Correlação Linear

Associação de duas amostras (quantitativa)

Felipe Figueiredo

Instituto Nacional de Traumatologia e Ortopedia

Correlação Linear
Felipe Figueiredo

Sumário

Correlação Linear
Felipe Figueiredo

Variância (Revisão)

Correlação Linear
Felipe Figueiredo

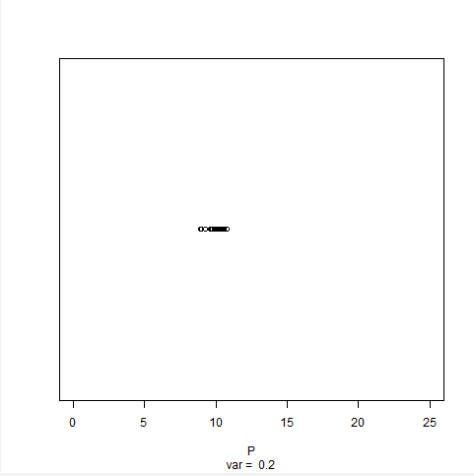
- Relembrando: a variância (assim como o desvio-padrão) é uma medida da dispersão da amostra
- Medida sumária que resume o quanto os dados se desviam da média
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

Interpretação

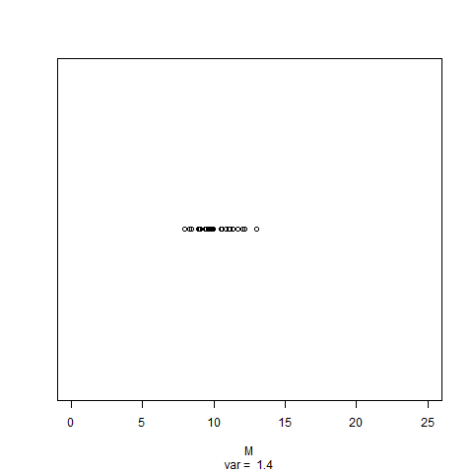
Quanto maior a variância, maior a dispersão em relação ao centro.

Visualização - Variância “pequena”

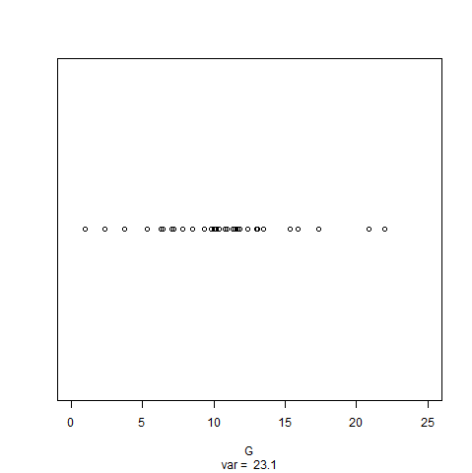
Correlação Linear
Felipe Figueiredo



Visualização - Variância “média”



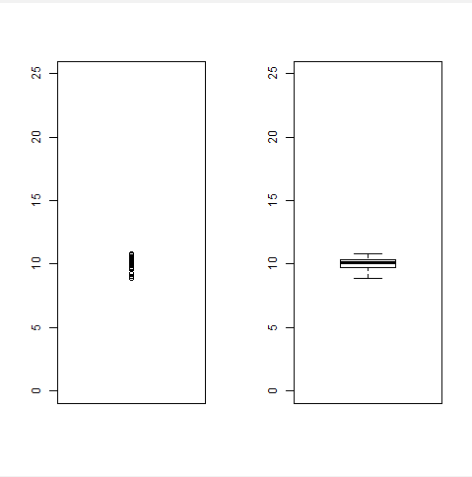
Visualização - Variância “grande”



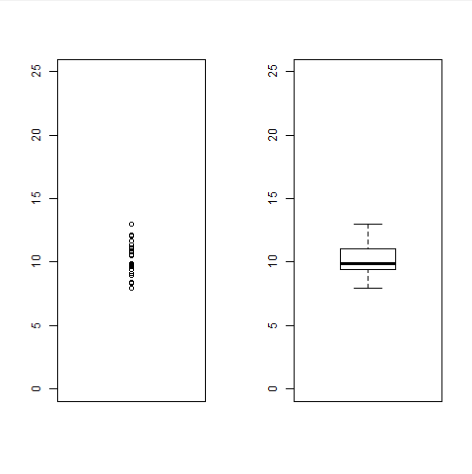
Variância em cada eixo

- Como vamos medir a variância **entre** duas variáveis, devemos considerar a variância de cada uma delas
- Exemplos anteriores, variância no eixo horizontal
- Vamos rever, agora na vertical
- (e aproveitar para incrementar a visualização de **uma** variância)

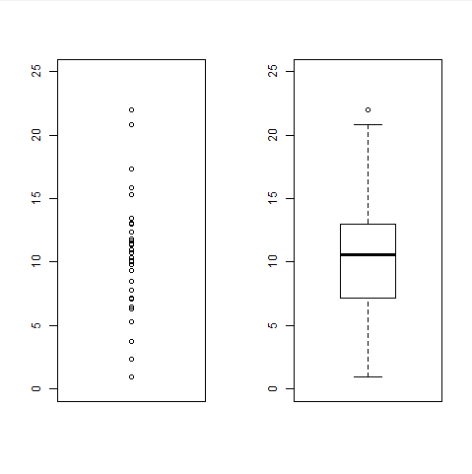
Visualização - Variância “pequena” - boxplot



Visualização - Variância “média” - boxplot



Visualização - Variância “grande” - boxplot



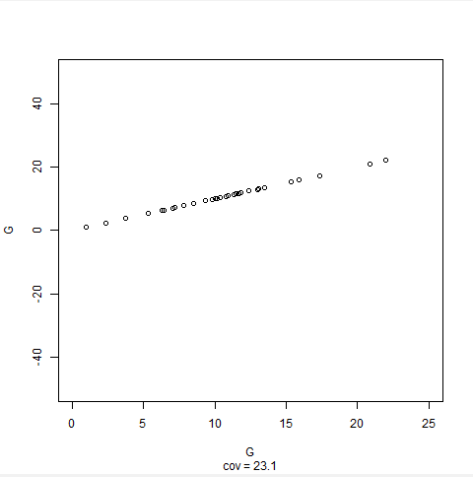
Covariância entre duas amostras

Definition

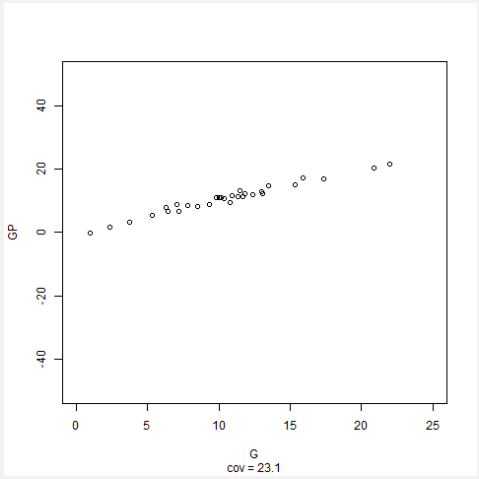
A covariância entre duas variáveis X e Y é uma medida de quanto ambas variam juntas (uma em relação à outra).

- Um único número que representa a variação conjunta entre as duas variáveis
- Pode assumir qualquer valor (positivo, negativo, etc)
- Magnitude absoluta (desconsiderando o sinal) indica o grau de dependência
- Obs: duas variáveis independentes tem covariância igual a zero!

Covariância



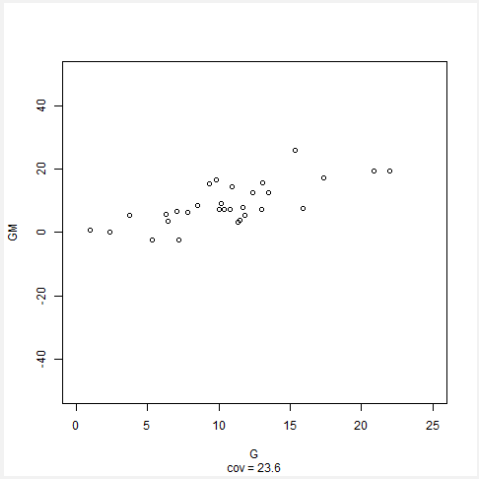
Covariância



Covariância

Correlação Linear

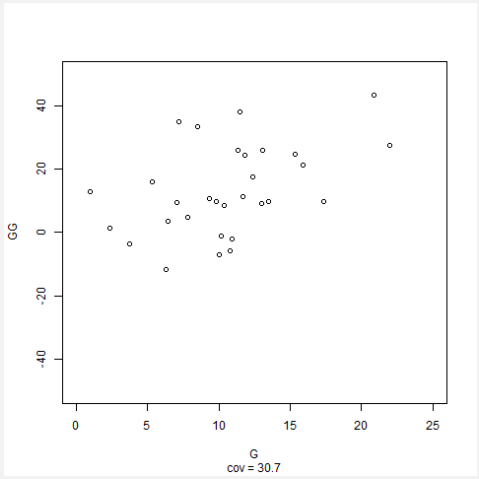
Felipe Figueiredo



Covariância

Correlação Linear

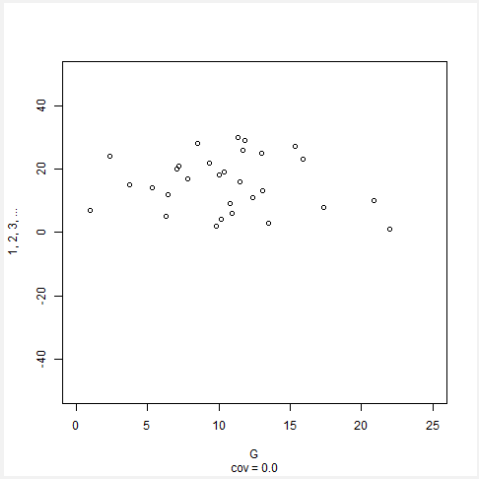
Felipe Figueiredo



Covariância - amostras independentes

Correlação Linear

Felipe Figueiredo



Luz.. Câmera... Ação!

Normalização

Teoria x Prática

A covariância é útil para entender a força (e direção) da associação entre duas variáveis, mas *o fato de ela não ter uma escala definida a torna difícil de usar na prática*.

Precisamos de uma medida semelhante de associação, mas que fique em uma escala restrita (digamos, sempre entre -1 e 1).

O nome desta solução é coeficiente de correlação *r*.

Tipos de variáveis envolvidas

- Considere duas amostras X e Y, de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
  - X: variável independente (ou variável explanatória)
  - Y: variável dependente (ou variável resposta)

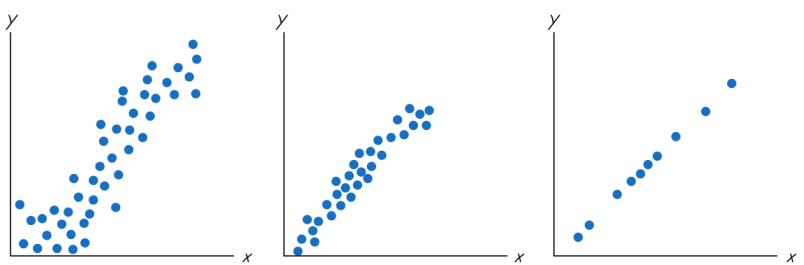
Medidas de associação

- Como definir (e mensurar!) o grau de associação entre duas amostras?
- Se uma amostra é dependente de outra, é razoável assumir que isso possa ser observável por estatísticas sumárias
- Como resumir esta informação em uma única grandeza numérica?

Medidas de associação

- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

Exemplo



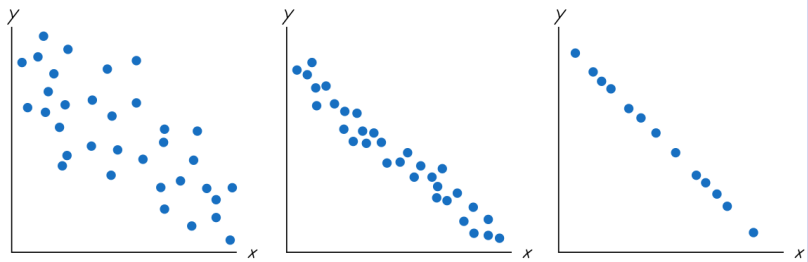
(a) Positive correlation between x and y

(b) Strong positive correlation between x and y

(c) Perfect positive correlation between x and y

(Fonte: Triola)

Exemplo



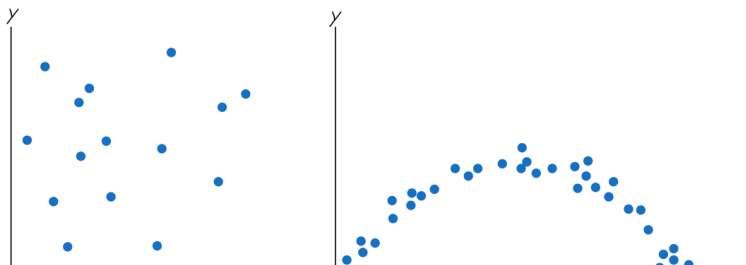
(d) Negative correlation between x and y

(e) Strong negative correlation between x and y

(f) Perfect negative correlation between x and y

(Fonte: Triola)

Exemplo



(g) No correlation between x and y

(h) Nonlinear relationship between x and y

(Fonte: Triola)

Coeficiente de correlação

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1.
- Mede a associação linear entre duas variáveis.
  - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Correlação

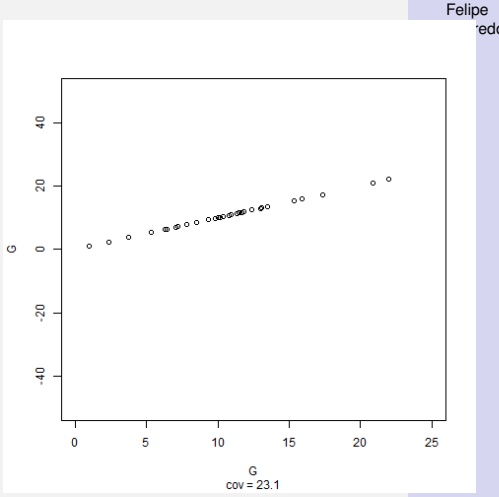
- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

IC e Teste de significância

- Se tivéssemos os dados de toda a população, poderíamos calcular o parâmetro ρ
- Na prática, só podemos calcular a estatística r da amostra
- Utilizamos r como estimador para ρ, e testamos a significância estatística da forma usual

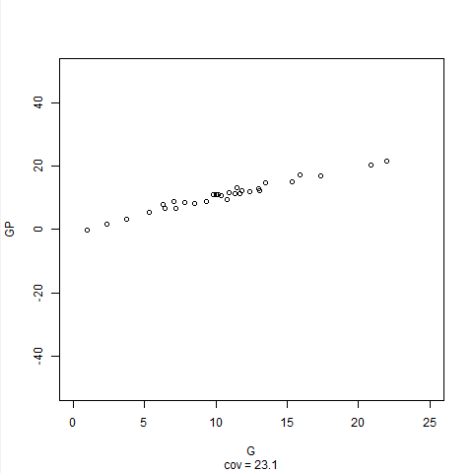
Exemplo


```
Pearson's product-moment correlation
data: G and G
t = 355110000, df = 28, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to
0
95 percent confidence interval:
1 1
sample estimates:
cor
1
```



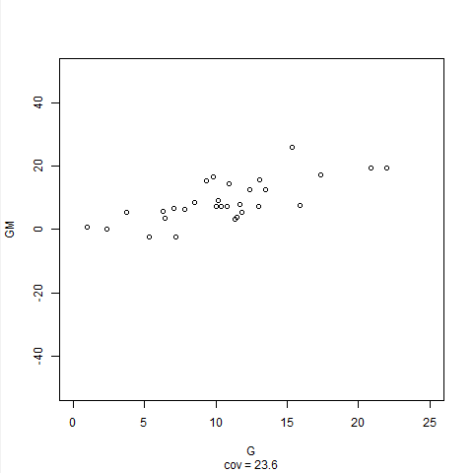
Exemplo


```
Pearson's product-moment correlation
data: G and GP
t = 28.803, df = 28, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to
0
95 percent confidence interval:
0.9653236 0.992e2253
sample estimates:
cor
0.9835406
```



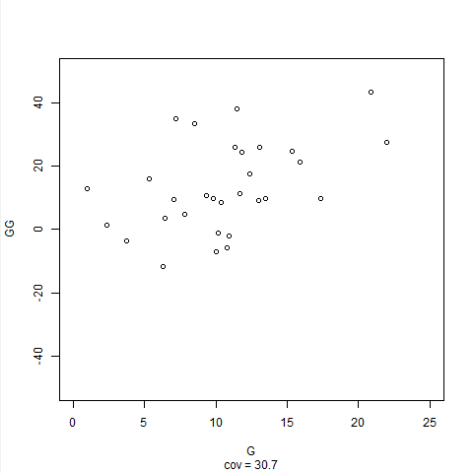
Exemplo


```
Pearson's product-moment correlation
data: G and GM
t = 5.6488, df = 28, p-value = 4.727e-06
alternative hypothesis: true correlation is not equal to
0
95 percent confidence interval:
0.5013686 0.8631382
sample estimates:
cor
0.7298133
```



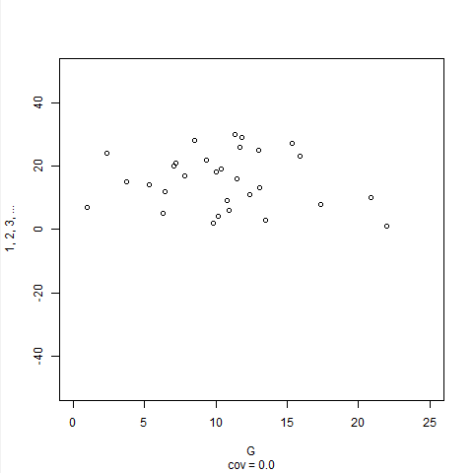
Exemplo


```
Pearson's product-moment correlation
data: G and GG
t = 2.6943, df = 28, p-value = 0.01179
alternative hypothesis: true correlation is not equal to
0
95 percent confidence interval:
0.1117472 0.6996458
sample estimates:
cor
0.4537489
```



Exemplo - amostras independentes


```
Pearson's product-moment correlation
data: G and seq(1, 30)
t = -0.64301, df = 28, p-value = 0.5254
alternative hypothesis: true correlation is not equal to
0
95 percent confidence interval:
-0.4608704 0.2505266
sample estimates:
cor
-0.1206304
```



Exemplo

Example

Pesquisadores queriam entender por que a insulina varia tanto entre indivíduos. Imaginaram que a composição lipídica das células do músculo afetam a sensibilidade do músculo para a insulina. Para isto, eles injetaram insulina em 13 jovens adultos, e determinaram quanta glicose eles precisariam injetar nos sujeitos para manter o nível de glicose sanguínea constante. A quantidade de glicose injetada para manter o nível sanguíneo constante é, então, uma medida da sensibilidade à insulina.
(Fonte: Motulsky, 1995)

Exemplo

Example

Os pesquisadores fizeram uma pequena biópsia nos músculos para aferir a fração de ácidos graxos poli-insaturados que tem entre 20 e 22 carbonos (%C20-22). Como variável resposta, mediram o índice de sensibilidade à insulina.

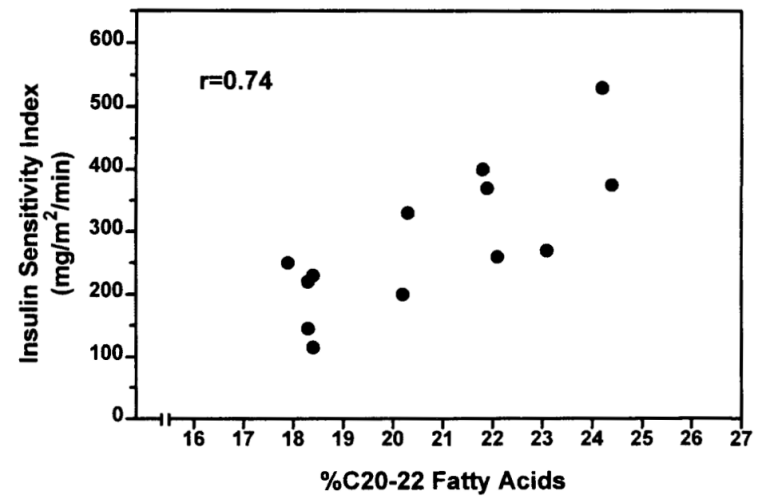
Valores tabelados a seguir.

Exemplo

Table 17.1. Correlation Between %C20–22 and Insulin Sensitivity

% C20–22 Polyunsaturated Fatty Acids	Insulin Sensitivity (mg/m²/min)
17.9	250
18.3	220
18.3	145
18.4	115
18.4	230
20.2	200
20.3	330
21.8	400
21.9	370
22.1	260
23.1	270
24.2	530
24.4	375

Exemplo: Diagrama de dispersão dos dados



Obs: na verdade, $r = 0.77$.

Exemplo

- O tamanho da amostra foi $n = 13$
- (Antigamente) consultáva-se o valor crítico de r na tabela
- H_0 : não há relação entre as variáveis na população ($H_0 : \rho = 0$).
- Observe: Quais são as informações necessárias para se consultar a tabela?

TABLE A-6 Critical Values of the Pearson Correlation Coefficient r

n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590

Exemplo

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

Exemplo

- Pode-se também calcular o p-valor para o coeficiente de correlação r.
- Para este exemplo, teríamos $p = 0.0021$.
- Interpretação: se não houver relação entre as variáveis (H_0), existe apenas 0.21% de chance de observamos uma correlação tão forte com um estudo deste tamanho

Exemplo

Por que as duas variáveis são tão correlacionadas? Considere 4 possibilidades:

1. o conteúdo lipídico das membranas determina a sensibilidade à insulina
2. A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
3. tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
4. as duas variáveis não são correlacionados na população, e a estimativa observada nessa amostra é mera coincidência

Interpretando o *r*

- Nunca devemos ignorar a última possibilidade (erro tipo I)!
- o p-valor indica quão rara é essa coincidência
- neste caso, em apenas 0.21% dos experimentos não haveria uma correlação real, e estaríamos cometendo um erro de interpretação

Interpretação

- Se a correlação é 0, então X e Y não variam juntos (independentes)
- Se a correlação é positiva, então quando uma aumenta, a outra aumenta em proporção direta (linear)
- Se a correlação é negativa, então quando uma aumenta, a outra diminui em proporção inversa (linear)

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Causa x efeito

- Se há uma relação de causalidade entre as duas variáveis, a correlação será não nula (positiva ou negativa)
- Quanto maior for a relação de dependência entre as variáveis, maior será o módulo da correlação.
- Se as variáveis não são relacionadas, a correlação será nula.

Causalidade?

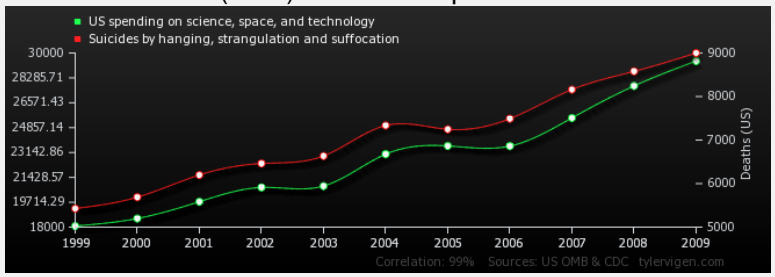
- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Exemplo

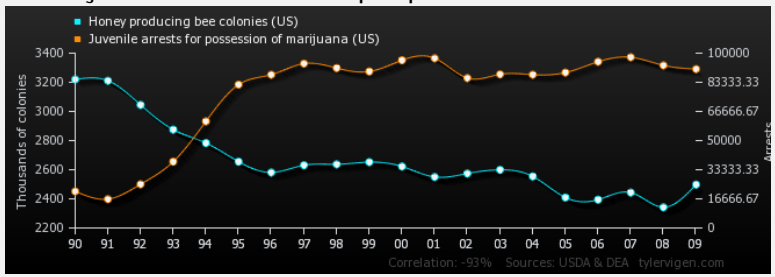
Gasto com C&T (EUA) x Suicídios por enforcamento



Correlação: 0.992082
(Fonte: Spurious correlations)

Exemplo

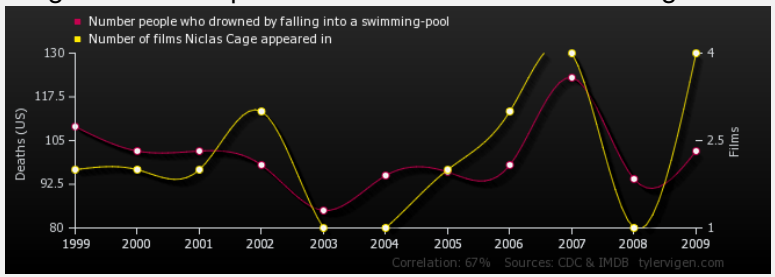
Produção de mel x Prisões por posse de maconha



Correlação: -0.933389
(Fonte: Spurious correlations)

Exemplo

Afogamentos em piscina x Filmes com Nicholas Cage



Correlação: 0.666004
(Fonte: Spurious correlations)

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?

Leitura pós-aula e exercícios selecionados

Leitura obrigatória

- Capítulo 17, pular as seções:
  - cálculo do r, do IC, do p-valor
  - correlação de Spearman, e seu cálculo
  - Interpretação do r^2

Exercícios

Capítulo 17, problemas **1**, **3** e **5**.
Problema **6**, usar:
$r = 0.8868,\ IC95\% = [0.4856, 0.9794],\ p = 0.0033.$ r^2=?

Leitura recomendada

Capítulo 17: Interpretação do r^2 e Correlação de Spearman