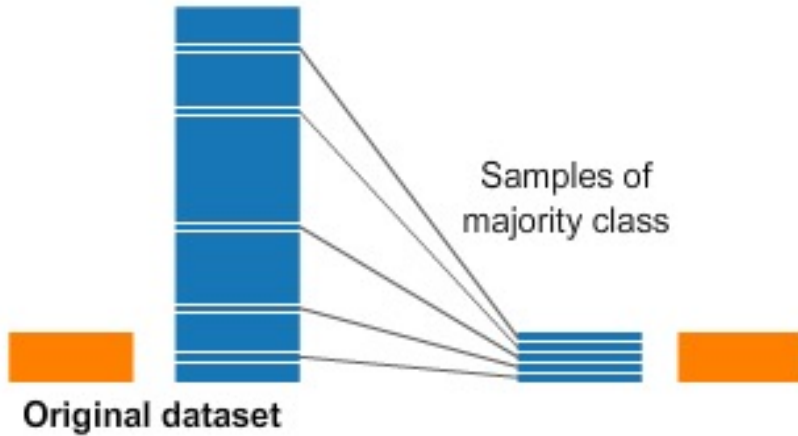


# ESTATÍSTICA PARA CIÊNCIA DE DADOS E MACHINE LEARNING

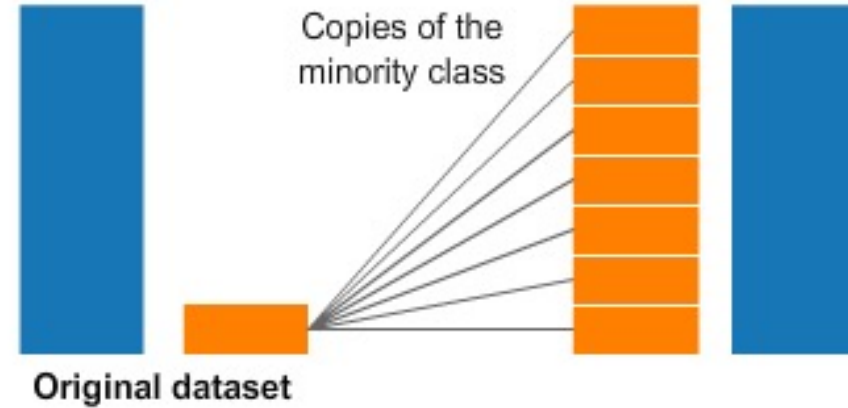


# SUBAMOSTRAGEM E SOBREAMOSTRAGEM

## Undersampling

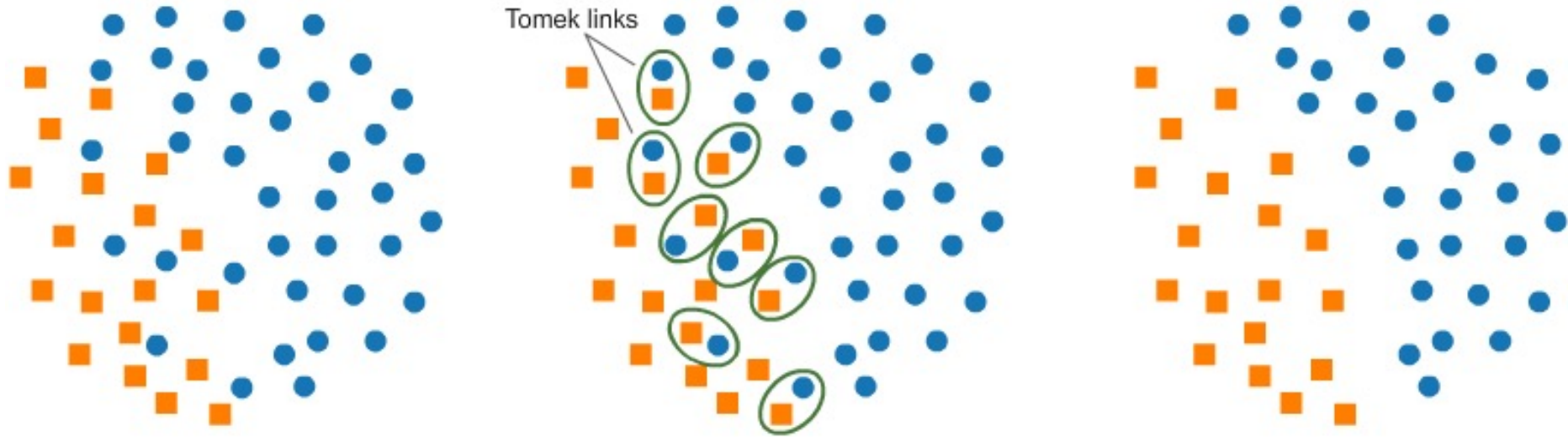


## Oversampling



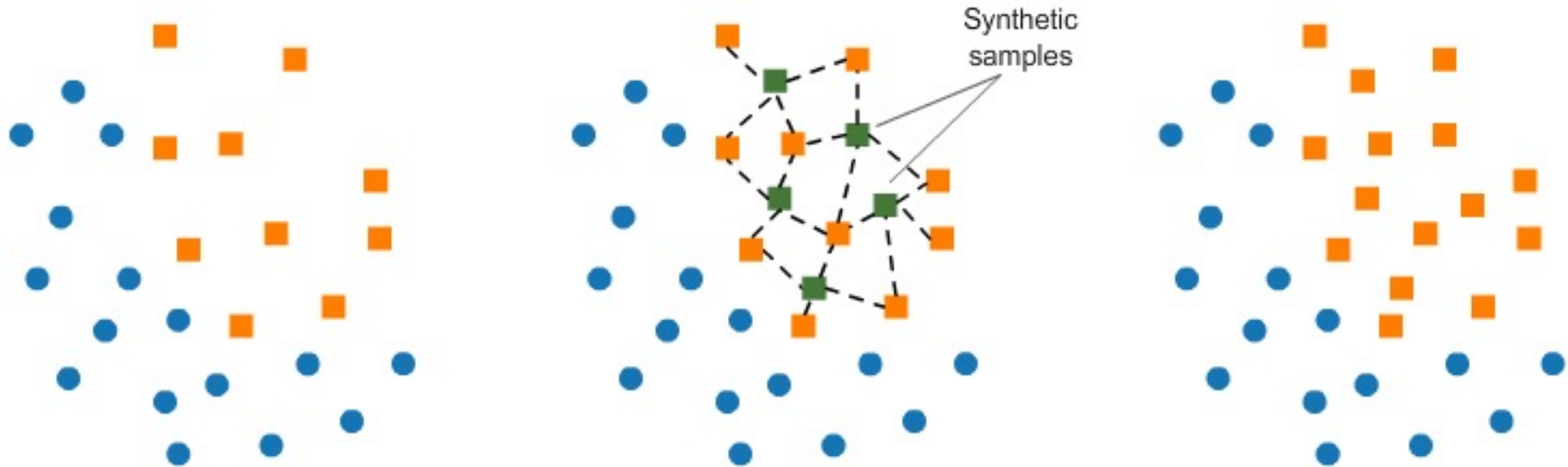
Fonte: <https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1>

# SUBAMOSTRAGEM – TOMEK LINKS



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

# SOBREAMOSTRAGEM – SMOTE




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

# VARIÂNCIA, DESVIO PADRÃO E COEFICIENTE DE VARIAÇÃO

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

$$2^2 = 4$$

$$10^2 = 100$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{150 + 151 + 152 + 152 + 153 + 154 + 155 + 155 + 155}{9} = 153$$


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

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

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

$$28 / 9 = 3,11$$

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

“Erro” se substituirmos pelo  
valor da média

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

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



o quão longe os  
valores estão do  
“valor  
esperado”

# DISTRIBUIÇÃO NORMAL

19:00



18:50



19:10



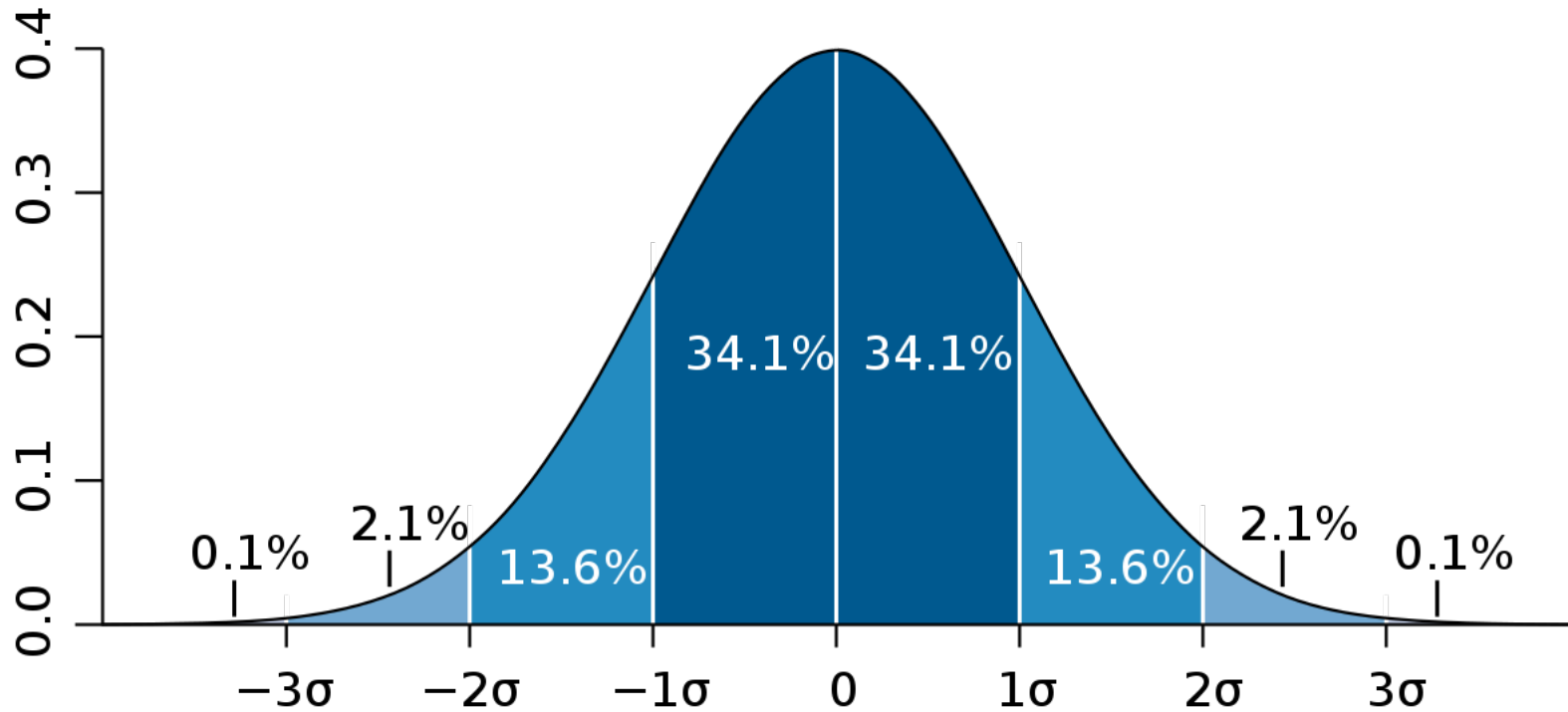
18:40



19:20

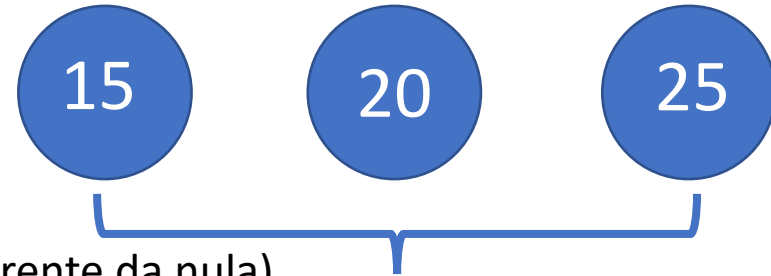


# DISTRIBUIÇÃO NORMAL



# TESTES DE HIPÓTESES

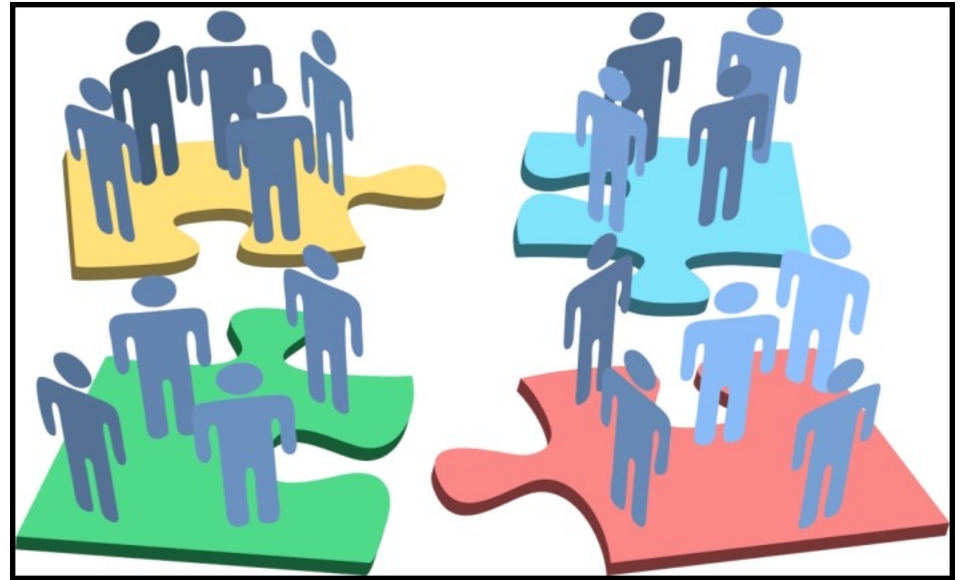
- Resposta sim ou não, para confirmar ou rejeitar uma afirmação
- Hipótese: ideia a ser testada
- Hipótese nula ( $H_0$ )
  - Afirmação que já existia
  - Presumir que é verdadeira até que se prove o contrário
- Hipótese alternativa ( $H_1$ )
  - O que está tentando provar (tudo o que é diferente da 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-value  $\geq$  alpha: não rejeita  $H_0$  (não temos evidências)
  - p-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





# 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
- $H_0$ : não há diferença estatística
- $H_1$ : 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 <b>151</b>	<b>152</b>	<b>156</b>

Média geral: 153

F crítico = 3,88 (consultar tabela)

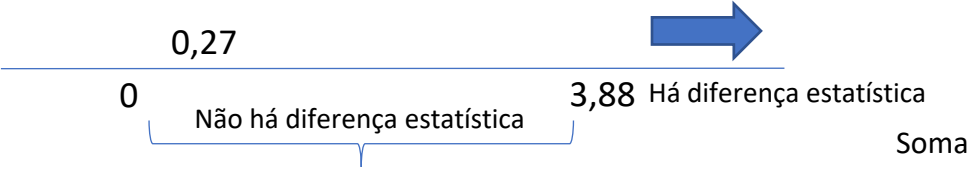
Quadrado		
Grupo A	Grupo B	Grupo C
$(151 - 153)^2 = 4$	$(152 - 153)^2 = 1$	$(156 - 153)^2 = 9$

SSG (sum of squares group):  $14 \times 5 = 70$   
DFG (degrees of freedom groups):  $3 - 1 = 2$

SSE (sum of squares error): 1506  
DFE = linhas - 1 x grupos  
DFE =  $(5 - 1) \times 3 = 12$

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

$$F = \frac{\frac{SSG}{DFG}}{\frac{SSE}{DFE}}$$
$$F = \frac{\frac{70}{2}}{\frac{1506}{12}} = 0.2788$$



# COVARIÂNCIA, COEFICIENTE DE CORRELAÇÃO E COEFICIENTE DE DETERMINAÇÃO

Tamanho (m <sup>2</sup> )	Preço	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) * (y_i - \bar{y})$
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) 12,92 (dp)	73.250 (média) 13.865,42 (dp)			<b>535.500</b> (soma)

$$C(x, y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n - 1}$$

$$C(x, y) = \frac{535.500}{3} = 178500,00$$

$$Cr(x, y) = \frac{Cov(x, y)}{Std(x) * Std(y)}$$

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

> 0, variáveis se movem juntas

< 0, variáveis se movem em direções opostas

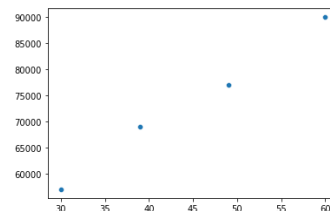
= 0, variáveis são independentes

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

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

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

98% da variável dependente  
consegue ser explicada pelas  
variáveis explanatórias

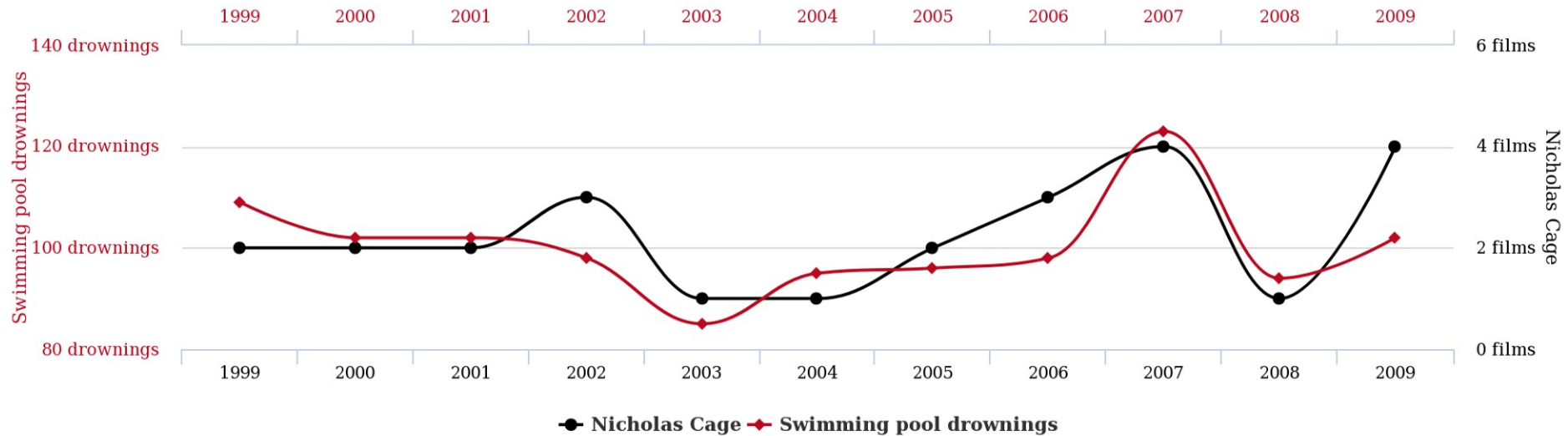


# COEFICIENTE DE CORRELAÇÃO

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

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

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