Disciplina: 5170 - Probabilidade e Estatística

Aula 06

Prof. George Lucas M. Pezzott
glmpezzott2@uem.br

Departamento de Estatística - UEM
Sala: 222 - Bloco: E-90

Segunda-feira - 19:30 $\sim$ 21:10 - D67 - Sala 108
Sexta-feira - 21:20 $\sim$ 23:00 - D67 - Sala 108

Análise bidimensional

- Em muitos problemas práticos, para compreender melhor a população, as variáveis devem ser analisadas conjuntamente. A análise isolada de cada variável pode não ser suficiente.

- Assim, uma **análise bivariada (ou bidimensional)** é quando analisamos conjuntamente duas variáveis relacionadas ao mesmo problema.
  - **Escolaridade** *versus* **Intenção de voto** em pesquisas eleitorais;
  - **Tempo de Experiência** *versus* **Salário** de uma empresa.
  - **Peso** *versus* **Altura** em indivíduos;

- A análise bivariada é uma extensão da análise de uma única variável e é também um caso particular da análise multivariada.

Vamos considerar os seguintes casos para uma análise bivariada:

- Variável Qualitativa $\times$ Variável Qualitativa.
- Variável Qualitativa $\times$ Variável Quantitativa.
- Variável Quantitativa $\times$ Variável Quantitativa.

Variável Qualitativa $\times$ Variável Qualitativa

- **Exemplo:** variável X: tipo de residência (Própria, Alugada, Outra) e variável Y: *status* do proprietário com relação a linhas de créditos (bom ou mau pagador).

Indivíduo	X: tipo de residência	Y: *Status*
1	Própria	Bom
2	Própria	Bom
3	Alugada	Bom
4	Própria	Mau
5	Alugada	Mau
6	Alugada	Bom
7	Outra	Bom
8	Alugada	Mau
9	Outra	Bom
⋮	⋮	⋮
2000	Própria	Bom

Variável Qualitativa × Variável Qualitativa

- Para cada variável, podemos fazer uma tabela de frequências.

Residência	Frequência	Frequência relativa (%)
Própria	429	21,45
Alugada	1171	58,55
Outro	400	20,00
Total	2000	100,00

Variável Qualitativa × Variável Qualitativa

- Para cada variável, podemos fazer uma tabela de frequências.

Status	Frequência	Frequência relativa (%)
Mau	713	35,65
Bom	1287	64,35
Total	2000	100,00

Variável Qualitativa × Variável Qualitativa

- Contudo, podemos fazer uma tabela de dupla entrada, conhecida como **Tabela de Contingência**:

  Tabela: Distribuição das frequências cruzadas das variáveis 'tipo de residência' e '*status*' de pagador.

	Status		
Residência	Mau	Bom	Total
Própria	165	264	429
Alugada	399	772	1171
Outro	149	251	400
Total	713	1287	2000

Variável Qualitativa $\times$ Variável Qualitativa

- A **Tabela de Contingência** apresenta a distribuição conjunta das variáveis.
- Nas primeiras linha e coluna são colocadas as respostas das duas variáveis. O corpo da tabela exibe as frequências observadas dos pares de respostas entre os indivíduos;
- A Tabela permite encontrar as distribuições da cada variável, também chamadas de **distribuições marginais**, que são os totais nas linhas e colunas.

Variável Qualitativa $\times$ Variável Qualitativa

- Para melhor visualização da associação entre as variáveis, podemos calcular as frequências relativas de uma variável em cada resposta da outra variável.

Tabela: Distribuição das frequências relativas condicionais das variáveis 'tipo de residência' e '*status*' de pagador.

	Status		
Residência	Mau	Bom	Total
Própria	165/713	264/1287	429/2000
Alugada	399/713	772/1287	1171/2000
Outro	149/713	251/1287	400/2000
Total	713/713	1287/1287	2000/2000

Variável Qualitativa × Variável Qualitativa

Tabela: Distribuição das frequências relativas condicionais das variáveis 'tipo de residência' e '*status*' de pagador.

	Status		
Residência	Mau	Bom	Total
Própria	23,13	20,52	21,45
Alugada	55,96	59,98	58,55
Outro	20,89	19,50	20,00
Total	100,00	100,00	100,00

- **Interpretação**: Aparentemente, não há associação entre o tipo de residência e o status de bom pagador.

Variável Qualitativa $\times$ Variável Qualitativa

- Por exemplo, observando as frequências relativas em cada resposta de tipo de residência.
  Tabela: Distribuição das frequências relativas condicionais das variáveis 'tipo de residência' e '*status*' de pagador.

	Status		
Residência	Mau	Bom	Total
Própria	165/429	264/429	429/429
Alugada	399/1171	399/1171	1171/1171
Outro	149/400	251/400	400/400
Total	713/2000	1287/2000	2000/2000

Variável Qualitativa × Variável Qualitativa

- Por exemplo, observando as frequências relativas em cada resposta de tipo de residência.
  Tabela: Distribuição das frequências relativas condicionais das variáveis 'tipo de residência' e '*status*' de pagador.

	Status		
Residência	Mau	Bom	Total
Própria	38,46	61,54	100,00
Alugada	34,07	65,93	100,00
Outro	37,25	62,75	100,00
Total	35,65	64,35	100,00

  - **Interpretação**: Aparentemente, não há associação entre o tipo de residência e o status de bom pagador.

Variável Qualitativa × Variável Qualitativa

- Exemplo de possível associação entre as variáveis
  Tabela: Distribuição das frequências relativas condicionais das variáveis 'tipo de residência' e '*status*' de pagador.

	Status		
Residência	Mau	Bom	Total
Própria	7,46	92,54	100,00
Alugada	10,07	89,93	100,00
Outro	60,25	39,75	100,00
Total	35,65	64,35	100,00

  - **Interpretação**: Aparentemente, há associação entre o tipo de residência e o status de bom pagador, pois dependendo do tipo de residência do indivíduo, há certa tendência na resposta de seu *status*.

Variável Qualitativa $\times$ Variável Quantitativa

- Quando as duas variáveis a serem analisadas são de natureza diferente e uma delas é quantitativa e a outra qualitativa.
- **Exemplo:** X: Escolaridade e Y: Salário na empresa.

Indivíduo	Escolaridade	Salário(R$)
1	Ens. Médio	1212,34
2	Ens. Médio	1511,45
3	Ens. Fundamental	998,32
4	Ens. Superior	4500,00
5	Pós-graduação	6133,90
6	Ens. Superior	3999,01
⋮	⋮	⋮
100	Ens. Médio	1914,20

Variável Qualitativa $\times$ Variável Quantitativa

- Quando as duas variáveis a serem analisadas são de natureza diferente e uma delas é quantitativa e a outra qualitativa.
- **Tabela:** Estatísticas descritivas para a variável Y: Salário na empresa separadas por X: Escolaridade.

	Ens. Fund.	Ens. Médio	Ens. Sup.	Pós-grad.
n	10.00	50.00	30.00	20.00
$\bar{x}$	976.74	2497.46	4538.68	5935.91
Md	970.54	2539.98	4513.38	5970.70
s	78.72	195.28	314.75	528.02
CV(100%)	8.06%	7,82 %	6.93%	8.90%
Min	848.57	1965.32	3783.63	4874.26
Max	1127.29	2905.25	5023.72	6890.46
Amplitude	278.72	939.93	1240.10	2016.20
Q_1	928.07	2335.89	4310.86	5663.84
Q_3	1009.75	2636.01	4830.55	6252.08
d_Q	81.68	300.12	519.69	588.24

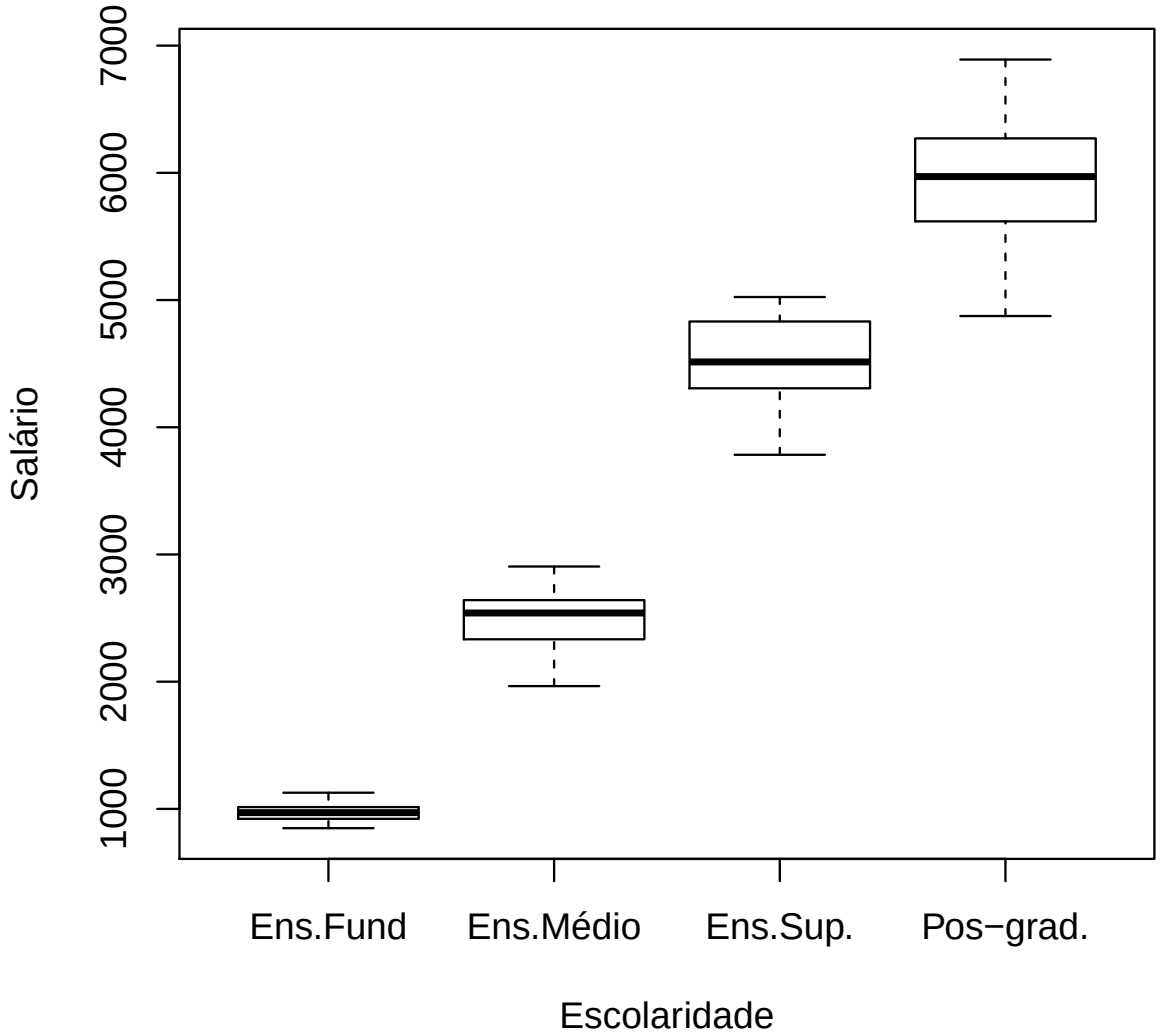


Figura: Box-plot dos salários *versus* Escolaridade.

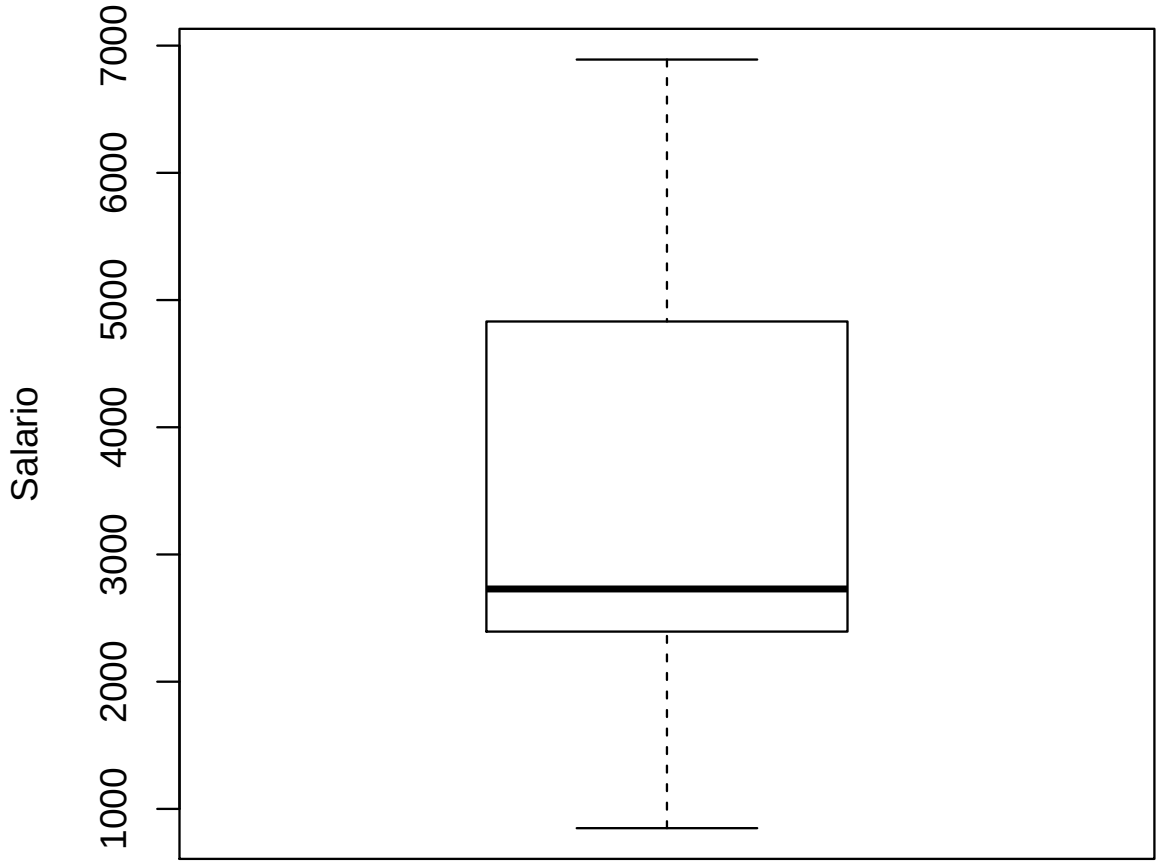


Figura: Box-plot dos salarios

Variável Quantitativa $\times$ Variável Quantitativa

- Quando as duas variáveis a serem analisadas são de mesma natureza sendo ambas quantitativas, podemos estar interessados em analisar o grau de "correlação" entre elas.
- Numa população de pessoas, podemos dizer que as variáveis peso e altura são correlacionadas positivamente, pois a maioria dos indivíduos altos também é pesada, enquanto a maioria dos indivíduos baixos é leve.
- De forma análoga, o faturamento de uma empresa e o nível de utilização do seu sistema computacional devem ter correlação positiva.
- Já a quantidade de memória RAM e o tempo de Processamento devem ter correlação negativa;

Análise Bivariada Coeficiente de Correlação

- Dizemos que duas variáveis, X e Y, estão **positivamente correlacionadas** quando elas caminham num mesmo sentido, ou seja, elementos com valores pequenos de X tendem a ter valores pequenos de Y e elementos com valores grandes de X tendem a ter valores grandes de Y.
- Por outro lado, elas estão **negativamente correlacionadas** quando elas caminham em sentidos opostos, ou seja, elementos com valores de X pequenos tendem a ter valores grandes de Y e elementos com valores grandes de X tendem a ter valores pequenos de Y.

- Uma forma de visualizarmos se duas variáveis apresentam-se correlacionadas é através de um **diagrama de dispersão**
- No diagrama de dispersão, os valores das variáveis são representados por pontos, num sistema cartesiano.
- Cada ponto é o par (x, y) de cada observação na amostra referentes às variáveis X e Y.
- Segue um exemplo.

Análise Bivariada

Tabela: Volume de vendas (Y) na última semana *versus* anos de experiência (X) de 10 vendedores da loja A.

Vendedor	Volume de Vendas	Anos de experiência
1	9	6
2	6	5
3	4	3
4	3	1
5	3	4
6	5	3
7	8	6
8	2	2
9	7	4
10	4	2

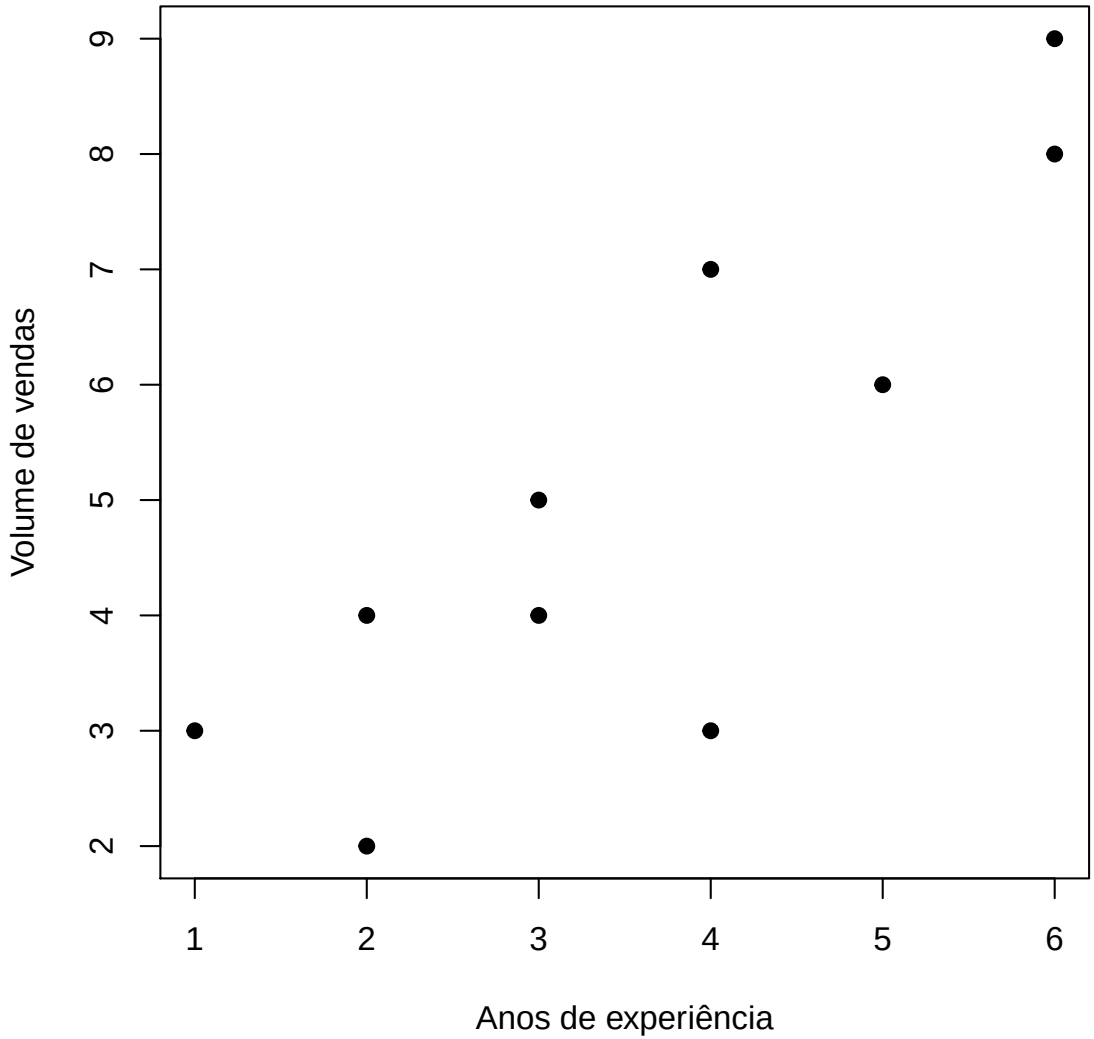


Figura: Diagrama de dispersão dos anos de experiência *versus* o volume em vendas.

Correlação

- Será que quanto mais anos de experiência maior será o volumes em vendas?
- O **coeficiente de correlação** é uma medida numérica da "força" e "direção" da relação (ou associação) entre duas variáveis quantitativas contínuas.

Covariância

A **covariância**, $Cov(X, Y)$, entre duas variáveis quantitativas contínuas é uma medida do quanto uma das variáveis se modifica quando a outra se modifica.

- A covariância $Cov(X, Y)$ pode ser estimada a partir dos dados $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ por:

$$cov_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

em que $\overline{x}$ e $\overline{y}$ são as respectivas médias dos dados.

Covariância

- Se a covariância é igual a zero, entendemos que, conforme as observações de uma das variáveis crescem, as observações da outra variável não tendem a crescer ou decrescer, ou seja, não há uma relação linear entre as duas variáveis.
- Se a covariância é maior que zero, entendemos que, conforme as observações de uma das variáveis crescem, as observações da outra variável tendem a crescer também.
- Se a covariância é menor que zero, entendemos que, conforme as observações de uma das variáveis crescem, as observações da outra variável tendem a decrescer.

Covariância

- As variáveis X e Y são expressas em diferentes unidades de medida. A experiência é expressa em anos, e as vendas é em produtos (carros, casas,...).
- No entanto, podemos calcular a covariância e os coeficientes de correlação entre variáveis com unidades de medida diferentes.

Tabela: Valores para o cálculo da covariância.

i	x_i	$(x_i - \bar{x})$	y_i	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	$(6 - 3,6) = 2.40$	9	$(9 - 5.1) = 3.90$	9.36
2	5	$(5 - 3,6) = 1.40$	6	$(6 - 5.1) = 0.90$	1.26
3	3	$(3 - 3,6) = -0.60$	4	$(4 - 5.1) = -1.10$	0.66
4	1	$(1 - 3,6) = -2.60$	3	$(3 - 5.1) = -2.10$	5.46
5	4	$(4 - 3,6) = 0.40$	3	$(3 - 5.1) = -2.10$	-0.84
6	3	$(3 - 3,6) = -0.60$	5	$(5 - 5.1) = -0.10$	0.06
7	6	$(6 - 3,6) = 2.40$	8	$(8 - 5.1) = 2.90$	6.96
8	2	$(2 - 3,6) = -1.60$	2	$(2 - 5.1) = -3.10$	4.96
9	4	$(4 - 3,6) = 0.40$	7	$(7 - 5.1) = 1.90$	0.76
10	2	$(2 - 3,6) = -1.60$	4	$(4 - 5.1) = -1.10$	1.76
-		$\bar{x} = 3,6$		$\bar{y} = 5,1$	Total = 30,4

Covariância

- A covariância entre X e Y é dada por:

$$Cov(X, Y) = \frac{\sum_{i=1}^{10}(x_i - 3,6)(y_i - 5,1)}{10 - 1} = 3,377778$$

- Como a covariância é maior do que zero, entendemos que pessoas com mais anos de experiência tendem a vender mais produtos.
- Como interpretamos a magnitude da covariância? Ou seja, esse valor obtido, 3,377778, pode nos indicar o "tamanho"da associação entre essas variáveis?

Correlação de Pearson

- Uma medida de **relação linear entre as variáveis quantitativas** pode ser calculada com o **coeficiente de correlação de Pearson**.

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}},$$

em que $Cov(X,Y)$ representa a covariância entre as variáveis quantitativas X e Y; e os termos $Var(X)$ e $Var(Y)$ são as respectivas variâncias.

- Essa quantidade $\rho_{X,Y}$ sempre assume valores entre -1 e 1, sua interpretação como medida de associação é simples.

Correlação de Pearson

- Valores de $\rho_{X,Y}$ iguais a zero evidenciam que não há associação entre as variáveis X e Y.
- Valores próximo de zero (sejam eles negativos ou positivos) indicam uma associação muito fraca entre as variáveis.
- Valores de $\rho_{X,Y}$ próximos a -1 ou 1 indicam uma associação forte entre X e Y.

Correlação de Pearson

Tabela: Interpretação do coeficiente de correlação segundo Zou et al. (2003)

$\rho_{X,Y}$	Direção e força da associação
-1,0	Perfeita e negativa
-0,8	Forte e negativa
-0,5	Moderada e negativa
-0,2	Fraca e negativa
0	Ausência de associação
0,2	Fraca e positiva
0,5	Moderada e positiva
0,8	Forte e positiva
1,0	Perfeita e positiva

- Coeficientes maiores que zero indicam correlação positiva (quanto maior X, maior Y).
- Coeficientes menores que zero indicam correlação negativa (quanto maior X, menor Y).

Correlação de Pearson

No exemplo da idade e níveis de triglicérides:

- O desvio padrão das experiência é $s_X = 1,712698$ anos.
- O desvio padrão das vendas é $s_Y = 2,330951$ produtos.
  O coeficiente de correlação entre essas variáveis é:

$$\rho_{X,Y} = \frac{3,377778}{1,712698 \times 2,33095} = 0,8460915$$

- A partir da medida de correlação de Pearson, os dados evidenciam uma forte e positiva associação entre os anos de experiência e os volumes em vendas dos vendedores.

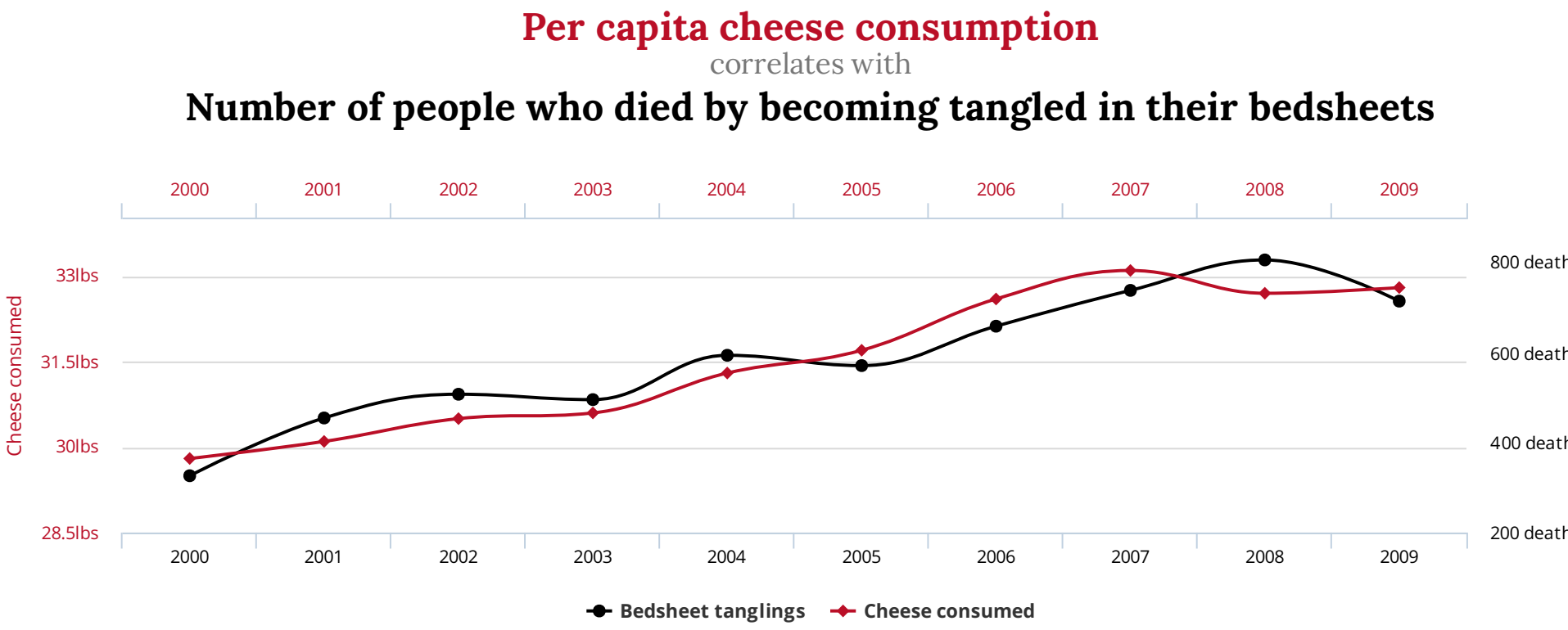
Análise Bivariada

Cuidado!

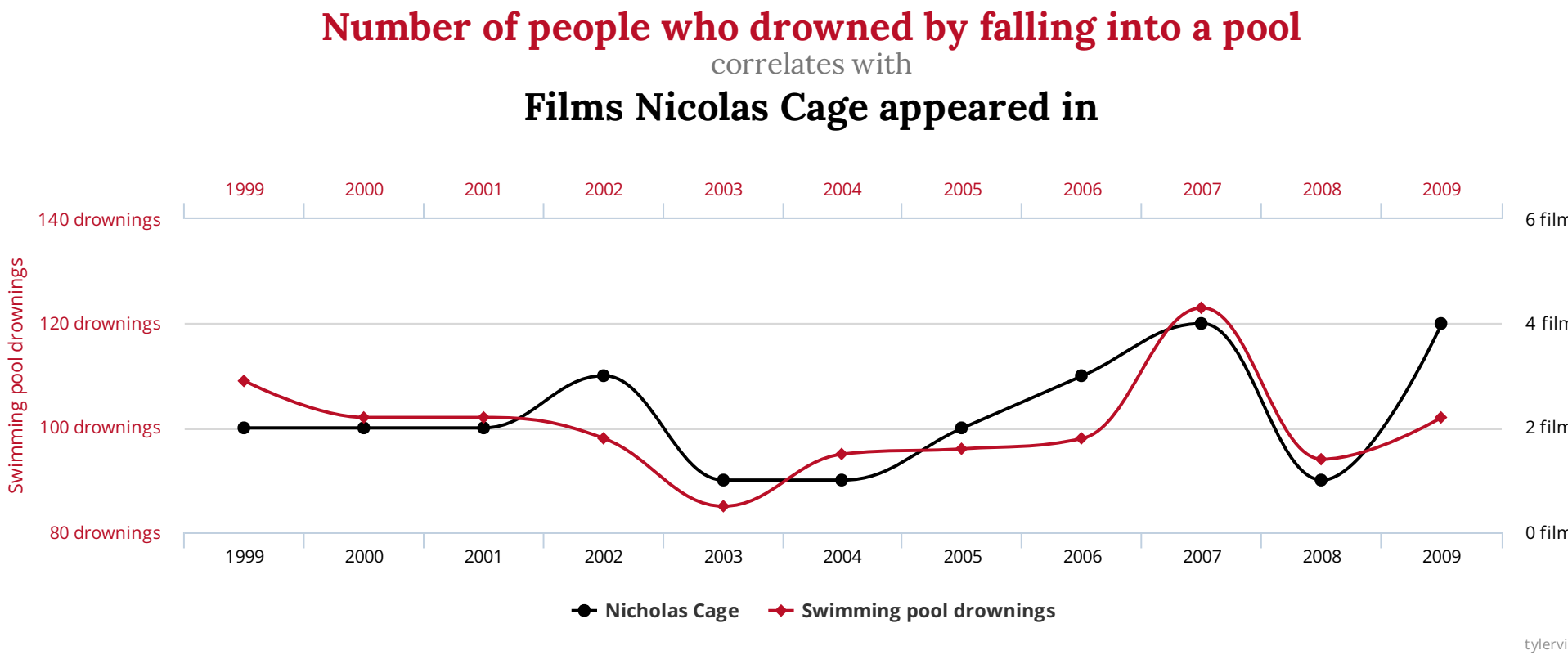
Correlação não implica causalidade.

- Exemplos:
  http://www.tylervigen.com/spurious-correlations

Correlação: 0,9471



Correlação: 0,66604



- Outros análises podem ser realizadas para medir associação entre duas variáveis quantitativas. Por exemplo, uma tabela de contingência pode ser construída considerando a construção de classes.
- **Exemplo:** Anos de experiência e salário (R$) de funcionários de um banco.

	Salário (R$)				
Tempo (anos)	800|–1000	1000|–2000	2000 |– 40000	$\geq$ 4000	Total
1 |– 5	10	8	2	0	20
5 |– 8	0	6	4	1	11
8 |–12	0	3	5	8	16
$\geq$ 12	0	3	6	8	17
Total	10	20	17	17	64