



Preditiva.ai

Estatística Descritiva

Análise Bidimensional

Já sabemos como resumir e analisar cada variável de um conjunto de dados, mas:

“E se tivermos que analisar o comportamento de **2 variáveis simultaneamente?**”

Por exemplo:

- Qual a taxa de turnover por nível de formação?
- Qual o percentual de compra do novo livro do Flávio Augusto por região do Brasil?

A **Análise Bidimensional** é o nome dado a um conjunto de técnicas utilizadas para:

Analisar o **comportamento conjunto** de **duas variáveis**

Considerando os **diferentes tipos de variáveis**, podemos ter 3 situações:

1. Duas variáveis **quantitativas**
2. Uma variável **qualitativa** e outra variável **quantitativa**
3. Duas variáveis **qualitativas**

Análise Bidimensional

2 Variáveis Quantitativas



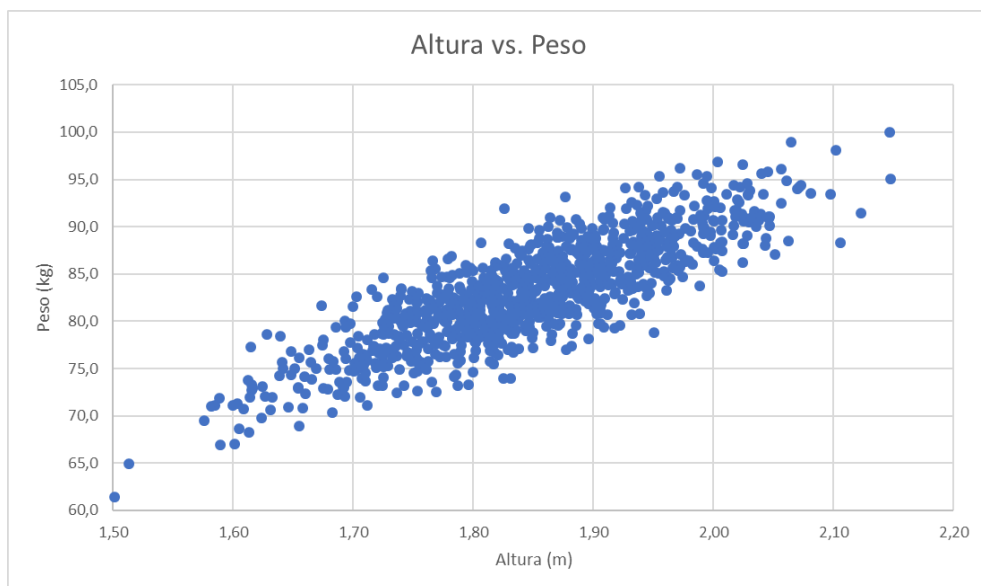
Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas



A análise de 2 variáveis quantitativas inicia-se com o **Gráfico de Dispersão**. Nele uma das variáveis fica no eixo X e a outra no eixo Y.

Exemplo: Em uma pesquisa com os suecos, foram obtidos os dados de peso e altura de cada habitante. O gráfico de dispersão com essas variáveis é apresentado abaixo.



Podemos perceber que existe uma **relação entre altura e peso**, ou seja, quanto **maior a altura**, maior o peso.

A essa relação damos o nome de **correlação**, logo **Altura e Peso estão correlacionados**.

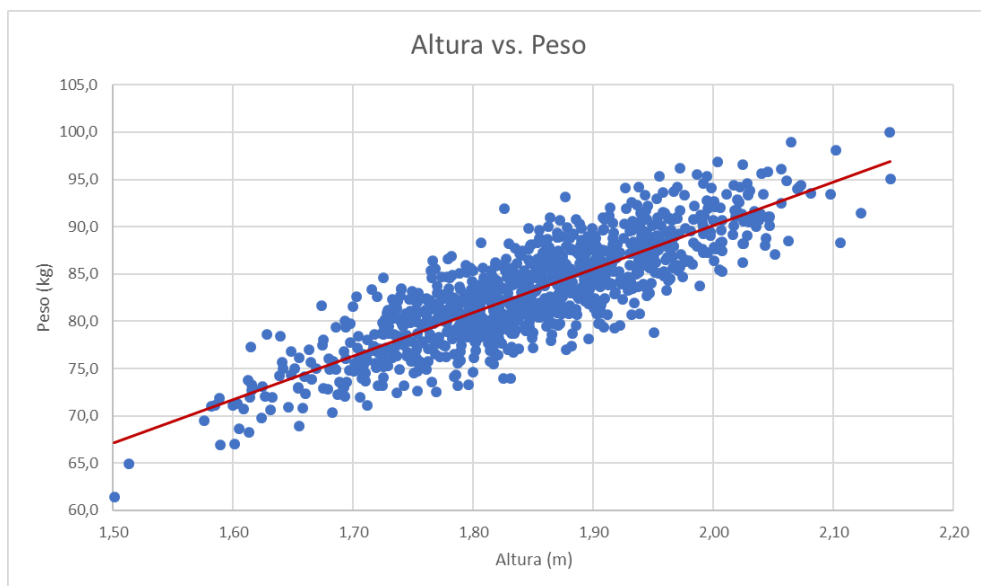
Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas



O **Gráfico de Dispersão** fornece a **distribuição conjunta** da Altura e Peso, e junto com ela a visualização de uma possível **correlação entre essas duas variáveis**.

Nesse caso, a **correlação entre Altura e Peso aparenta ser linear**, ou seja, é possível **definir uma equação linear** em que dada a Altura encontramos o Peso, e vice-versa.



Entender a **correlação entre variáveis** é algo bastante **poderoso**! Mas é preciso ter cautela. Veremos a seguir...

Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas



A existência de **correlação entre duas variáveis** indica que elas estão de alguma forma **associadas**, mas nem sempre isso quer dizer que **uma variável “causa” a outra**.

Em nosso exemplo:

- As pessoas são mais **pesadas** porque são mais **altas**?
- As pessoas são mais **altas** porque são mais **pesadas**?
- As pessoas são mais **altas** e mais **pesadas** devido a outro fator não observado? Genético, por exemplo.

Essa é a grande diferença entre **correlação** e **causalidade**!

Ou seja, **nem toda correlação é causalidade** mas **toda causalidade gera uma correlação**.

Vamos ver esse assunto com mais detalhes adiante.

Uma forma de **medirmos a força da correlação** entre duas variáveis quantitativas, como Altura e Peso, é calculando o **Coeficiente de Correlação de Pearson**:

$$\text{corr}(X, Y) = \frac{1}{n} \cdot \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{dp(X)} \right) \cdot \left(\frac{y_i - \bar{y}}{dp(Y)} \right)$$

O **Coeficiente de Correlação de Pearson** varia entre -1 e +1 e indica:

- **Correlação positiva forte:** coeficiente próximo a 1
- **Correlação inexistente:** coeficiente próximo a zero
- **Correlação negativa forte:** coeficiente próximo a -1

Estatística Descritiva

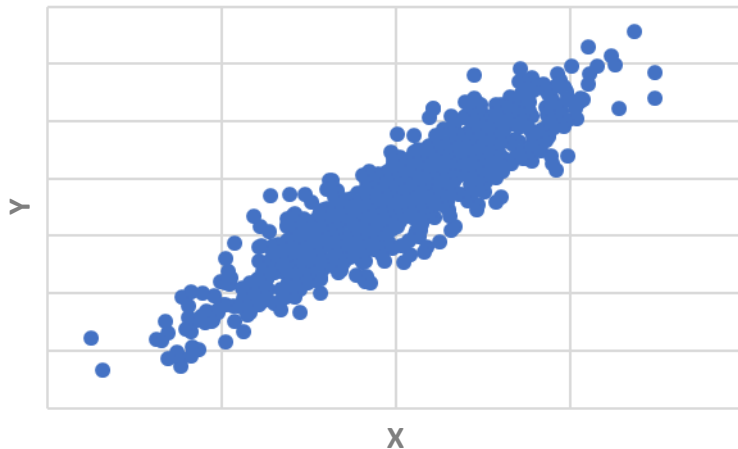
Análise Bidimensional: 2 variáveis quantitativas



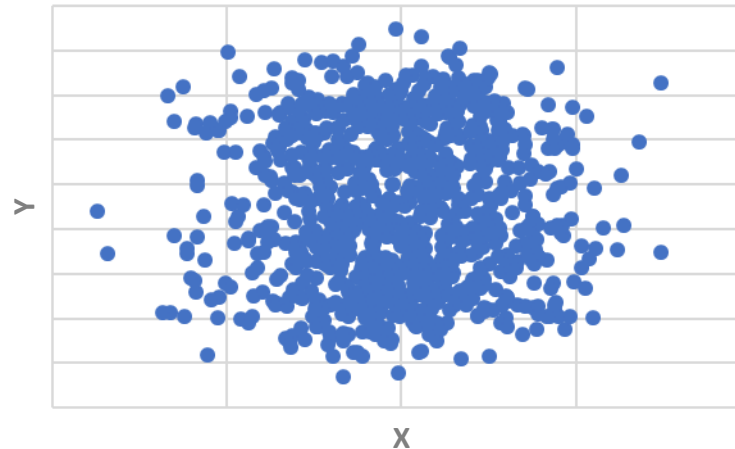
O **Coeficiente de Correlação de Pearson** varia entre -1 e +1 e indica:

- **Correlação positiva:** quando a variável X **aumenta**, a variável Y também **aumenta**
- **Correlação inexistente:** quando a variável X **aumenta**, a variável Y **não se altera**
- **Correlação negativa:** quando a variável X **aumenta**, a variável Y **diminui**

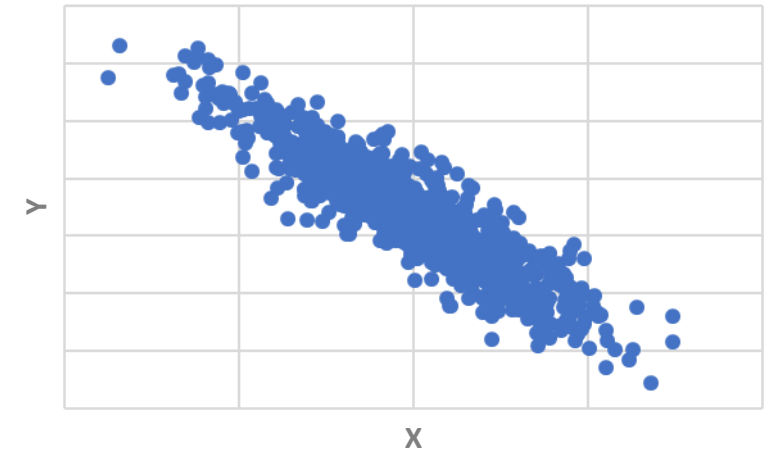
Correlação Positiva



Correlação Neutra ou Inexistente



Correlação Negativa



Estatística Descritiva

Análise Bidimensional: 2 variáveis quantitativas

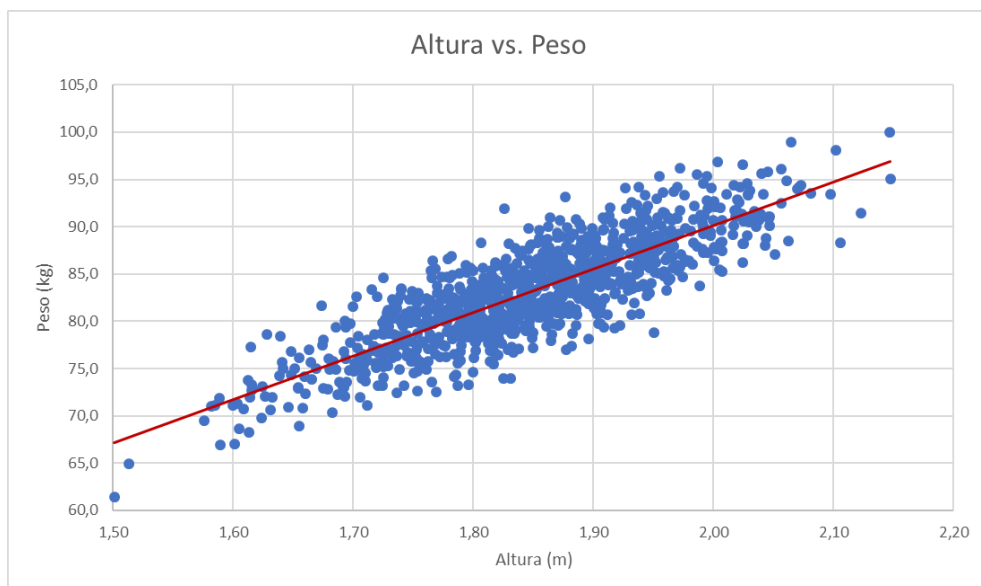


No Excel: **CORREL(X ; Y)**

X: observações da variável 1

Y: observações da variável 2

Intervalo	Força
$-1,0 < r < -0,7$	Fortemente Negativa
$-0,6 < r < 0,6$	Fraca
$0,7 < r < 1,0$	Fortemente Positiva



Neste exemplo, vamos considerar **X = Altura** e **Y = Peso**.

Utilizando **CORREL(Altura;Peso) = 0,85**

Logo, **existe uma forte correlação positiva** entre Altura e Peso.

Análise Bidimensional

2 Variáveis Qualitativas, 1 Binária



Outro tipo de medida de associação **muito útil é o Information Value (IV)**. Essa medida é responsável por fornecer o “**poder de separação**” que uma variável qualitativa de duas ou mais categorias possui sobre outra variável de duas categorias (variável binária).

Exemplos de variáveis binárias:

1. Bons clientes x Maus clientes
2. Comprou x Não comprou
3. Doente x Não doente
4. Verdadeiro x Falso
5. Entre outros...

Em vários problemas de Analytics, estamos interessados em descobrir quais fatores, isto é, **quais variáveis são responsáveis por separar as categorias das variáveis binárias**.

Ex: Qual variável separa um Bom cliente de um Mau cliente para um empréstimo?
Renda? Profissão? Escolaridade?

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária



Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Escolaridade** é útil para separar **Bons e Maus** clientes de um financiamento:

Variável qualitativa			Variável qualitativa binária							
Escolaridade	Frequência	% Freq	Qte Bons	Qte Maus	% Bons	% Maus	% Taxa de Bons	Odds	LN(Odds)	IV
Ensino Fundamental	123	5%	7	116	1,3%	5,6%	5,7%	0,2	-1,441	0,06
Ensino Médio	800	31%	87	713	16,5%	34,5%	10,9%	0,5	-0,737	0,13
Graduação	557	21%	107	450	20,3%	21,8%	19,2%	0,9	-0,070	0,00
Pós Graduação	1.114	43%	326	788	61,9%	38,1%	29,3%	1,6	0,484	0,11
	2.594	100%	527	2067	100%	100%	20,3%			0,31

Frequência: Quantidade de clientes em cada um dos níveis de escolaridade. Por exemplo, na categoria Ensino Fundamental temos **123** clientes.

Qte Bons: Quantidade de clientes classificados como bons em cada um dos níveis de escolaridade. Por exemplo, na categoria Ensino Fundamental, dos **123** clientes, temos **7** classificados como bons.

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária



Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Escolaridade** é útil para separar **Bons e Maus** clientes de um financiamento:

Variável qualitativa			Variável qualitativa binária							
Escolaridade	Frequência	% Freq	Qte Bons	Qte Maus	% Bons	% Maus	% Taxa de Bons	Odds	LN(Odds)	IV
Ensino Fundamental	123	5%	7	116	1,3%	5,6%	5,7%	0,2	-1,441	0,06
Ensino Médio	800	31%	87	713	16,5%	34,5%	10,9%	0,5	-0,737	0,13
Graduação	557	21%	107	450	20,3%	21,8%	19,2%	0,9	-0,070	0,00
Pós Graduação	1.114	43%	326	788	61,9%	38,1%	29,3%	1,6	0,484	0,11
	2.594	100%	527	2067	100%	100%	20,3%			0,31

% Bons: Percentual de bons clientes na categoria em relação ao total de bons clientes. Por exemplo, na categoria Ensino Fundamental temos **7** bons clientes de um total de **527** bons clientes, logo **1,3%** ($7/527$) dos bons clientes estão na categoria Ensino Fundamental.

% Maus: Mesmo conceito do **% Bons**, mas aplicado aos maus clientes.

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Escolaridade** é útil para separar **Bons e Maus** clientes de um financiamento:

Variável qualitativa	Variável qualitativa binária									
Escolaridade	Frequência	% Freq	Qte Bons	Qte Maus	% Bons	% Maus	% Taxa de Bons	Odds	LN(Odds)	IV
Ensino Fundamental	123	5%	7	116	1,3%	5,6%	5,7%	0,2	-1,441	0,06
Ensino Médio	800	31%	87	713	16,5%	34,5%	10,9%	0,5	-0,737	0,13
Graduação	557	21%	107	450	20,3%	21,8%	19,2%	0,9	-0,070	0,00
Pós Graduação	1.114	43%	326	788	61,9%	38,1%	29,3%	1,6	0,484	0,11
	2.594	100%	527	2067	100%	100%	20,3%			0,31

Taxa de bons: Percentual de bons clientes em relação ao total de clientes na categoria. Por exemplo, na categoria Ensino Fundamental temos **7** bons clientes de um total de **123** clientes na categoria, logo **5,7%** (7/123) dos clientes dessa categoria são bons clientes.

Odds: Razão entre %Bons e %Maus. Por exemplo, na categoria Ensino Fundamental temos **1,3%** no %Bons e **5,6%** no %Maus, logo **0,2** (1,3%/5,6%) é a chance de encontrarmos um bom cliente nessa categoria. Ou seja, a proporção na categoria Ensino Fundamental é de 1 bom cliente para 5 maus clientes.

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária



Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Escolaridade** é útil para separar **Bons e Maus** clientes de um financiamento:

Variável qualitativa			Variável qualitativa binária							
Escolaridade	Frequência	% Freq	Qte Bons	Qte Maus	% Bons	% Maus	% Taxa de Bons	Odds	LN(Odds)	IV
Ensino Fundamental	123	5%	7	116	1,3%	5,6%	5,7%	0,2	-1,441	0,06
Ensino Médio	800	31%	87	713	16,5%	34,5%	10,9%	0,5	-0,737	0,13
Graduação	557	21%	107	450	20,3%	21,8%	19,2%	0,9	-0,070	0,00
Pós Graduação	1.114	43%	326	788	61,9%	38,1%	29,3%	1,6	0,484	0,11
	2.594	100%	527	2067	100%	100%	20,3%			0,31

LN(Odds): Logarítmo Natural da Odds. Por exemplo, na categoria Ensino Fundamental temos $\text{LN}(0,2) = -1,44$.

IV: Produto da diferença entre %Bons e %Maus pelo LN(Odds). Por exemplo, na categoria Ensino Fundamental temos $(1,3\% - 5,6\%) * -1,44 = 0,06$.

Estatística Descritiva



Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária

Vamos calcular o **Information Value (IV)** para um exemplo no qual queremos avaliar se a **variável Escolaridade** é útil para separar **Bons e Maus** clientes de um financiamento:

Variável qualitativa			Variável qualitativa binária							
Escolaridade	Frequência	% Freq	Qte Bons	Qte Maus	% Bons	% Maus	% Taxa de Bons	Odds	LN(Odds)	IV
Ensino Fundamental	123	5%	7	116	1,3%	5,6%	5,7%	0,2	-1,441	0,06
Ensino Médio	800	31%	87	713	16,5%	34,5%	10,9%	0,5	-0,737	0,13
Graduação	557	21%	107	450	20,3%	21,8%	19,2%	0,9	-0,070	0,00
Pós Graduação	1.114	43%	326	788	61,9%	38,1%	29,3%	1,6	0,484	0,11
	2.594	100%	527	2067	100%	100%	20,3%			0,31

Para fins de medida de associação, estamos interessados na soma dos **IV's** de cada categoria da variável.

Neste exemplo, o **IV Total** = 0,06 + 0,13 + 0,00 + 0,11 = **0,31**

Estatística Descritiva

Análise Bidimensional: 2 Variáveis Qualitativas, 1 Binária



Após calcularmos o **IV**, como avaliamos se a variável possui um **alto poder de separação**?

Abaixo apresentamos uma referência bastante **utilizada na prática**:

IV Total	Poder de separação
< 0,02	Muito fraco
0,02 a 0,1	Fraco
0,1 a 0,3	Médio
0,3 a 0,5	Forte
> 0,5	Muito bom pra ser verdade...Verifique!

Em nosso exemplo, a variável **Escolaridade** obteve um $IV = 0,35$, ou seja, ela possui um **forte** poder de separação entre os bons e maus clientes.

Logo, ao perguntar a escolaridade de um cliente, você já pode ter uma boa estimativa se ele pagará em dia ou atrasará.

Análise Bidimensional

1 Variável Qualitativa e 1 Variável Quantitativa



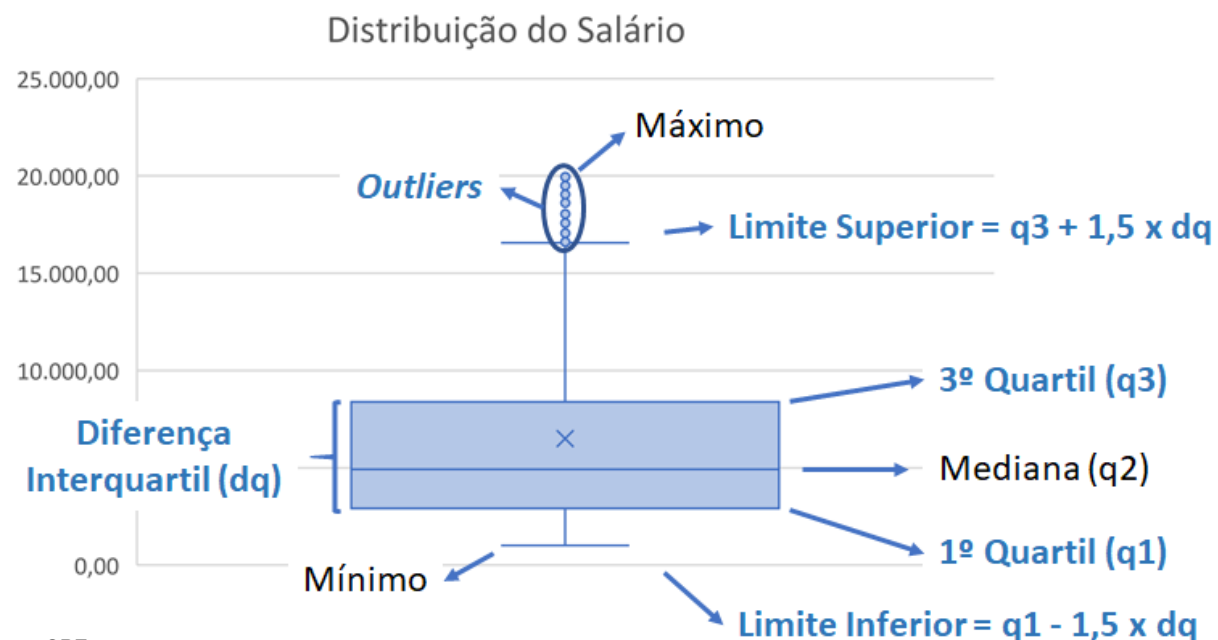
Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa



Quando temos 1 variável qualitativa e 1 variável quantitativa queremos analisar a **distribuição da variável quantitativa para cada categoria**, ou nível, da variável qualitativa.

Uma técnica bastante utilizada e prática é a construção de **boxplots** para a variável quantitativa, separando as distribuições em cada nível da variável qualitativa.

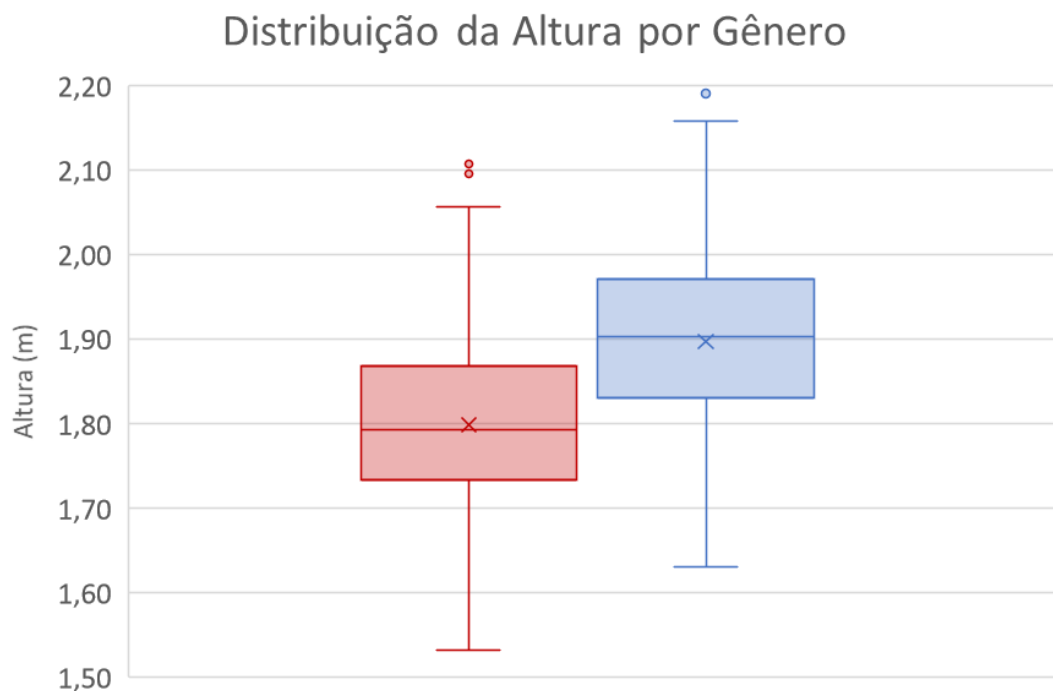


Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa



Voltemos a nossa pesquisa com os suecos, mas agora vamos analisar a variável Altura por gênero. No gráfico, o **boxplot vermelho** corresponde a distribuição da altura das mulheres e o **boxplot azul** corresponde a distribuição da altura dos homens.



Insights

1. As **medidas de posição** da altura dos homens são ligeiramente superiores as das mulheres.
2. As duas distribuições de altura possuem **dispersão semelhante**.
3. **Poucas observações discrepantes (outliers)**.

Uma outra forma de analisarmos os dados é criar uma tabela com as **medidas resumo** separadas por gênero.

Gênero	N	Média de altura	Variância de altura	Desvio Padrão de altura	Mínimo de altura	Máximo de altura	1º Quartil de altura	Mediana de altura	3º Quartil de altura
Feminino	600	1,60	0,010	0,10	1,33	1,91	1,53	1,59	1,67
Masculino	400	1,85	0,010	0,10	1,45	2,14	1,78	1,85	1,92
Total	1.000	1,70	0,025	0,16	1,33	2,14	1,57	1,68	1,82

Utilizando essa tabela podemos:

1. Identificar que a quantidade de mulheres **é maior** do que a quantidade de homens.
2. Confirmamos as informações extraídas anteriormente utilizando os **boxplots**.



Para medir a associação entre variáveis qualitativas e quantitativas utilizamos o **Coeficiente de Determinação**, também conhecido como **R²**. Neste exemplo, calcularíamos:

$$R^2 = 1 - \frac{\overline{var(Altura)}}{var(Altura)}$$

$$\text{Sendo, } \overline{var(Altura)} = \frac{\sum_{i=1}^k n_i \cdot var_i(Altura)}{\sum_{i=1}^k n_i}$$

(Média das variâncias para cada gênero)

Intuitivamente, o **R²** mede quanto da **variância total** é explicada pela **introdução da variável qualitativa** e é uma medida que **varia entre 0 e 1**.

Dessa forma:

- **R²** igual a zero: indica a **inexistência** de associação entre as variáveis
- **R²** igual a 1: indica **forte associação** entre as variáveis

Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa



Calculando o R^2 para este exemplo:

$$\overline{var(Altura)} = \frac{600 \cdot 0,010 + 400 \cdot 0,010}{600 + 400} = 0,010 \quad R^2 = 1 - \frac{0,010}{0,025} = 0,6 = \mathbf{60\%}$$

Gênero	N	Média de altura	Variância de altura	Desvio Padrão de altura	Mínimo de altura	Máximo de altura	1º Quartil de altura	Mediana de altura	3º Quartil de altura
Feminino	600	1,60	0,010	0,10	1,33	1,91	1,53	1,59	1,67
Masculino	400	1,85	0,010	0,10	1,45	2,14	1,78	1,85	1,92
Total	1.000	1,70	0,025	0,16	1,33	2,14	1,57	1,68	1,82

Estatística Descritiva

Análise Bidimensional: 1 Qualitativa e 1 Quantitativa



Calculando o R^2 para este exemplo:

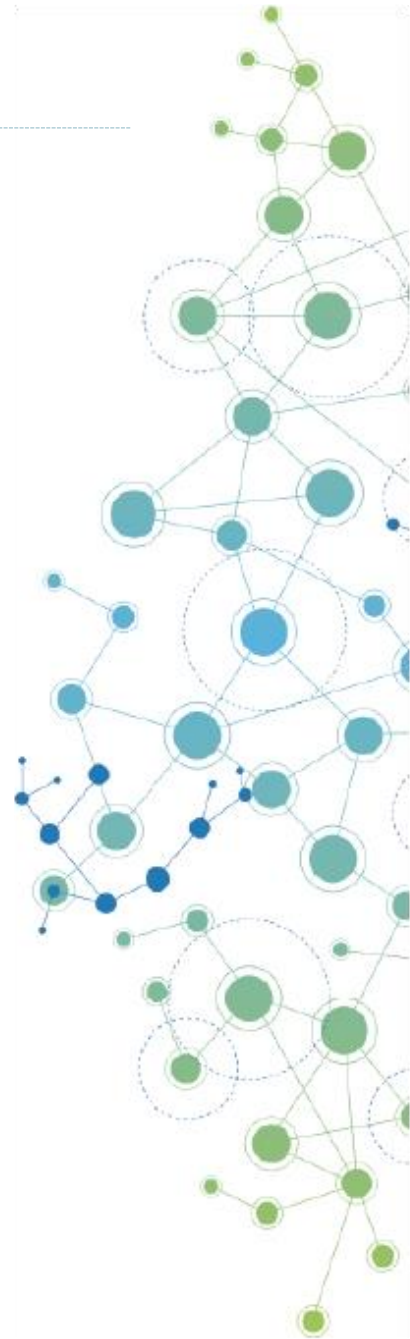
$$\overline{var(Altura)} = \frac{600 \cdot 0,010 + 400 \cdot 0,010}{600 + 400} = 0,010 \quad R^2 = 1 - \frac{0,010}{0,025} = 0,6 = \mathbf{60\%}$$

Nesse exemplo, vemos que a associação entre Altura e Gênero existe, e o R^2 igual a 60% indica que essa associação **é forte**.

Ou seja, **o gênero explica 60% da diferença** de altura entre homens e mulheres.

Análise Bidimensional

Correlação vs. Causalidade



É fundamental dominarmos a **diferença entre esses 2 conceitos** para não cairmos em algumas **armadilhas** de Analytics.

Vejamos a definição destes 2 termos:

- **Correlação**: relação de **dependência** ou **associação** entre duas variáveis.
- **Causalidade**: relação entre um evento A e um evento B, sendo que o evento B é **consequência** do evento A.

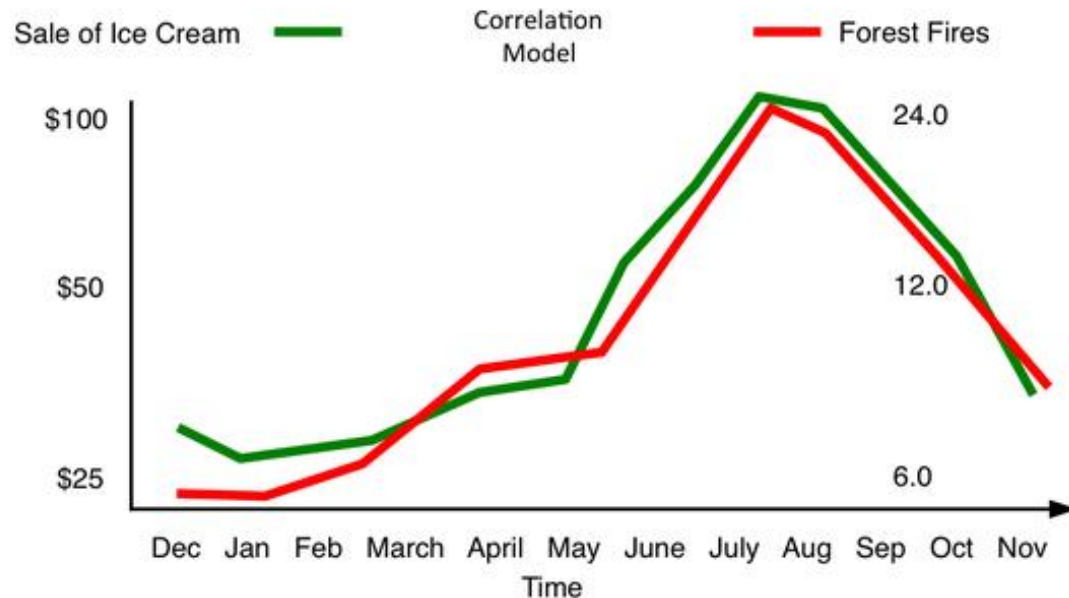
Ou seja, **Correlação** está relacionada com a **dependência ou associação** e a **Causalidade** relacionada a **consequência**.

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade



Vamos avaliar a **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas**:



Fonte: <https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect>

Você acha que a **venda de sorvetes** pode **causar** **incêndios nas florestas**?

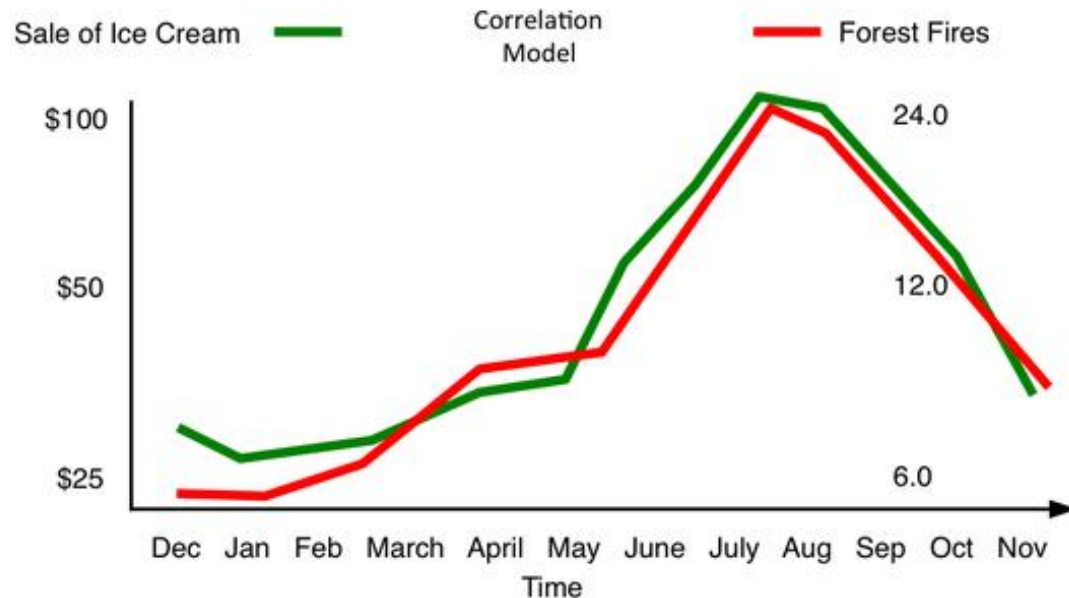
Neste caso há uma **3ª variável não avaliada** e que faz mais sentido ser a **causadora** do aumento no **consumo de sorvete** e dos **incêndios nas florestas**: **o clima quente!**

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade



Vamos avaliar a **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas**:



Fonte: <https://www.decisionskills.com/blog/how-ice-cream-kills-understanding-cause-and-effect>

A **Correlação** entre **Venda de Sorvetes** e **Incêndio nas Florestas** é conhecida como **Correlação Espúria**.

As **Correlações Espúrias** podem ser uma armadilha para **falsas conclusões**.

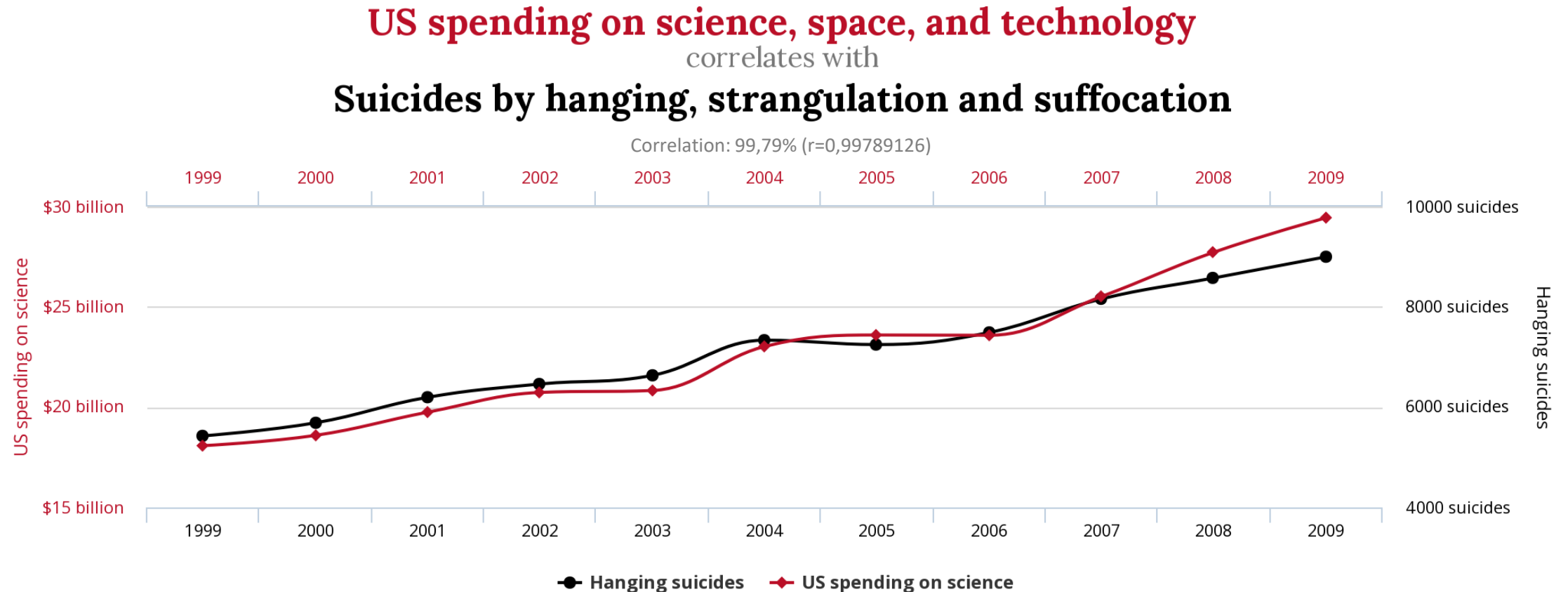
Vejamos alguns outros exemplos.

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade



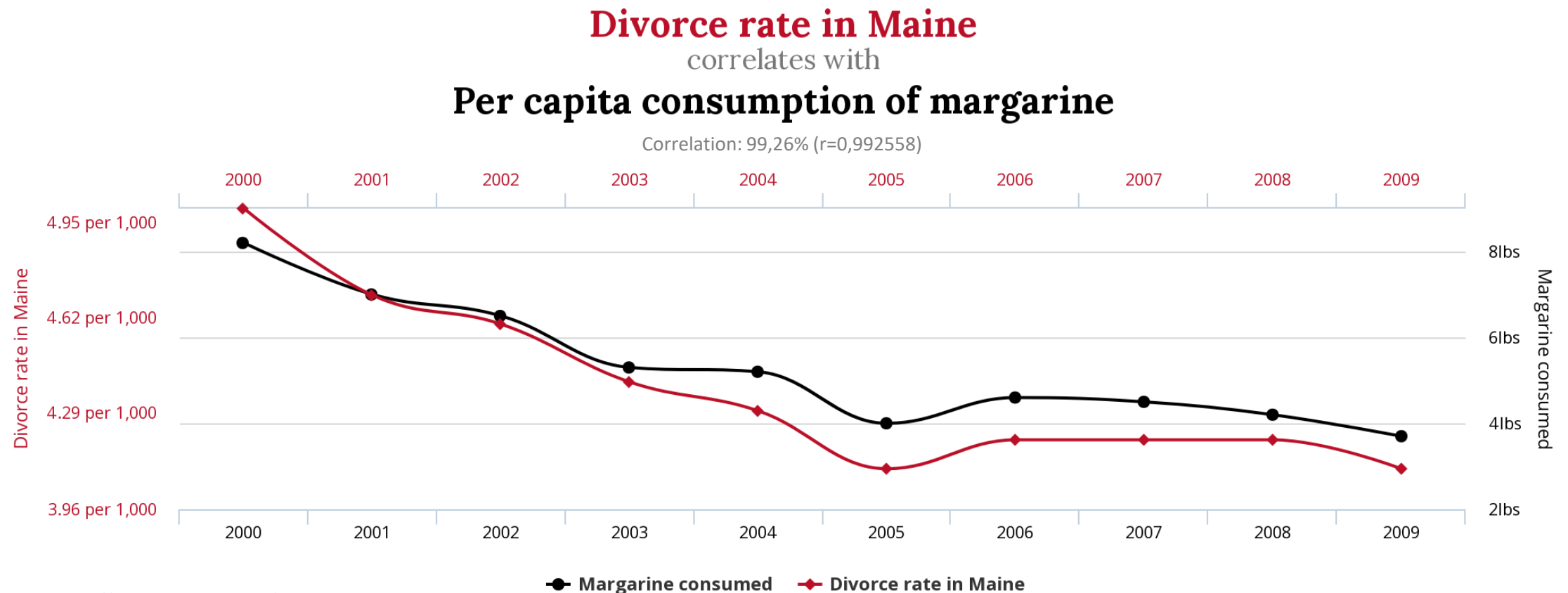
Exemplo 1: Gasto em Pesquisa no EUA vs. Suicídios



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

tylervigen.com

Exemplo 2: Divórcios em Maine vs. Consumo de margarina



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

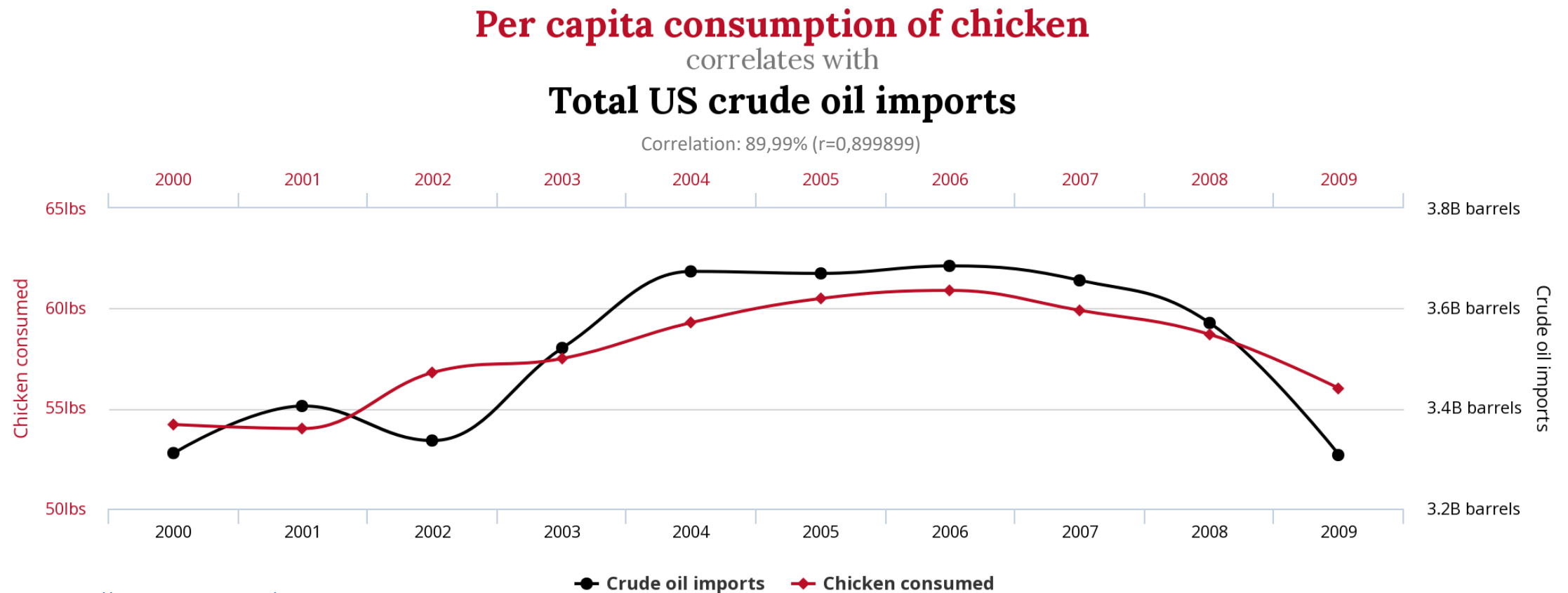
tylervigen.com

Estatística Descritiva

Análise Bidimensional: Correlação vs. Causalidade



Exemplo 3: Consumo de frango vs. Importação de petróleo



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

tylervigen.com

Conclusões:

- Se 2 variáveis estão **correlacionadas**, **pode ou não** haver **causalidade**
- Se houver **correlação** e não houver **causalidade** entre essas 2 variáveis, possivelmente há uma **3ª variável que não foi observada**
- **Mantenha-se cético: busque fortes evidências** para assumir a **causalidade**
- Antes de assumir a **causalidade** responda as seguintes perguntas:
 - **Por que** a variável A **causa** a variável B?
 - **Como** a variável A **causa** a variável B?

