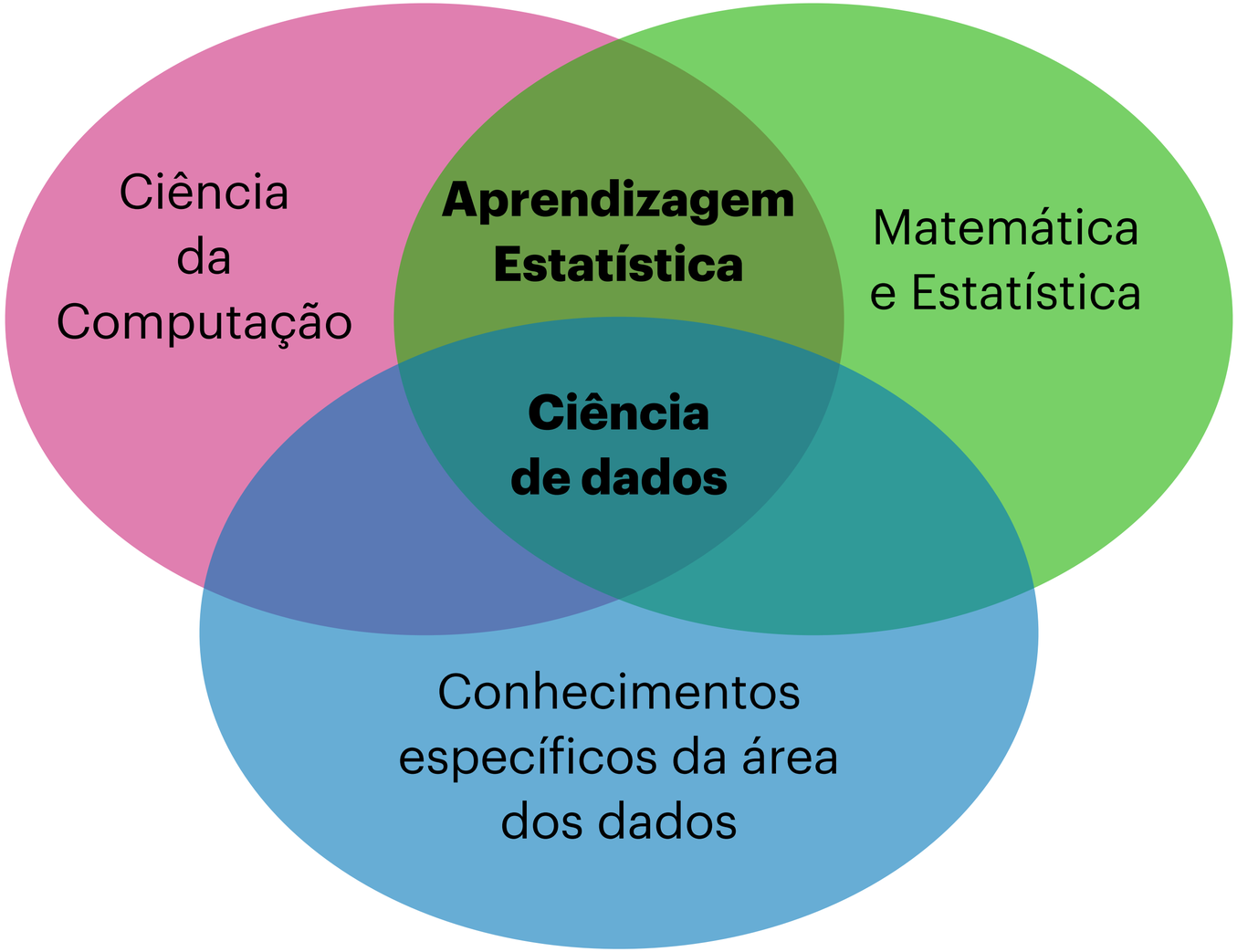
Aprendizagem estatística em altas dimensões

Florencia Leonardi

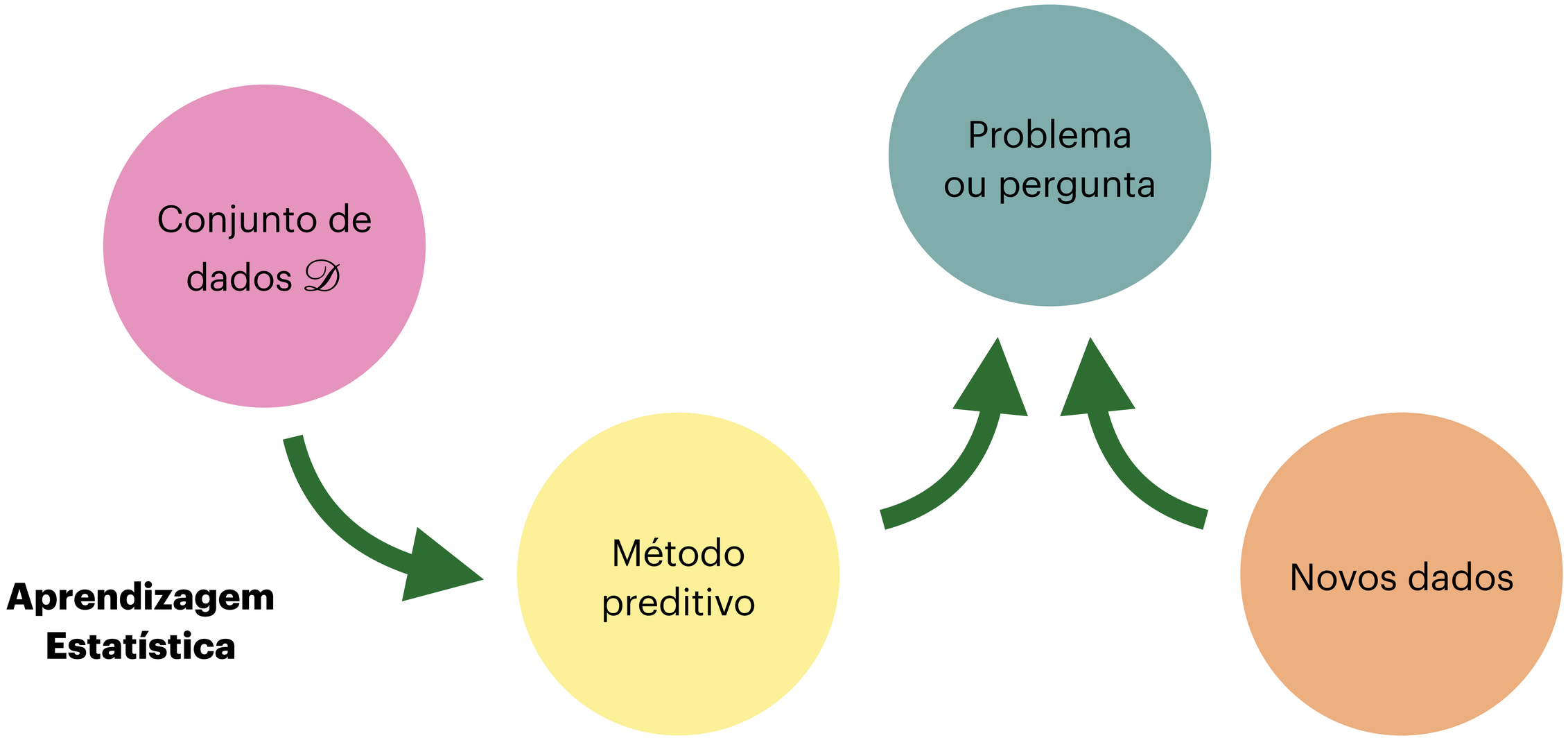
Conteúdo

- Visão geral da aprendizagem estatística
- Revisão do conceito de variáveis aleatórias, esperança e variância
- Estimadores da esperança e variância
- Viés de um estimador
- Erro quadrático médio
- Principais problemas na aprendizagem estatística supervisionada

Aprendizagem estatística



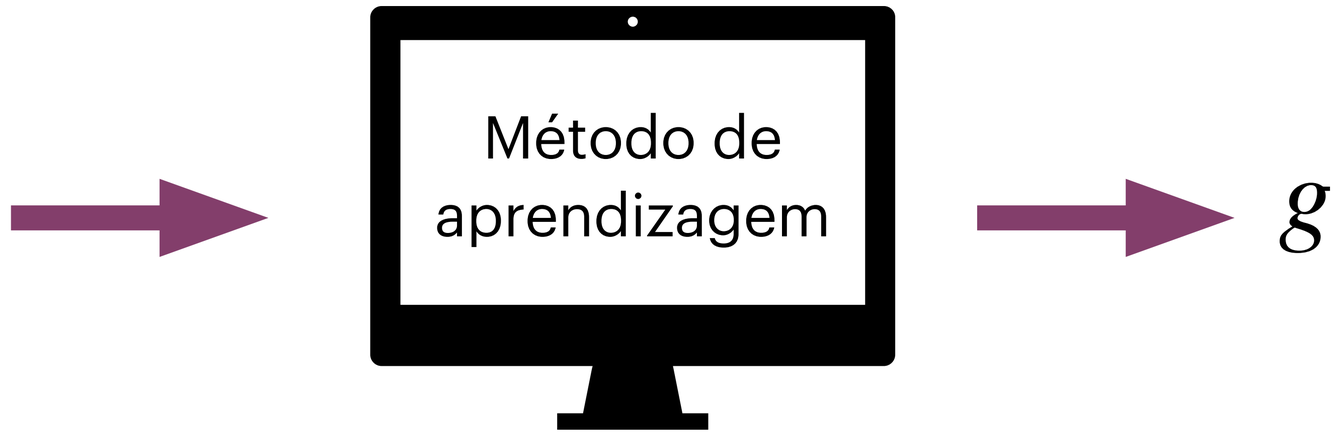
Aprendizagem estatística



Aprendizagem estatística supervisionada

Aprendizagem

$$\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$



$$g$$

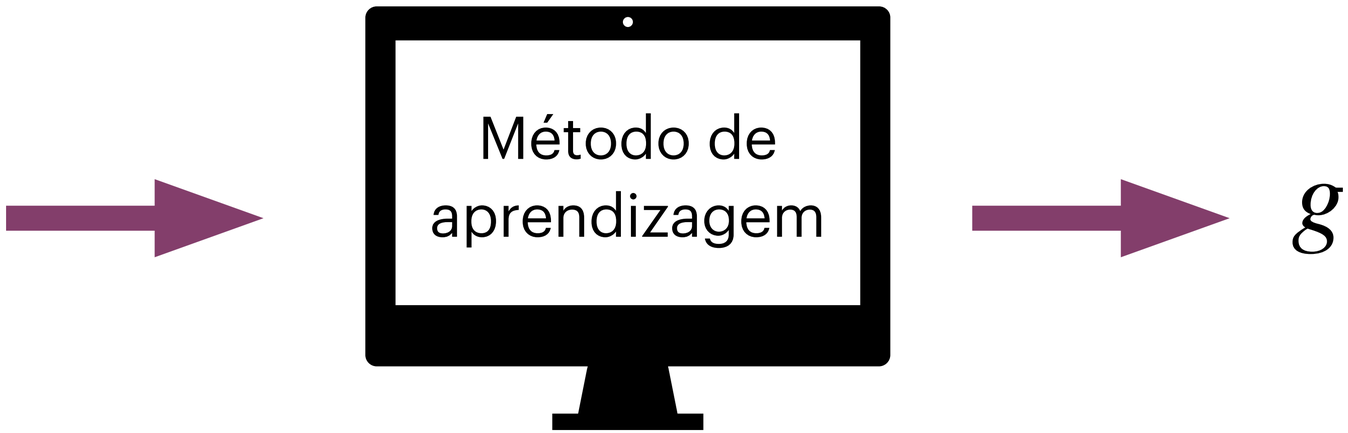
Predição

Novos dados: x → $\hat{y} = g(x)$

Aprendizagem estatística não supervisionada

Aprendizagem

$$\mathscr{D} = \{x_1, \ldots, x_n\}$$



$$g$$

Predição

Novos dados: x → $g(x)$

Aprendizagem estatística supervisionada

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$$x_i \in \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$$

$$y_i \in \mathcal{Y}$$

Predição

$$x \rightarrow \hat{y} = g(x)$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$y_i \in \mathcal{Y} = \{c_1, \ldots, c_K\} \rightarrow$$ **Problema de classificação**

$$y_i \in \mathcal{Y} = \mathbb{R} \rightarrow$$ **Problema de regressão**

Aprendizagem estatística supervisionada

$$x_i \in \mathcal{X} = \mathbb{R}^{16\times16}$$

$$y_i \in \mathcal{Y} = \{0,1,\ldots,9\}$$

Problema de classificação

Aprendizagem estatística supervisionada

Price	Condo	Size	Rooms		Negotiation type
930	220	47	2		rent
1000	148	45	2		rent
1000	100	48	2	...	rent
990000	870	121	3		sale
410000	630	51	2		sale
820000	1000	109	3		sale

$$y_i \in \mathcal{Y} = \mathbb{R}$$

Problema de regressão

$x_i \in \mathcal{X}_1 \times \ldots, \times \mathcal{X}_{15}$ (variáveis contínuas e categóricas)

Aprendizagem estatística supervisionada

- Assumiremos que os dados $\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ são realizações independentes de uma distribuição de probabilidade conjunta $p(x, y)$

- Na ciência de dados em geral e na aprendizagem estatística em particular tenta-se assumir o mínimo de hipóteses possíveis sobre a forma específica de $p(x, y)$

- Mas muitas vezes, para poder estudar as propriedades teóricas dos métodos de aprendizagem estatística, precisamos assumir certas hipóteses adicionais sobre o "modelo" teórico que gerou os dados, neste caso dado por $p(x, y)$

Variáveis aleatórias

- Uma variável aleatória é um objeto matemático para descrever as realizações (observações) de experimentos aleatórios
- As variáveis aleatórias podem assumir valores num conjunto contínuo, por exemplo $\mathbb{R}$ ou o intervalo $[0,1]$, ou num conjunto enumerável como por exemplo $\mathbb{N}$ ou $\{0,1,\ldots,K-1\}$
- As variáveis aleatórias são descritas completamente pela sua função densidade de probabilidade (também chamada de função de probabilidade no caso discreto), que nos indica quais regiões ou valores tem maior probabilidade de ocorrência para essa variável
- Tanto no caso contínuo quanto no caso discreto usaremos a notação $p(x)$ para a função de densidade ou a função de probabilidade de uma variável aleatória X

Variáveis aleatórias

Suponhamos que temos este pote com 30 balas de morango e 70 balas de limão

Escolhemos uma bala sem olhar, e depois anotamos o sabor dela, com o seguinte código:

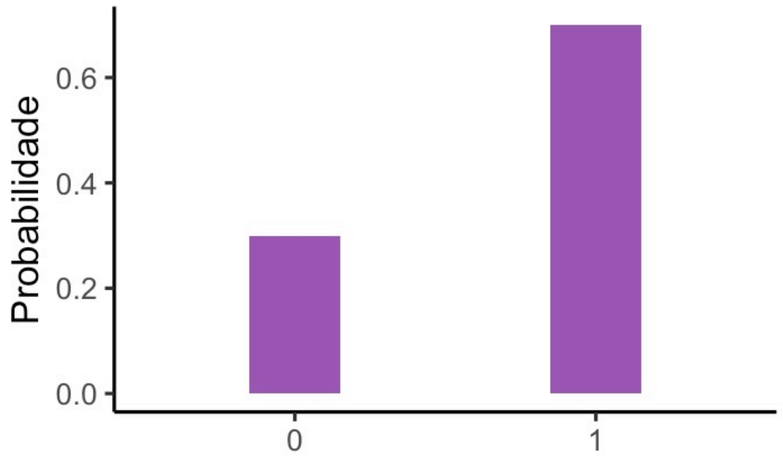
0: se for de sabor morango
1: se for de sabor limão

Qual é a probabilidade dessa bala ser de limão, ou seja termos como resultado o código 1?

Variáveis aleatórias

$$X \in \{0,1\}$$

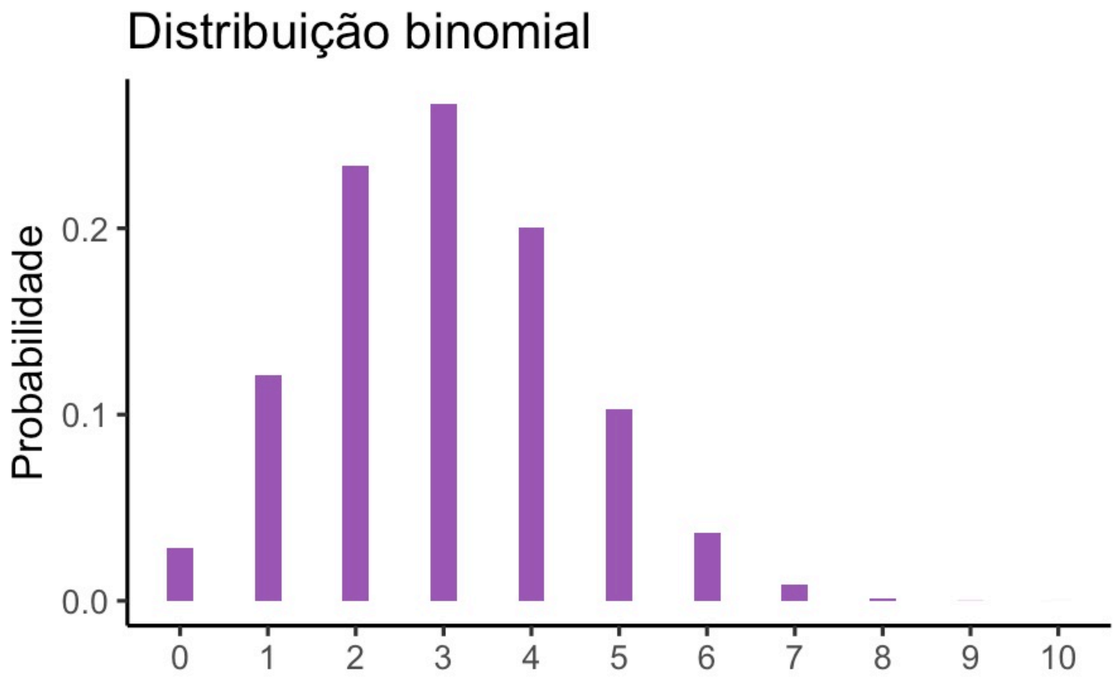
$$p(x) = \begin{cases} 0.3 & \text{se } x = 0; \\ 0.7 & \text{se } x = 1. \end{cases}$$

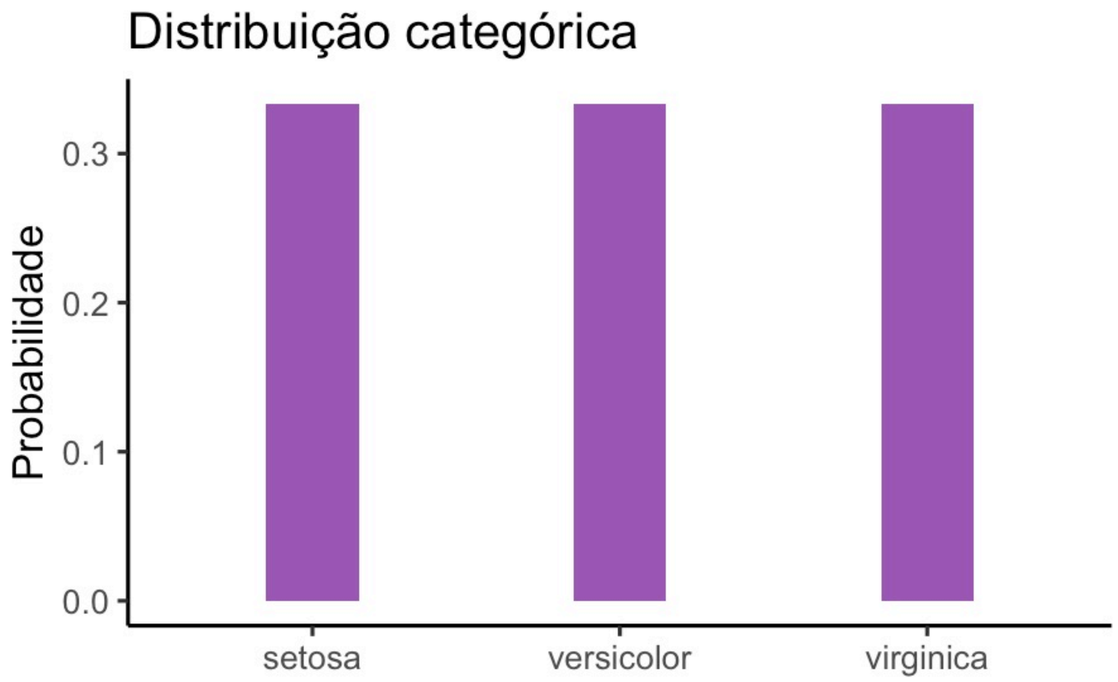


Observemos que todas as probabilidades pertencem ao intervalo $[0,1]$ e a soma das probabilidades é 1

→ Esta propriedade vale para todas as variáveis aleatórias discretas

Variáveis aleatórias

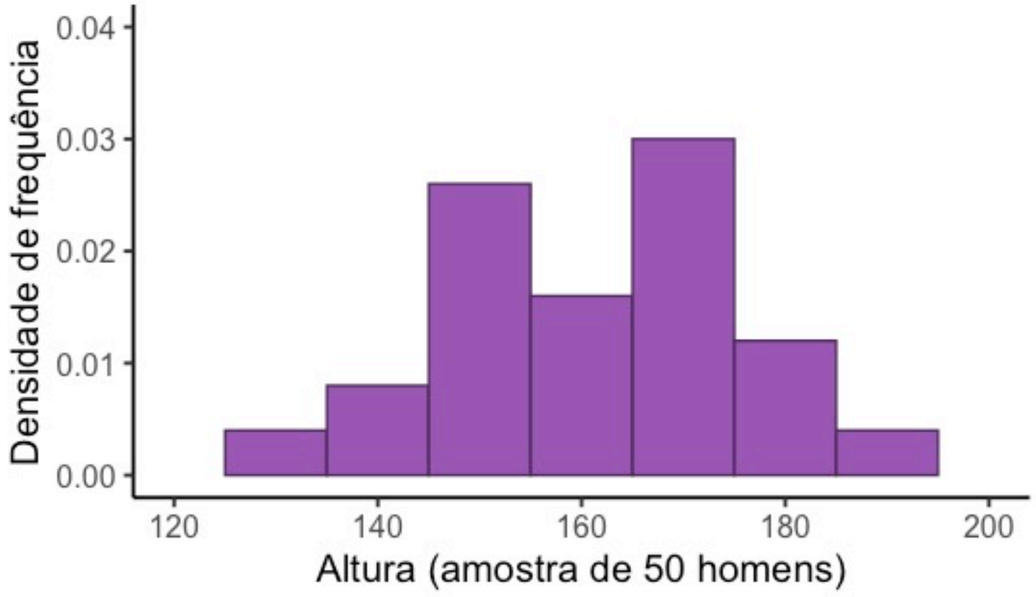


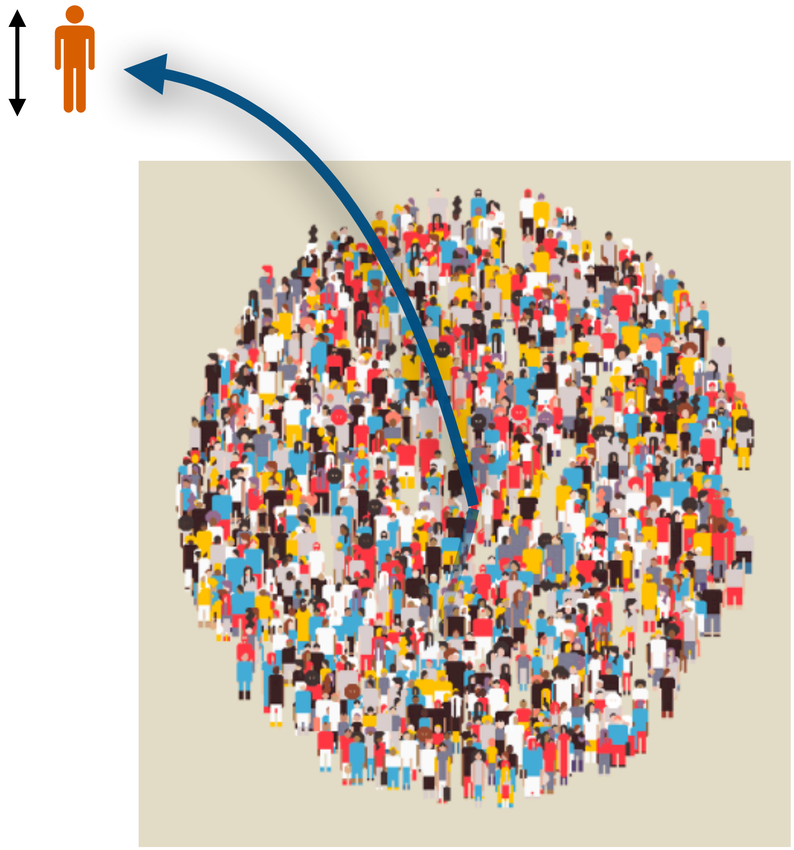


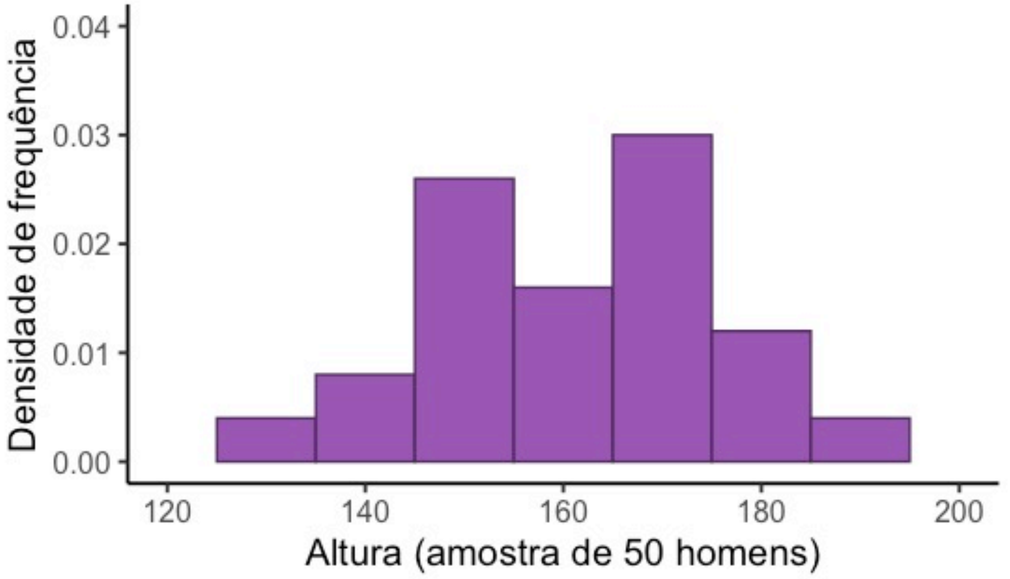
Variáveis aleatórias

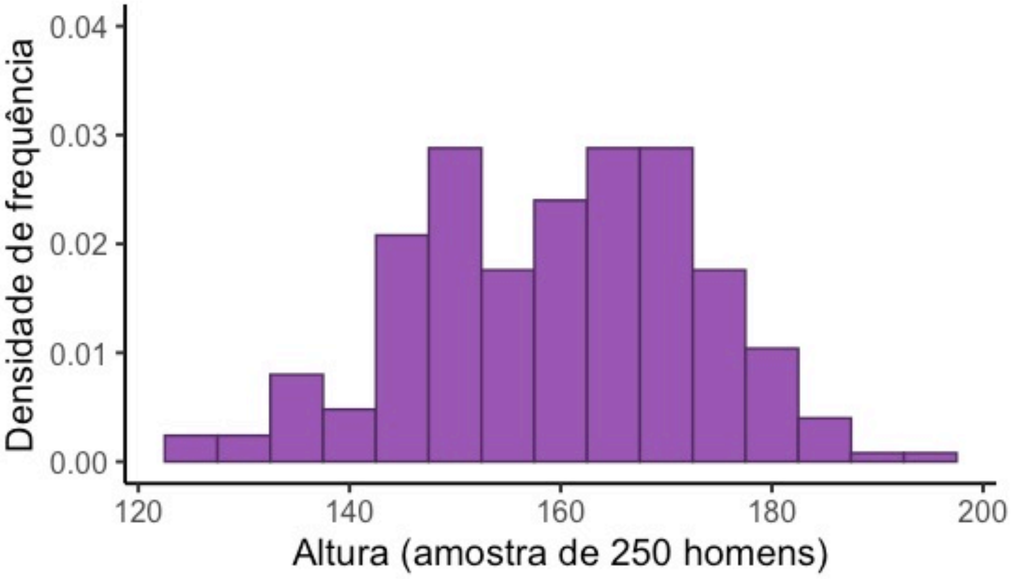
Agora suponhamos que escolhemos um indivíduo ao acaso de uma população muito grande e anotamos a altura do indivíduo

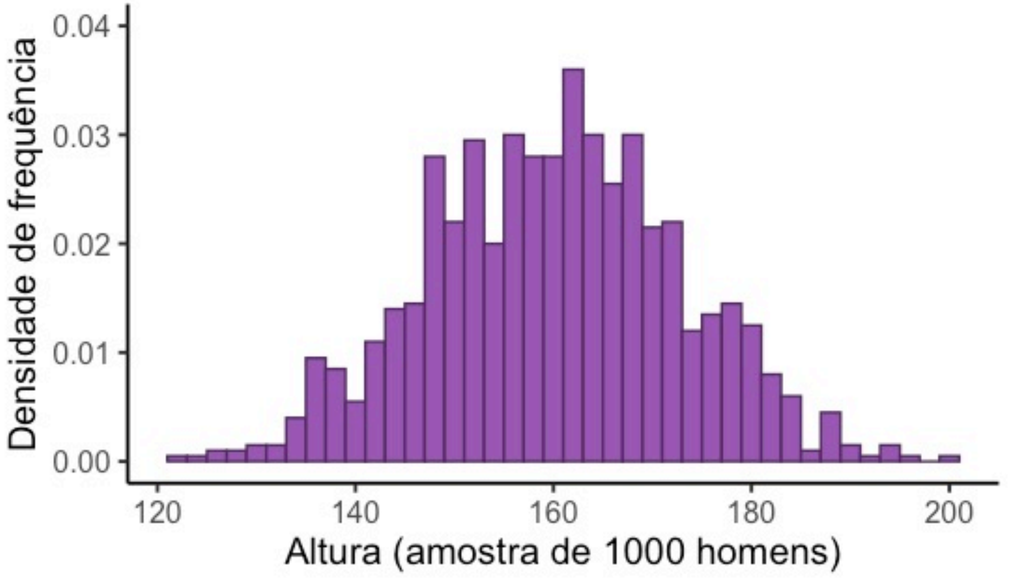
Qual é a probabilidade dele ter altura menor a 1,65m?

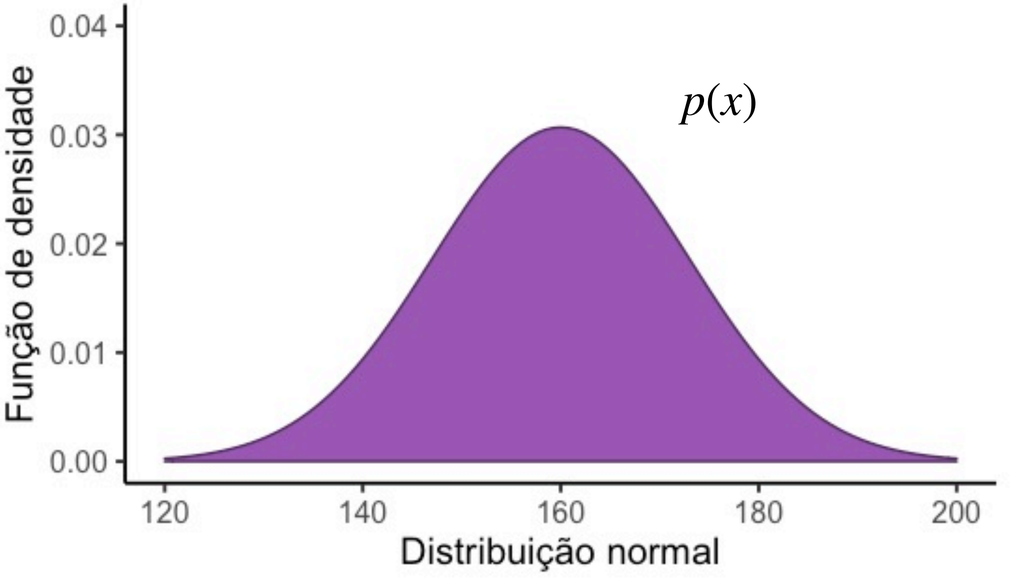





Densidade de frequência
0.04
0.03
0.02
0.01
0.00
120
140
160
180
200
Altura (amostra de 50 homens)


Densidade de frequência
0.04
0.03
0.02
0.01
0.00
120
140
160
180
200
Altura (amostra de 250 homens)


Densidade de frequência
0.04
0.03
0.02
0.01
0.00
120
140
160
180
200
Altura (amostra de 1000 homens)

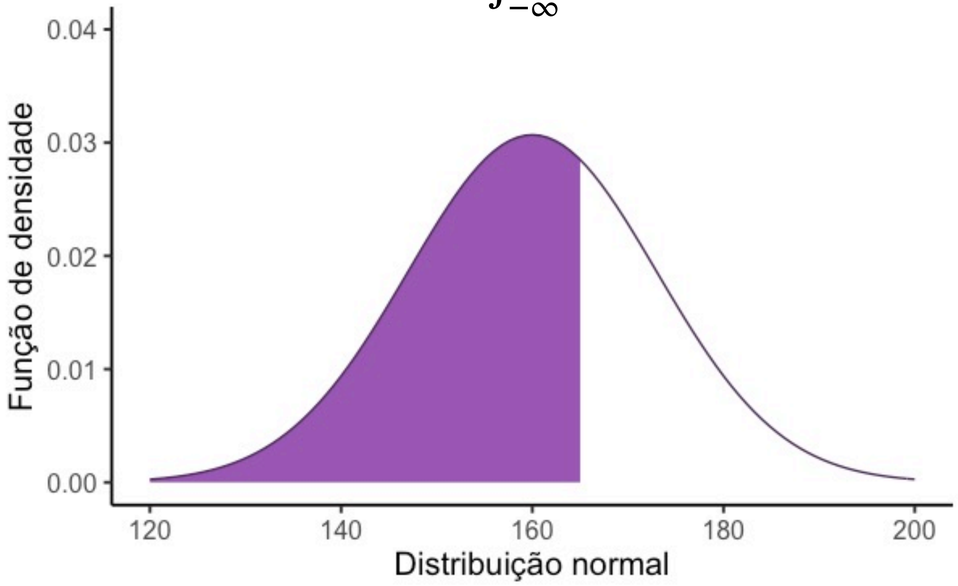

Função de densidade
0.04
0.03
0.02
0.01
0.00
$p(x)$
120
140
160
180
200
Distribuição normal

Variáveis aleatórias

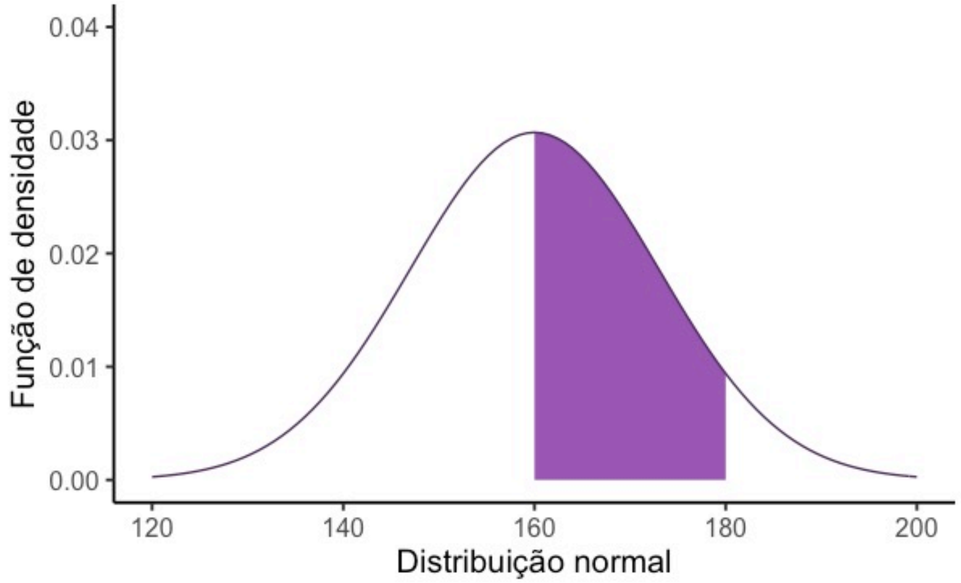
As variáveis aleatórias contínuas são descritas pela sua função densidade de probabilidade $p(x)$

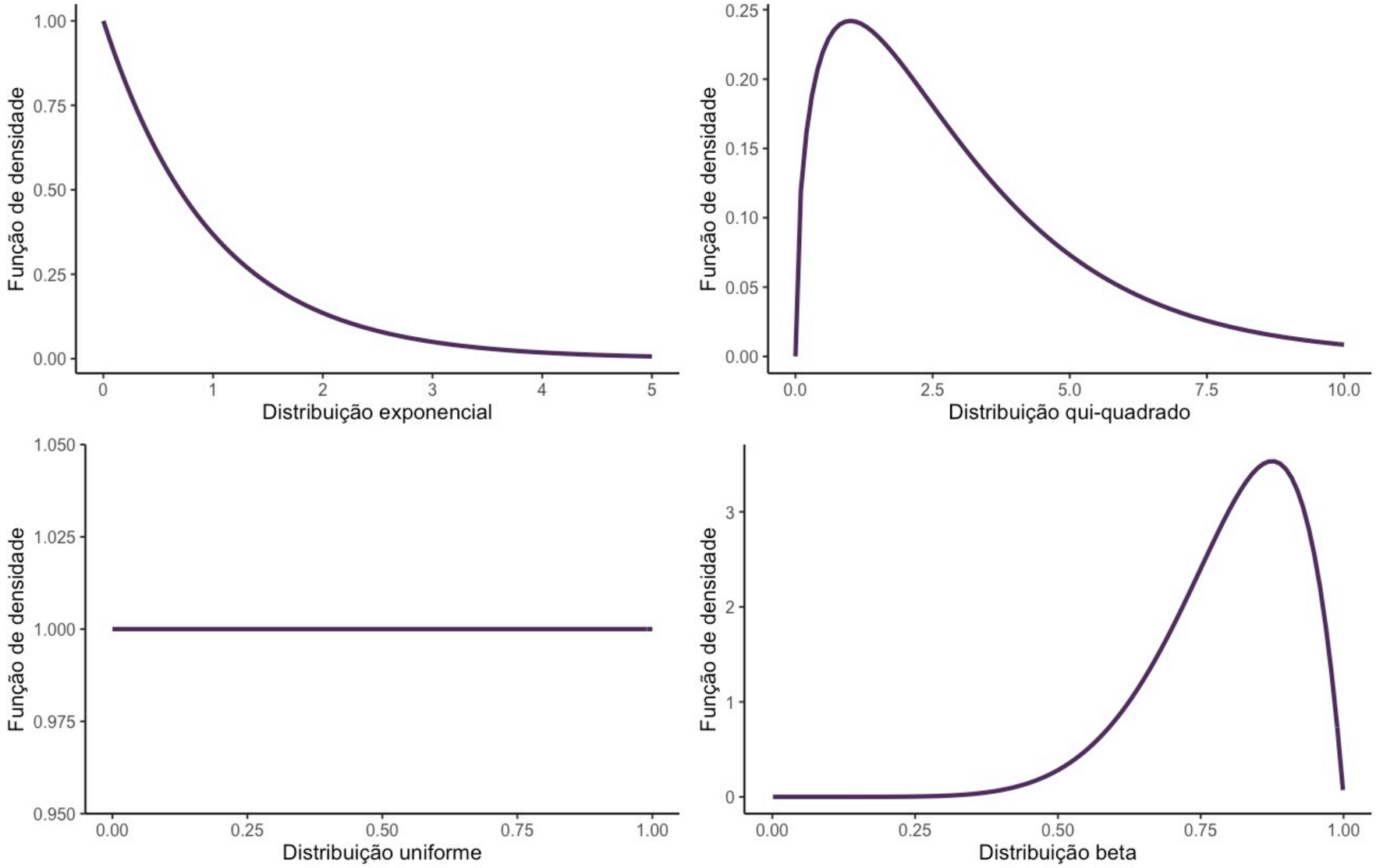
As probabilidades são calculadas como áreas embaixo desta curva

$$P(X < 165) = \int_{-\infty}^{165} p(x)dx = 0.65$$



$$P(160 < X < 180) = \int_{160}^{180} p(x)dx = 0.44$$

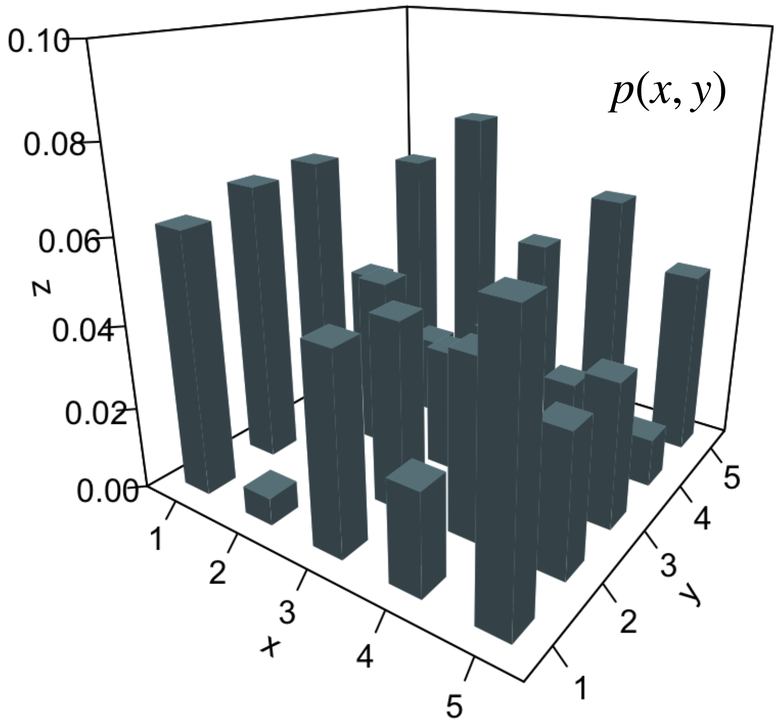




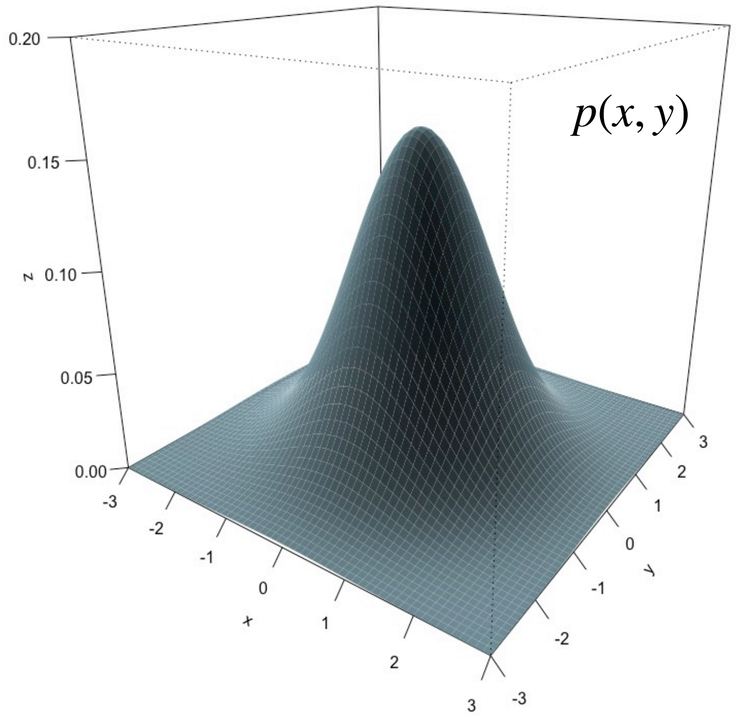
As funções de densidade são não negativas e a área total embaixo da curva e dentro do domínio é 1

Distribuições conjuntas

Estas noções podem ser generalizadas para funções de probabilidade ou funções de densidade definidas sobre vetores (discretos ou contínuos)



$\mathcal{X} = \mathcal{Y} = \{1,2,\ldots,5\}$



$\mathcal{X} = \mathcal{Y} = \mathbb{R}$

Probabilidade condicional e independência

A função de densidade condicional de Y dado X está dada por

$$p(y \mid x) = \frac{p(x, y)}{p(x)} \quad \text{para todo } x \text{ tal que } p(x) > 0$$

Dizemos que as variáveis X e Y são independentes se a sua função de densidade $p(x, y)$ fatorar como

$$p(x, y) = p(x)p(y)$$

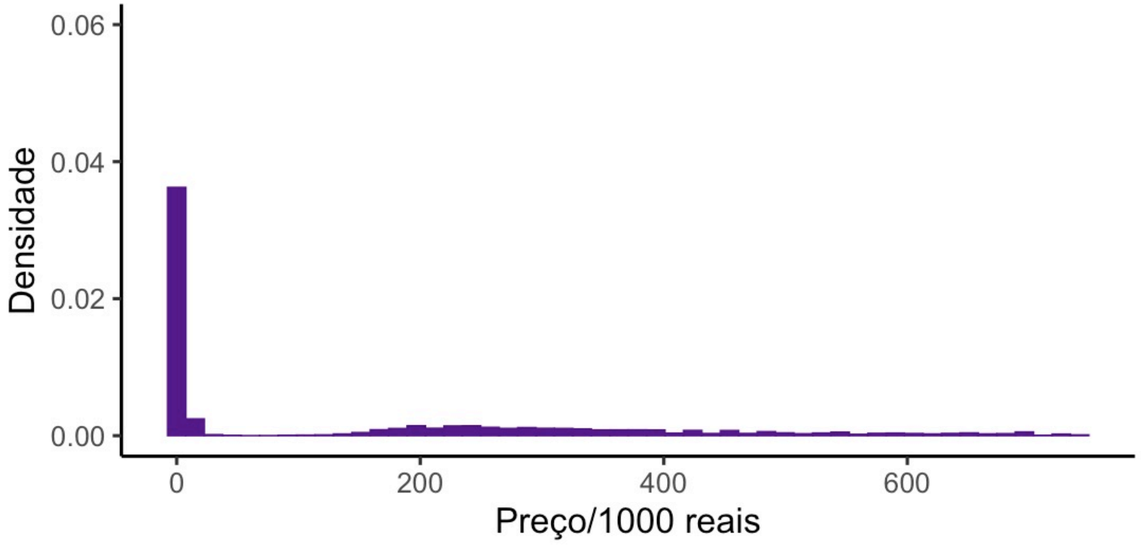
→ Veja que neste caso temos que $p(y \mid x) = p(y)$

Probabilidade condicional e independência

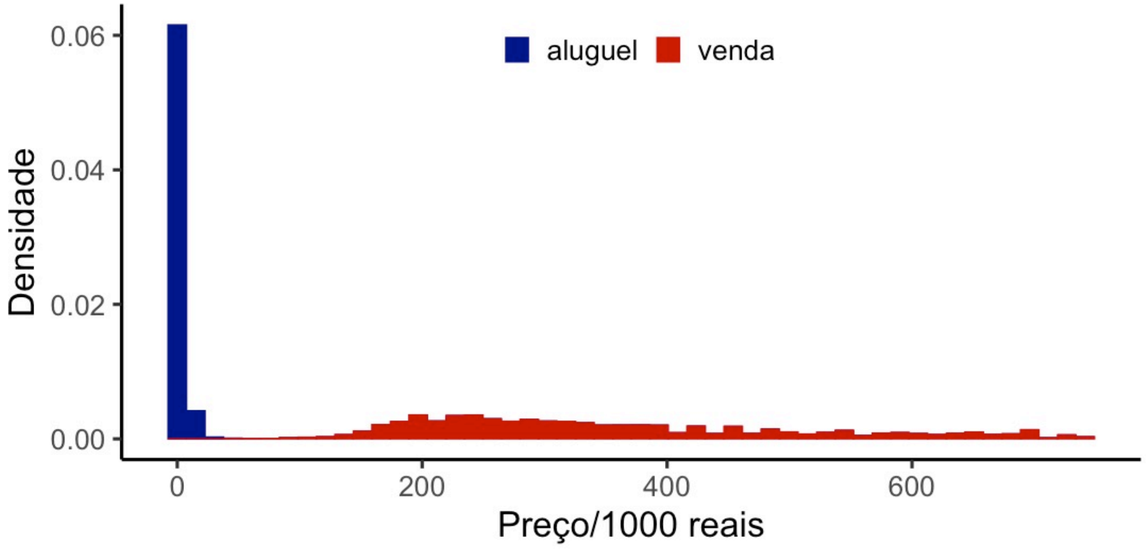
X : preço do imóvel

Y : tipo de operação (aluguel/venda)

Price	Condo	Size	Rooms	...	Negotiation type
930	220	47	2		rent
1000	148	45	2		rent
1000	100	48	2		rent
990000	870	121	3		sale
410000	630	51	2		sale
820000	1000	109	3		sale



$p(x)$



$p(x \mid y)$

Média e variância

- Quando observamos dados de uma variável aleatória, queremos de certa forma resumir a informação contida nessa amostra em algumas poucas medidas
- Duas das medidas “resumo” mais utilizadas são a média e a variância
- A média representa o “centro de massa” da distribuição da variável e a variância representa a dispersão dos dados ao redor desse centro
- Esses dois valores nos dão informação importante sobre essa variável aleatória e o que pode ser esperado em novas observações

Média e variância

Variável discreta

$$\mathbb{E}(X) = \sum xp(x)$$

Variável contínua

$$\mathbb{E}(X) = \int xp(x)dx$$

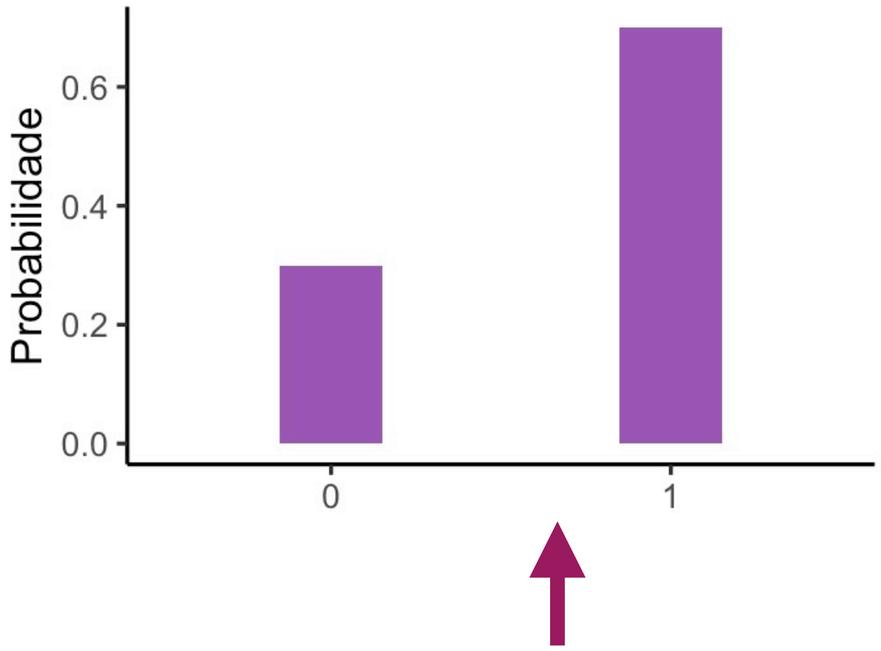
$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2\end{aligned}$$

Média e variância

$$X \in \{0,1\}$$

$$p(x) = \begin{cases} 0.3 & \text{se } x = 0; \\ 0.7 & \text{se } x = 1. \end{cases}$$

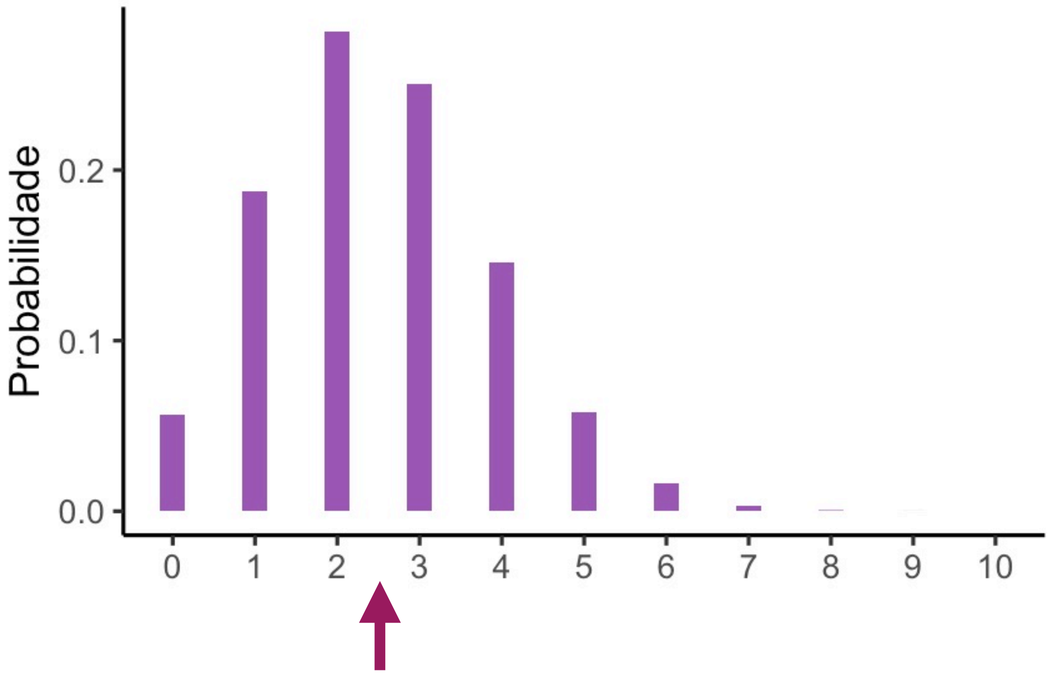
$$\begin{aligned} \mathbb{E}(X) &= \sum xp(x) \\ &= 0 \times 0.3 + 1 \times 0.7 \end{aligned}$$



$$\mathbb{E}(X) = 0.7$$

Média e variância

Binomial de parâmetros $n = 10$, $p = 0{,}25$

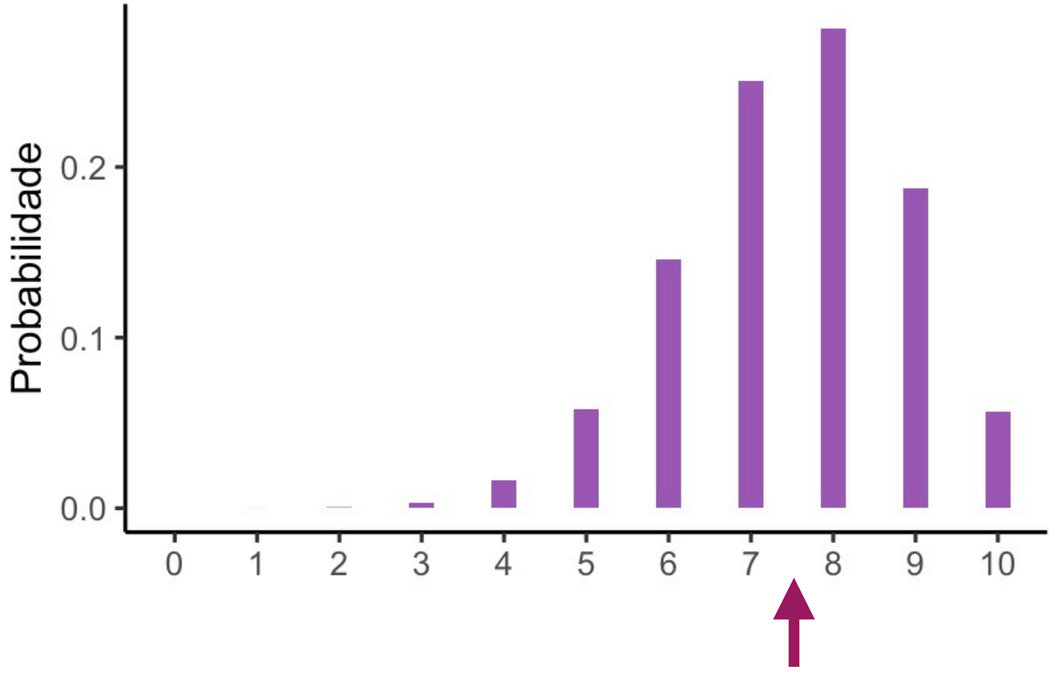


$\mathbb{E}(X) = 2.5$

$\text{Var}(X) = 1.875$

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n$$

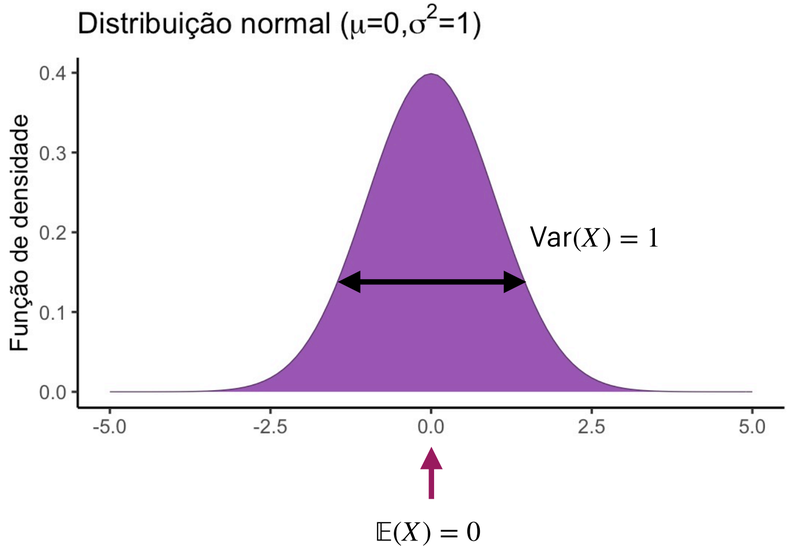
Binomial de parâmetros $n = 10$, $p = 0{,}75$

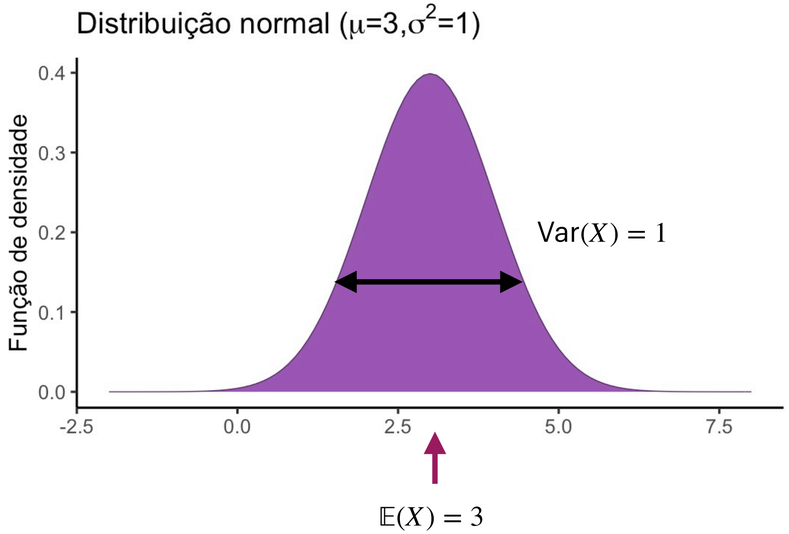


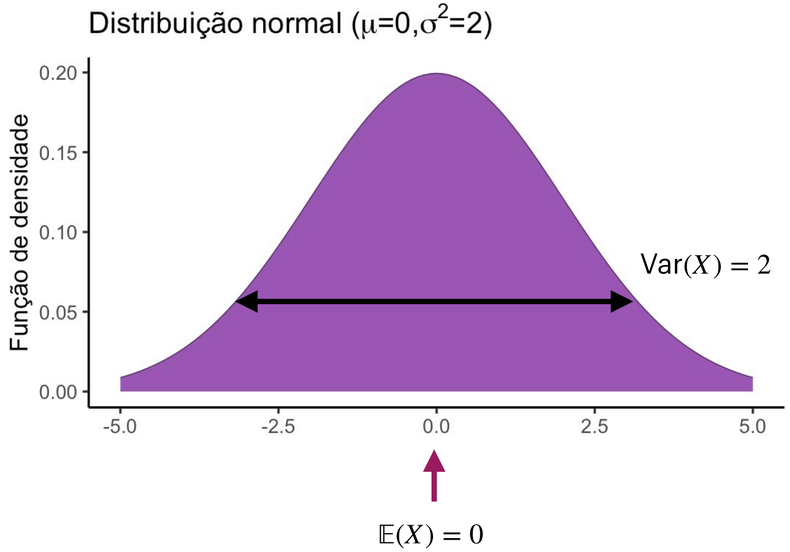
$\mathbb{E}(X) = 7.5$

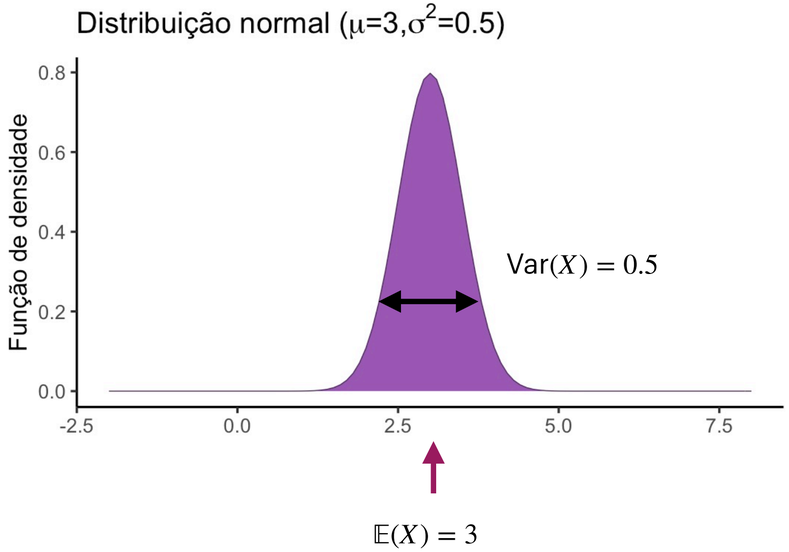
$\text{Var}(X) = 1.875$

Média e variância









Estimadores

Suponhamos que queremos estimar uma quantidade desconhecida $\theta = T(X)$, por exemplo $\theta = \mathbb{E}(X)$ ou $\theta = \text{Var}(X)$ de uma variável aleatória X.

Para fazer isso, vamos nos basear numa amostra $\mathscr{D} = \{x_1, \ldots, x_n\}$ de variáveis independentes e com a mesma distribuição de X e vamos usar um estimador $\hat{\theta} = t(x_1, \ldots, x_n)$ que é uma função da amostra $\mathscr{D}$

Se $\theta = \mathbb{E}(X)$ $\qquad$ $\hat{\theta} = \widehat{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$

Se $\theta = \text{Var}(X)$ $\qquad$ $\hat{\theta} = \widehat{\text{Var}}(X) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$

Viés de um estimador

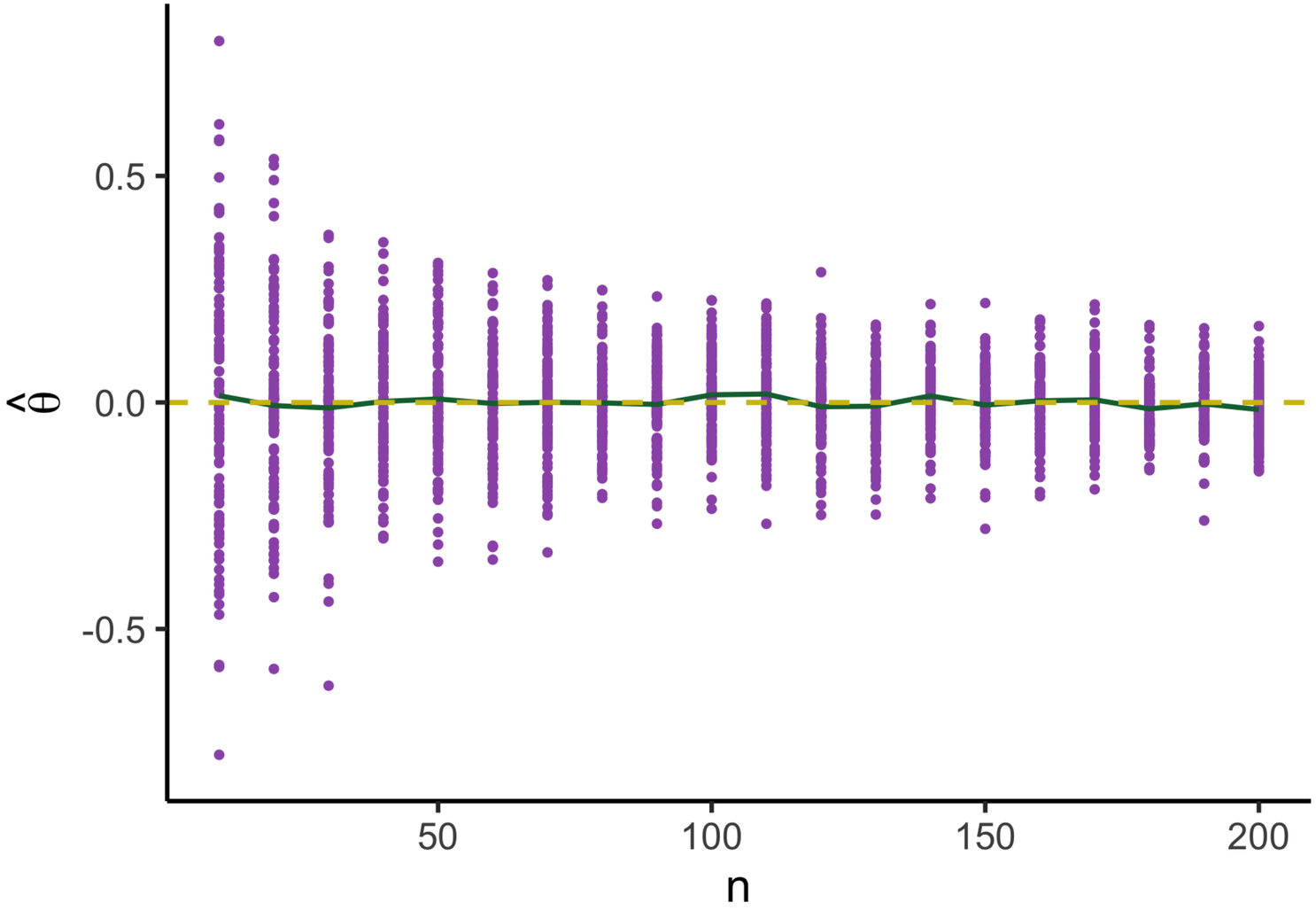
Um fato importante sobre um estimador $\hat{\theta}$ é que ele também é uma variável aleatória, logo podemos também estar interessados em calcular $\mathbb{E}(\hat{\theta})$ e $\text{Var}(\hat{\theta})$.

Uma noção importante é a de viés do estimador $\hat{\theta}$ em relação ao parâmetro θ, que é definido por:

$$\text{Viés}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Se $\text{Viés}(\hat{\theta}) \neq 0$ então $\hat{\theta}$ é dito estimador “viesado”

Viés de um estimador



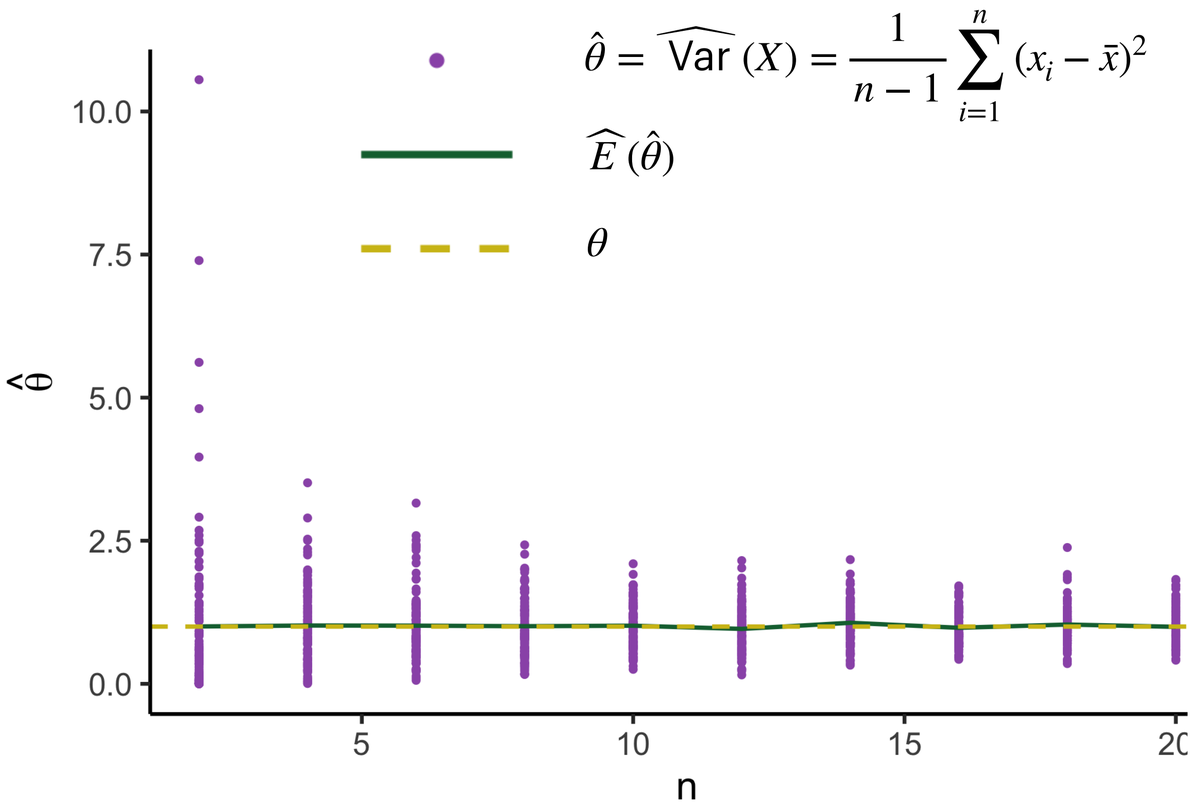
$$x_1, \ldots, x_n \sim \mathcal{N}(0,1)$$

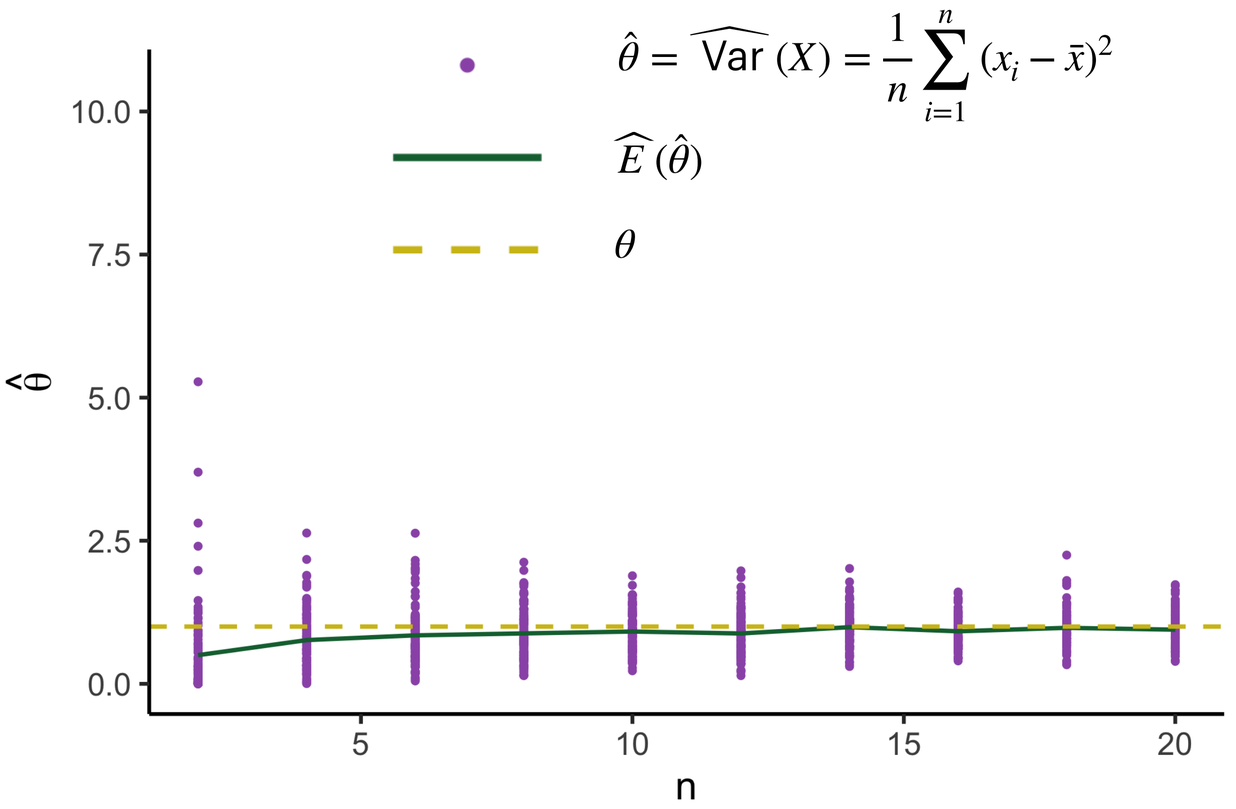
$\hat{\theta} = \widehat{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\widehat{E}(\hat{\theta})$

θ

Viés de um estimador





Variância de um estimador

Outra quantidade de interesse em relação a um estimador $\hat{\theta}$ é sua variância, que está definida como:

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] = \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2$$

Se $x_1, \ldots, x_n$ são independentes e $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ então $\text{Var}(\hat{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(x_i)$

Se $\text{Var}(x_i) = \sigma^2$ para todo $i = 1,\ldots,n$ então $\text{Var}(\hat{\theta}) = \frac{1}{n}\sigma^2$

Erro quadrático médio

Uma forma de avaliar o desempenho de $\hat{\theta}$ como estimador de θ é através do erro quadrático médio, definido por

$$\text{EQM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Observe que $\text{EQM}(\hat{\theta})$ é o valor esperado das distâncias ao quadrado entre $\hat{\theta}$ e θ e portanto avalia “quão próximo”, em média, $\hat{\theta}$ está de θ

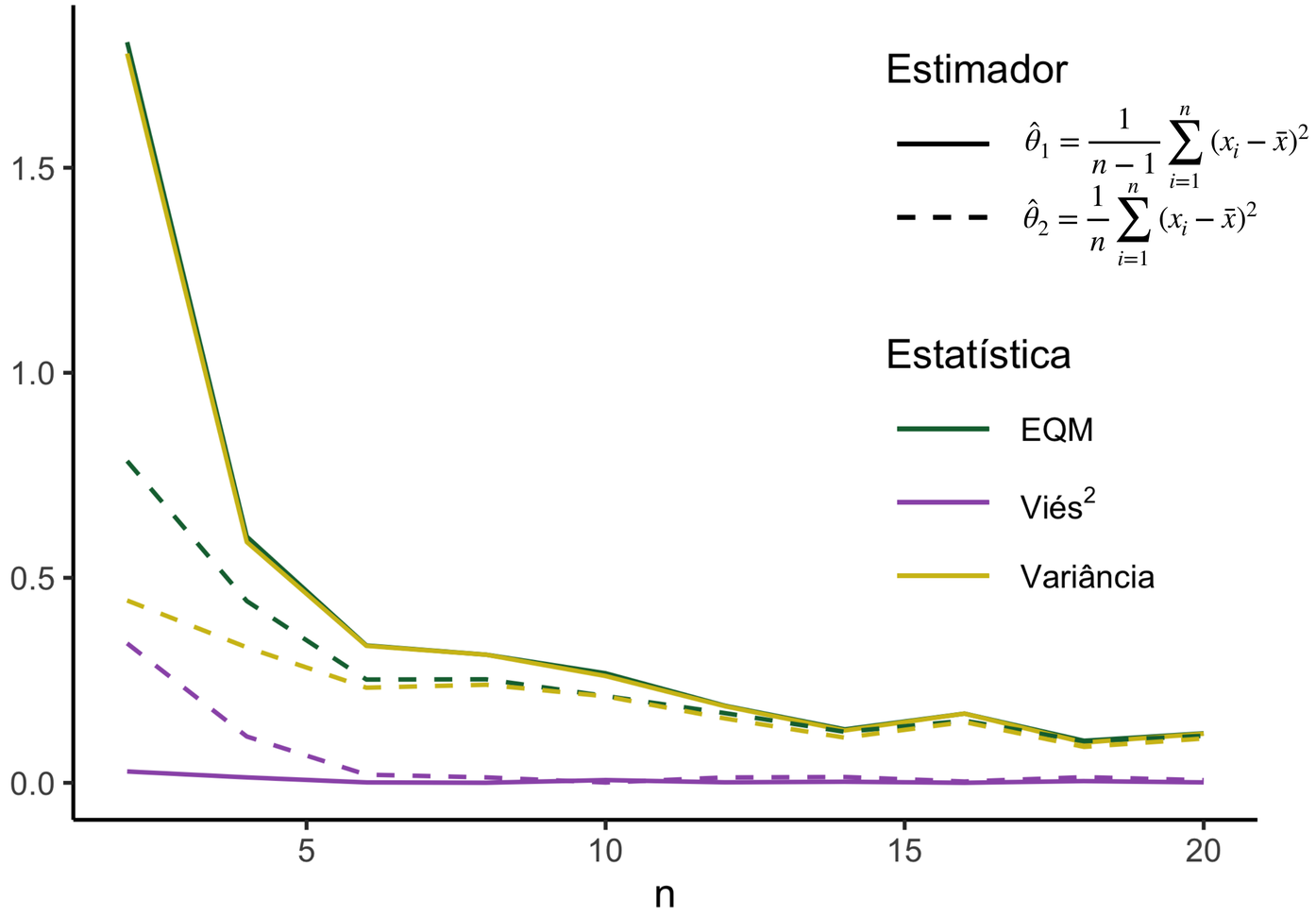
Erro quadrático médio

O erro quadrático médio satisfaz a seguinte propriedade:

$$\text{EQM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \text{Viés}(\hat{\theta})^2$$

Erro quadrático médio

$$EQM(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] \;=\; Var(\hat{\theta}) + Viés(\hat{\theta})^2$$



Erro quadrático médio

$$\text{EQM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] \;=\; \text{Var}(\hat{\theta}) + \text{Viés}(\hat{\theta})^2$$

Demonstração:

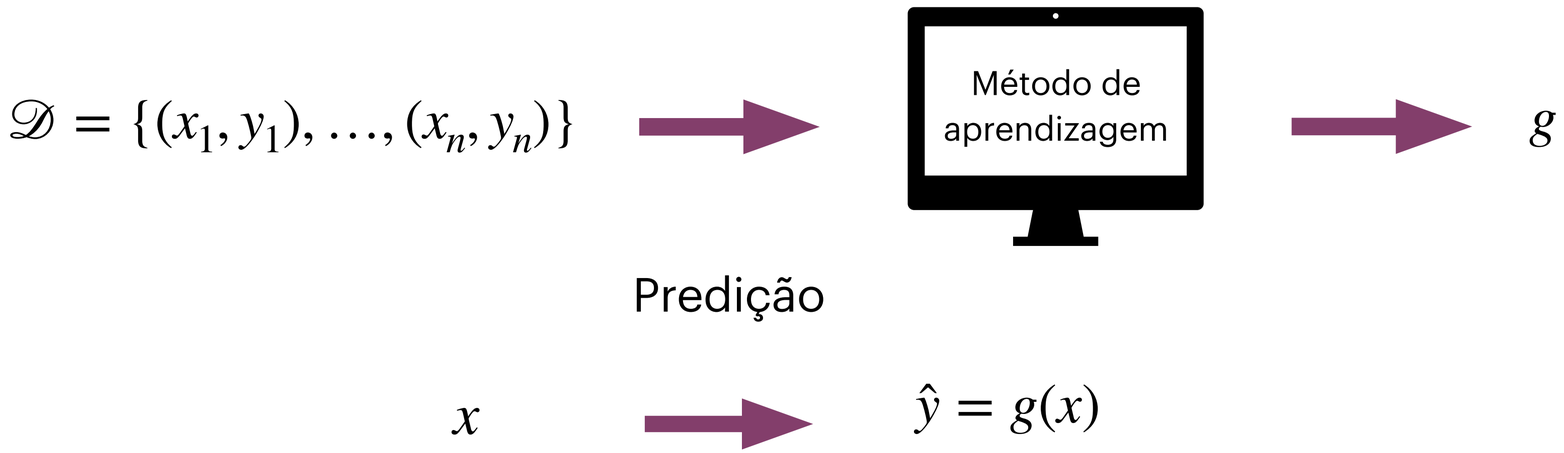
$$(\hat{\theta} - \theta)^2 \;= (\hat{\theta} - \mathbb{E}(\hat{\theta}) \;+\; \mathbb{E}(\hat{\theta}) - \theta)^2$$

$$= (\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) + (\mathbb{E}(\hat{\theta}) - \theta)^2$$

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + \; 2\mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) \; + (\mathbb{E}(\hat{\theta}) - \theta)^2$$

$$= \underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2]}_{\text{Var}(\hat{\theta})} + \underbrace{2(\mathbb{E}(\hat{\theta}) - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)}_{0} \; + \underbrace{(\mathbb{E}(\hat{\theta}) - \theta)^2}_{\text{Viés}(\hat{\theta})^2}$$

Objetivo da aprendizagem estatística supervisionada



- O objetivo é escolher g de tal forma que a predição $\hat{y}$ esteja "próxima" de y
- Para medir a proximidade de $\hat{y}$ e y usamos uma *função de custo* L escolhida de acordo com o problema
- O objetivo da aprendizagem será então minimizar $\mathbb{E}(L(y, g(x)))$ onde (x, y) é uma observação de teste (não pertencente à amostra)

Função de custo

É escolhida antes de qualquer análise e de acordo com o problema

$$y_i \in \mathcal{Y} = \{c_1, \ldots, c_K\}$$ → **Problema de classificação**

Exemplos:

$$L(y, \hat{y}) = \mathbf{1}\{y \neq \hat{y}\}$$

$$L(y, \hat{y}) = -\sum_{k=1}^{K} \mathbf{1}\{y = c_k\} \log \hat{y}_k, \qquad \hat{y} = (p(c_1), \ldots, p(c_K))$$

Função de custo

$$y_i \in \mathcal{Y} = \mathbb{R}$$

Problema de regressão

$$L(y, \hat{y}) = (y - \hat{y})^2$$

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{se } |y - \hat{y}| \leq \delta, \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{c.c} \end{cases}$$

$$L(y, \hat{y}) = |y - \hat{y}|$$

$$L(y, \hat{y}) = \log(\cosh(y - \hat{y}))$$

Desafios na aprendizagem estatística supervisionada

Dada a função de custo L e a amostra $\mathscr{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$:

- Como estimar $\mathbb{E}(L(y, g(x))$ para uma função g escolhida com base em $\mathscr{D}$?

- Como escolher g de forma a minimizar $\mathbb{E}(L(y, g(x))$?

Esses serão os principais problemas que iremos abordar neste curso, para diferentes famílias de funções $\mathscr{G}$