

# ESTATÍSTICA PARA CIÊNCIA DE DADOS E MACHINE LEARNING

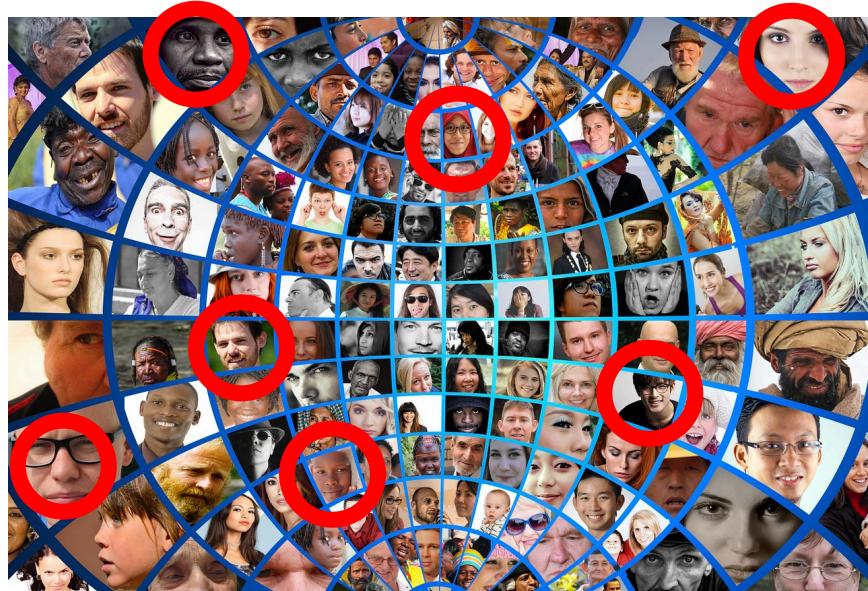


<p><b>Parte 1 – População e amostra</b></p> <ul style="list-style-type: none"> <li>• Técnicas de amostragem</li> <li>• Subamostragem e sobreamostragem</li> </ul>	<p><b>Parte 2 – Dados relativos e absolutos</b></p> <ul style="list-style-type: none"> <li>• Percentuais</li> <li>• Índices, coeficientes e taxas</li> </ul>	<p><b>Parte 3 – Distribuição de frequência</b></p> <ul style="list-style-type: none"> <li>• Cálculos passo a passo</li> <li>• Aplicação em regras de associação</li> </ul>
<p><b>Parte 4 – Medidas de posição e dispersão</b></p> <ul style="list-style-type: none"> <li>• Média, moda e mediana</li> <li>• Quartis e percentis</li> <li>• Variância e desvio padrão</li> <li>• Avaliação de algoritmos de classificação</li> </ul>	<p><b>Parte 5 – Distribuições estatísticas</b></p> <ul style="list-style-type: none"> <li>• Distribuição normal</li> <li>• Distribuições não normais</li> <li>• Naïve Bayes Multinomial e Bernoulli</li> <li>• Padronização (z-score)</li> <li>• Pesos em redes neurais artificiais</li> </ul>	<p><b>Parte 6 – Probabilidade</b></p> <ul style="list-style-type: none"> <li>• Probabilidade básica</li> <li>• Distribuições de probabilidade</li> <li>• Probabilidade e machine learning</li> </ul>
<p><b>Parte 7 – Intervalos de confiança e testes de hipóteses</b></p> <ul style="list-style-type: none"> <li>• Cálculos passo a passo</li> <li>• Distribuição T Student</li> <li>• ANOVA e Qui Quadrado</li> <li>• Seleção de atributos</li> <li>• Avaliação de algoritmos</li> <li>• Estatística paramétrica e não paramétrica</li> </ul>	<p><b>Parte 8 – Correlação e regressão</b></p> <ul style="list-style-type: none"> <li>• Cálculos passo a passo</li> <li>• Regressão linear simples e múltipla</li> </ul>	<p><b>Parte 9 – Visualização</b></p> <ul style="list-style-type: none"> <li>• Gráficos</li> <li>• Mapas com latitude e longitude</li> </ul>

# CONTEÚDO

- População e amostra
- Tabela de números aleatórios
- Amostragem aleatória simples
- Amostragem sistemática
- Amostragem por grupos
- Amostragem estratificada
- Amostragem de reservatório
- Dados desbalanceados
  - Classificação
  - Naïve bayes
  - Subamostragem (undersampling)
  - Sobreamostragem (oversampling)

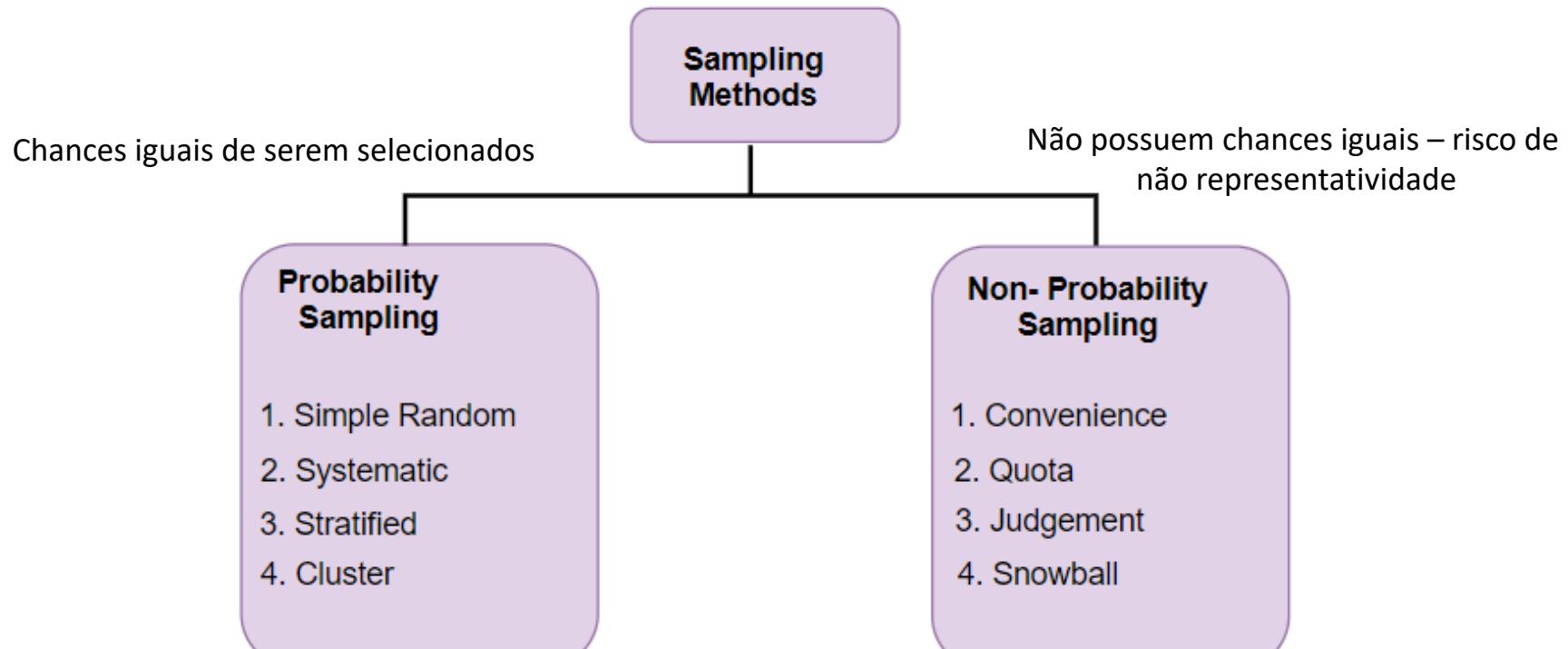
# POPULAÇÃO E AMOSTRA



A amostra é sempre menor que a população  
Mais rápido para processar  
Menor tempo para analisar  
Os números que são obtidos da amostra são as estatísticas  
A amostra precisa ser **randômica e representativa**



# TIPOS DE AMOSTRAGEM



Fonte: <https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/>

# AMOSTRAGEM – TABELA DE NÚMEROS ALEATÓRIOS

14	15	92	65	35	89	79	32	38	46	26	43	38	32	79	50
41	97	16	93	99	37	51	05	82	09	74	94	45	92	30	78
06	28	62	08	99	86	28	03	48	25	34	21	17	06	79	82
08	65	13	28	23	06	64	70	93	84	46	09	55	05	82	23
53	59	40	81	28	48	11	17	45	02	84	10	27	01	93	85
05	55	96	44	62	29	48	95	49	30	38	19	64	42	88	10
66	59	33	44	61	28	47	56	48	23	37	86	78	31	65	27
19	09	14	56	48	56	69	23	46	03	48	61	04	54	32	66
13	39	36	07	26	02	49	14	12	73	72	45	87	00	66	03
58	81	74	88	15	20	92	09	62	82	92	54	09	17	15	36
78	92	59	03	60	01	13	30	53	05	48	82	04	66	52	13
46	95	19	41	51	16	09	43	30	57	27	03	65	75	95	91
09	21	86	11	73	81	93	26	11	79	31	05	11	85	48	07
23	79	99	62	74	95	67	35	18	85	75	27	24	89	12	27
93															

População: 80

Amostra: 5

14, 15, 65, 35, 79

População: 400

Amostra: 5

122, 272, 188, 274, 237

# AMOSTRAGEM SISTEMÁTICA



População: 28 casas     $28 / 5 = 5,6$  (arredondamos para 6)

Amostra: 5 casas

# AMOSTRAGEM POR GRUPOS



População: 28 casas      Selecionar randomicamente um dos grupos  
4 grupos

# AMOSTRA ESTRATIFICADA

- População: 90 pessoas
- 54 mulheres e 36 homens
- Amostra de 10% da população: 9 pessoas
- Mulheres representam 60% da população
- Homens representam 40% da população
- Quantidade de mulheres
  - $54 * 10 / 100 = 5,4$  (arredondamos para 5)
- Quantidade de homens
  - $36 * 10 / 100 = 3,6$  (arredondamos para 4)
- Cálculos para saber quantos elementos de cada grupo devem ser selecionados. Utilizar na sequência outra técnica para seleção das amostras

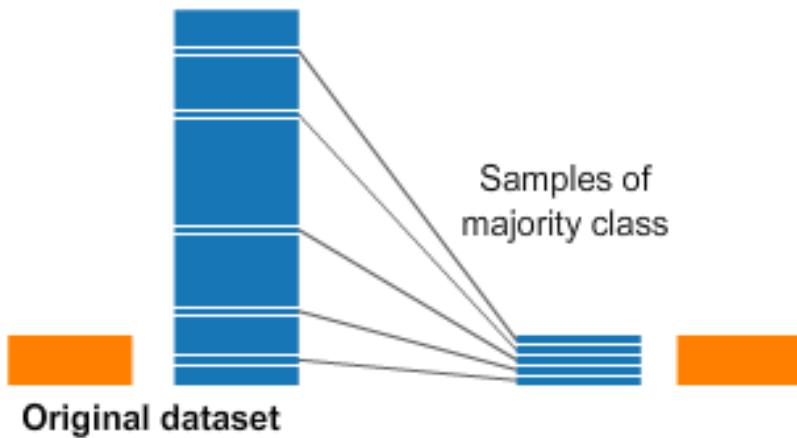
# AMOSTRA DE RESERVATÓRIO

- **Data stream** de itens com tamanho desconhecido que pode ser acessado somente uma vez
- Algoritmo para sortear um item do stream, porém, cada item deve possuir a mesma probabilidade de seleção
- Exemplo dos chapéus: <https://www.youtube.com/watch?v=A1iwzSew5QY>

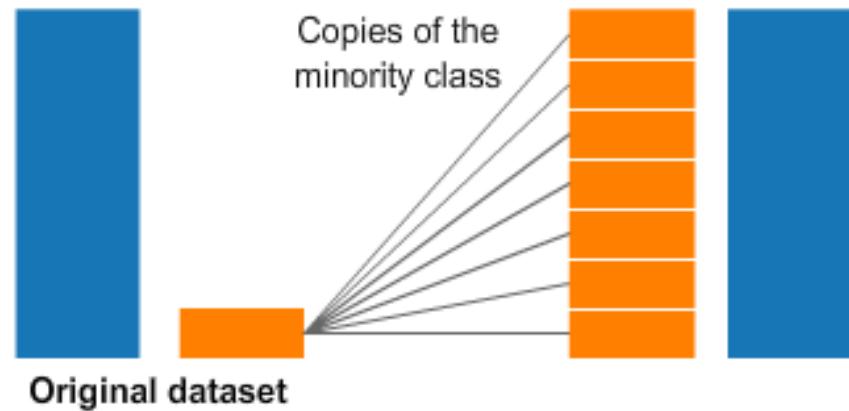


# SUBAMOSTRAGEM E SOBREAMOSTRAGEM

**Undersampling**

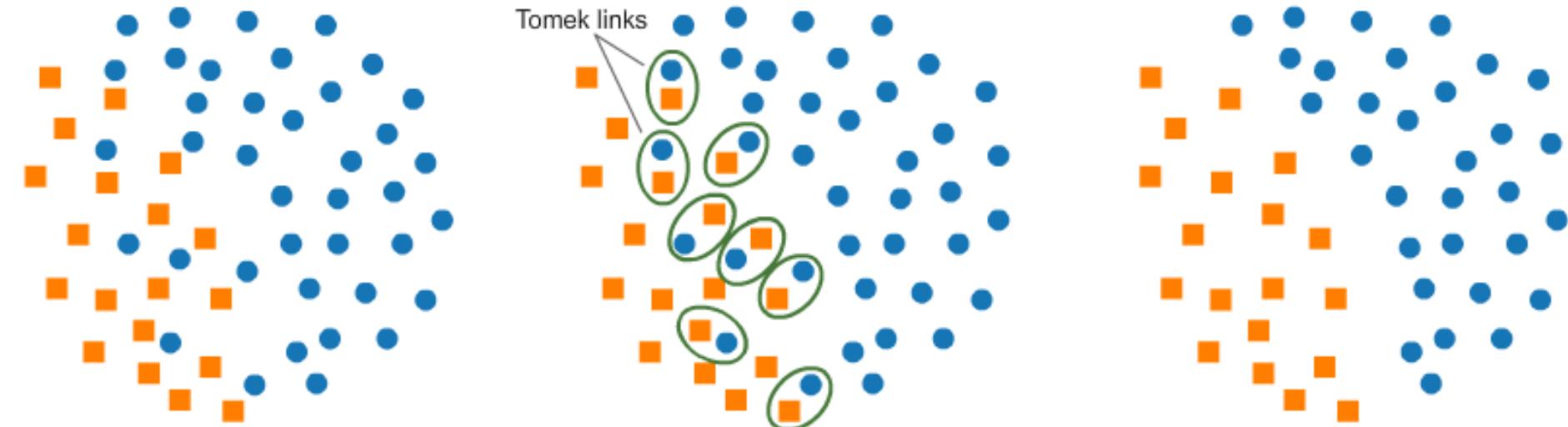


**Oversampling**



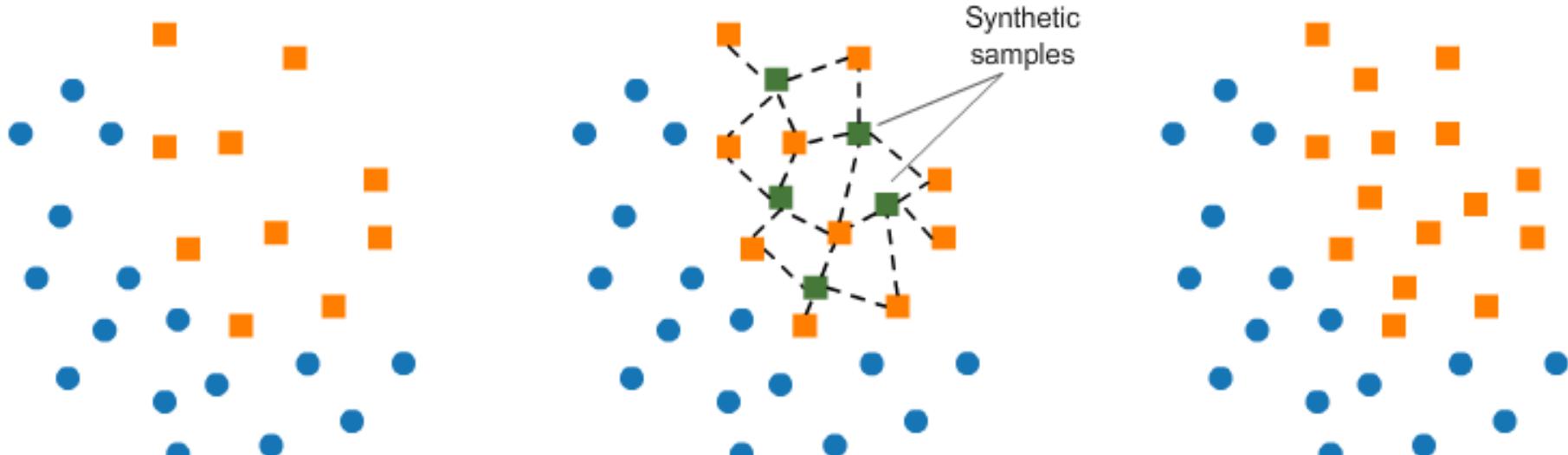
Fonte: <https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1>

# SUBAMOSTRAGEM – TOMEK LINKS



Fonte: <https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1>

# SOBREAMOSTRAGEM – SMOTE



Fonte: <https://www.kaggle.com/ratja/resampling-strategies-for-imbalanced-datasets#t1>

# DADOS ABSOLUTOS E RELATIVOS

- Dados absolutos:
  - Coleta direta da fonte sem nenhum outro tipo de manipulação (somente contagem, ordenação)
- Dados relativos
  - Fácil entendimento para ajudar nas comparações entre quantidades
- Porcentagem
- Índices
- Coeficientes
- Taxas

# DADOS ABSOLUTOS E RELATIVOS – PORCENTAGEM

- Destacar a participação da “parte no todo” – comparativos

Emprego	Nova Jersey	Florida	% Nova Jersey	% Florida
Administrador de banco de dados	97.350	77.140	33.30	36.56
Programador	82.080	71.540	28.08	33.90
Arquiteto de redes	112.840	62.310	38.62	29.54
Total	<b>292.270</b>	<b>210.990</b>	<b>100</b>	<b>100</b>

# DADOS ABSOLUTOS E RELATIVOS – ÍNDICES

- Razões entre duas grandezas
- Resumir em um só número o comportamento geral de uma variável
- $\text{Índice cefálico} = \frac{\text{largura} \times \text{comprimento}}{100}$
- $\text{Densidade demográfica} = \frac{\text{população} \times \text{superfície}}{100}$
- $\text{Produção per capita} = \frac{\text{valor total da produção}}{\text{população}}$
- $\text{Renda per capita} = \frac{\text{renda}}{\text{população}}$
- Índice Bovespa: <https://blog.magnetis.com.br/o-que-e-indice-bovespa/>

# DADOS ABSOLUTOS E RELATIVOS – COEFICIENTES E TAXAS

- Razões entre o número de ocorrências e o número total
- Taxa: coeficientes multiplicados por uma potência de 10 (10, 100, 1000)
- *Coeficiente de natalidade* =  $\frac{\text{número de nascimentos}}{\text{população}}$
- *Taxa de natalidade* = *coeficiente de natalidade* x 1000
- *Coeficiente de mortalidade* =  $\frac{\text{número de óbitos}}{\text{população}}$
- *Taxa de mortalidade* = *coeficiente de mortalidade* x 1000
- *Coeficiente de evasão* =  $\frac{\text{número de alunos evadidos}}{\text{número inicial de matrículas}}$
- *Taxa de evasão* = *coeficiente de evasão* x 100

# DADOS ABSOLUTOS E RELATIVOS – COEFICIENTES E TAXAS

Ano graduação	Matrículas março	Matrículas novembro	Taxa de evasão
1º	70	65	7.14
2º	50	48	4.00
3º	47	40	14.89
4º	23	22	4.34
<b>Total</b>	<b>190</b>	<b>175</b>	<b>7.89</b>

$$\text{Coeficiente de evasão} = \frac{\text{número de alunos evadidos}}{\text{número inicial de matrículas}}$$

$$\text{Taxa de evasão} = \text{coeficiente de evasão} \times 100$$

# DISTRIBUIÇÃO DE FREQUÊNCIA

- Cálculos passo a passo
- Histograma
- Regras de associação – algoritmo Apriori
- Distribuição de frequência e regras de associação

# DISTRIBUIÇÃO DE FREQUÊNCIA

Tabela primitiva

160	165	167	164	160	166	160	161	150	152
173	160	155	164	168	162	161	168	163	156
155	169	151	170	164	155	152	163	160	155
157	156	158	158	161	154	161	156	172	153

Tabela ordenada (rol)

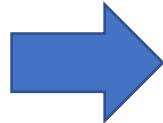
150	151	152	152	153	154	155	155	155	155
156	156	156	157	158	158	160	160	160	160
160	161	161	161	161	162	163	163	164	164
164	165	166	167	168	168	169	170	172	173

$X_{\min}$ : 150

$X_{\max}$ : 173

# DISTRIBUIÇÃO DE FREQUÊNCIA

Estatura (cm)	Frequência
150	1
151	1
152	2
153	1
154	1
155	4
156	3
157	1
158	2
160	5
161	4
162	1
163	2
164	3
165	1
166	1
167	1
168	2
169	1
170	1
172	1
173	1
Total	40



Estatura (cm)	Frequência
150  -- 154	5
154  -- 158	9
158  -- 162	11
162  -- 166	7
166  -- 170	5
170  --  173	3
Total	40

# DISTRIBUIÇÃO DE FREQUÊNCIA

Estatura (cm)	Frequência
150  -- 154	5
154  -- 158	9
158  -- 162	11
162  -- 166	7
166  -- 170	5
170  --  173	3
<b>Total</b>	<b>40</b>

**Classe:** intervalos de variação da variável representados simbolicamente por i

**Limite de classe**

Exemplo:  $l_1 = 150$  e  $L_2 = 158$

**Amplitude de um intervalo de classe (hi)**

$$hi = L_i - l_i \quad (154 - 150 = 4)$$

**Amplitude total da distribuição (AT)**

$$AT = L_{(max)} - L_{(min)} = 173 - 150 = 23$$

**Amplitude amostral (AA)**  $X_{min}: 150$

$$AA = X_{(max)} - X_{(min)} = 173 - 150 = 23 \quad X_{max}: 173$$

**Ponto médio de uma classe ( $x_i$ )**

$$Xi = (L_i + l_i) / 2 = (158 + 154) / 2 = 156 \text{ cm}$$

**Frequência**

$$f_2 = 9 \text{ (número de elementos na classe 2)}$$

# DISTRIBUIÇÃO DE FREQUÊNCIA

Estatura (cm)	Frequência
150   -- 154	5
154   -- 158	9
158   -- 162	11
162   -- 166	7
166   -- 170	5
170   --   173	3
<b>Total</b>	<b>40</b>

Determinar o número de classes

Fórmula de Sturges ( $i = 1 + 3.3 \log n$ )

$$1 + 3.3 * \log(40)$$

$$1 + 3.3 * 1.6 = 6.28$$

Determinar a amplitude do intervalo de classe

$h = AA / i$  (sempre arredondar para cima)

$$23 / 6 = 3,83 \text{ (arredondado} = 4)$$

## Amplitude amostral (AA)

$$AA = X_{(\max)} - X_{(\min)} = 173 - 150 = 23$$

# MEDIDAS DE POSIÇÃO E DISPERSÃO

- Estatística descritiva – descrever e summarizar um conjunto de dados
- Média, mediana e moda
- Média aritmética, geométrica, harmônica e quadrática
- Quartis e percentis
- Variância, desvio padrão e coeficiente de variação
- Avaliação de algoritmos de machine learning
- Seleção de atributos com variância

# MÉDIA ARITMÉTICA, MODA E MEDIANA – DADOS NÃO AGRUPADOS

150	151	152	152	153	154	155	155	155	155
156	156	156	157	158	158	160	160	160	160
160	161	161	161	161	162	163	163	164	164
164	165	166	167	168	168	169	170	172	173

Média

$$\bar{x} = \frac{\sum x_i}{n}$$

160.375

Moda

160

Mediana (ímpar)

$$Mediana = \frac{n}{2}$$

$$Mediana = \frac{9}{2}$$

$$Mediana = 4,5$$

Mediana = 5 (arredondado)

Mediana (par)

$$m = \frac{n}{2}$$

$$m = 20$$

$$m = \frac{160 + 160}{2}$$

$$m = 160$$

150	151	152	152	153	154	155	155	155
-----	-----	-----	-----	-----	-----	-----	-----	-----



# MÉDIA ARITMÉTICA, MODA E MEDIANA – DADOS NÃO AGRUPADOS

150	151	152	152	153	154	155	155	155	155
156	156	156	157	158	158	160	160	160	160
160	161	161	161	161	162	163	163	164	164
164	165	166	999	168	168	900	170	172	173

Média

$$\bar{x} = \frac{\sum x_i}{n}$$

199.225

Moda

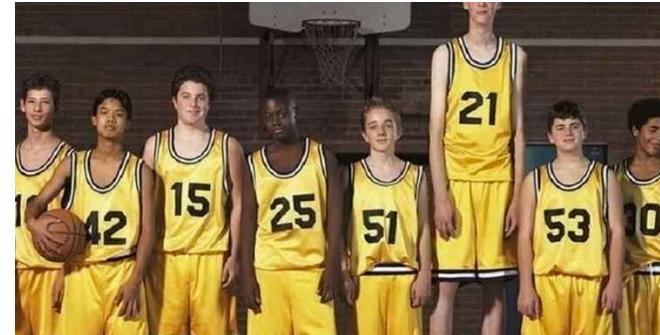
160

Mediana (par)

$$m = \frac{n}{2}$$

$m = 20$

$$m = \frac{160 + 160}{2}$$
$$m = 160$$



Fonte: [https://www.correiobrasiliense.com.br/app/noticia/ciencia-e-saude/2018/10/24/interna\\_ciencia\\_saude,714758/estudo-aponta-que-pessoas-altas-tem-mais-ricos-de-desenvolver-cancer.shtml](https://www.correiobrasiliense.com.br/app/noticia/ciencia-e-saude/2018/10/24/interna_ciencia_saude,714758/estudo-aponta-que-pessoas-altas-tem-mais-ricos-de-desenvolver-cancer.shtml)

# MÉDIA ARITMÉTICA PONDERADA

Bimestre	Nota	Peso
1º	9	1
2º	8	2
3º	7	3
4º	3	4

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\frac{9 + 8 + 7 + 3}{4} = 6,75$$

$$M_p = \frac{p_1 \cdot x_1 + p_2 \cdot x_2 + \dots + p_n \cdot x_n}{p_1 + p_2 + \dots + p_n}$$

$$\frac{9 * 1 + 8 * 2 + 7 * 3 + 3 * 4}{1 + 2 + 3 + 4} = 5,80$$

$$\frac{9 + 8 + 8 + 7 + 7 + 7 + 3 + 3 + 3 + 3}{1 + 2 + 3 + 4}$$

# MÉDIA ARITMÉTICA, MODA E MEDIANA – DADOS AGRUPADOS

Estatura (cm)	$f_i$	$x_i$	$f_i \cdot x_i$	$F_i$
150  -- 154	5	152	760	5
154  -- 158	9	156	1404	14
158  -- 162	11	160	1760	25
162  -- 166	7	164	1148	32
166  -- 170	5	168	840	37
170  --  174	3	172	516	40
Total	40		6428	

Ponto médio de uma classe ( $x_i$ )

$$X_i = (L_i + l_i) / 2 = (158 + 154) / 2 = 156 \text{ cm}$$

$$\frac{\sum f_i}{2} = \frac{40}{2} = 20$$

$$\bar{x} = \frac{\sum f_i \cdot x_i}{\sum f_i}$$

$$\bar{x} = \frac{6428}{40}$$

$$\bar{x} = 160,7$$

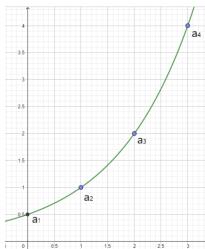
Média  
Moda  
160

$$Md = l + \frac{(\frac{\sum f_i}{2} - F_{ant}) \cdot h}{f_i}$$

$$Md = 158 + \frac{(20 - 14) \cdot 4}{11}$$

$$Md = 160,18$$

# MÉDIA GEOMÉTRICA, HARMÔNICA E QUADRÁTICA



150	151	152	152	153	154	155	155	155	155
156	156	156	157	158	158	160	160	160	160
160	161	161	161	161	162	163	163	164	164
164	165	166	167	168	168	169	170	172	173

$$\bar{g} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_n}$$

Aplicações na geometria, para comparar lados de prismas

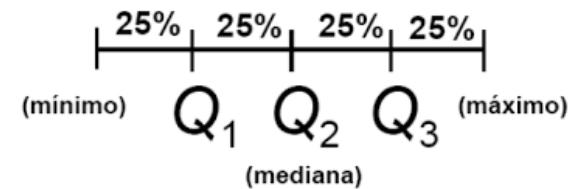
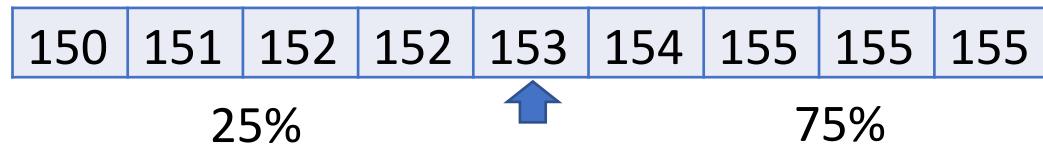
Matemática financeira que envolvem taxa percentual acumulada

$$\bar{h} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Avaliar desempenho em aprendizagem de máquina

$$QM = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

Aplicações na física  
Modelos de regressão



Mediana (ímpar)

$$\text{Mediana} = \frac{n}{2}$$

$$\text{Mediana} = \frac{9}{2}$$

$$\text{Mediana} = 4,5$$

$m = 5$  (arredondado)

Mediana (par)

$$m = \frac{n}{2}$$

$$m = 2$$

$$m = \frac{151 + 152}{2}$$

$$m = 151,5$$

Mediana (par)

$$m = \frac{n}{2}$$

$$m = 2$$

$$m = \frac{155 + 155}{2}$$

$$m = 155$$

# QUARTIS – DADOS AGRUPADOS

Estatura (cm)	$f_i$	$x_i$	$f_i \cdot x_i$	$F_i$
150  -- 154	5	152	760	5
154  -- 158	9	156	1404	14
158  -- 162	11	160	1760	25
162  -- 166	7	164	1148	32
166  -- 170	5	168	840	37
170  --  174	3	172	516	40
<b>Total</b>	<b>40</b>		<b>6428</b>	

$$\frac{\sum f_i}{4} = \frac{40}{4} = 10$$

$$\frac{3 \sum f_i}{4} = \frac{120}{4} = 30$$

$$Q1 = l + \frac{\left(\frac{\sum f_i}{4} - F_{ant}\right) \cdot h}{f_i}$$

$$Q1 = 154 + \frac{(10 - 5) \cdot 4}{9}$$

$$Q1 = 156,22$$

$$Q3 = l + \frac{\left(\frac{3 \sum f_i}{4} - F_{ant}\right) \cdot h}{f_i}$$

$$Q3 = 162 + \frac{(30 - 25) \cdot 4}{7}$$

$$Q3 = 164,85$$

# AMPLITUDE TOTAL E DIFERENÇA INTERQUARTIL

150	151	152	152	153	154	155	155	155
		151,5		Mediana		155		
		Q1		Q2		Q3		

## Amplitude total (AT)

$$AT = X_{(\max)} - X_{(\min)} = 155 - 150 = 5$$

## Diferença interquartil

$$Q3 - Q1 = 155 - 151,5 = 3,5$$

## Outliers

$$\text{Cerca inferior} = Q1 - (1.5 * DI) = 146,25$$

$$\text{Cerca superior} = Q3 + (1.5 * DI) = 160,25$$

# VARIÂNCIA, DESVIO PADRÃO E COEFICIENTE DE VARIAÇÃO

150	151	152	152	153	154	155	155	155
-----	-----	-----	-----	-----	-----	-----	-----	-----

$$2^2 = 4$$
$$10^2 = 100$$



o quanto longe os valores estão do “valor esperado”

$$\bar{x} = \frac{\sum x_i}{n}$$

$\frac{150 + 151 + 152 + 152 + 153 + 154 + 155 + 155 + 155}{9} = 153$

*Desvio*

$$\text{Desvio} = 3 \ 2 \ 1 \ 1 \ 0 \ 1 \ 2 \ 2 \ 2$$

$$3^2 + 2^2 + 1^2 + 1^2 + 0^2 + 1^2 + 2^2 + 2^2 + 2^2$$

$$9 + 4 + 1 + 1 + 0 + 1 + 4 + 4 + 4$$

$$28 / 9 = 3,11$$

$$\text{Desvio padrão} = \sqrt{3,11} = 1,76$$

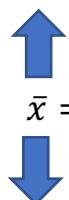
“Erro” se substituirmos pelo valor da média

$$CV = \frac{\sigma}{\bar{X}} \cdot 100$$

$$CV = \frac{1,76}{153} \cdot 100 = 1,15\%$$

# DESVIO PADRÃO – DADOS AGRUPADOS

Estatura (cm)	$f_i$	$x_i$	$f_i \cdot x_i$	$x_i^2$	$f_i \cdot x_i^2$	$F_i$
150  -- 154	5	152	760	23104	115520	5
154  -- 158	9	156	1404	24336	219024	14
158  -- 162	11	160	1760	25600	281600	25
162  -- 166	7	164	1148	26896	188272	32
166  -- 170	5	168	840	28224	141120	37
170  --  174	3	172	516	29584	88752	40
Total	40		6428		1034288	



$$\bar{x} = 160,7$$

$$dp = \sqrt{\frac{\sum f_i \cdot x_i^2}{\sum f_i} - \left( \frac{\sum f_i \cdot x_i}{\sum f_i} \right)^2}$$

$$dp = \sqrt{25857,2 - (160,7)^2}$$



$$dp = \sqrt{\frac{1034288}{40} - \left( \frac{6428}{40} \right)^2}$$

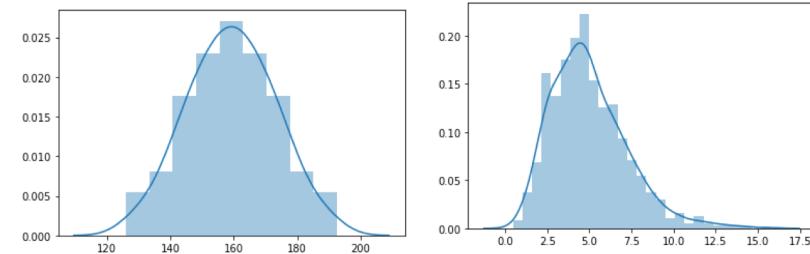
$$dp = \sqrt{25857,2 - 25824,49}$$

$$dp = 5,71$$

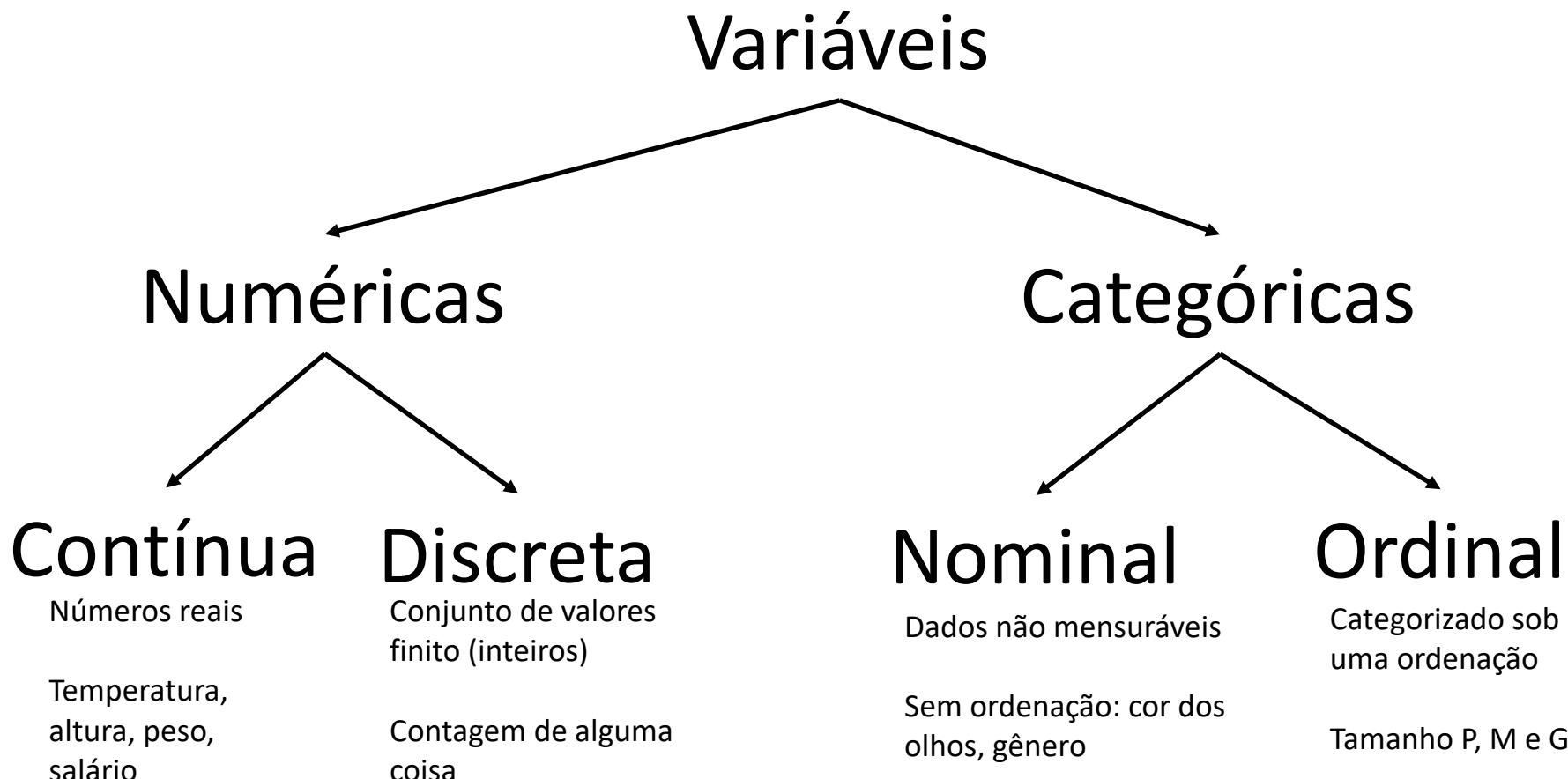
# DISTRIBUIÇÕES ESTATÍSTICAS

- Estatística inferencial
- Como os dados estão dispostos
- Distribuições
  - Normal (padronizada)
  - Gama
  - Exponencial
  - Uniforme
  - Bernoulli
  - Binomial
  - Poisson
- Naïve Bayes (Bernoulli e Multinomial)
- Padronização + kNN
- Tratamento de dados enviesados
- Inicialização de pesos em redes neurais
- Testes de normalidade (estatística paramétrica e não paramétrica)

Estatura (cm)	Frequência
150  -- 154	5
154  -- 158	9
158  -- 162	11
162  -- 166	7
166  -- 170	5
170  --  173	3
<b>Total</b>	<b>40</b>



# TIPOS DE VARIÁVEIS



# DISTRIBUIÇÃO NORMAL

18:50



18:40



19:00



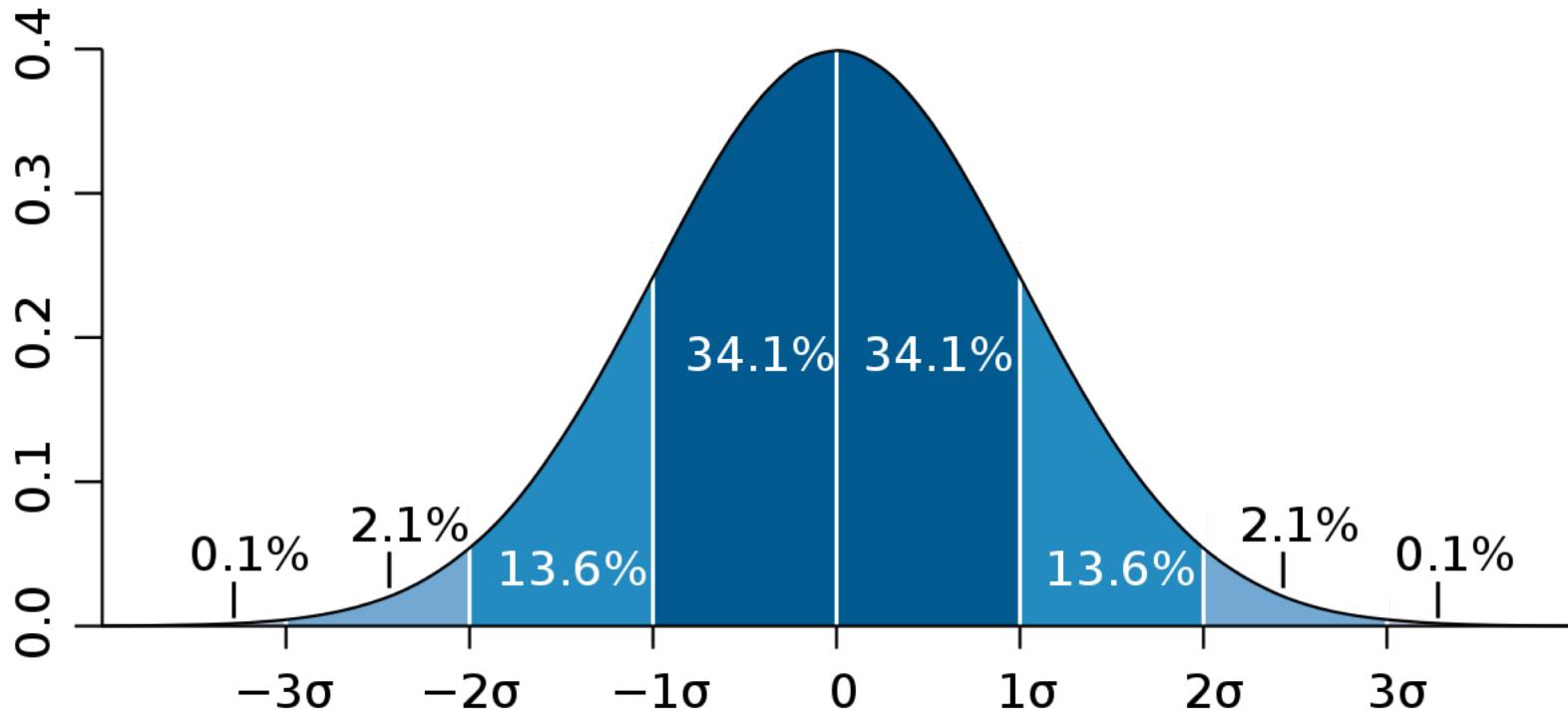
19:10



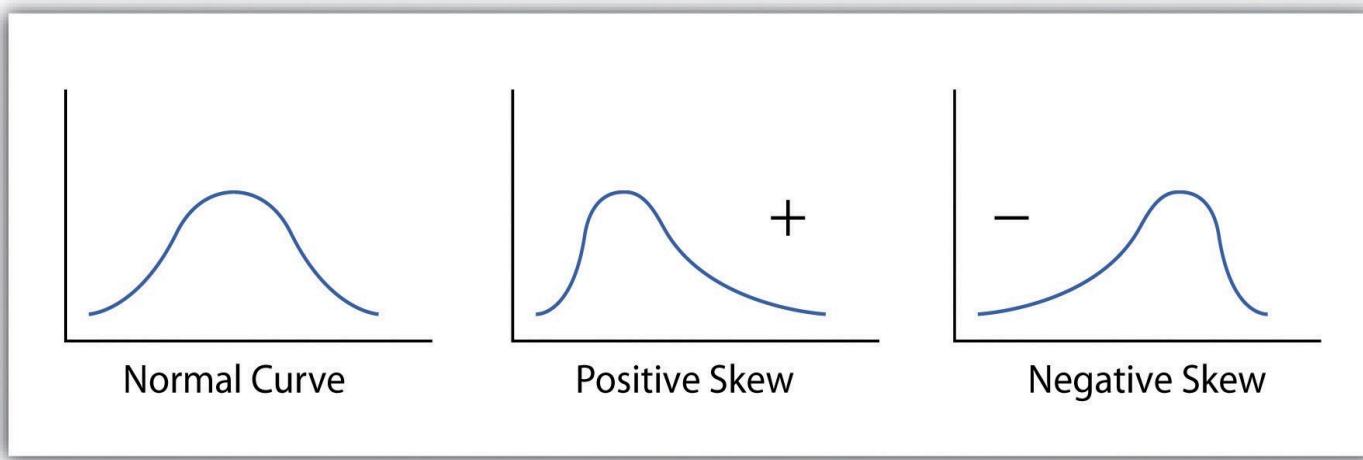
19:20



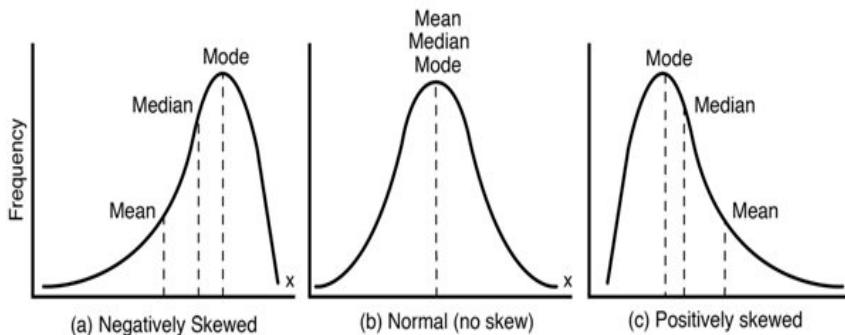
# DISTRIBUIÇÃO NORMAL



# ENVIESAMENTO



Fonte: <https://becominghuman.ai/how-to-deal-with-skewed-dataset-in-machine-learning-afd2928011cc>



Fonte: [https://www.researchgate.net/publication/294890337\\_ACOUSTIC\\_EMISSION\\_TESTS\\_ON\\_THE\\_ANALYSIS\\_OF\\_CRACKED\\_SHAFTS\\_OF\\_DIFFERENT\\_CRACK\\_DEPTHS/figures?lo=1](https://www.researchgate.net/publication/294890337_ACOUSTIC_EMISSION_TESTS_ON_THE_ANALYSIS_OF_CRACKED_SHAFTS_OF_DIFFERENT_CRACK_DEPTHS/figures?lo=1)

# DISTRIBUIÇÃO NORMAL PADRONIZADA

- Distribuições normais não possuem a mesma média e desvio padrão
- Difícil comparar resultados entre duas ou mais bases de dados
- Transformar a distribuição
  - Média: 0
  - Desvio padrão: 1

$$Z_{score} = \frac{x - \text{média}}{\text{desvio padrão}}$$

# DISTRIBUIÇÃO NORMAL PADRONIZADA

$$Z_{score} = \frac{x - \text{média}}{\text{desvio padrão}}$$

$$x = \frac{60 - 38,33}{20,20} = 1,07$$

$$x = \frac{35 - 38,33}{20,20} = -0,16$$

$$x = \frac{20 - 38,33}{20,20} = -0,90$$

Idade
60
35
20

Média = 38,33

Desvio padrão = 20,20

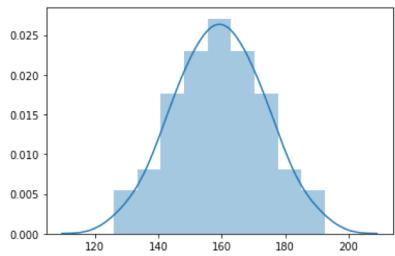
Idade
1,07
-0,16
-0,90

Média = 0,003

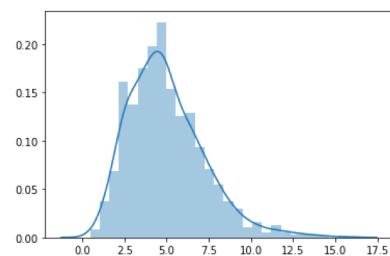
Desvio padrão = 0,995

# DISTRIBUIÇÕES

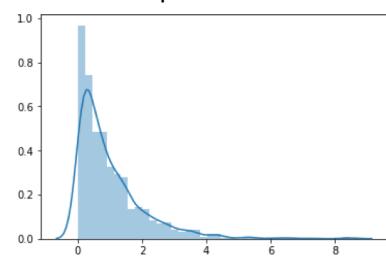
Normal



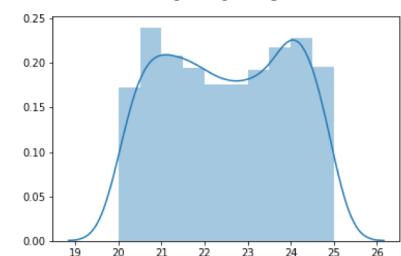
Gama



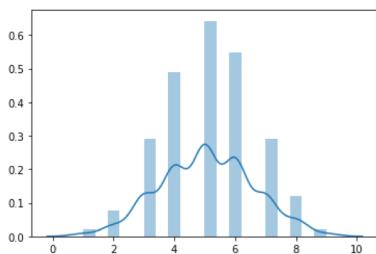
Exponencial



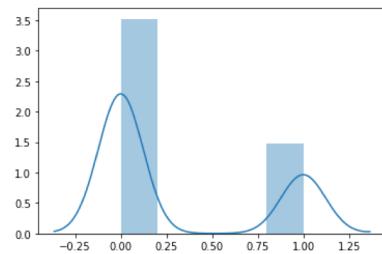
Uniforme



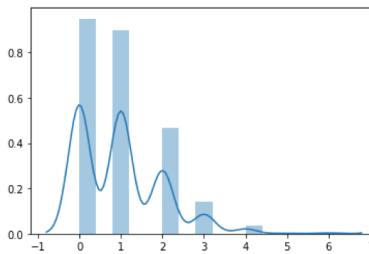
Binomial



Bernoulli



Poisson



# PROBABILIDADE

- Básico sobre probabilidade
  - Permutação
  - Combinação
  - Interseção, união e complemento
  - Eventos dependentes e independentes
- Redes Bayesianas
- Classificador ótimo de Bayes e Naïve Bayes
- Probabilidade e distribuições
  - Normal
  - Binomial
  - Poisson

# PROBABILIDADE

- São as chances de um evento ocorrer
- Representado com números entre 0 e 1
- Probabilidade de jogar uma moeda ( $1/2 = 0.5 = 50\%$ )
- O ato de jogar a moeda é chamado de tentativa (trial) – experimento
- Cada jogada da moeda é independente da outra
- Probabilidade de jogar um dado
- $P = 1/6 = 16\%$



# EXPERIMENTO, EVENTO E ESPAÇO AMOSTRAL

- Cada tentativa de jogar a moeda é chamado de um **experimento**
- Cada resultado (cara ou coroa) é chamado de **evento**
- A soma de todos os possíveis eventos é chamado de **espaço amostral**
- Exemplo – dados
  - Cada “jogada” é um experimento
  - Eventos: 1, 2, 3, 4, 5, 6
  - Espaço amostral: {E1, E2, E3, E4, E5, E6}



# PROBABILIDADE – EXEMPLOS

- Calcular a probabilidade de obter o número 5
- Evento:  $E_5 = 5$  (um evento)
- Espaço amostral:  $\{E_1, E_2, E_3, E_4, E_5, E_6\}$
- Probabilidade:  $P = \frac{\text{evento}}{\text{espaço amostral}}$
- $P = 1 / 6$
- $P = 0,16$  (16%)



## PROBABILIDADE – EXEMPLOS

- Temos uma mala com 6 bolas: 3 vermelhas, 2 amarelas e 1 azul
- Qual a probabilidade de selecionar uma bola amarela?
- Evento = 2
- Espaço amostral = 6
- $P = 2 / 6 (33\%)$

# PERMUTAÇÃO

- Arranjar objetos em uma sequência
- Quais são as permutações possíveis para as letras A, B e C?
- Fatorial!
- $3! = 3 \times 2 \times 1 = 6$  (permutações)
- ABC, ACB, BAC, BCA, CAB, CBA

# PERMUTAÇÃO PARA SUBCONJUNTOS

- Criar uma senha com 5 caracteres, que pode ser composto por letras e números de 0 até 9
- Números e letras não podem ser repetidos
- Evento: 5 caracteres
- Espaço amostral: 26 letras + 10 dígitos = 36
- $P_{(nr)} = \frac{n!}{(n-r)!}$
- $P_{(nr)} = \frac{36!}{(36-5)!} = 45.239.040$

# PERMUTAÇÃO PARA SUBCONJUNTOS

- Criar uma senha com 5 caracteres, que pode ser composto por letras e números de 0 até 9
- Números e letras PODEM ser repetidos
- Evento: 5 caracteres
- Espaço amostral: 26 letras + 10 dígitos = 36
- $n^r = 36^5 = 60.466.176$

# COMBINAÇÃO

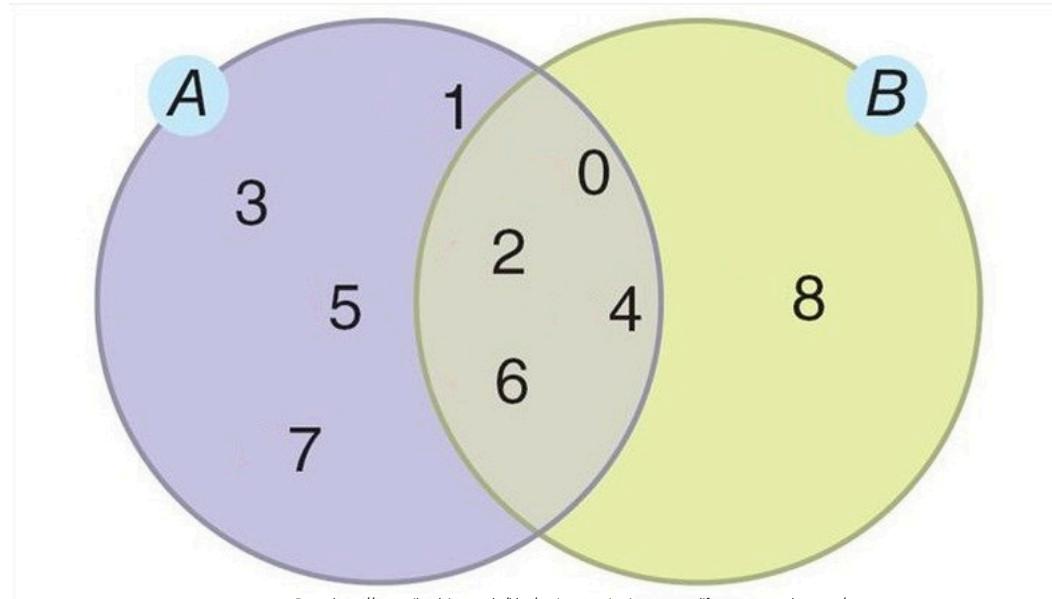
- Número possível de arranjos em uma coleção (a ordem não importa como na permutação)
- Quantas combinações de 2 letras podem ser feitas com ABCDEF?
- Sem considerar repetições
- $C_{(nr)} = \frac{n!}{r!(n-r)!}$
- $C_{(nr)} = \frac{6!}{2!(6-2)!} = 15$

# COMBINAÇÃO

- Número possível de arranjos em uma coleção (a ordem não importa como na permutação)
- Quantas combinações de 2 letras podem ser feitas com ABCDEF?
- Considerando repetições
- $C_{(nr)} = \frac{(n+r-1)!}{r!(n-1)!}$
- $C_{(nr)} = \frac{(6+2-1)!}{2!(6-2)!} = 105$

# INTERSEÇÃO

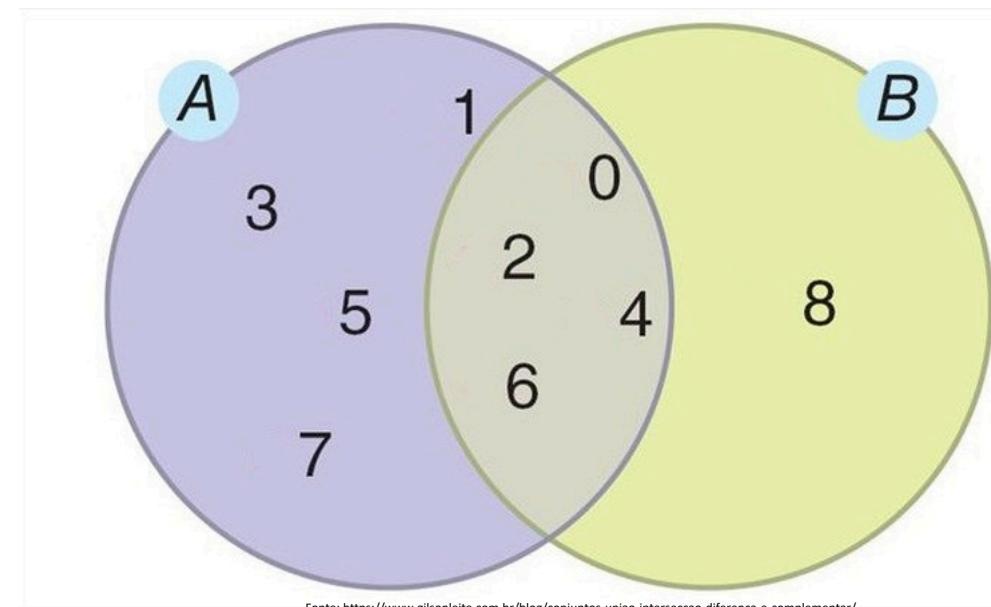
- $A \cap B$
- $B \cap A$
- $P(A \cap B) = 4 / 9 = 0,44$  (44%)



Fonte: <https://www.gilsonleite.com.br/blog/conjuntos-uniao-interseccao-diferenca-e-complementar/>

# UNIÃO

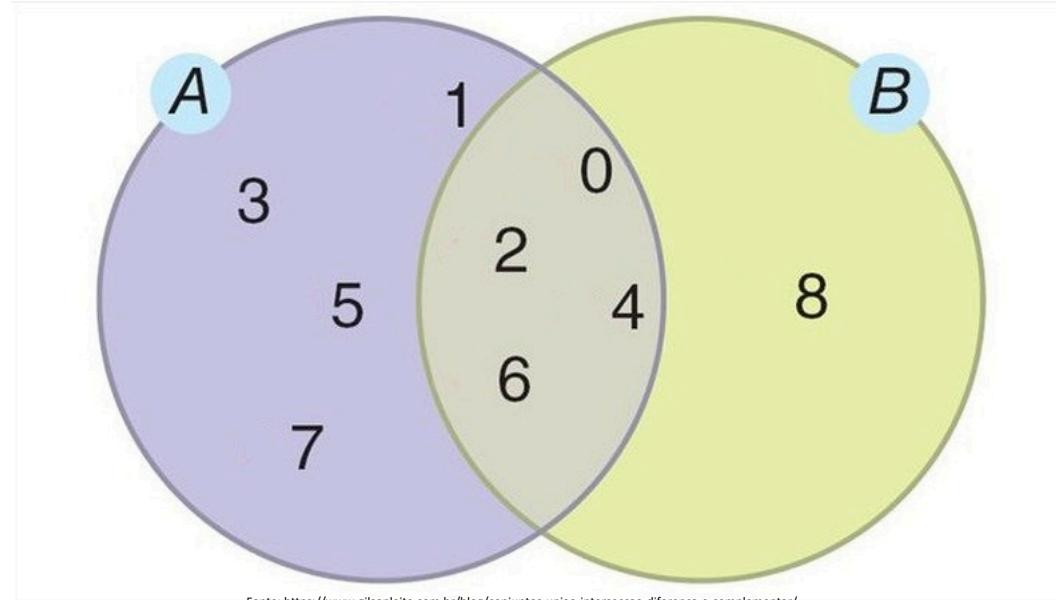
- $A \cup B$
- $B \cup A$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) = 8/9 + 5/9 - 4/9$
- $P(A \cup B) = 0,88 + 0,55 - 0,44$
- $P(A \cup B) = 0,99$



Fonte: <https://www.gilsonleite.com.br/blog/conjuntos-uniao-interseccao-diferenca-e-complementar/>

# COMPLEMENTO

- Estão em um conjunto mas não estão em outro
- Complementar de B em relação A é  $A - B$
- $A - B$
- $P(\bar{A}) = 1 - P(A)$
- $P(\bar{A}) = 1 - 8/9$
- $P(\bar{A}) = 0,12$
- $P(\bar{B}) = 1 - P(B)$
- $P(\bar{B}) = 1 - 5/9$
- $P(\bar{B}) = 0,45$



# EVENTOS INDEPENDENTES

- O resultado de um evento não influencia na resposta de outro evento
- Jogar uma moeda 2 vezes (as chances são independentes)
- Calcular a probabilidade de obter dois “coroas” em duas tentativas
- $P = \frac{1}{2} \cdot \frac{1}{2}$
- $P = \frac{1}{4} = 0.25 \text{ (25\%)}$



# EVENTOS DEPENDENTES

- O resultado do primeiro evento influencia no resultado do segundo evento
- Um baralho possui 52 cartas e 13 dessas são de “espada”
- Qual a probabilidade de tirar 2 cartas de espada?

$$\bullet P = \frac{13}{52} \cdot \frac{12}{51}$$

$$\bullet P = \frac{156}{2652} = 0.05 \text{ (5,88%)}$$

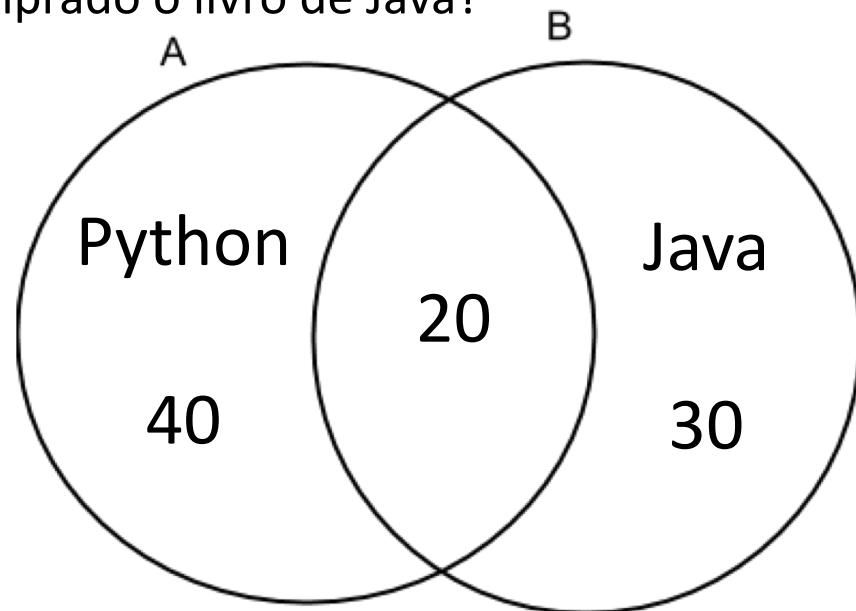


# PROBABILIDADE CONDICIONAL

- Calcular a probabilidade do evento A, dado que o evento B ocorreu
- $P(a|b) = x$ , pode ser lido como: “*Dado o evento b, a probabilidade do evento a é x*”
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(a|b) = P(a \cap b) / P(b)$ , ou  $P(a|b)P(b) = P(a,b)$ .  $P(a,b)$  é a probabilidade do evento conjunto do evento  $a \wedge b$
- Exemplo
  - $P(\text{Cárie} | \text{Dor}) = 0.8$ , indica que caso um paciente esteja com dor (de dente) e nenhuma outra informação esteja disponível, então, a probabilidade do paciente ter uma cárie é de 0.8

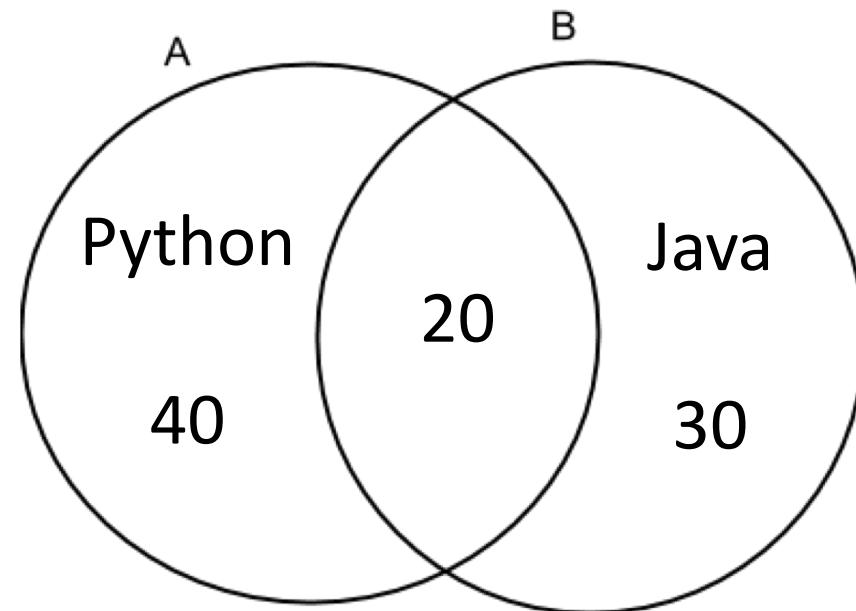
# PROBABILIDADE CONDICIONAL

- Em um grupo de 90 pessoas, 40 compraram o livro de Python, 30 compraram o livro de Java e 20 compraram o livro de Python e de Java. Se escolhermos uma pessoa que comprou o livro de Python, qual a probabilidade desta pessoa ter comprado o livro de Java?
- $P(B|A) = \frac{P(A \cap B)}{P(A)}$
- $P(B|A) = \frac{20}{40} = 0.5 \text{ (50\%)}$



# PROBABILIDADE CONDICIONAL – ADIÇÃO

- Se selecionarmos um elemento randomicamente, qual a probabilidade de comprar o livro de Python ou de Java?
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) = \frac{40}{90} + \frac{30}{90} - \frac{20}{90}$
- $P(A \cup B) = 0,4 + 0,3 - 0,2$
- $P(A \cup B) = 0,5$



# REDES BAYESIANAS

- Probabilidade condicional
  - $P(a/b) = x$ , pode ser lido como: “*Dado o evento b, a probabilidade do evento a é x*”
- Regra fundamental
  - $P(a/b) = P(a,b)/P(b)$ , ou  $P(a/b)P(b) = P(a,b)$ .  $P(a,b)$  é a probabilidade do evento conjunto do evento  $a \wedge b$
- Exemplo 1
  - $P(\text{Cárie}/\text{Dor}) = 0.8$ , indica que caso um paciente esteja com dor (de dente) e nenhuma outra informação esteja disponível, então, a probabilidade do paciente ter uma cárie é de 0.8

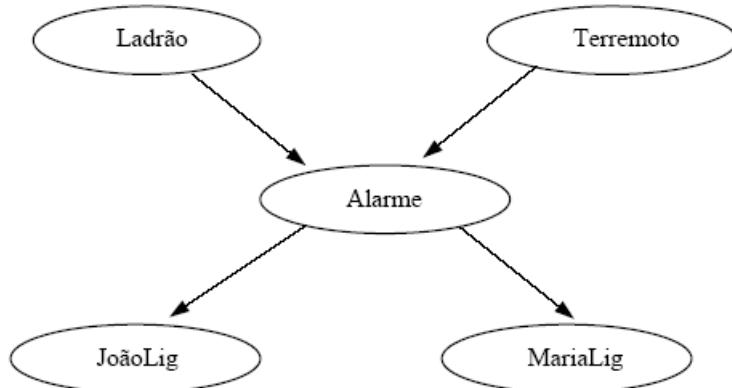
# REDES BAYESIANAS

- Exemplo 2
  - Um médico sabe que a meningite causa torcicolo em 50% dos casos. Porém, o médico também conhece algumas probabilidades incondicionais que dizem que, um caso de meningite atinge 1/50000 pessoas e, a probabilidade de alguém ter torcicolo é de 1/20. ”
  - $T$  e  $M$ , é probabilidade incondicional de um paciente ter torcicolo e a probabilidade incondicional de um paciente ter meningite
    - $P(T|M) = 0.5$  (probabilidade de ter torcicolo tendo meningite)
    - $P(M) = 1/50000$
    - $P(T) = 1/20$
  - Aplicando a fórmula
    - $P(M|T) = (P(T|M)P(M))/P(T) = (0.5 \times 1/50000)/(1/20) = 0.0002$

## REDES BAYESIANAS – PROBLEMA DO ALARME

*Você possui um novo alarme contra ladrões em casa. Este alarme é muito confiável na detecção de ladrões, entretanto, ele também pode disparar caso ocorra um terremoto. Você tem dois vizinhos, João e Maria, os quais prometeram telefonar-lhe no trabalho caso o alarme dispare. João sempre liga quando ouve o alarme, entretanto, algumas vezes confunde o alarme com o telefone e também liga nestes casos. Maria, por outro lado, gosta de ouvir música alta e às vezes não escuta o alarme*

# REDES BAYESIANAS – PROBLEMA DO ALARME



Ladrão	Terremoto	$P(\text{Alarme} \text{Ladrão}, \text{Terremoto})$	
		Verdadeiro	Falso
Verdadeiro	Verdadeiro	0.95	0.050
Verdadeiro	Falso	0.95	0.050
Falso	Verdadeiro	0.29	0.71
Falso	Falso	0.001	0.999

# REDES BAYESIANAS – PROBLEMA DO ALARME

L	P(L)
V	.001

Ladrão

T	P(T)
V	.002

Terremoto

M	P(M)
V	.70
F	.01

Alarme

J	P(J)
V	.90
F	.05

JoãoLig

MariaLig

L	T	P(A)
V	V	.95
V	F	.95
F	V	.29
F	F	.001

# REDES BAYESIANAS

- Considere que se deseja calcular a probabilidade do alarme ter tocado, mas, nem um ladrão nem um terremoto aconteceram, e ambos, João e Maria ligaram, ou  $P(J \wedge M \wedge A \wedge \neg L \wedge \neg T)$ .
- $P(J \wedge M \wedge A \wedge \neg L \wedge \neg T) = P(J|A)P(M|A)P(A|\neg L \wedge \neg T)P(\neg L)P(\neg T)$
- $= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
- $= 0.00062$

# PROBABILIDADE – EXERCÍCIOS

- Dado o lançamento de um dado
- Probabilidade de obter um número par
  - Evento = 2, 4 e 6 (3)
  - Espaço amostral = 1, 2, 3, 4, 5, 6
  - $P = 3 / 6 = 0,5$  (50%)
- Probabilidade de obter um número menor do que 6
  - Evento = 1, 2, 3, 4 e 5 (5)
  - Espaço amostral = 1, 2, 3, 4, 5, 6
  - $P = 5 / 6 = 0,83$  (83%)



# PROBABILIDADE – EXERCÍCIOS

- Um baralho possui 52 cartas e 13 dessas são de “espada”. Qual a probabilidade de tirar 5 cartas de espadas sem reposição?

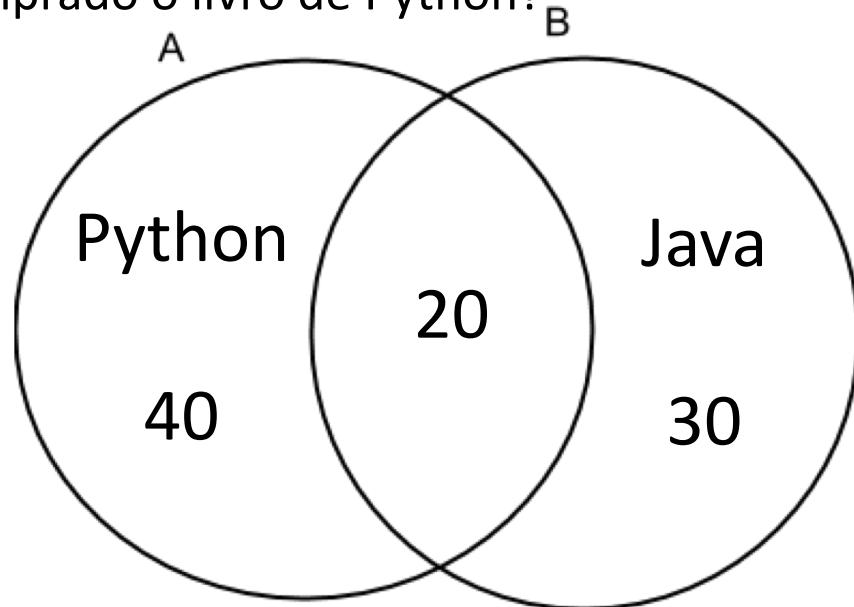
- $P = \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} = 0,0005 \text{ (0,05\%)}$

- Evento dependente



# PROBABILIDADE – EXERCÍCIOS

- Em um grupo de 90 pessoas, 40 compraram o livro de Python, 30 compraram o livro de Java e 20 compraram o livro de Python e de Java. Se escolhermos uma pessoa que comprou o livro de Java, qual a probabilidade desta pessoa ter comprado o livro de Python?
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(B|A) = \frac{20}{30} = 0.66 \text{ (66\%)}$



# NAÏVE BAYES

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	=> 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	=> 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	=> 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.000	Baixo
Ruim	Alta	Nenhuma	=> 15.000 a <= 35.000	Alto

História = Boa

Dívida = Alta

Garantias = Nenhuma

Renda = > 35

Soma: 0,0079 + 0,0052 + 0,0514 = **0,0645**

$$P(\text{Alto}) = 6/14 * 1/6 * 4/6 * 6/6 * 1/6$$

$$P(\text{Alto}) = 0,0079$$

$$P(\text{Alto}) = 0,0079 / 0,0645 * 100 = \mathbf{12,24\%}$$

$$P(\text{Moderado}) = 3/14 * 1/3 * 1/3 * 2/3 * 1/3$$

$$P(\text{Moderado}) = 0,0052$$

$$P(\text{Moderado}) = 0,0052 / 0,0645 * 100 = \mathbf{8,06\%}$$

$$P(\text{Baixo}) = 5/14 * 3/5 * 2/5 * 3/5 * 5/5$$

$$P(\text{Baixo}) = 0,0514$$

$$P(\text{Baixo}) = 0,0514 / 0,0645 * 100 = \mathbf{79,68\%}$$

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15000 3	=> 15000 4	<= 35000 4
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

# CLASSIFICADOR ÓTIMO DE BAYES

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	≥ 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	≥ 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	≥ 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.000	Baixo
Ruim	Alta	Nenhuma	≥ 15.000 a <= 35.000	Alto

$x_t = \langle \text{História} = \text{Boa}, \text{Dívida} = \text{Alta}, \text{Garantias} = \text{Nenhuma}, \text{Renda} = > 35 \rangle$

$$P(c_j|x_t) = \frac{P(x_t|c_j) \cdot P(c_j)}{P(x_t)}$$



Probabilidade  
a posteriori

Probabilidades a priori das classes

$$P(\text{alto}) = \frac{6}{14} = 0,43 \text{ (43\%)}$$

$$P(\text{moderado}) = \frac{3}{14} = 0,22 \text{ (22\%)}$$

$$P(\text{baixo}) = \frac{5}{14} = 0,35 \text{ (35\%)}$$

Probabilidades condicionais

$$P(x_t|_{\text{alto}}) \quad P(x_t|_{\text{moderado}}) \quad P(x_t|_{\text{baixo}})$$

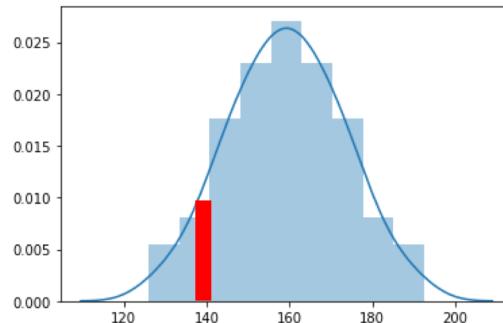
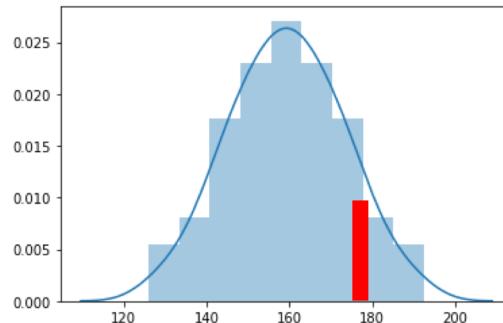
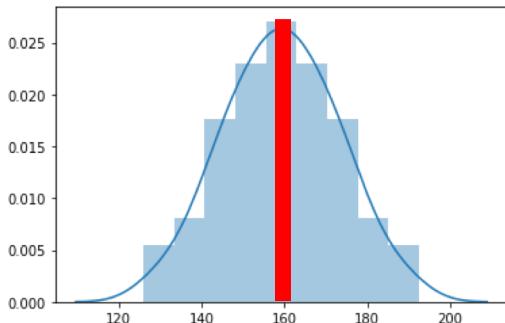
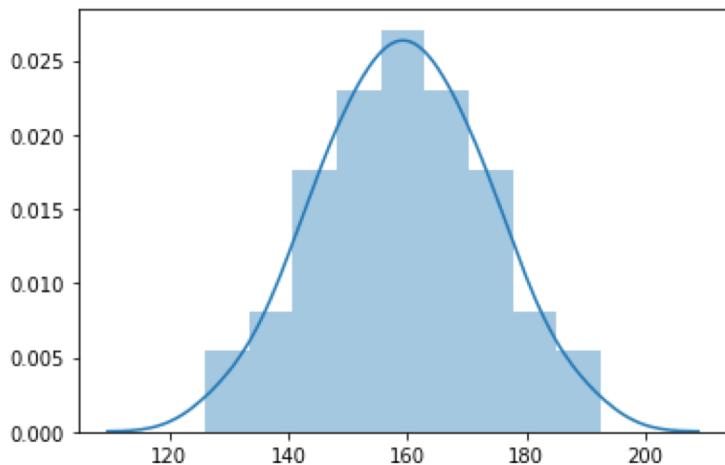
$$\begin{aligned} P &= \langle \text{História} = \text{Boa}, \text{Dívida} = \text{Alta}, \text{Garantias} = \text{Nenhuma}, \text{Renda} = > 35 | \text{alto} \rangle \\ P &= \langle \text{História} = \text{Ruim}, \text{Dívida} = \text{Alta}, \text{Garantias} = \text{Nenhuma}, \text{Renda} = > 35 | \text{alto} \rangle \\ P &= \langle \text{História} = \text{Desconhecida}, \text{Dívida} = \text{Alta}, \text{Garantias} = \text{Nenhuma}, \text{Renda} = > 35 | \text{alto} \rangle \end{aligned}$$

$$3 \times (3 \times 2 \times 2 \times 3) = 108 \text{ probabilidades condicionais!}$$

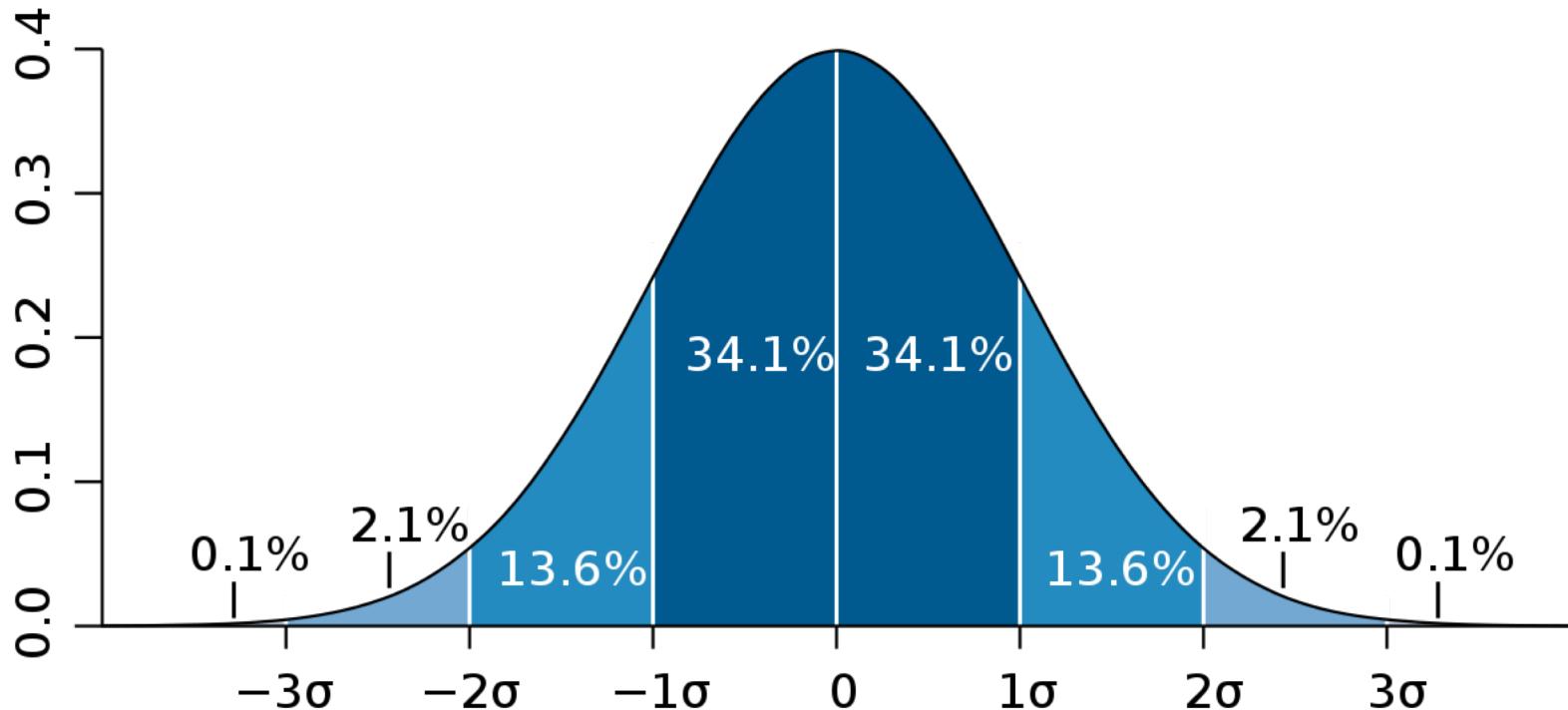
Ótimo de Bayes: características condicionalmente dependentes

Naïve Bayes: características condicionalmente independentes

# PROBABILIDADE – DISTRIBUIÇÃO NORMAL



# DISTRIBUIÇÃO NORMAL

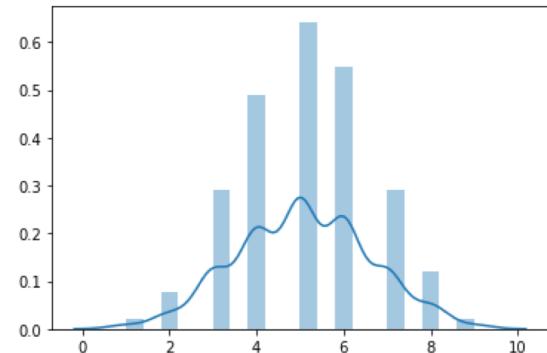


# PROBABILIDADE – DISTRIBUIÇÃO BINOMIAL

- Respostas sucesso ou fracasso e experimentos independentes
- Moedas e baralho?
- Probabilidade de selecionar “coroa” 5 vezes
- Parâmetros
  - $X = 5$  (número de sucessos)
  - $p = 0,5$  (probabilidade de sucesso)
  - $n = 10$  (quantidade de tentativas – trials)

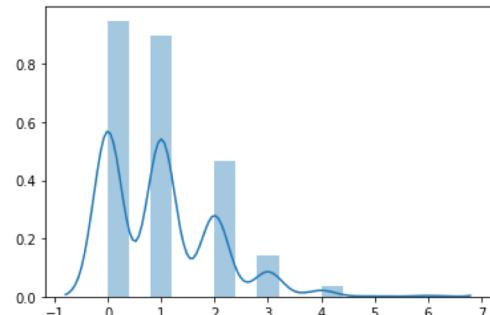


$$P(X=x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$



# PROBABILIDADE – DISTRIBUIÇÃO DE POISSON

- Ocorrência de eventos no decorrer do tempo (não considera o número de experimentos)
- Os eventos devem ser independentes
- Considera o número de “sucessos” baseado no tempo
- Parâmetros da fórmula
  - X: número de eventos calculados
  - Número de Euler (2.71828)
  - Número médio de eventos
- O número médio de carros vendidos por dia é 10. Qual a probabilidade de vender 14 carros amanhã?
  - X = 14
  - Média: 10

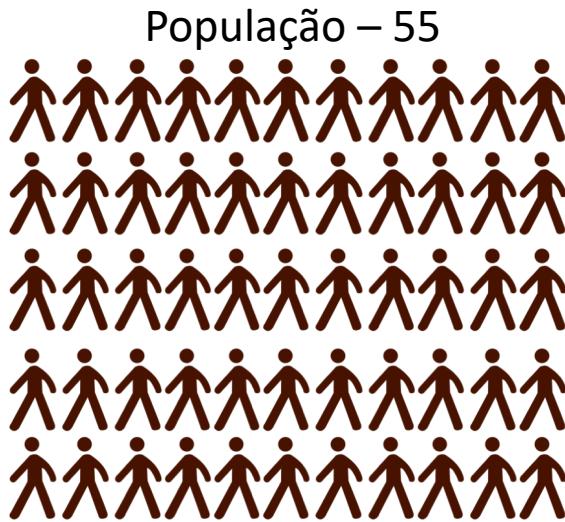


$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

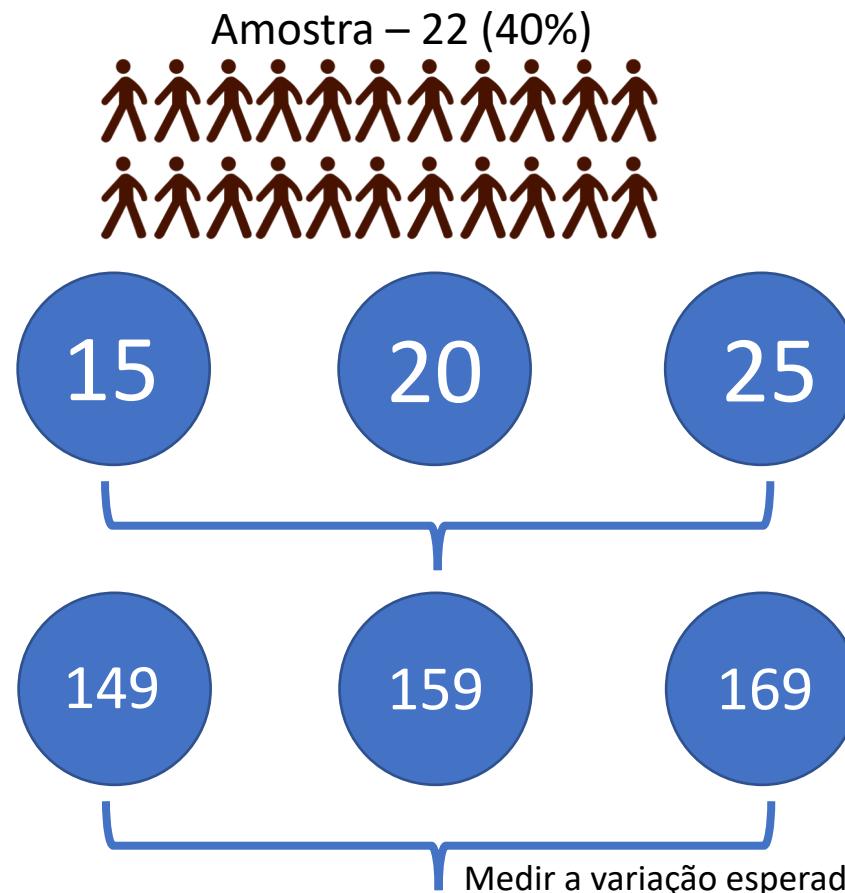
# INTERVALOS DE CONFIANÇA E TESTES DE HIPÓTESES

- Intervalos de confiança – cálculos passo a passo
- Distribuição T Student
- Intervalos de confiança em machine learning
- Testes de hipóteses
  - Teste Z
  - Teste T
  - Qui quadrado
  - ANOVA
  - Qui quadrado e ANOVA para seleção de atributos
- Testes de Wilcoxon, Friedman e Nemenyi
- Aplicações para avaliação de algoritmos/trabalhos científicos

# INTERVALOS DE CONFIANÇA



Indica que os experimentos estarão dentro do intervalo de confiança com certeza de 95%. Em 95% dos casos, a população “real” estará neste intervalo



# INTERVALOS DE CONFIANÇA – CÁLCULOS

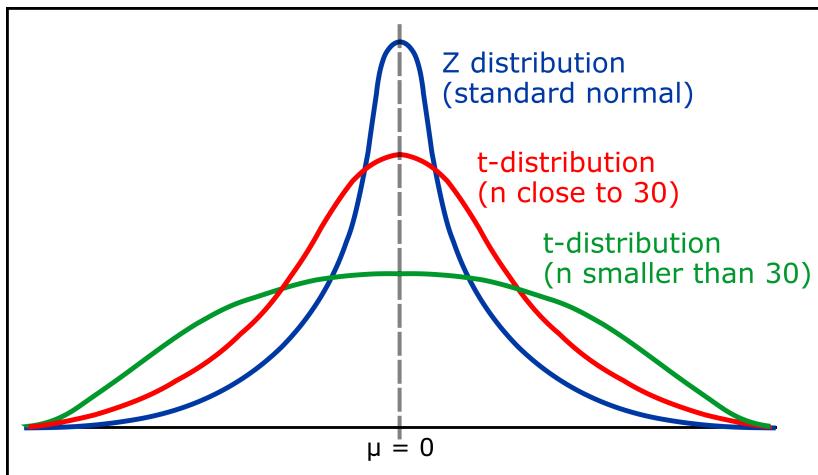
- $[\bar{x} - Za_{/2} \frac{\sigma}{\sqrt{n}}]$
- $[\bar{x} + Za_{/2} \frac{\sigma}{\sqrt{n}}]$
- Parâmetros da fórmula
  - Média: 159,25
  - Desvio padrão: 13,65
  - n: quantidade de números
  - alpha: 1 – confiança ( $1 - 0,95 = 0,05$ )
- $\frac{a}{2} = \frac{0,05}{2} = 0,025, 1 - 0,025 = 0,975$
- $[159,25 - 1,96 \frac{13,65}{\sqrt{100}}], [159,25 + 1,96 \frac{13,65}{\sqrt{100}}] = [156,57 \text{ } 161,92]$
- Diferença: 2,67 ( $159,25 - 156,57$  ou  $161,92 - 159,57$ )

# INTERVALOS DE CONFIANÇA

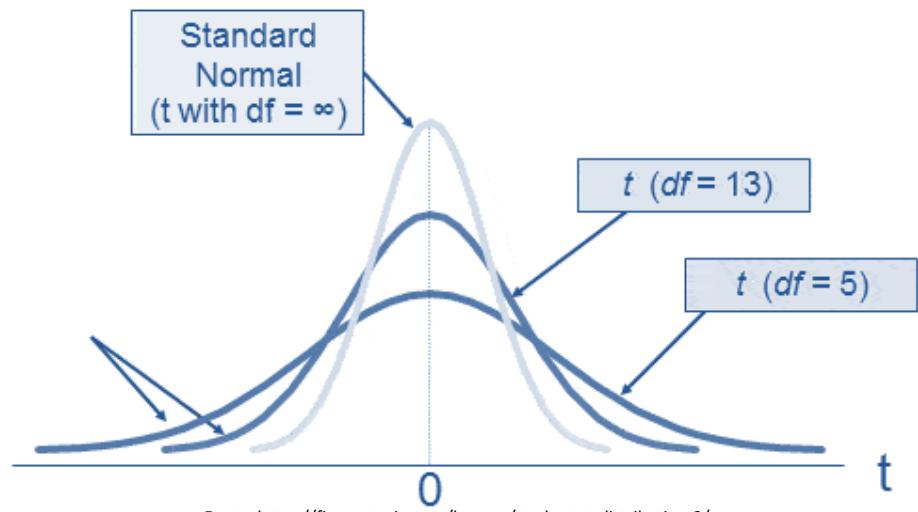
Nível de confiança	Alpha	Z-score
90%	10% (0,10)	1.645
95%	5% (0,05)	1.96
98%	2% (0,02)	2.33
99%	1% (0,01)	2.575

# DISTRIBUIÇÃO T STUDENT

- Poucos dados e variação não conhecida (30 números)
- Maior dispersão dos dados
- Graus de liberdade



Fonte: <https://andyjconnelly.wordpress.com/2017/05/16/uncertainty-and-repeats/>



Fonte: <https://financetrain.com/lessons/students-t-distribution-2/>

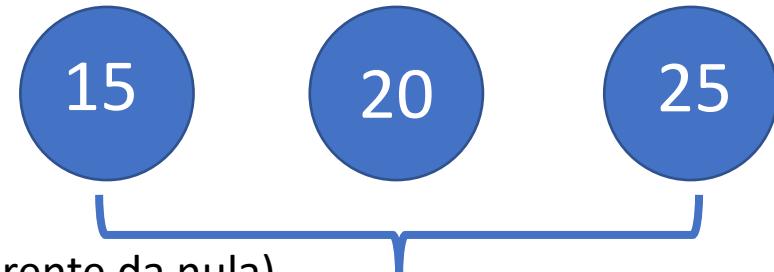
# DISTRIBUIÇÃO T STUDENT

- $[\bar{x} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}]$
- Parâmetros da fórmula
  - Média: 161,77
  - Desvio padrão da amostra (S): 12,78
  - n: quantidade de números
  - alpha: 1 - confiança ( $1 - 0,95 = 0,05$ )
- $t_{n-1} = 8$
- $\frac{\alpha}{2} = \frac{0,05}{2} = 0,025 = 2,306$  (consultar tabela)
- $[161,77 - 2,306 \frac{12,78}{\sqrt{9}}], [161,77 + 2,306 \frac{12,78}{\sqrt{9}}] = [151,94 \text{ } 171,59]$
- Diferença: 9,83 ( $161,77 - 151,94$  ou  $171,59 - 161,77$ )

149
160
147
189
175
168
156
160
152

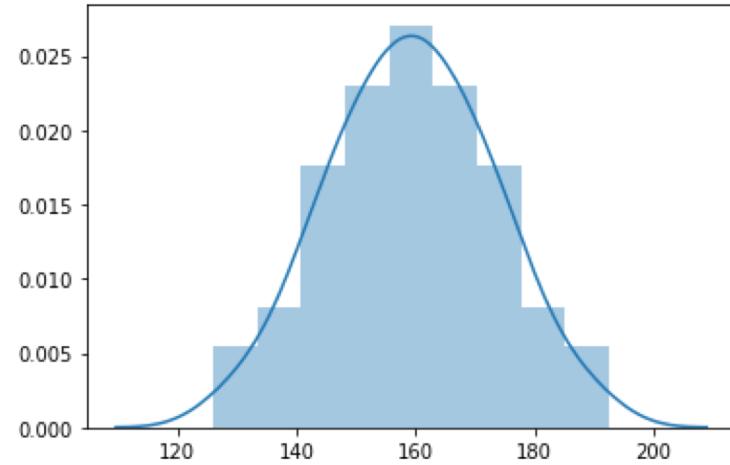
# TESTES DE HIPÓTESES

- Resposta sim ou não, para confirmar ou rejeitar uma afirmação
- Hipótese: ideia a ser testada
- Hipótese nula ( $H_0$ )
  - Afirmação que já existia
  - Presumir que é verdadeira até que se prove o contrário
- Hipótese alternativa ( $H_1$ )
  - O que está tentando provar (tudo o que é diferente da nula)
- Alpha
  - Probabilidade de rejeitar a hipótese nula, quanto menor mais seguro é o resultado (nível de significância) – em geral 0,01 ou 0,05
  - 5% de chances de concluir que existe uma diferença quando não há diferença real
- Valor de p (p-value)
  - $p\text{-value} \geq \alpha$ : não rejeita  $H_0$  (não temos evidências)
  - $p\text{-value} < \alpha$ : rejeita  $H_0$  (temos evidência)
- Erro Tipo I: rejeitar a hipótese nula quando não deveria
- Erro Tipo II: não rejeitar nula quando deveria ter rejeitado



# TESTE DE HIPÓTESE Z

- $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
- Parâmetros da fórmula
  - Média H1: 164,02
  - Média H0: 159,25
  - Desvio padrão H1: 14,05
  - n: quantidade de números
  - alpha: 0,05
- $Z = \frac{164,02 - 159,25}{\frac{14,05}{\sqrt{100}}} = \frac{4,77}{1,4} = 3,39$
- $Z = 0,999$  (buscar na tabela)
- Valor de p =  $1 - 0,999 = 0,001$



Valor de p é menor que alpha, o que indica que rejeitamos a hipótese nula (H0) e aceitamos a hipótese H1

A média atual de alturas é de 164,02

# QUI QUADRADO

Frequência observada	Visão computacional	Algoritmos de busca	Total
Homens	30	20	50
Mulheres	22	28	50
<b>Total</b>	<b>52</b>	<b>48</b>	<b>100</b>

Frequência esperada	Visão computacional	Algoritmos de busca	Total
Homens	26	24	50
Mulheres	26	24	50
<b>Total</b>	<b>52</b>	<b>48</b>	<b>100</b>

$$(52 \times 50) / 100 = 26$$

Fonte dos dados: <https://www.youtube.com/watch?v=4QfHVbpAoSg>

# QUI QUADRADO

Frequência ( $f_o$ )	Frequência esperada ( $f_e$ )	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
30	26	4	16	0,62
20	24	-4	16	0,67
22	26	-4	16	0,62
28	24	4	16	0,67

$$\text{Grau liberdade} = (r - 1)(c - 1)$$

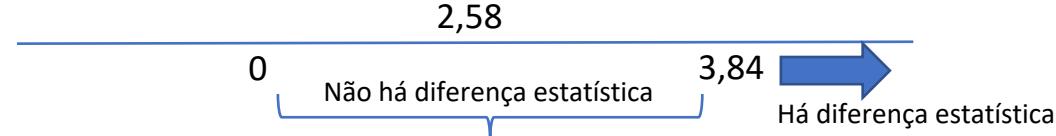
$$\text{Grau liberdade} = (2 - 1)(2 - 1) = 1$$

Alpha = 0,05

$\chi^2$  crítico = 3,84 (consultar tabela)

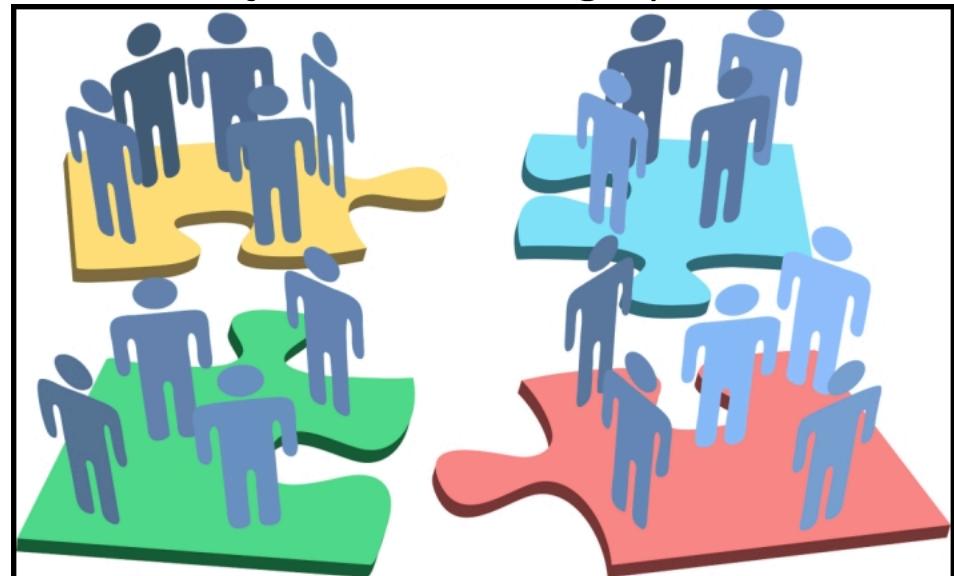
$$\chi^2 = 2,58$$

Qui quadrado



# ANOVA – ANÁLISE DE VARIAÇÃO

- Comparação entre 3 ou mais grupos (amostras independentes)
- Uma variável quantitativa e uma ou mais variáveis qualitativas
- Distribuição normal (estatística paramétrica)
- Variação entre os grupos comparando a variação dentro dos grupos
- H<sub>0</sub>: não há diferença estatística
- H<sub>1</sub>: existe diferença estatística



Fonte: <https://marcelocoruja.blogspot.com/2017/04/sociologia-importancia-dos-grupos-e-das.html>

# ANOVA – ANÁLISE DE VARIAÇÃO

	Grupo A	Grupo B	Grupo C
165	130	163	
152	169	158	
143	164	154	
140	143	149	
155	154	156	
Média	151	152	156

Média geral: 153

F crítico = 3,88 (consultar tabela)

0,27

Não há diferença estatística

Quadrado		
Grupo A	Grupo B	Grupo C
$(151 - 153)^2 = 4$	$(152 - 153)^2 = 1$	$(156 - 153)^2 = 9$
		Total: 14

SSG (sum of squares group):  $14 \times 5 = 70$

DFG (degrees of freedom groups):  $3 - 1 = 2$

$$F = \frac{\frac{SSG}{DFG}}{\frac{SSE}{DFE}}$$

$$F = \frac{\frac{70}{2}}{\frac{1506}{12}} = 0.2788$$

Quadrado erro		
$(valor - média)^2$	$(valor - média)^2$	$(valor - média)^2$
$(165 - 151)^2 = 196$	$(130 - 152)^2 = 484$	$(163 - 156)^2 = 49$
$(152 - 151)^2 = 1$	$(169 - 152)^2 = 289$	$(158 - 156)^2 = 4$
$(143 - 151)^2 = 64$	$(164 - 152)^2 = 144$	$(154 - 156)^2 = 4$
$(140 - 151)^2 = 121$	$(143 - 152)^2 = 81$	$(149 - 156)^2 = 49$
$(155 - 151)^2 = 16$	$(154 - 152)^2 = 4$	$(156 - 156)^2 = 0$
Soma	398	1002
		106



3,88 Há diferença estatística

# CORRELAÇÃO E REGRESSÃO

- Correlação: correspondência entre variáveis
- Regressão: previsões
- Covariância, correlação e determinação – cálculos passo a passo e implementação
- Regressão linear simples e múltipla
- Métricas de erro

# COVARIÂNCIA, COEFICIENTE DE CORRELAÇÃO E COEFICIENTE DE DETERMINAÇÃO

Tamanho (m <sup>2</sup> )	Preço	x <sub>i</sub> - $\bar{x}$	y <sub>i</sub> - $\bar{y}$	(x <sub>i</sub> - $\bar{x}$ ) * (y <sub>i</sub> - $\bar{y}$ )
30	57.000	-14,5	-16.250	235.625
39	69.000	-5,5	-4.250	23.375
49	77.000	4,5	3.750	16.875
60	90.000	15,5	16.750	259.625
44,5 (média)	73.250 (média)			
12,92 (dp)	13.865,42 (dp)			<b>535.500</b> (soma)

$$C(x,y) = \frac{\sum (xi - \bar{x}) * (yi - \bar{y})}{n - 1} \quad Cr(x,y) = \frac{Cov(x,y)}{Std(x) * Std(y)}$$

$$C(x,y) = \frac{535.500}{3} = 178500,00 \quad Cr(x,y) = \frac{178500,00}{12,92 * 13865,42} = 0,99$$

> 0, variáveis se movem juntas

< 0, variáveis se movem em direções opostas

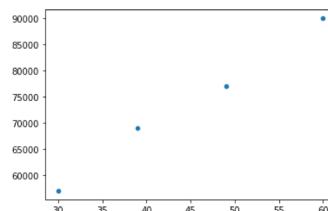
= 0, variáveis são independentes

$$Cd(x,y) = Cr^2$$

$$Cd(x,y) = 0,99^2$$

$$Cd(x,y) = 0,98$$

98% da variável dependente consegue ser explicada pelas variáveis explanatórias

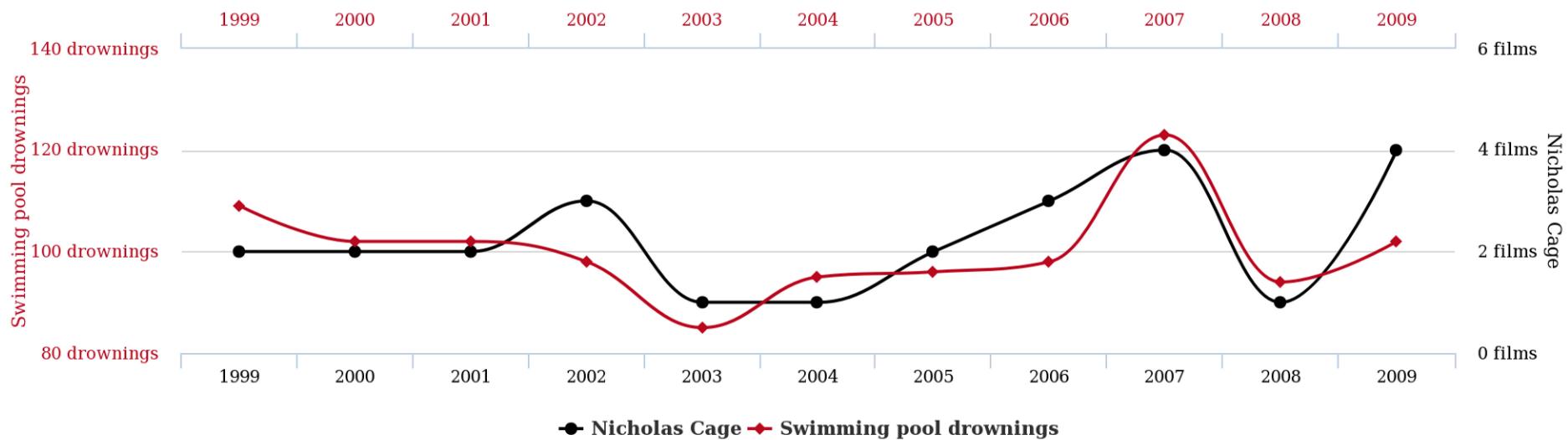


# COEFICIENTE DE CORRELAÇÃO

Correlação	Interpretação
0,00 a 0,19 ou 0,00 a -0,19	Correlação bem fraca
0,20 a 0,39 ou -0,20 a -0,39	Correlação fraca
0,40 a 0,69 ou -0,40 a -0,69	Correlação moderada
0,70 a 0,89 ou -0,70 a -0,89	Correlação forte
0,90 a 1,00 ou -0,90 a -1,00	Correlação muito forte

# CORRELAÇÃO NÃO É CAUSA

**Number of people who drowned by falling into a pool**  
correlates with  
**Films Nicolas Cage appeared in**



tylervigen.com

# MÉTRICAS DE ERROS

- Mean absolute error (MAE)
  - Diferenças absolutas entre as previsões e os valores reais
- Mean squared error (MSE)
  - Diferenças elevadas ao quadrado (erros penalizados)
- Root mean squared error (RMSE)
  - Interpretação facilitada

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

# VISUALIZAÇÃO

- Gráficos: dispersão, barra, pizza (setor), linha
- Boxplot
- Atributos categóricos
- Subgráficos
- Mapas

