





**FERNANDA PERES**  
PROFESSORA E  
CONSULTORA NA ÁREA DE  
ANÁLISE DE DADOS



# VARIÂNCIA E COVARIÂNCIA



# AGENDA

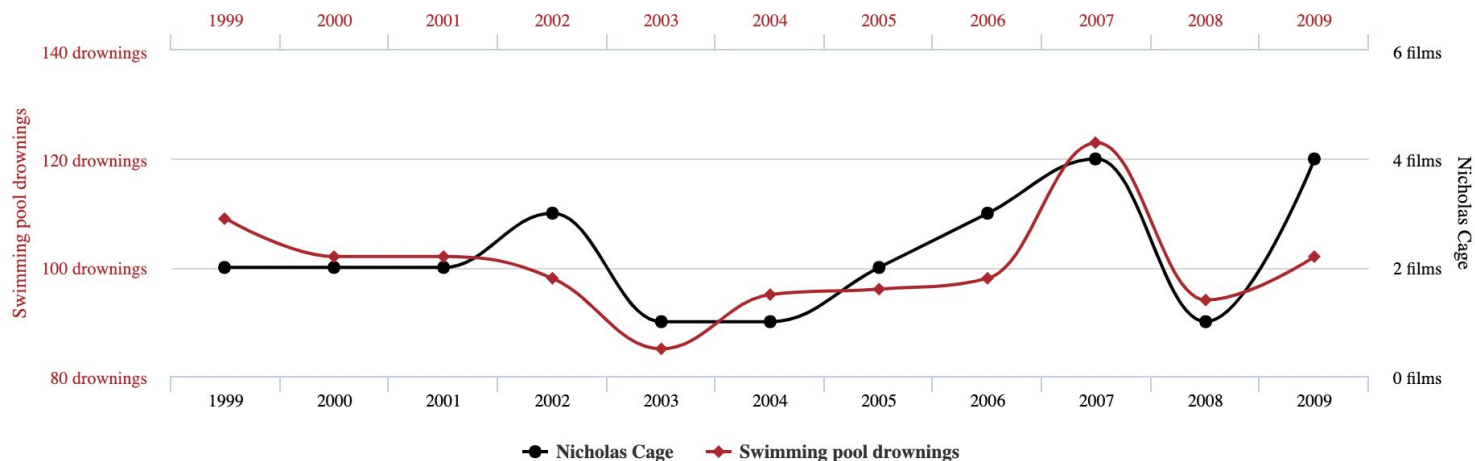
- **O que é variância?**
- **O que é covariância?**
- **Coeficiente de correlação**
- **Resumo da aula**

# Number of people who drowned by falling into a pool

correlates with

## Films Nicolas Cage appeared in

Correlation: 66.6% ( $r=0.666004$ )



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

tylervigen.com

<https://www.tylervigen.com/spurious-correlations>

A vertical bar with a gradient from light green at the top to light blue at the bottom.

# O QUE É VARIÂNCIA?

Em qual desses dois grupos as idades **variam mais**?

**Grupo A**

Idades (anos)
18
20
23
26
28

**Grupo B**

Idades (anos)
18
23
23
23
28

## Grupo A

Idades (anos)			
18 - 23	= -5	$\xrightarrow{x^2}$	= 25
20 - 23	= -3		= 9
23 - 23	= 0		= 0
26 - 23	= 3		= 9
28 - 23	= 5		= 25 +
			<hr/> 68

Média = 23 anos

$$\text{variância} = \frac{\text{soma}}{(n - 1)}$$

$$\text{variância} = \frac{68}{4}$$

$$\text{variância} = 17 \text{ anos}^2$$



Em qual desses dois grupos as idades **variam mais**?

**Grupo A**

Idades (anos)
18
20
23
26
28


$$s^2 = 17 \text{ anos}^2$$

**Grupo B**

Idades (anos)
18
23
23
23
28

$$s^2 = 12,5 \text{ anos}^2$$

Quanto maior a variância, mais  
**dispersos** os dados estão.


$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{(n - 1)}$$



Variância **amostral**

$$\sigma^2 = \frac{\sum_i^n (x_i - \mu_x)^2}{N}$$



Variância **populacional**

A vertical bar on the left side of the slide, transitioning from green at the top to blue at the bottom.

# **DESVIO-PADRÃO**

Corresponde à raiz quadrada da variância.

Está na **mesma unidade de medida** que os dados originais.

$$s = \sqrt{s^2}$$

Em qual desses dois grupos as idades **variam mais**?

**Grupo A**

Idades (anos)
18
20
23
26
28

$$s^2 = 17 \text{ anos}^2$$

$$s = 4,12 \text{ anos}$$

**Grupo B**

Idades (anos)
18
23
23
23
28

$$s^2 = 12,5 \text{ anos}^2$$

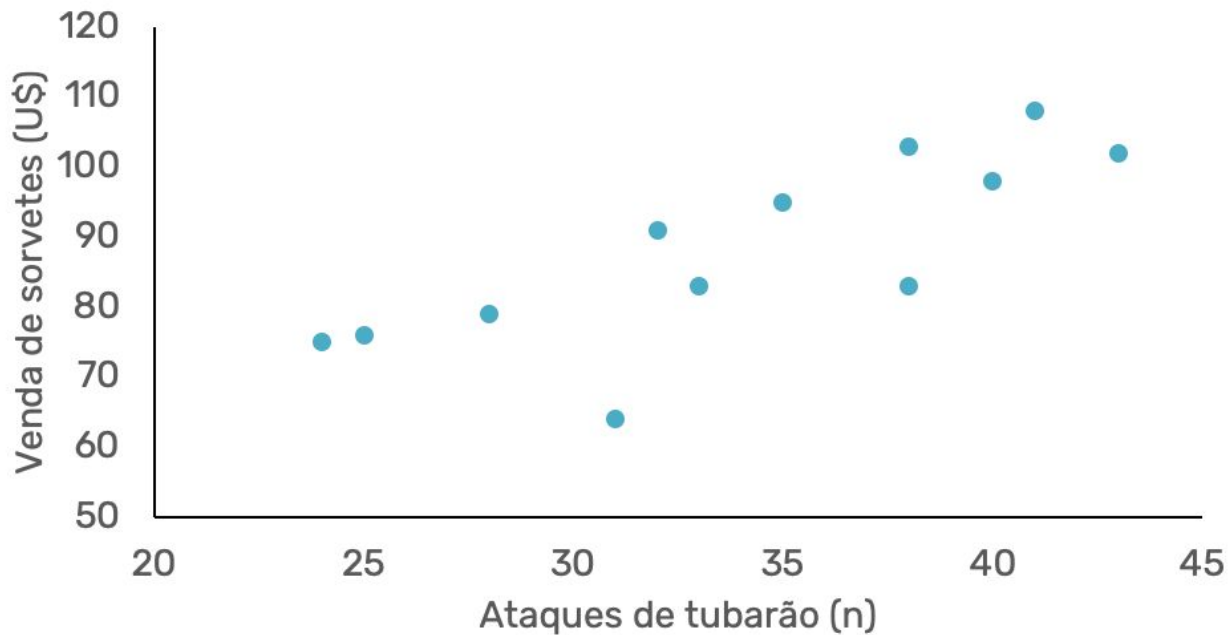
$$s = 3,54 \text{ anos}$$

Quanto maior o desvio-padrão,  
mais **dispersos** os dados estão.

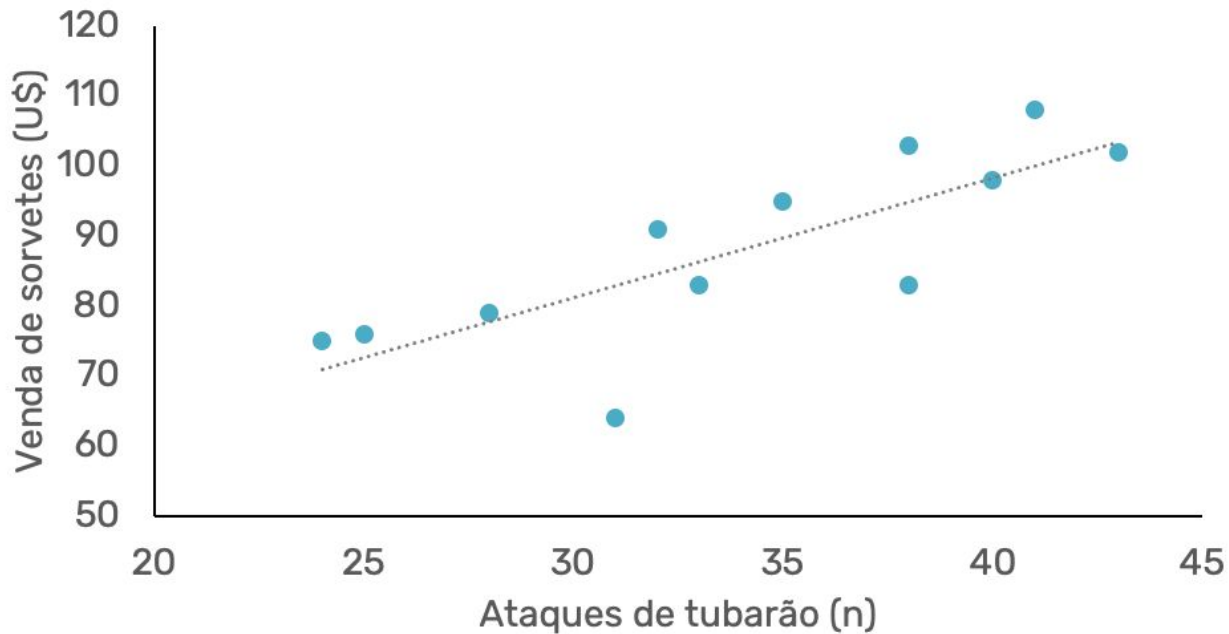
A vertical bar with a gradient from light green at the top to light blue at the bottom.

# O QUE É COVARIÂNCIA?

Uma medida que descreve a **associação linear** entre duas variáveis.




Uma medida que descreve a **associação linear** entre duas variáveis.





Como quantificar essa associação?


$$s_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$



Covariância **amostral**

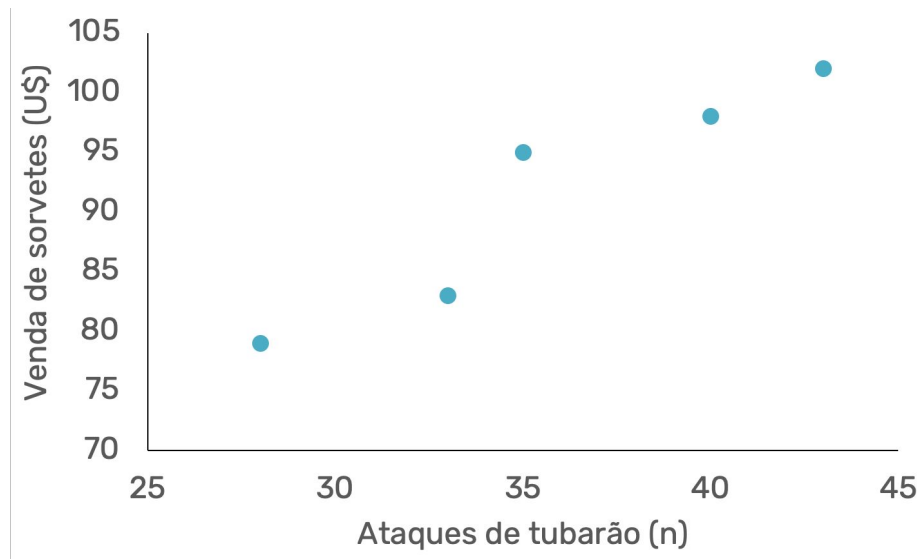
$$\sigma_{xy} = \frac{\sum_i^n (x_i - \mu_x)(y_i - \mu_y)}{N}$$



Covariância **populacional**

$$s_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

<b>Ataques de tubarão (x)</b>	<b>Venda de sorvetes (y)</b>
28	79
33	83
35	95
40	98
43	102



$$s_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

<b>Ataques de tubarão (x)</b>	<b>Venda de sorvetes (y)</b>
28	79
33	83
35	95
40	98
43	102
$\bar{x} = 35,8$	$\bar{y} = 91,4$

$$s_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

<b>Ataques de tubarão (x)</b>	<b><math>x_i - \bar{x}</math></b>	<b>Venda de sorvetes (y)</b>	<b><math>y_i - \bar{y}</math></b>
28		79	
33		83	
35		95	
40		98	
43		102	
$\bar{x} = 35,8$		$\bar{y} = 91,4$	

$$s_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)}$$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
-7,8	-12,4	
-2,8	-8,4	
-0,8	3,6	
4,2	6,6	
7,2	10,6	

Soma = 221,4

$$s_{xy} = \frac{\text{soma}}{n-1}$$

$$s_{xy} = \frac{221,4}{4}$$

$$s_{xy} = 55,35$$

$$s_{xy} = 55,35$$

Como interpretar a covariância?

É **positiva**: indica associação positiva entre as variáveis

Mas **não** conseguimos, com ela, interpretar a **magnitude** da associação



Para isso, precisamos calcular o **coeficiente de correlação**



# O COEFICIENTE DE CORRELAÇÃO DE PEARSON (R)

$$r = \frac{s_{xy}}{s_x \times s_y}$$

$$r = \frac{55,35}{5,89 \times 9,91}$$

$$r = \frac{55,35}{58,37}$$

$$r = 0,948$$

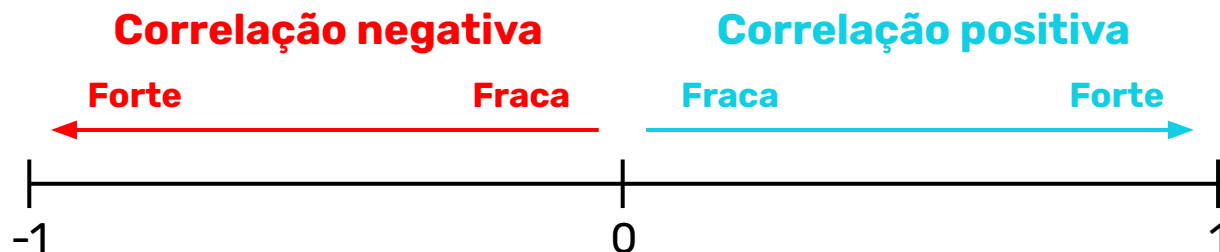
<b>Ataques de tubarão (x)</b>	<b>Venda de sorvetes (y)</b>
28	79
33	83
35	95
40	98
43	102

$$s_{xy} = 55,35$$

$$s_x = 5,89$$

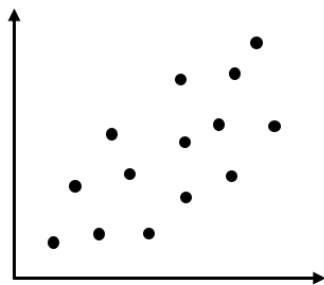
$$s_y = 9,91$$



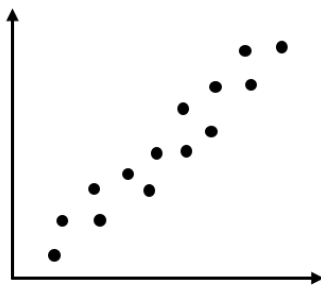


Valor absoluto de r (desconsiderando o sinal)	Interpretação
$0,00 \leq r < 0,30$	Correlação desprezível
$0,30 \leq r < 0,50$	Correlação fraca
$0,50 \leq r < 0,70$	Correlação moderada
$0,70 \leq r < 0,90$	Correlação forte
$0,90 \leq r < 1,00$	Correlação muito forte
$r = 1,00$	Correlação perfeita

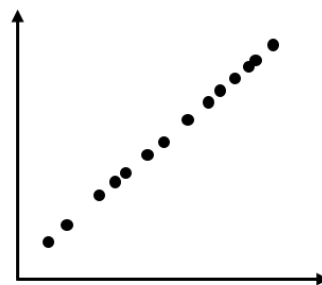
(Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin; 2003)



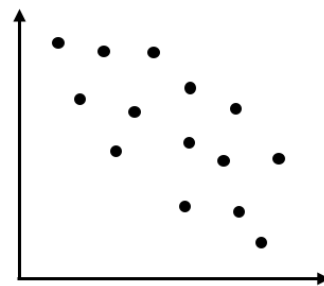
**Correlação  
positiva fraca**



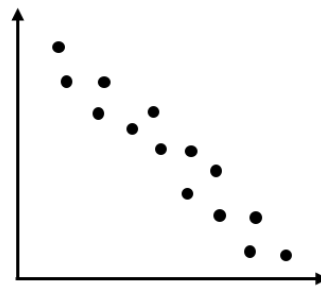
**Correlação  
positiva forte**



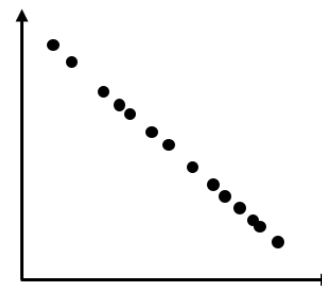
**Correlação  
positiva perfeita**



**Correlação  
negativa fraca**

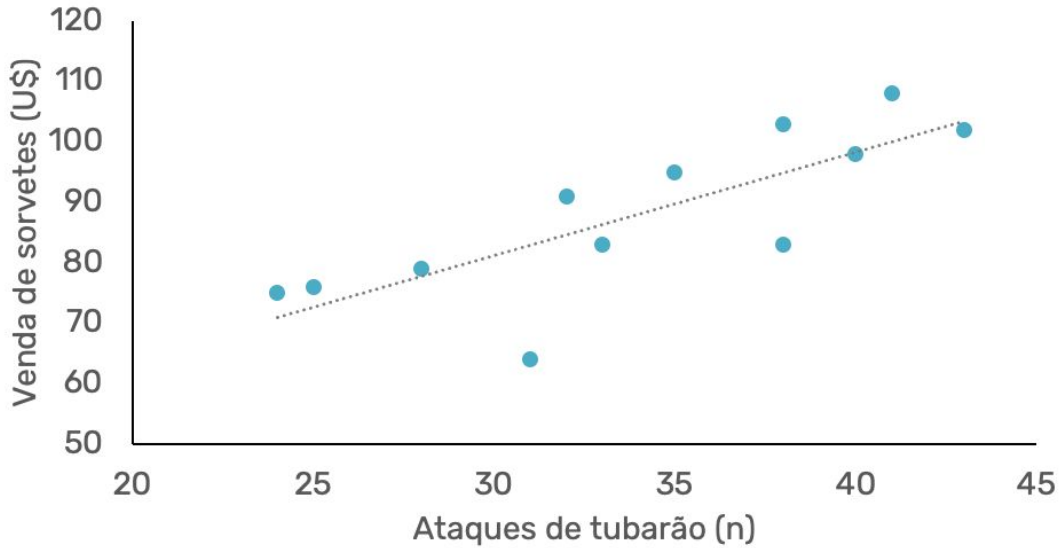


**Correlação  
negativa forte**

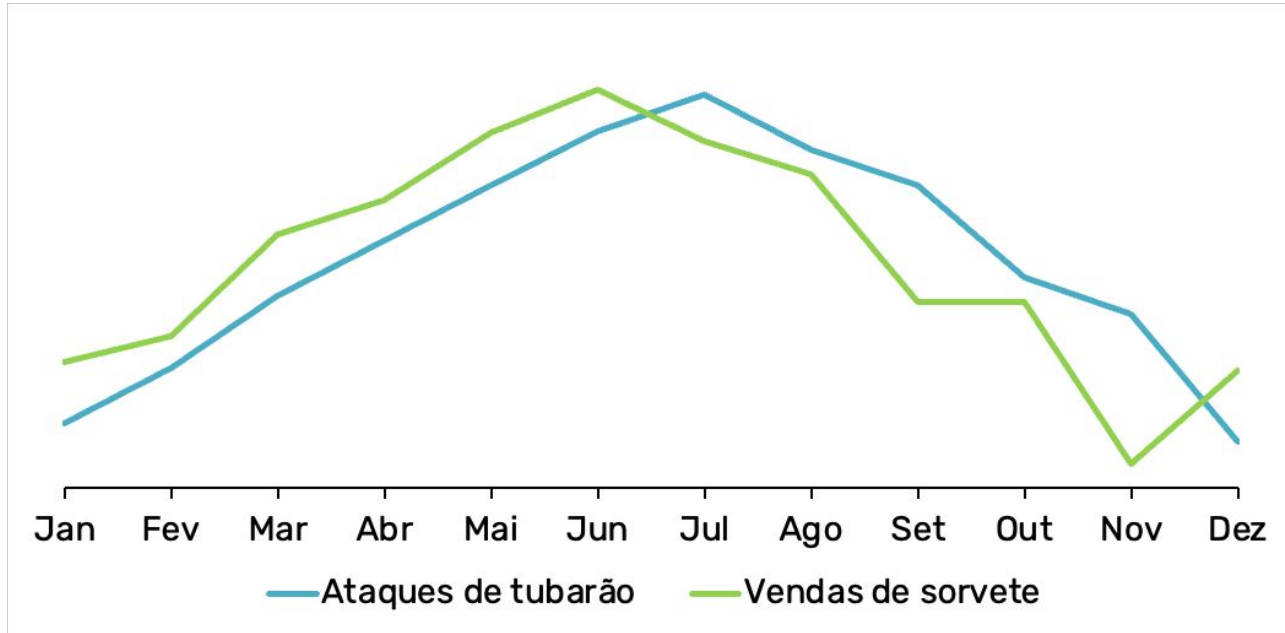


**Correlação  
negativa perfeita**

**Importante:**  
correlação  $\neq$  causalidade



**Importante:**  
correlação  $\neq$  causalidade

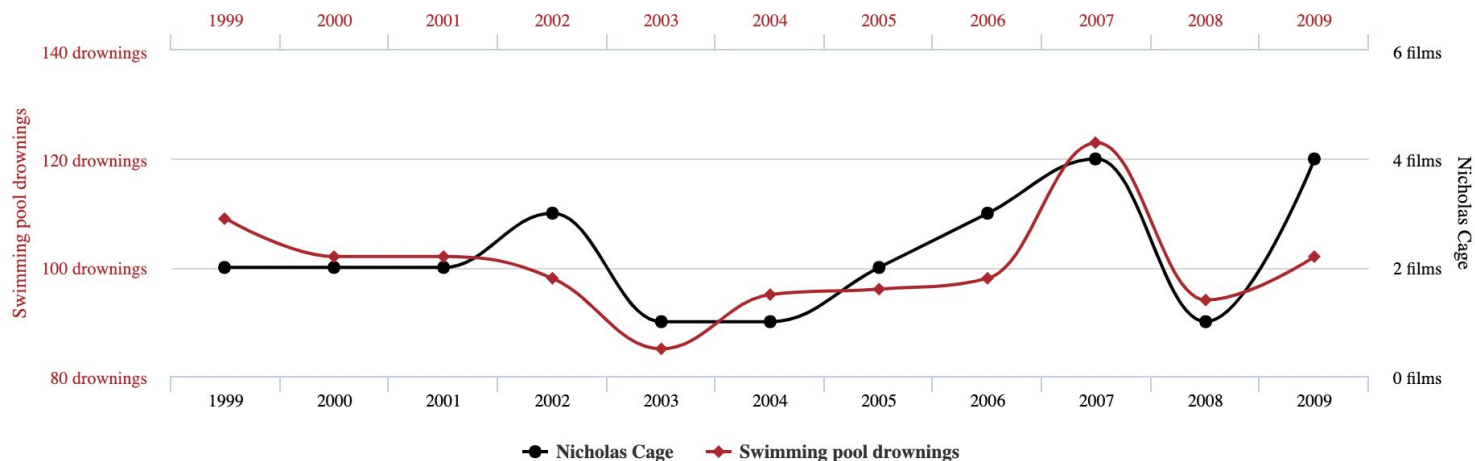


# Number of people who drowned by falling into a pool

correlates with

## Films Nicolas Cage appeared in

Correlation: 66.6% ( $r=0.666004$ )



Data sources: Centers for Disease Control & Prevention and Internet Movie Database


tylervigen.com

<https://www.tylervigen.com/spurious-correlations>

T

A vertical bar with a gradient from green at the top to blue at the bottom.

# RESUMO DA AULA

- 
- A vertical bar with a gradient from light green at the top to light blue at the bottom, located on the left side of the slide.
- **Variância** é uma forma de medir a dispersão dos dados
  - **Covariância** é uma forma de medir a associação entre duas variáveis
  - **Coeficiente de correlação** é uma medida padronizada de associação entre duas variáveis

