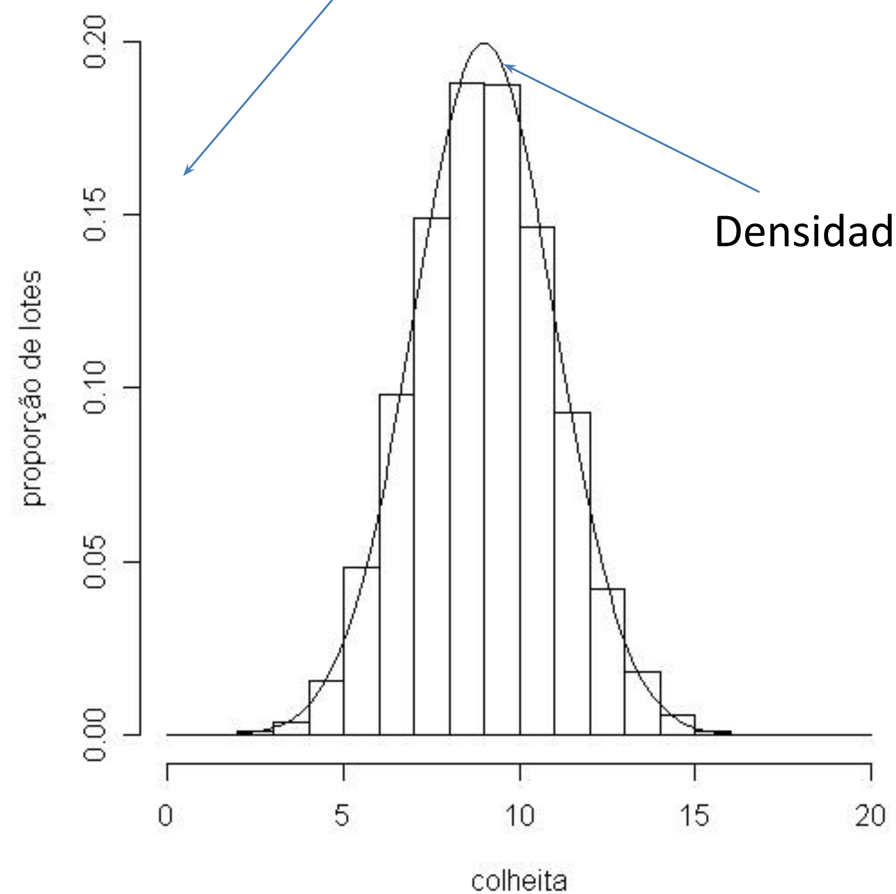
- O histograma vai ser parecido com a densidade subjacente. Por que?

Histograma ≈ densidade

Frequencia absoluta

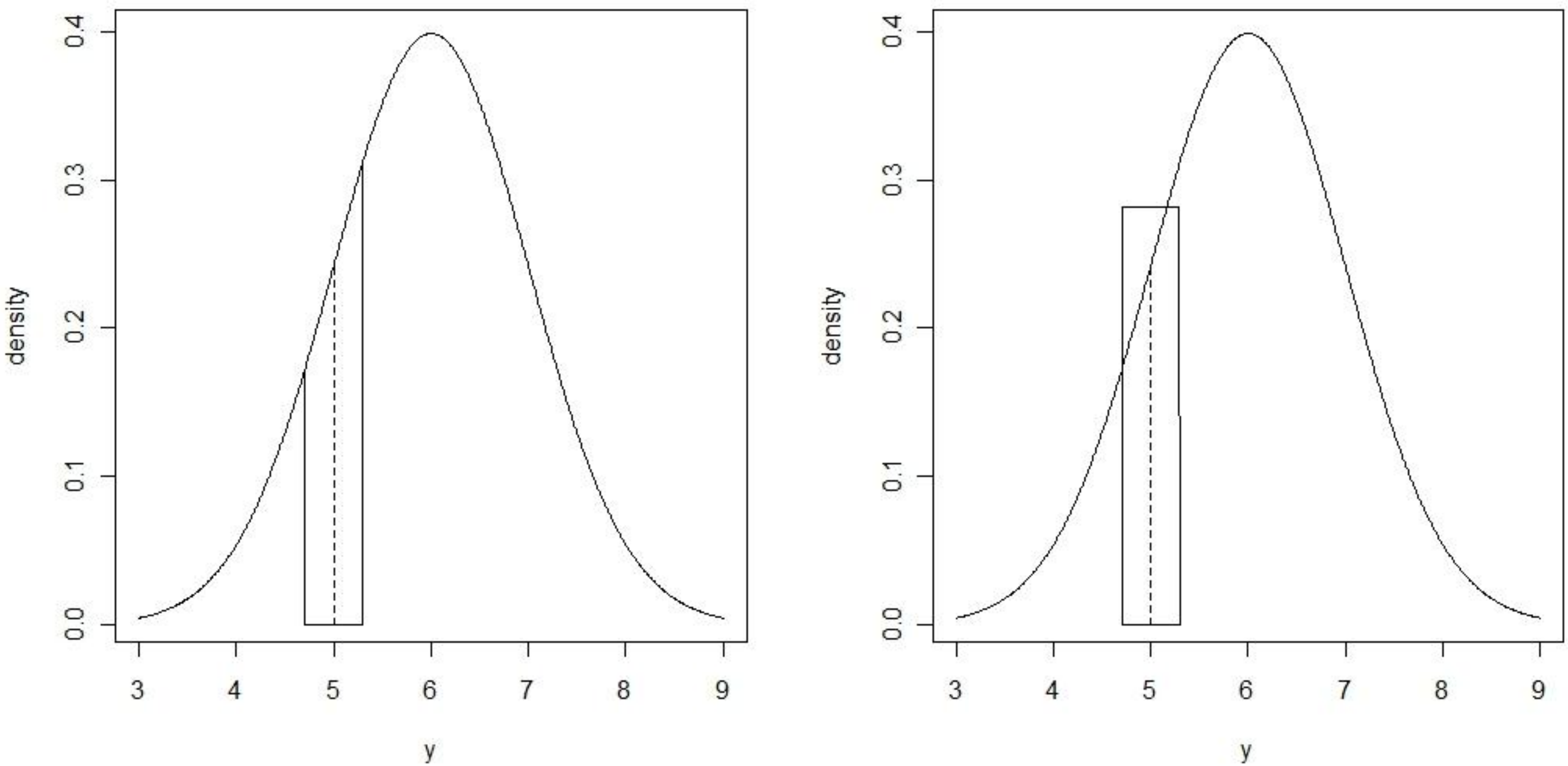


Histograma normalizado



Densidade

Similaridade de histograma e densidade



Histograma ≈ densidade

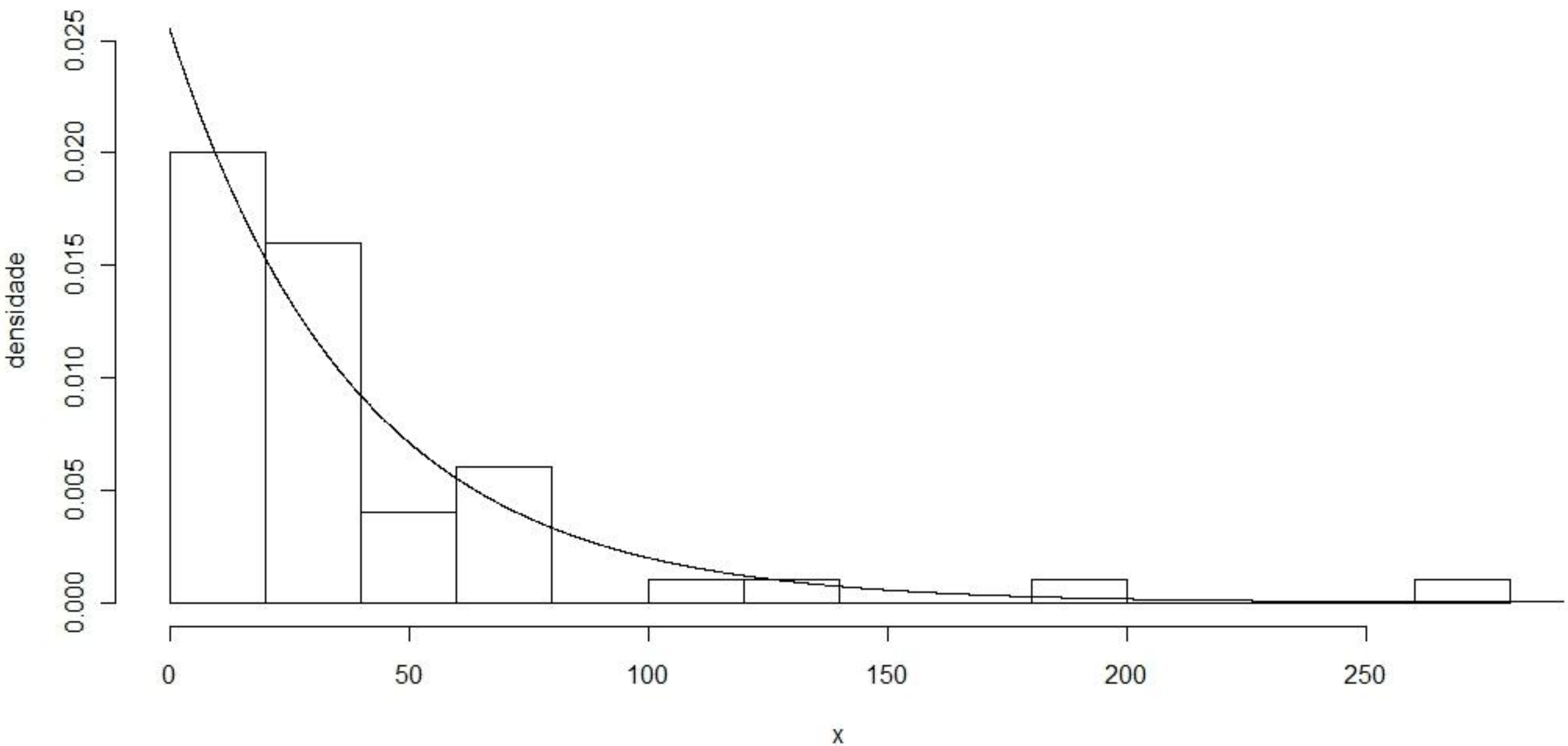
- Histograma normalizado da a dica de qual e' a densidade dos dados.
- Probab de cair num intervalo pequeno centrado em y_0 e de comprimento δ e' aproximadamente

$$\frac{\#\{Y_i's \in (y_0 - \delta/2,\ y_0 + \delta/2)\}}{n}$$

- Mas isto deve ser aproximadamente igual a

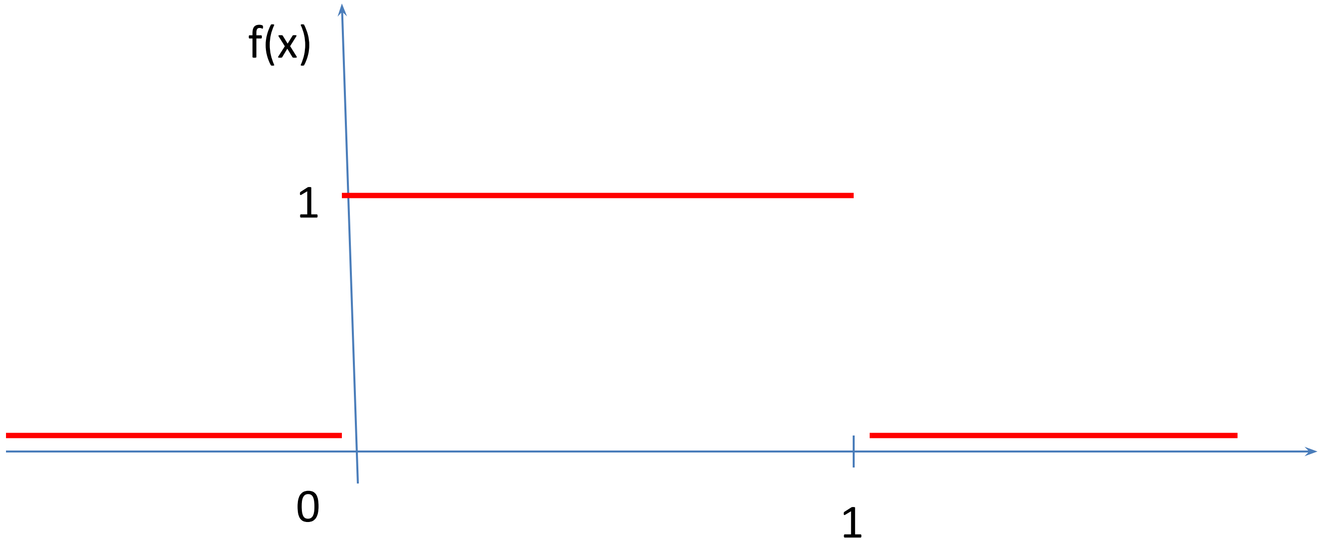
$$\mathbb{P}\left(Y \in (y_0 - \delta/2, y_0 + \delta/2)\right) = \int_{y_0-\delta/2}^{y_0+\delta/2} f^*(y)\ dy \approx f^*(y_0)\delta$$

Exemplo



Distribuição uniforme

- X ~ Uniforme em (0,1): U(0,1)

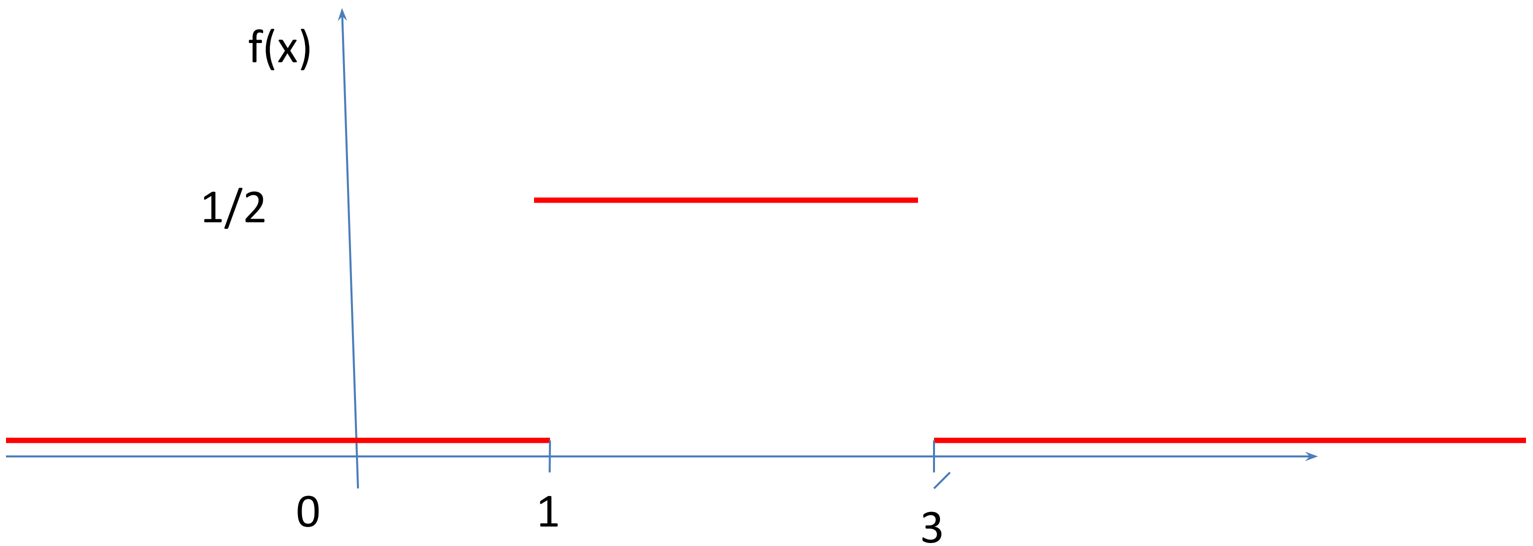


$$P(X \in [c,d]) = \int_c^d f(x)dx = \int_{[c,d]\cap[0,1]} f(x)dx = \int_{[c,d]\cap[0,1]} 1dx =$$

$$= \text{comprimento de}\,[c,d]\cap[0,1]$$

Distribuição uniforme

- Uniforme em (1,3) notação: U(1,3)



$$P(X \in [c,d]) = \int_c^d f(x)dx = \int_{[c,d]\cap[1,3]} f(x)dx = \int_{[c,d]\cap[1,3]} 0.5dx =$$

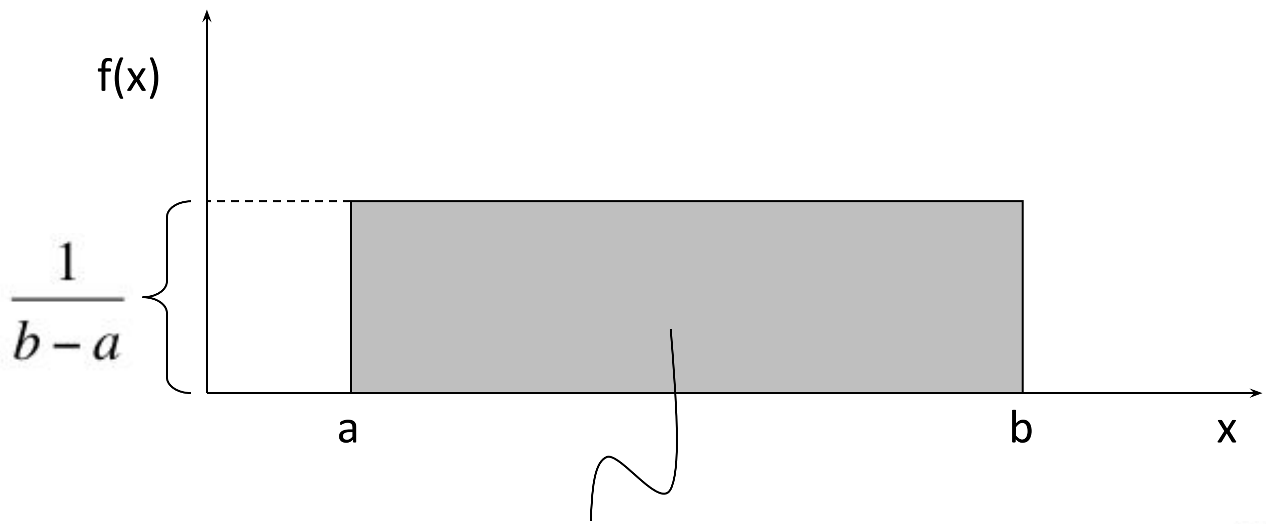
$$= 0.5 * \text{comprimento de } [c,d] \cap [1,3]$$

Distribuicao Uniforme em geral

- X ~ Unif(a,b) se a sua densidade e’ igual a

$$\boxed{f(x) = \frac{1}{b-a}, \ \ \textit{where} \ \ a \le x \le b}$$

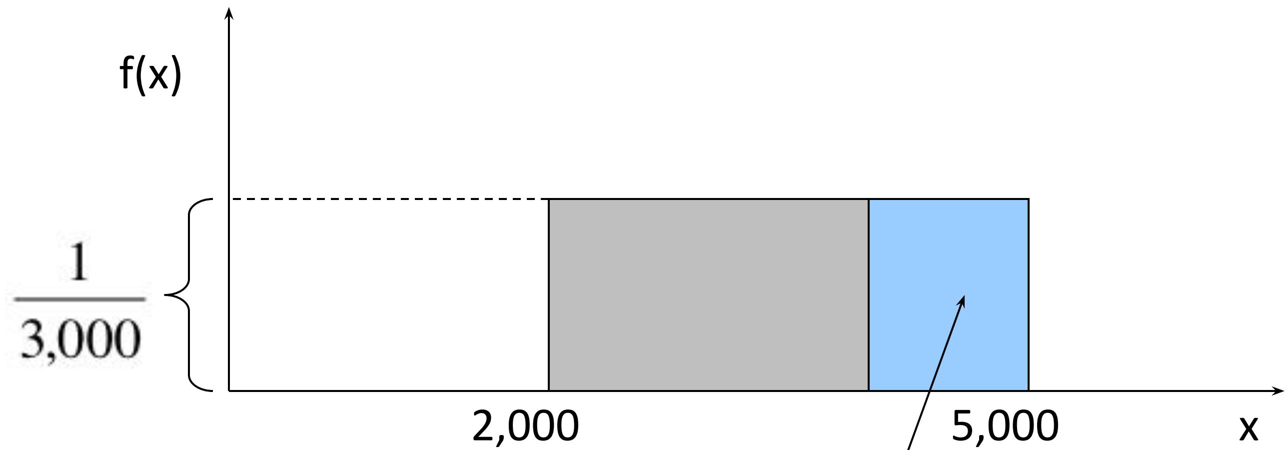
- E’ descrita pela funcao f(x) abaixo:



area = width x height = (b – a) x $\frac{1}{b-a}$ = 1

Exemplo

- The amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.



- ***What is the probability that the service station will sell <u>at least</u> 4,000 gallons?***
- Algebraically: what is **P(X ≥ 4,000)** ?
- **P(X ≥ 4,000)** = (5,000 – 4,000) x (1/3000) = .3333

Exponential Distribution

- Another important continuous distribution is the ***exponential distribution*** which has this probability density function:

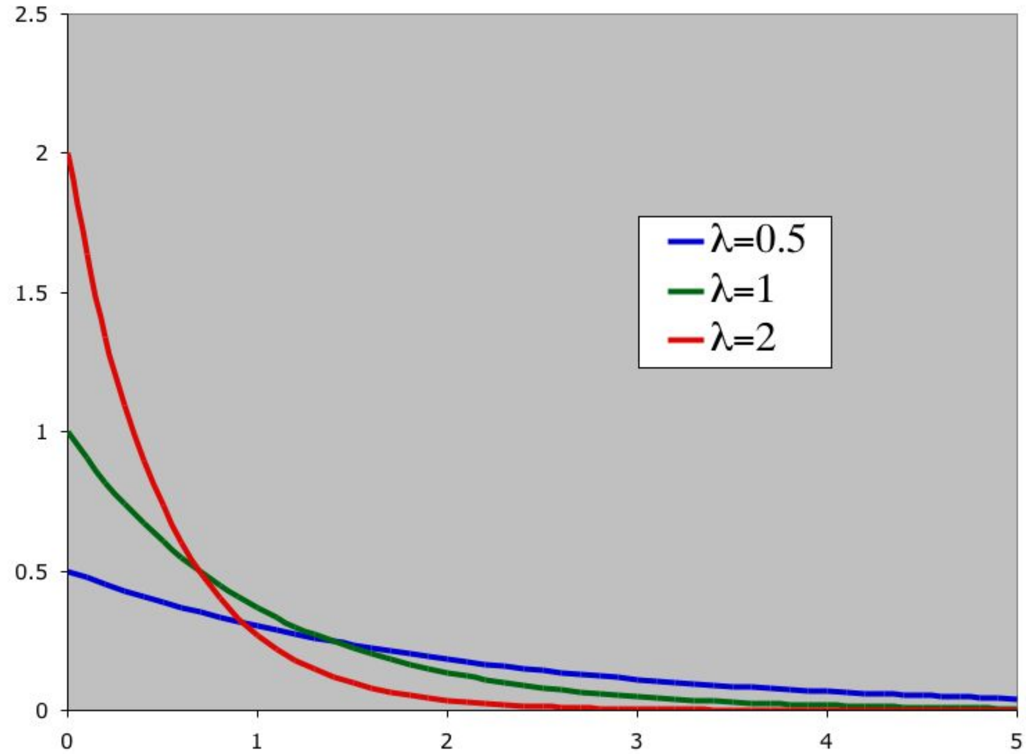
$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

- Note that x ≥ 0.
- λ is a constant > 0
- Time (for example) is a non-negative quantity; the exponential distribution is often used for time related phenomena such as the length of time between phone calls or between parts arriving at an assembly station.

Exponential Distribution

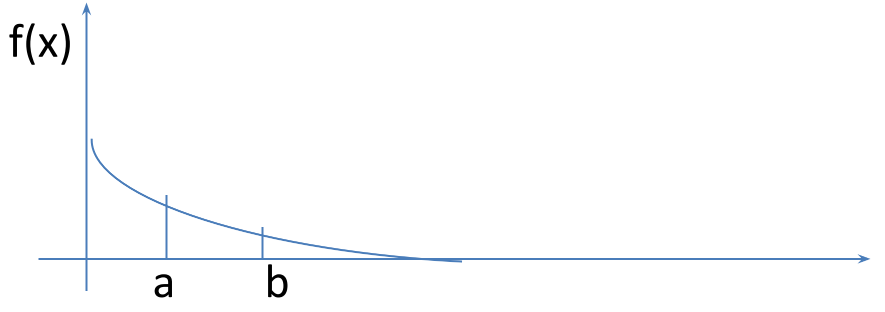
- The exponential distribution depends upon the value of λ
- Smaller values of λ "flatten" the curve increase the chance of larger values

- Exponential
- distributions for
- : λ **5**, **1**, **2**)



Calculando probab com exp(λ)

- Algumas probabilidades (com a > 0 e x > 0):



$$P(X \in (a,b)) = \int_a^b \lambda e^{-\lambda x} dx = e^{-\lambda a} - e^{-\lambda b}$$

$$P(X > x) = \int_x^\infty \lambda e^{-\lambda t} dt = e^{-\lambda x} - e^{-\lambda \infty} = e^{-\lambda x}$$

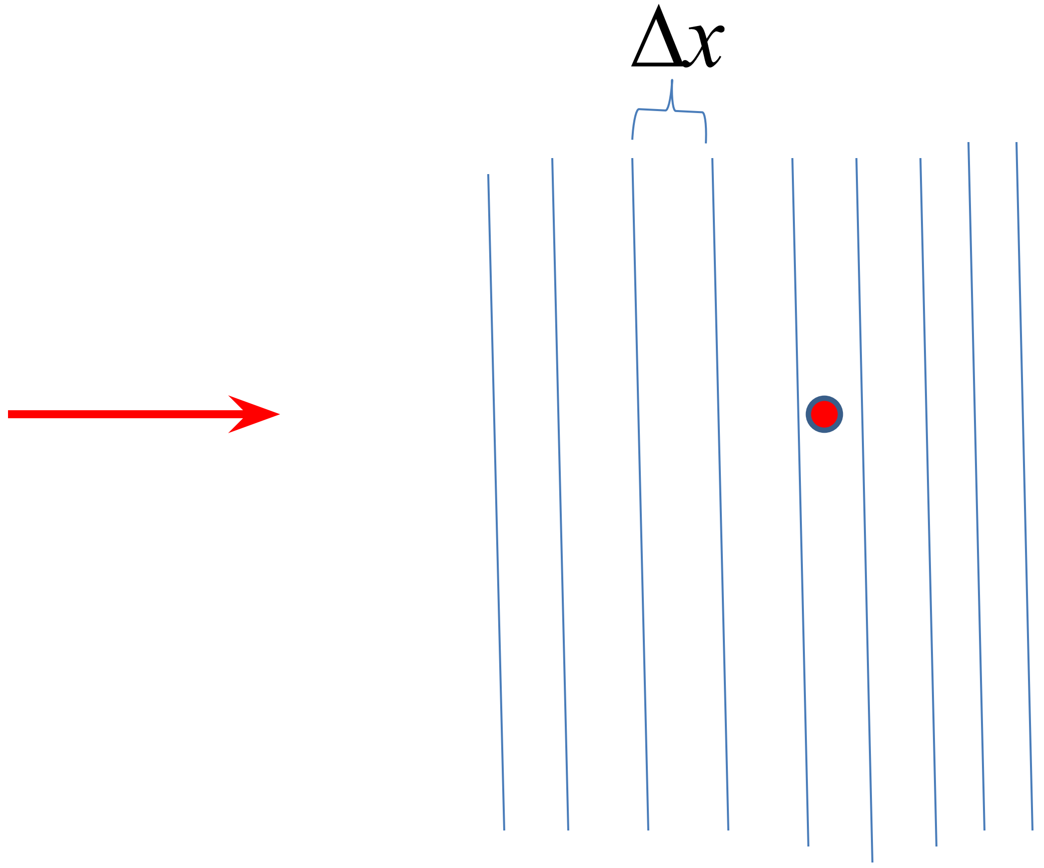
Quando uma v.a. e’ exp(λ)?

- Se X e’ uma v.a. com valores possiveis na semi-reta $(0, \infty)$
- Se $f(x) = \lambda e^{-\lambda x}$
- OU
- Se $P(X > x) = e^{-\lambda x}$ para $x > 0$

- ENTAO $X \sim \exp(\lambda)$

Como aparece a distribuição exponencial?

- Um modelo para a penetração de uma partícula em material metálico



Como aparece a distribuição exponencial?

- Um modelo para a penetração de uma partícula em material metálico
- O metal constituído de placas finas, cada uma com a mesma espessura Δx.
- Se a partícula passa através da *n*-ésima placa:
  - ela possui probabilidade $\rho \, \Delta x$ de ser absorvida na placa *n+1*
  - e probabilidade $1- \rho \, \Delta x$ de passar através da placa *n+1*
- $\rho > 0$ é uma constante que depende do material.
- E' como jogar uma moeda com P(sucesso)= $1- \rho \, \Delta x$ ate' obter o primeiro fracasso ...

Probab(penetrar k placas)

- $P(A_k)$ = P(penetrar pelo menos k placas) = ??

- E’ o mesmo que jogar a moeda para cima k vezes e obter k sucessos sucessivos

- $P(A_k)$ = P(k sucessos sucessivos) = $(1-\rho\Delta x)^k$

Probab(penetrar mais que x)

- P(A_k) = P(penetrar k placas ou mais) = $\left(1-\rho\Delta x\right)^k$
- Suponha que x = certa profundidade FIXA no material (um numero real > 0)
- P(penetrar mais que x) = ??
- Divida o intervalo (0, x) em muitos pequenos intervalos de comprimento Δx
- Isto e', se x = k * Δx k = x/Δx
- Portanto P(penetrar mais que x) ≈ P(penetrar k placas ou mais) =

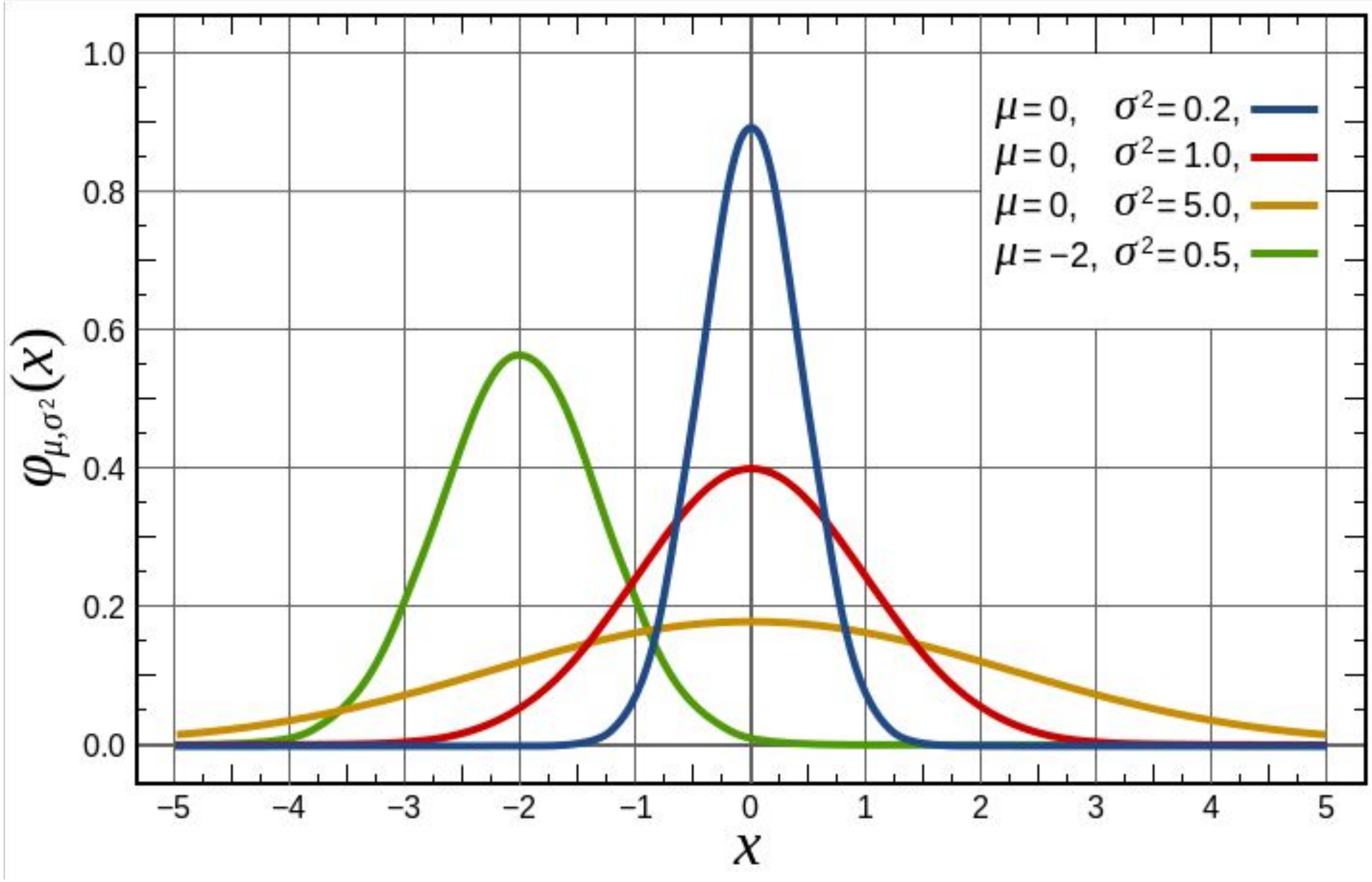
$$\left(1-\rho\Delta x\right)^{x/\Delta x} = \left[\left(1-\rho\Delta x\right)^{1/\Delta x}\right]^x \rightarrow \left[e^{-\rho}\right]^x = e^{-\rho x}$$

Conclusão: X ~ exp(λ)

- Seja X = profundidade penetrada
- Então $P(X > x) = e^{-\rho x}$
- Mas se $X \sim \exp(\rho)$ então $P(X > x) = e^{-\rho x}$
- Isto implica que a profundidade penetrada X e' uma v.a. com distribuição exponencial.

Distribuição normal (ou gaussiana)

- Densidade: $f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

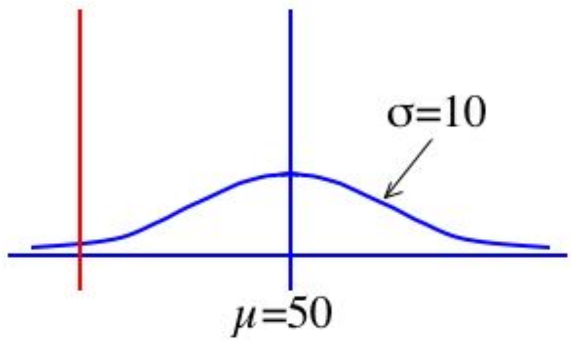


Notação:

$$X \sim N(\mu, \sigma^2)$$

Calculating Normal Probabilities…

- Sempre caimos na normal com mu=0 e sigma=1
- Example: X is ***normally distributed*** with a mean mu = 50 minutes and a standard deviation sigma=10:



- What is the probability that X is between 45 and 60?

- Algebraically speaking, what is **P(45 < X < 60)** ?

Cálculo de probabilidades gaussianas

Seja $X \sim N(\mu = 50, \sigma^2 = 10^2)$

Queremos $\mathbb{P}(40 < X < 60)$

EVENTO:

$$\begin{aligned}
[40 < X < 60] &= \{\omega \text{ tais que } 40 < X(\omega) < 60\} \\
&= \{\omega \text{ tais que } 40 - 50 < X(\omega) - 50 < 60 - 50\} \\
&= \left\{\omega \text{ tais que } \frac{40 - 50}{10} < \frac{X(\omega) - 50}{10} < \frac{60 - 50}{10}\right\}
\end{aligned}$$

Assim:

$$\mathbb{P}(40 < X < 60) = \mathbb{P}(-1 < \tfrac{X - 50}{10} < 1)$$

Cálculo de probabilidades gaussianas

Seja $X \sim N(\mu = 50, \sigma^2 = 10^2)$

Queremos $\mathbb{P}(40 < X < 60)$

EVENTO:

$$\begin{aligned}[40 < X < 60] &= \{\omega \text{ tais que } 40 < X(\omega) < 60\} \\ &= \{\omega \text{ tais que } 40 - 50 < X(\omega) - 50 < 60 - 50\} \\ &= \left\{\omega \text{ tais que } \frac{40-50}{10} < \frac{X(\omega)-50}{10} < \frac{60-50}{10}\right\}\end{aligned}$$

Assim:

$$\mathbb{P}(40 < X < 60) = \mathbb{P}(-1 < \tfrac{X-50}{10} < 1)$$

Seja $X \sim N(\mu = 50, \sigma^2 = 10^2)$

Queremos $\mathbb{P}(40 < X < 60)$

Temos $\mathbb{P}(40 < X < 60) = \mathbb{P}(-1 < \frac{X-50}{10} < 1)$

Isto é, <u>um evento</u> com X pode ser ESCRITO como <u>um evento</u> em termos de (X-mu)/sigma = Z

É o MESMO evento $\longrightarrow$ a probabilidade pode ser calculada com uma ou outra representação

E daÍ? Em que isto ajuda?

Propriedade de Gaussiana

Seja $X \sim N(\mu, \sigma^2)$

Defina a nova v.a. $Z = \frac{X-\mu}{\sigma}$

Por exemplo, se $X \sim N(\mu = 50, \sigma^2 = 10^2)$
então $Z = \frac{X-50}{10}$

Toda vez que um valor aleatório $X(\omega)$ for medido o outro valor (também aleatório) $Z(\omega) = \frac{X(\omega)-50}{10}$

Propriedade de Gaussiana

Seja $X \sim N(\mu, \sigma^2)$

Defina a nova v.a. $Z = \dfrac{X-\mu}{\sigma}$

TEOREMA: A v.a. Z possui distribuição N(0,1)

Assim, se X é uma gaussiana genérica $X \sim N(\mu, \sigma^2)$
a nova v.a. $Z = \frac{X-\mu}{\sigma}$ possui distribuição N(0,1)

Z é chamada de normal padronizada.

Problema numérico

Probabilidade =área sob a curva densidade =integral

Não temos fórmula exata para a integral da densidade normal

Cálculo numérico baseado na densidade N(0,1)

Para qualquer outra gaussiana, calculamos o evento equivalente em termos da N(0,1)

Na prática, em Python

Isto tudo é transparente para o usuário nos dias de hoje.

```
import numpy as np
import scipy.stats as stats

# p1 = P(N(50, 10^2) < 60)
p1 = stats.norm.cdf(60, 50, 10)

# método  cdf retorna apenas  probabilidade de X <= a
# Não retorna um intervalo diretamente
```

Outra probabilidade

E se quisermos P(a < X < b)?
Por exemplo, P(45 < X < 65) se X~N(mu, sigma^2)

Truque: Sejam

$$A = [X \leq 40] = \{\omega \text{ tais que } X(\omega) \leq 40\}$$
$$B = [40 < X < 60] = \{\omega \text{ tais que } 40 < X(\omega) < 60\}$$
$$C = [X < 60] = \{\omega \text{ tais que } X(\omega) < 60\}$$

Veja que C = A U B e que A e B são disjuntos (pois você não pode uma medição X(w) em A e B simultaneamente.

Então P(C) = P(A)+P(B) $\longrightarrow$ P(B) = P(C) - P(A)

Em Python

```
import numpy as np
import scipy.stats as stats

# p1 = P(N(mu, s^2) < x) = ??
p1 = stats.norm.cdf(x,  mu, s)

# Então
# P( 40 < N(50, 10^2) < 60) = P(N(50, 10^2) < 60) - P(N(50, 10^2) < 40)
# Isto é,
p2 =  stats.norm.cdf(60, 50, 10) - stats.norm.cdf(40, 50, 10)
```

Beta

Existem vários fenômenos cujas variáveis de interesse tem seus valores limitados acima e abaixo por números conhecidos a e b. Um exemplo típico é constituído por dados que aparecam em forma de proporção:

i) em cada empresa industrial brasileira(com produção acima de certo valor fixo) é medido o vetor (X_1,X_2) onde

X_1= proporção de gastos em salários na produção total durante 1985

X_2= proporção de gastos em energia na produção total durante 1985.

ii) a razão entre o comprimento do fêmur e o comprimento total da perna de um indivíduo.

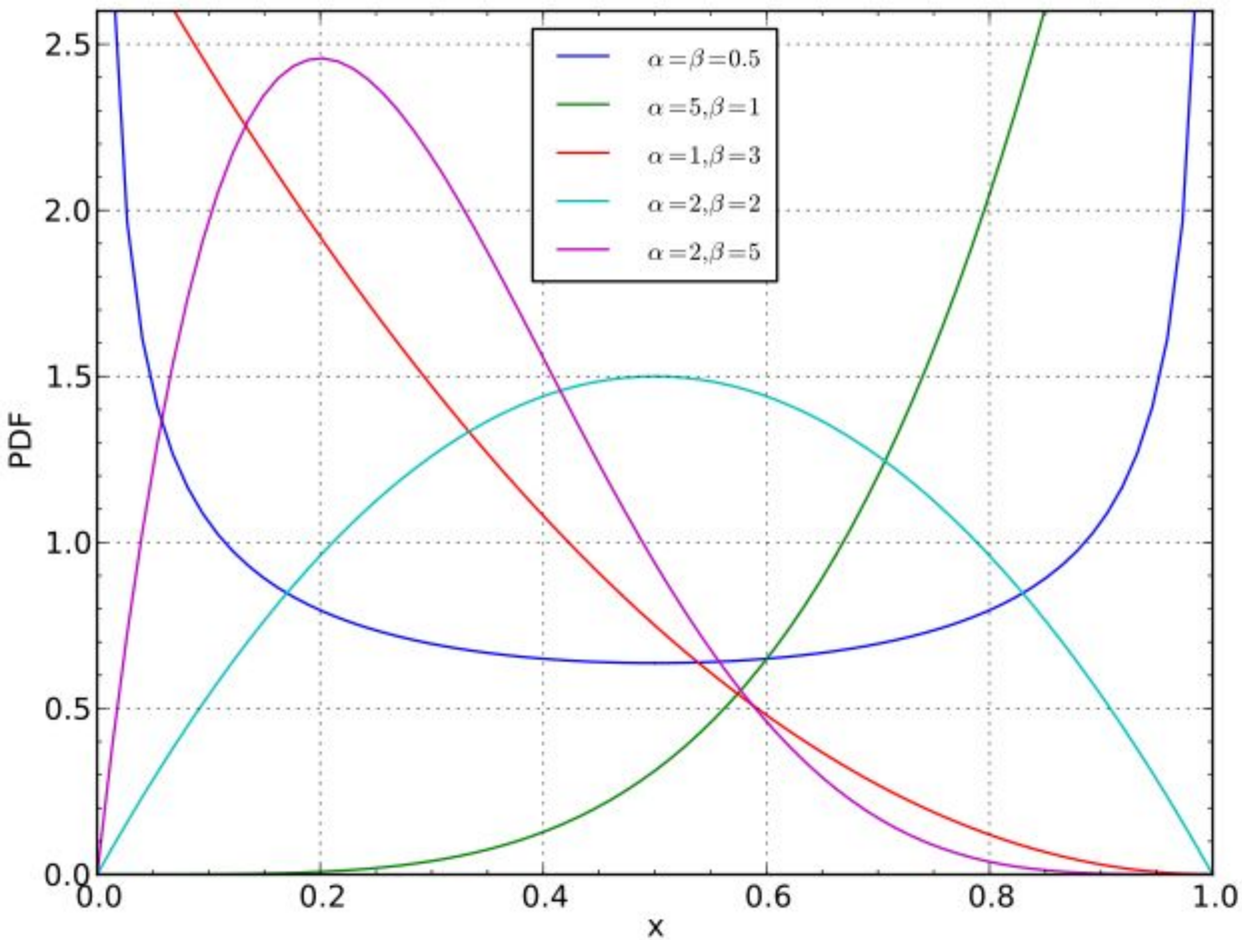
Uma classe de distribuições que inclue a distribuição uniforme e é rica o suficiente para fornecer modelos para a maioria das variáveis aleatórias que têm valores limitados é a classe de distribuições beta.

A variável aleatória Y tem distribuição beta com parâmetros r e s ($Y \sim \beta(r,s)$) onde r >0 e s >0 se Y tem densidade dada por

$$f(y) = \begin{cases} \frac{y^{r-1}(1-y)^{s-1}}{B(r,s)} & se \quad 0 \leq y \leq 1 \\ 0 & c.c. \end{cases} \tag{3.3}$$

onde $B(r,s) = \int_0^1 y^{r-1}(1-y)^{s-1} dy$ é a constante que torna f(y) uma densidade. B(r,s) é chamada

Flexibilidade da beta



Variação do preço de ações

Exemplo: Vamos representar por Y a variável que mede a proporção da variação de preços(médios) diários de ações quando estes preços caem. Isto é, se o preço(médio) de uma ação num dia é d_1e o preço desta ação no dia seguinte é $d_2 \leq d_1$ então $y = \dfrac{d_2 - d_1}{d_1}$. Se o preço sobe ($d_1 < d_2$), Y não é registrada para aquela ação. Um total de 2314 ações com preços que caíram foram observadas. Os dados estão resumidos na forma de distribuição de freqüência na tabela 10 e a figura 16 apresenta um histograma junto com a densidade de uma distribuição beta com parâmetros r=1,038 e s=10,63 . A maneira de escolher estes valores será assunto de um próximo capítulo. Nesta figura parece qeu a distribuição beta fornece um modelo razoável para a variável Y.

Histograma e densidade