# Correlação Linear

## Associação de duas amostras (quantitativa)

Felipe Figueiredo

Instituto Nacional de Traumatologia e Ortopedia

# Sumário

## Discussão da leitura obrigatória da aula passada

# Sumário

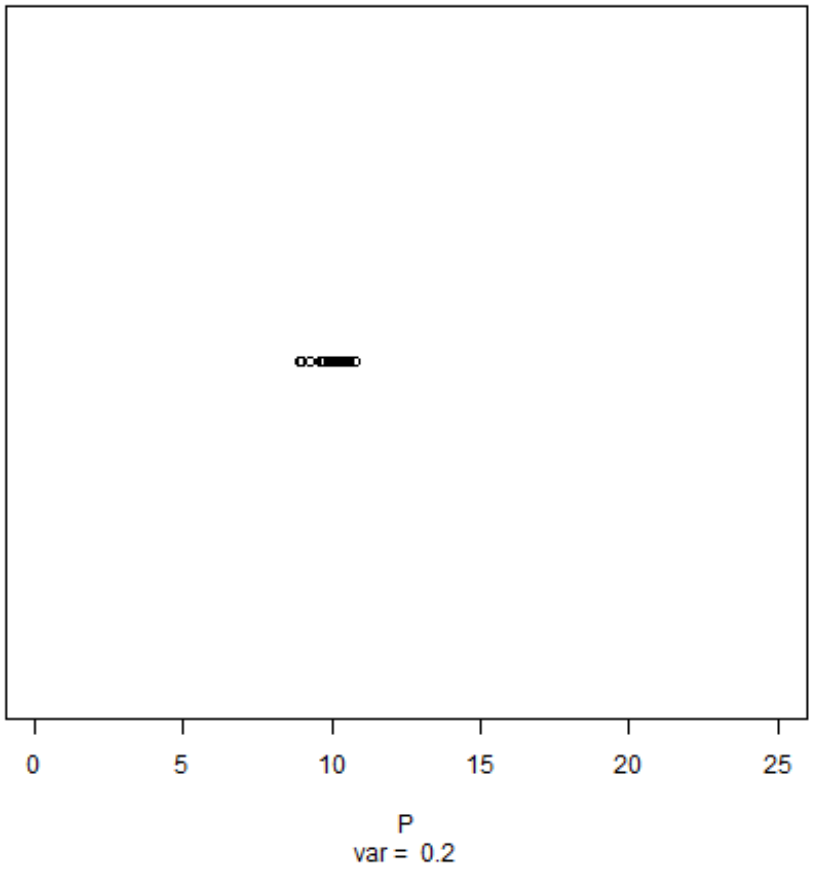
- A variância (assim como o DP) é uma medida da dispersão da amostra
- P: o quanto os dados se desviam da média?
- Medida sumária: um único número para a amostra

## Interpretação

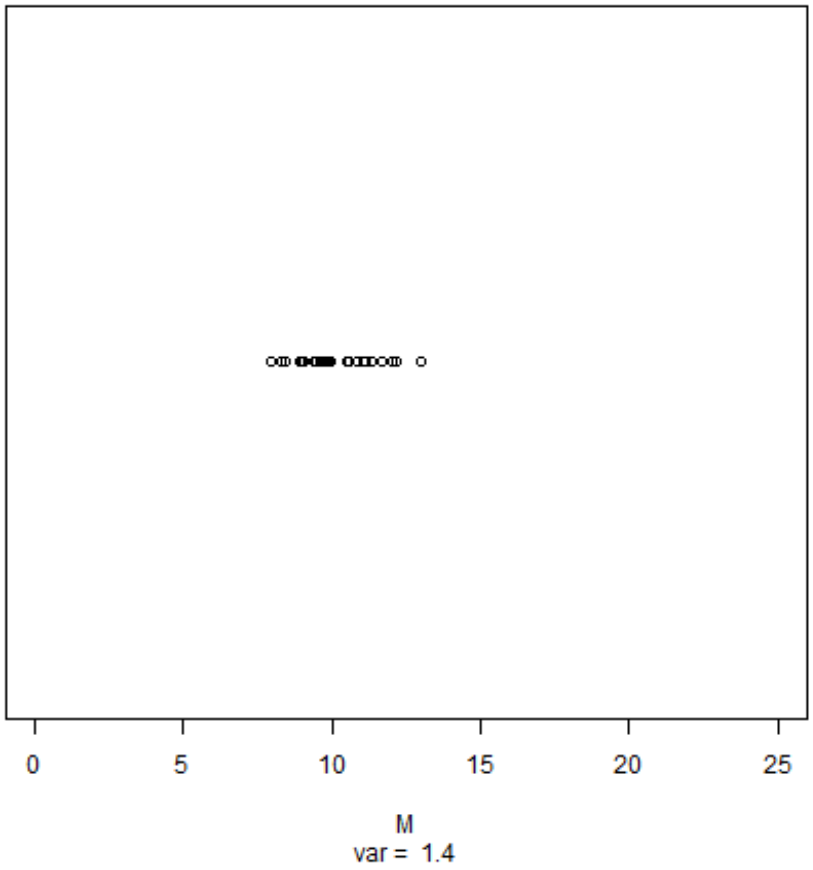
Quanto maior a variância...

... maior a dispersão em relação ao centro.

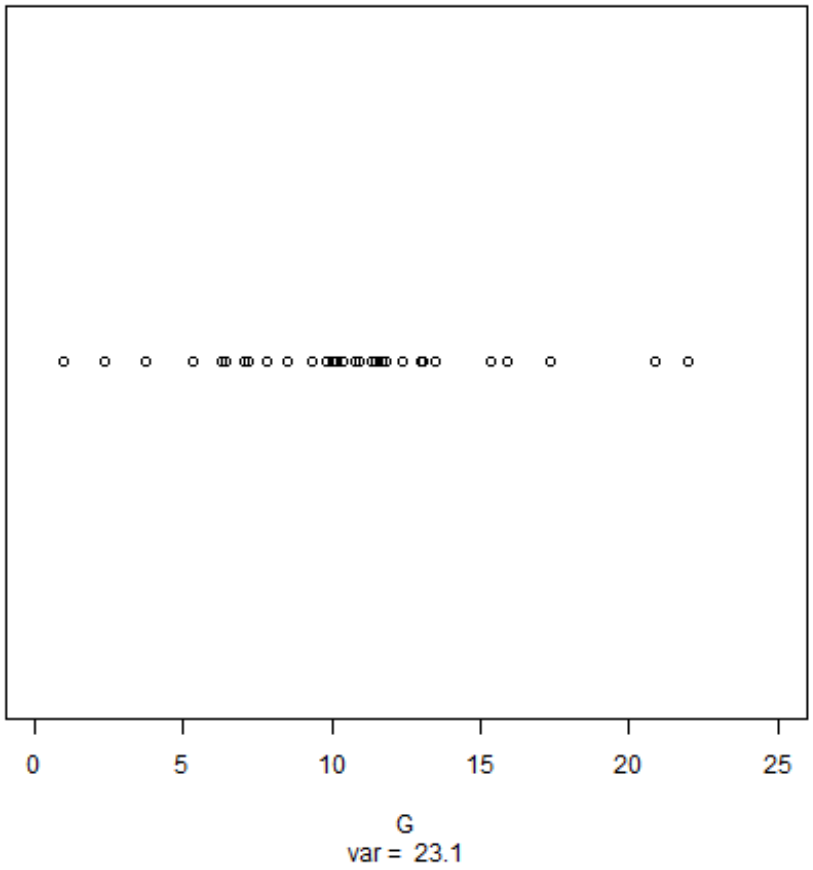
# Visualização – Dispersão “pequena”



# Visualização – Dispersão “média”



# Visualização – Dispersão “grande”

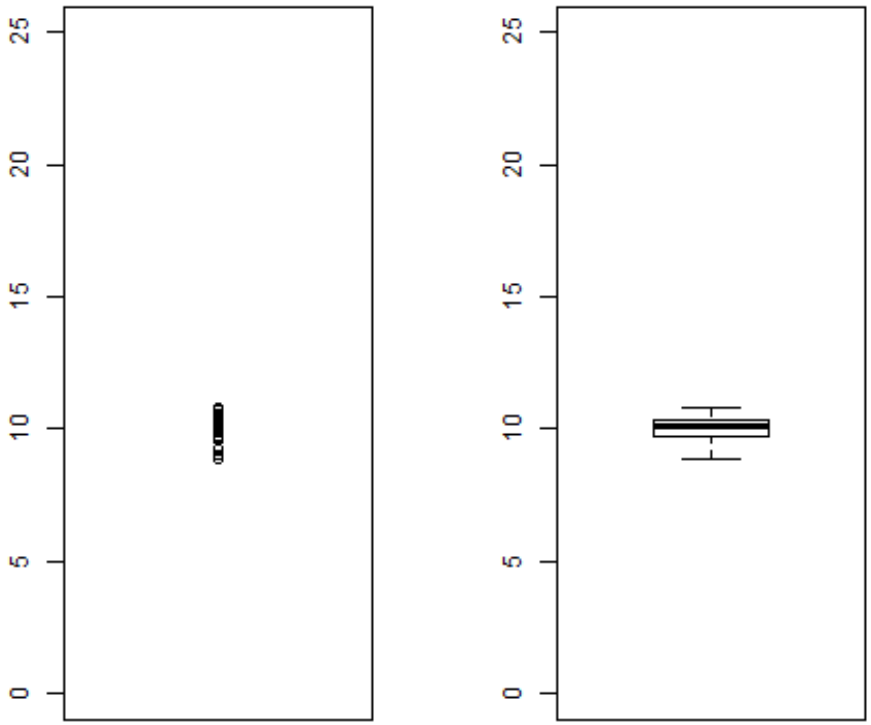


- Para medir a associação **entre** duas variáveis contínuas, devemos considerar a dispersão de cada uma delas

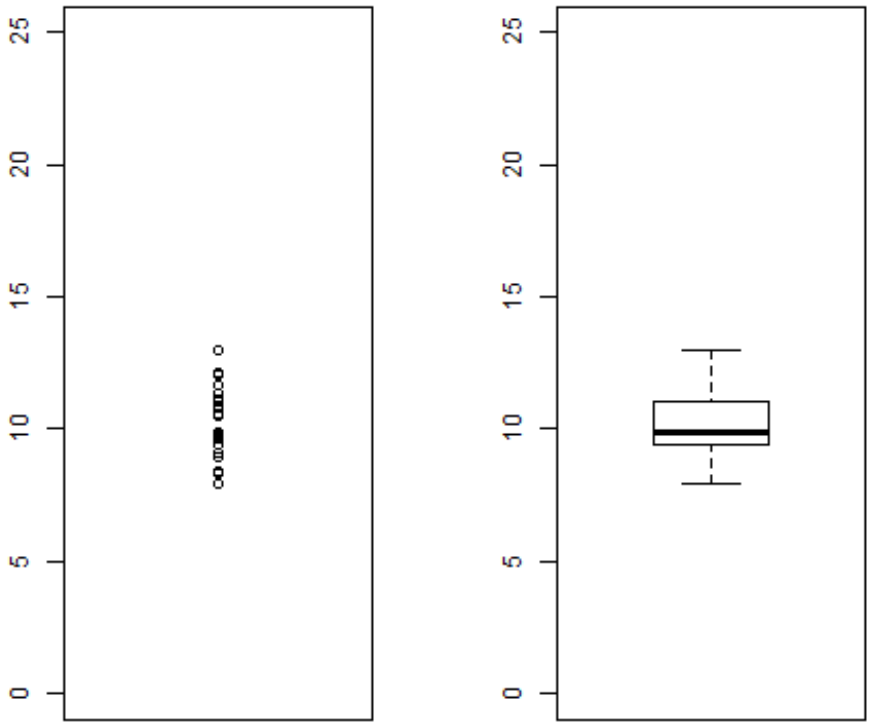
- Exemplos anteriores: dispersão no eixo horizontal
- Vejamos agora no eixo vertical

- (e aproveitar para incrementar a visualização de **uma** variância)

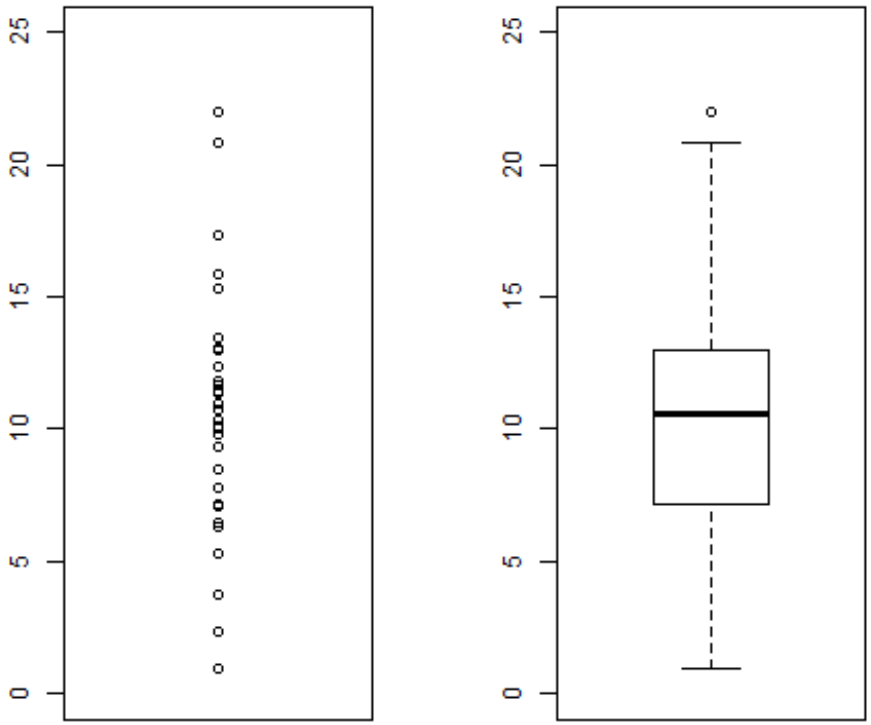
# Visualização – Dispersão “pequena” – boxplot



# Visualização – Dispersão “média” – boxplot



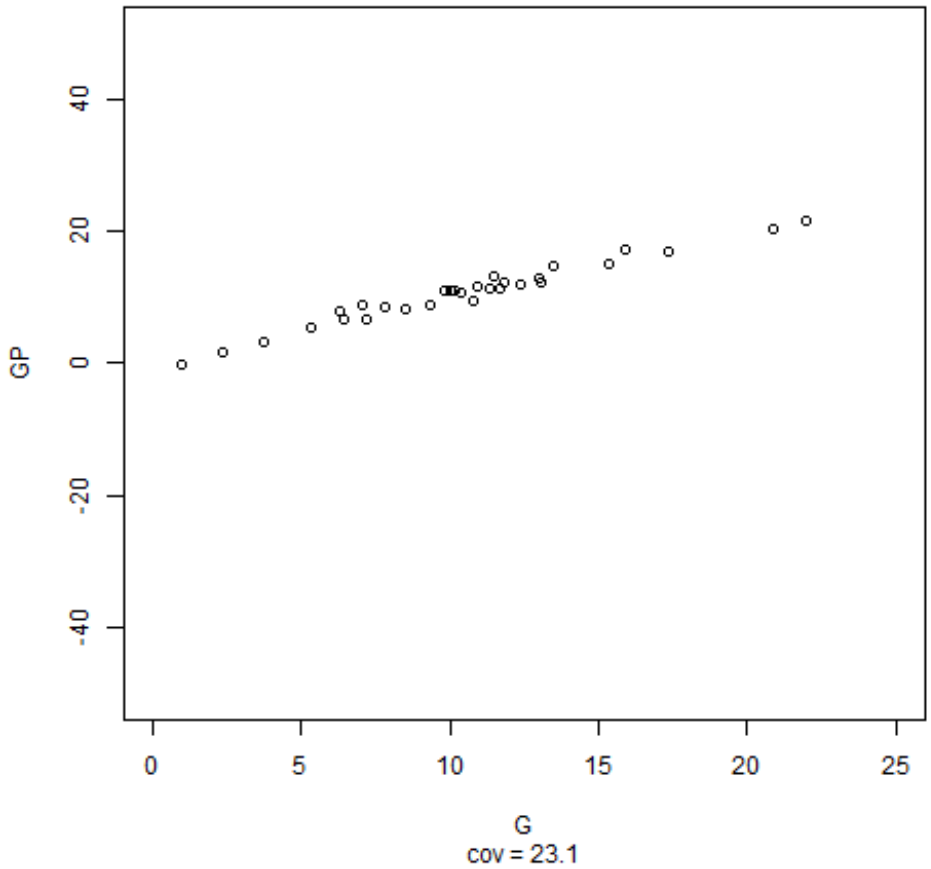
# Visualização – Dispersão “grande” – boxplot



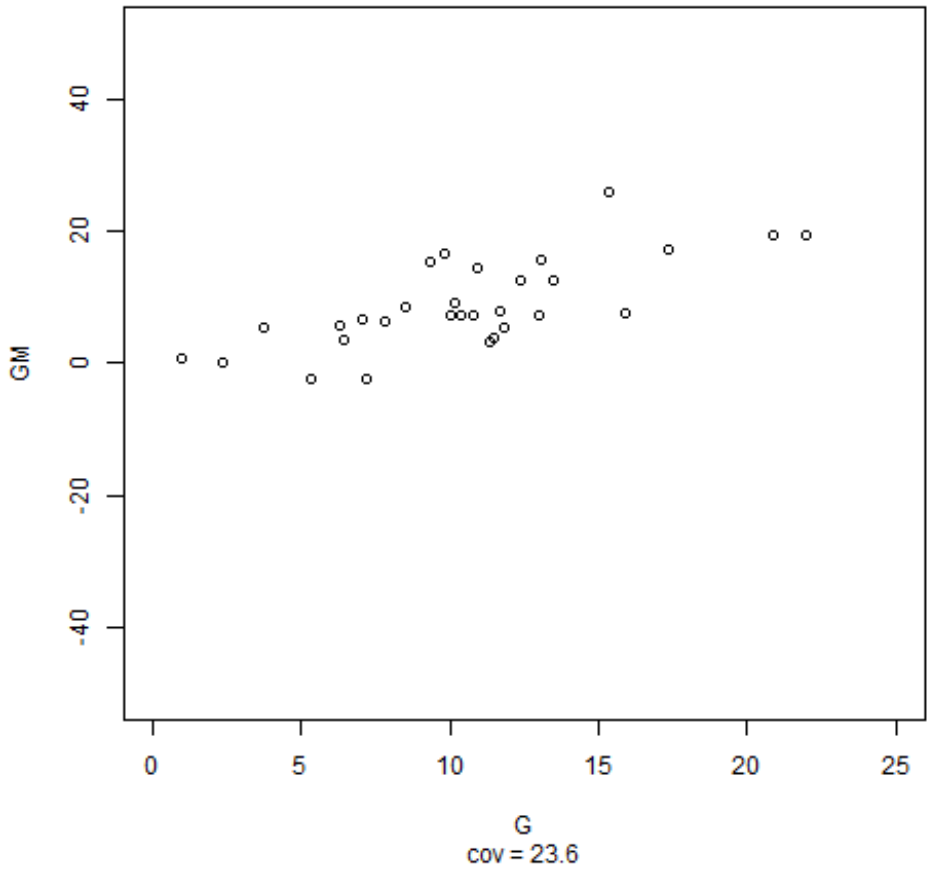
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

- Pareando duas amostras, podemos tentar observar:
  - a dispersão no eixo horizontal (difícil)
  - a dispersão no eixo vertical (difícil)
  - a “dispersão conjunta” entre ambas (fácil)

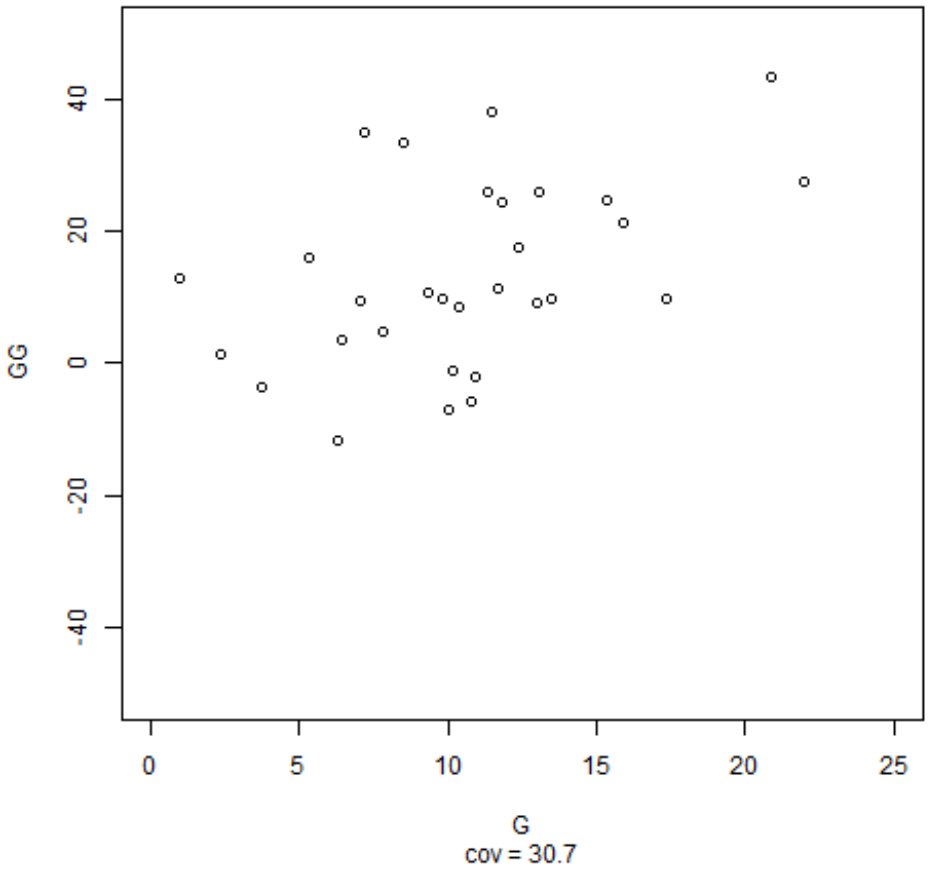
# Visualização – Dispersão “pequena”



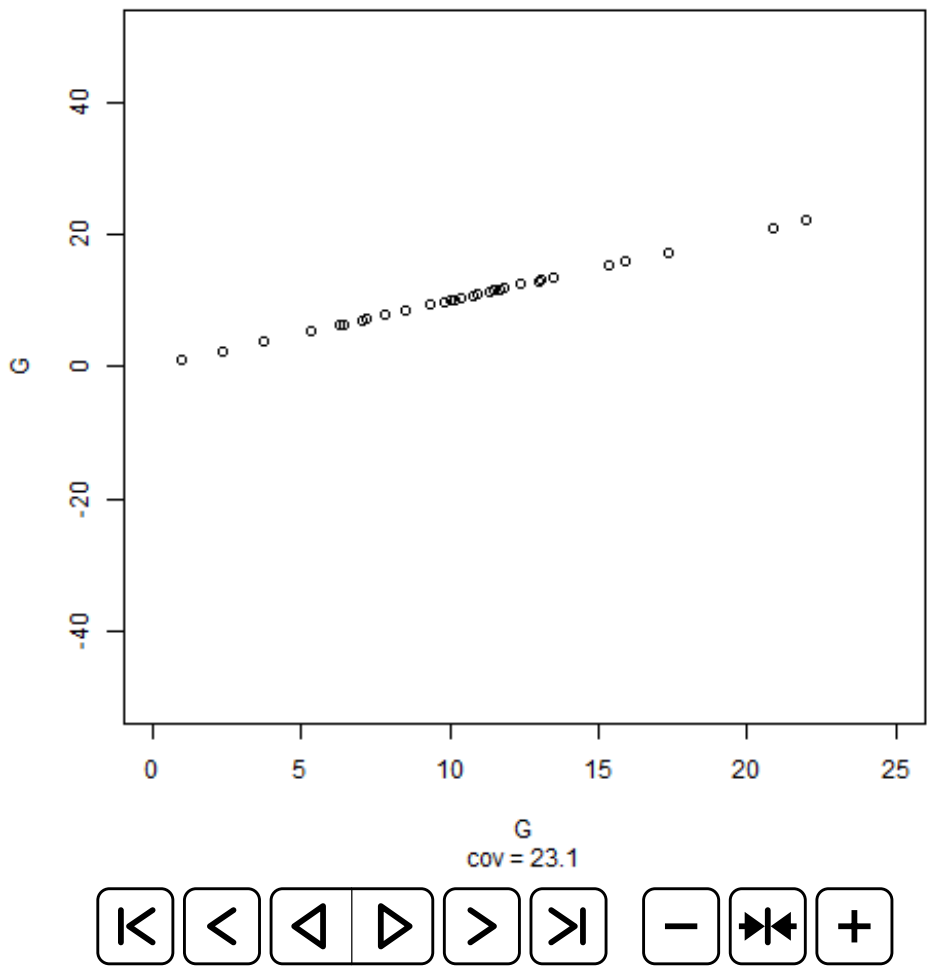
# Visualização – Dispersão “média”



# Visualização – Dispersão “grande”

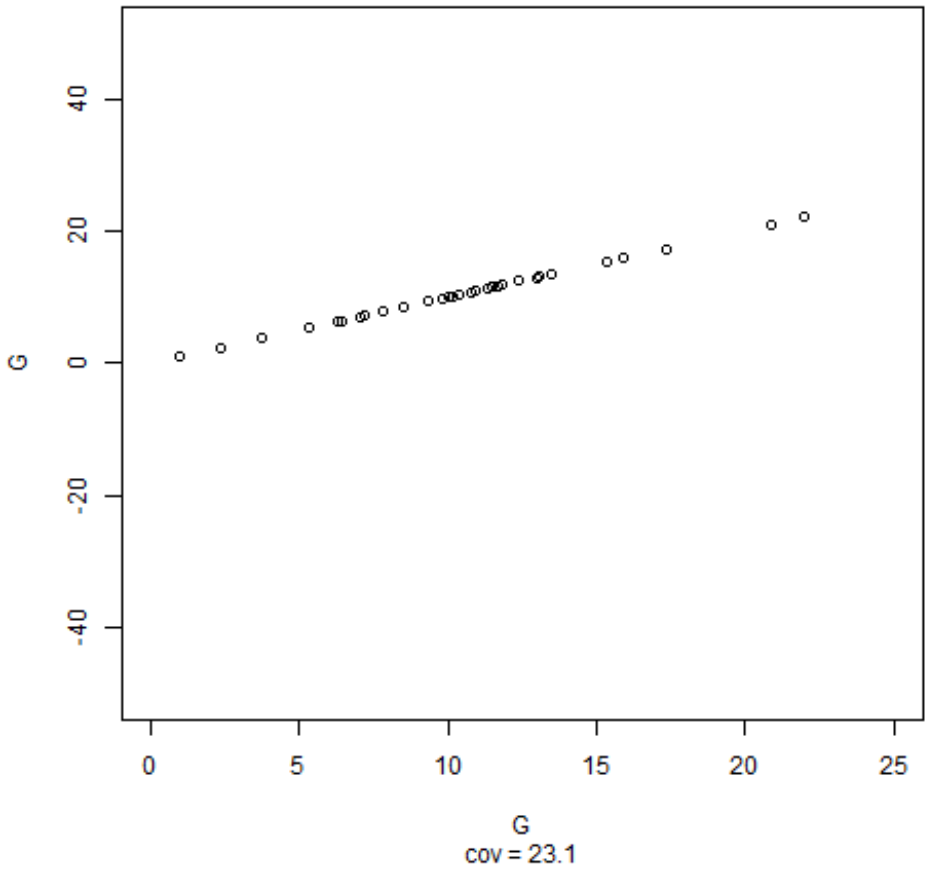


# Luz.. Câmera... Ação!

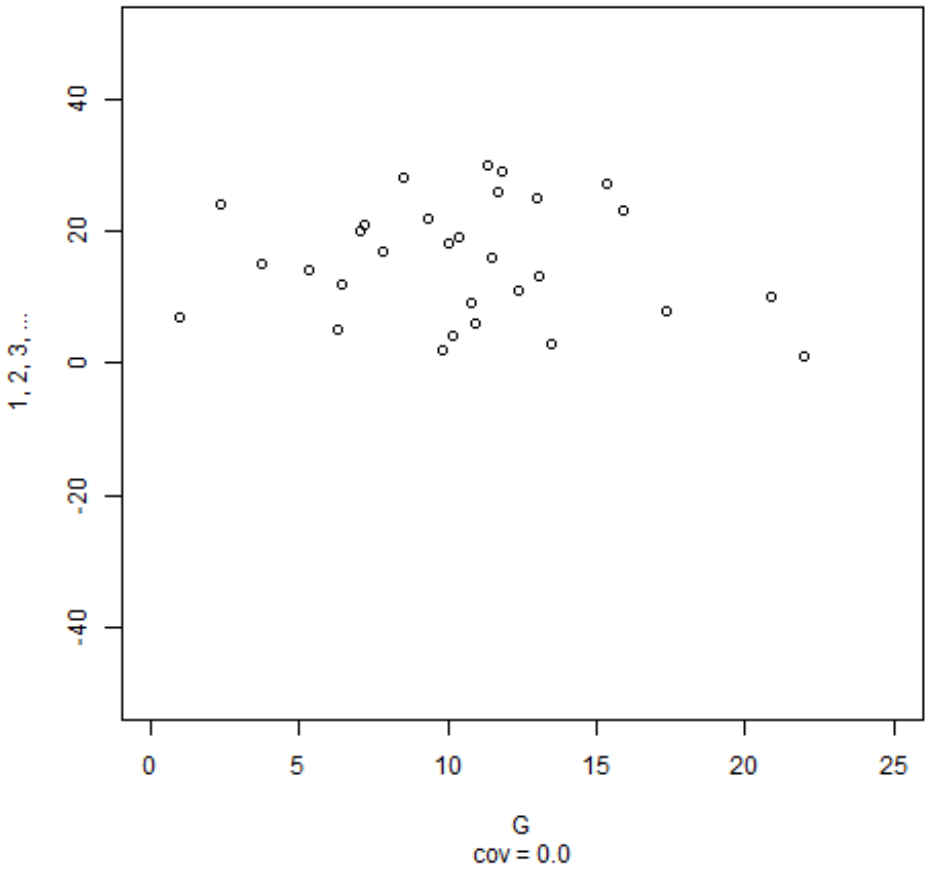


- Esta dispersão conjunta é a base para entender a associação
- Nos dois casos extremos temos:
  - duas variáveis perfeitamente associadas
  - duas variáveis não associadas

# Visualização – Dispersão conjunta “inexistente”



# Visualização – Dispersão amostras independentes



- O DP é uma medida a dispersão de uma variável contínua.
- Existe um análogo para duas variáveis, simultaneamente.

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

# Sumário

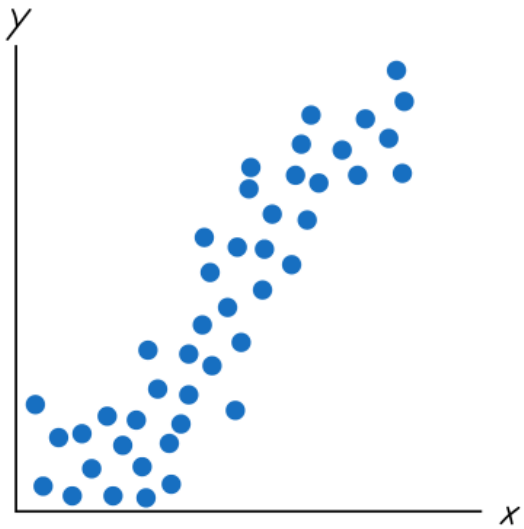
- Considere duas amostras X e Y, de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
  - X: variável independente (ou variável explanatória)
  - Y: variável dependente (ou variável resposta)

- 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?

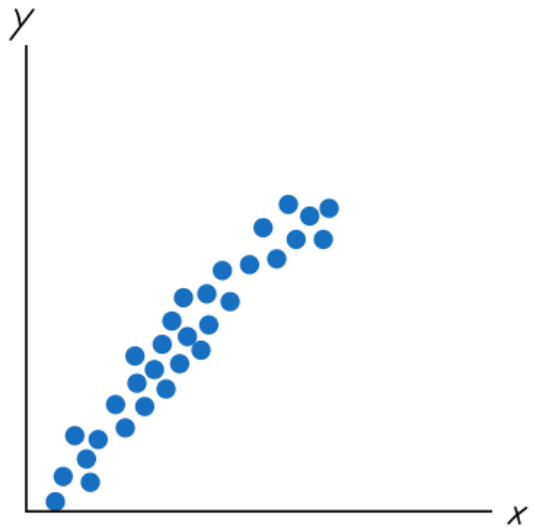
- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

# Exemplo

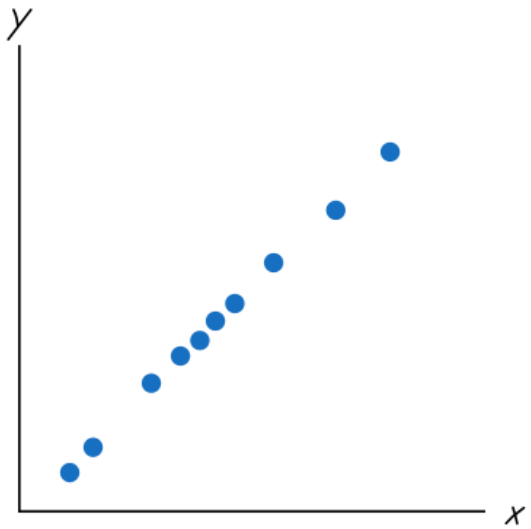





**(a)** Positive correlation between $x$ and $y$

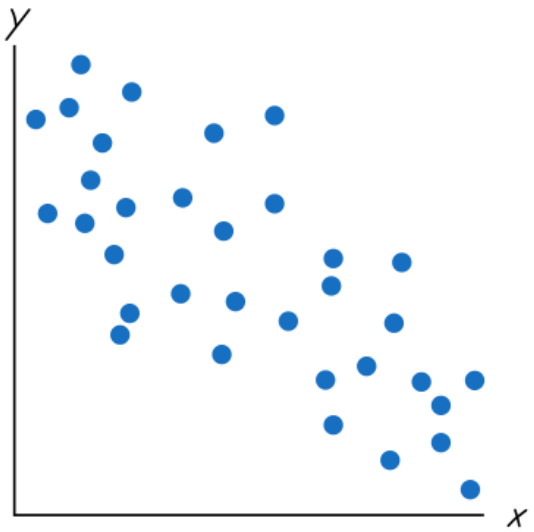



**(b)** Strong positive correlation between $x$ and $y$

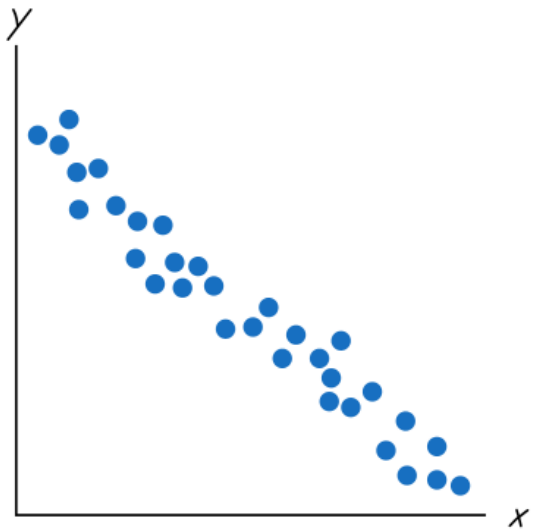



**(c)** Perfect positive correlation between $x$ and $y$

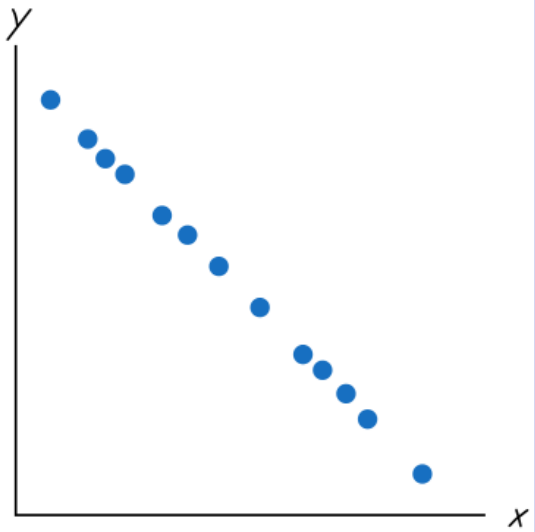
Fonte: Triola, 2004




**(d)** Negative correlation between $x$ and $y$

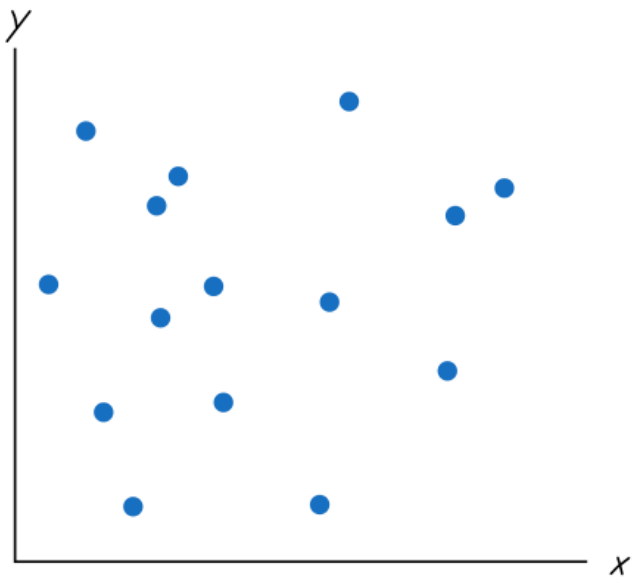



**(e)** Strong negative correlation between $x$ and $y$

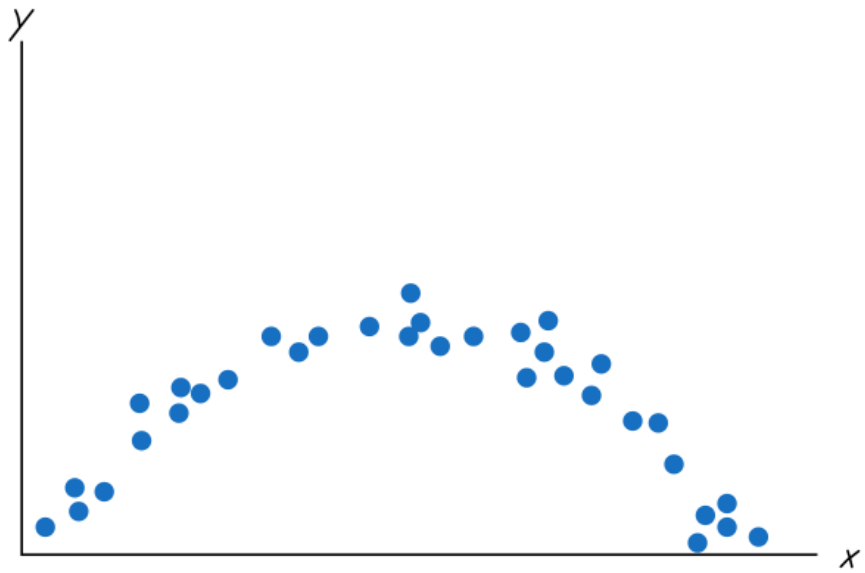



**(f)** Perfect negative correlation between $x$ and $y$

Fonte: Triola, 2004




**(g)** No correlation between $x$ and $y$




**(h)** Nonlinear relationship between $x$ and $y$

Fonte: Triola, 2004

# Sumário

Vamos começar considerando que há associação.

O que é a correlação neste caso?

## Definição

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

- É um número entre $-1$ e 1.
- Mede a associação linear entre duas variáveis.
  - Diretamente proporcional
  - Inversamente proporcional
  - ausência de proporcionalidade

## Interpretação

- 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.

## Interpretação

- 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.

## Interpretação

- 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.

## Interpretação

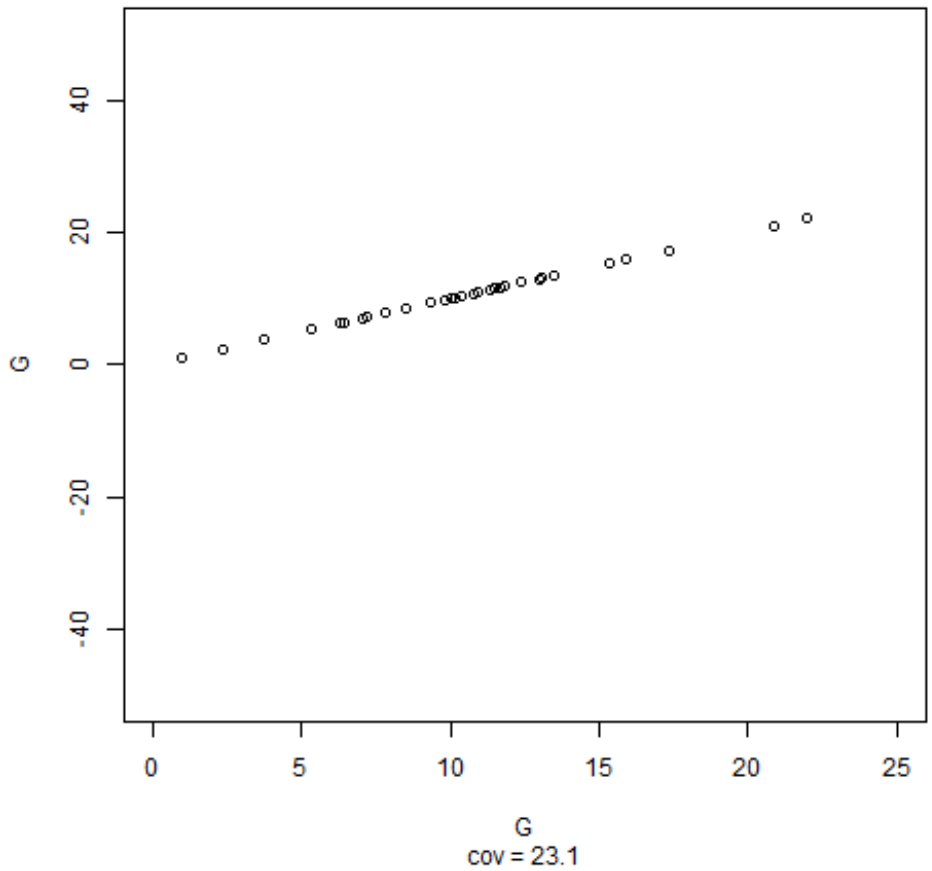
- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

- Se tivéssemos os dados de toda a população, poderíamos calcular o parâmetro $\rho$
- Na prática, só podemos calcular a estatística *r* da amostra
- Utilizamos *r* como estimador para $\rho$, e testamos a significância estatística da forma usual
  **Podemos apresentar o Intervalo de Confiança em torno de** *r*

- Se tivéssemos os dados de toda a população, poderíamos calcular o parâmetro $\rho$
- Na prática, só podemos calcular a estatística *r* da amostra
- Utilizamos *r* como estimador para $\rho$, e testamos a significância estatística da forma usual
  **Podemos apresentar o Intervalo de Confiança em torno de *r***

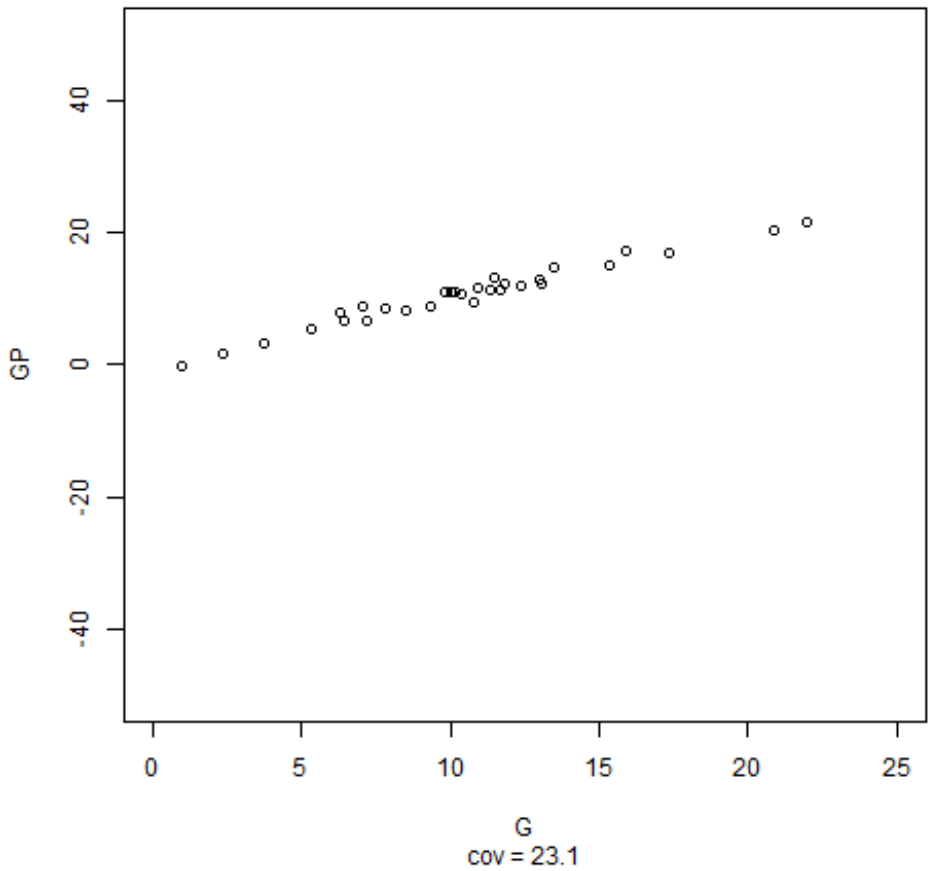
## Associação perfeita

```
Pearson's product-moment correlation

data:  G and G
t = 355106730, 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
```



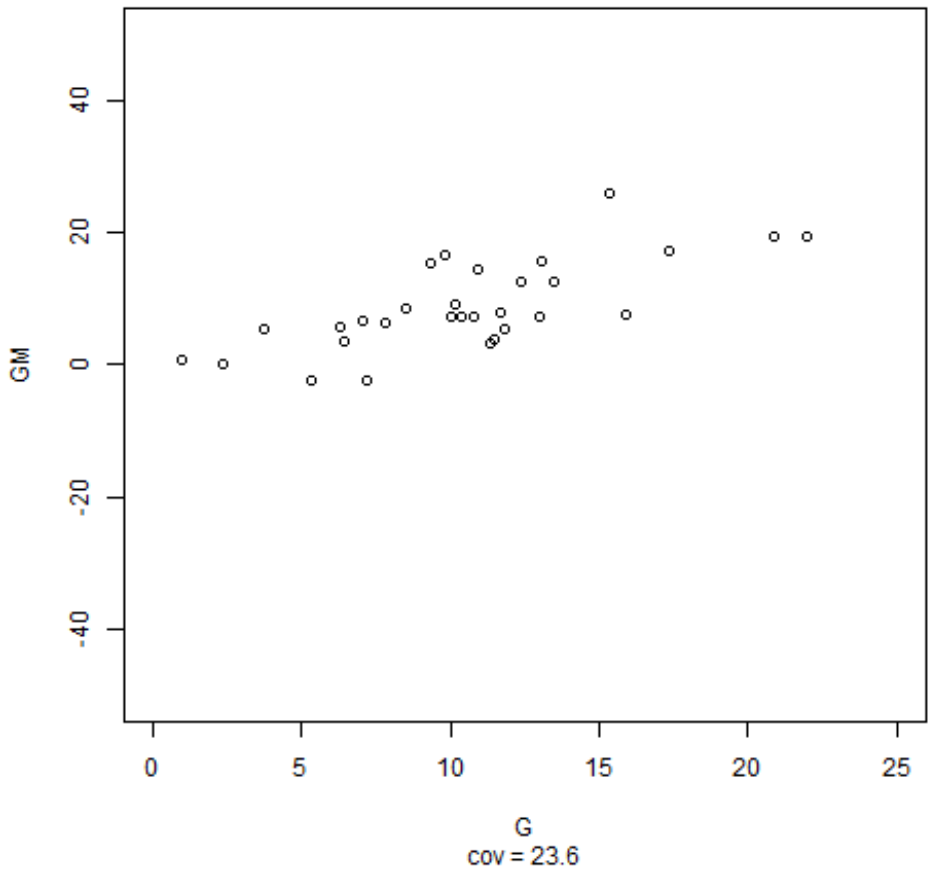
## Associação “forte”

```
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 0.9922253
sample estimates:
      cor
0.9835406
```



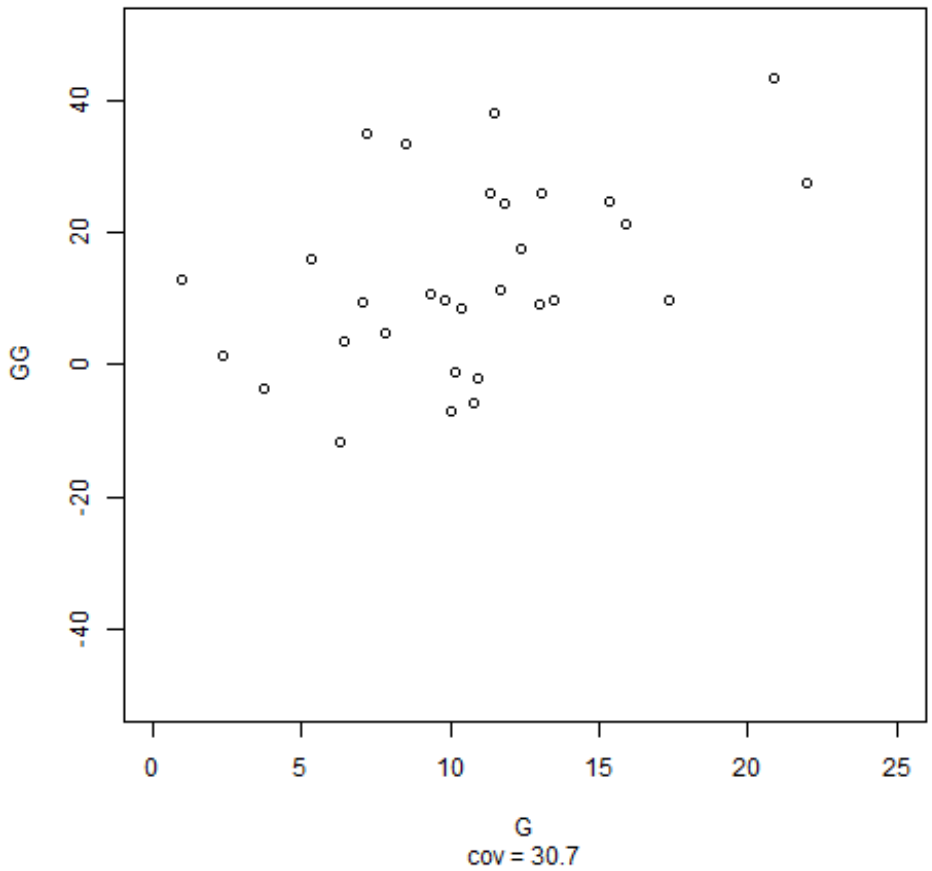
## Associação “moderada”

```
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
```



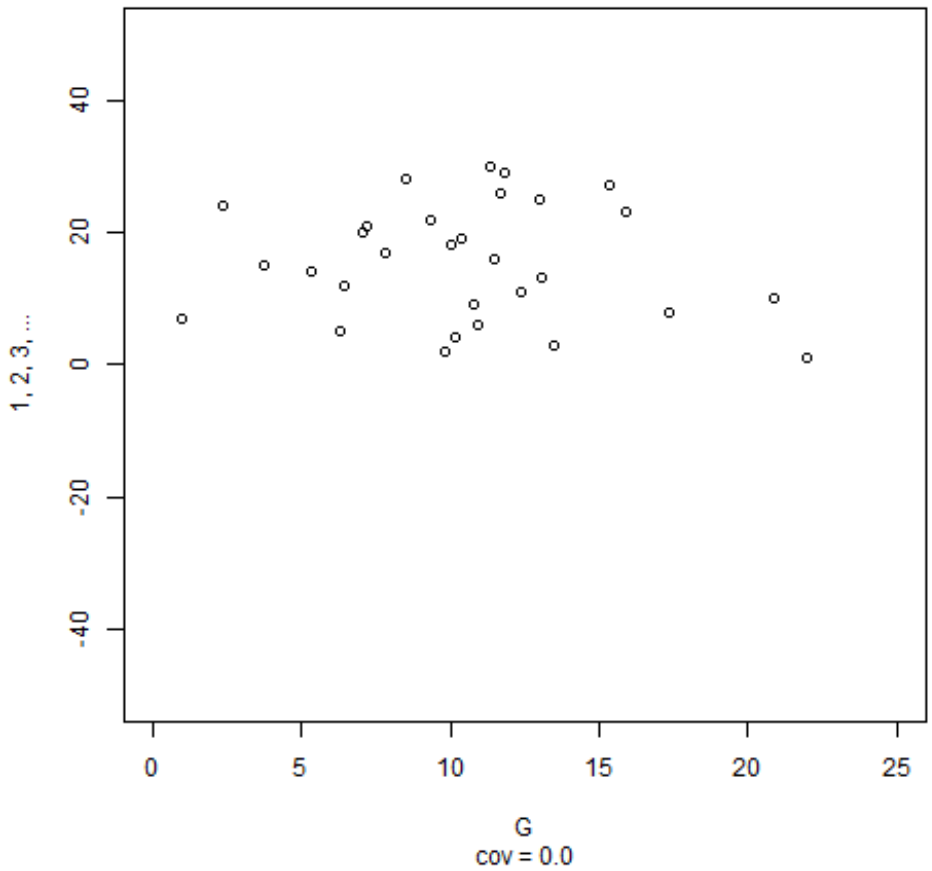
## Associação “fraca”

```
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
```



## Associação inexistente

```
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 17.1

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.

## Exemplo 17.1

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.

## Exemplo 17.1

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.

## Exemplo 17.1

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.

## Quais são as variáveis?

- Qual é a variável independente (X)?
- Qual é a variável dependente (Y)?

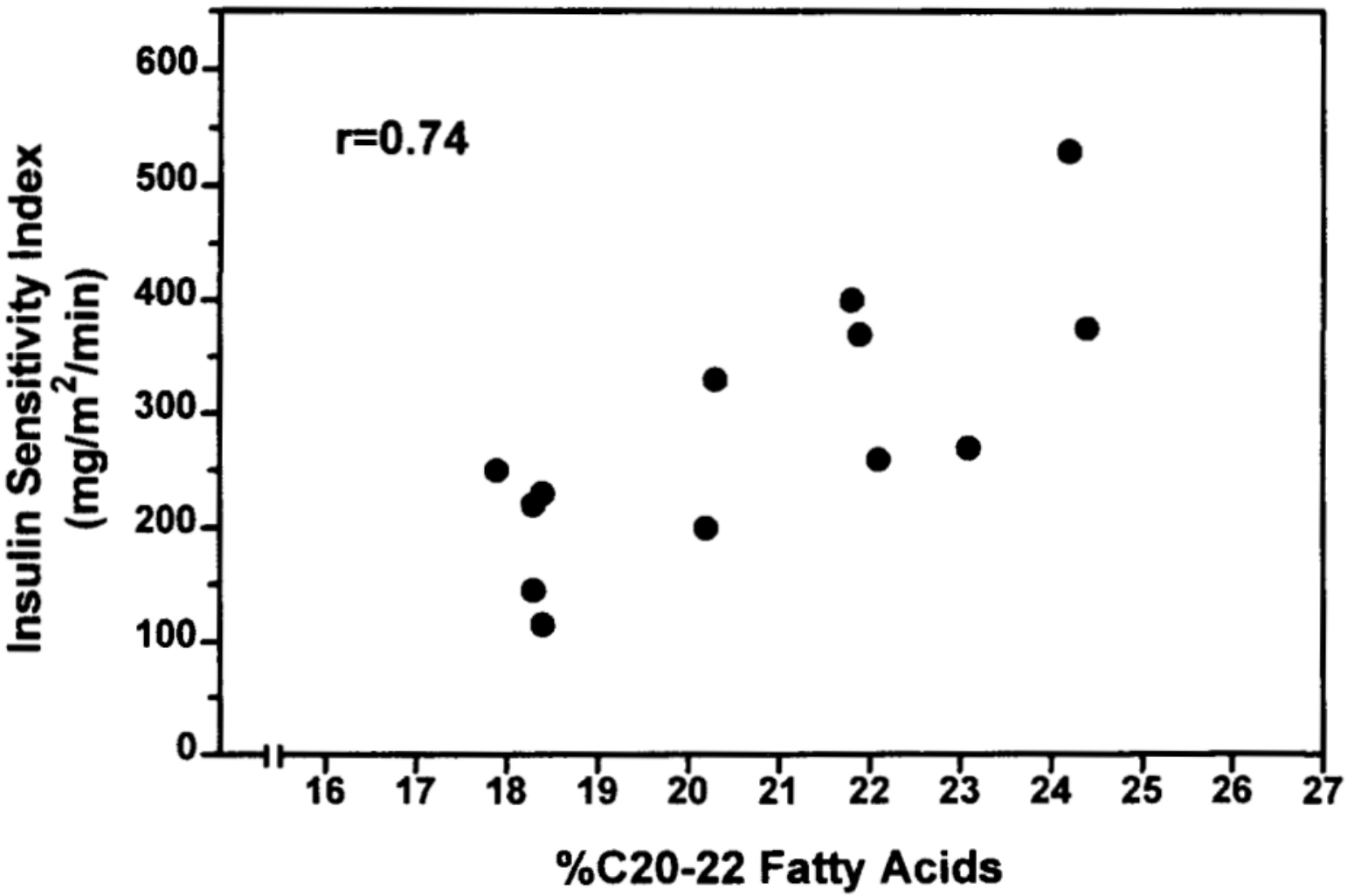
- Dependente: insulina (contínua)
- Independente: conteúdo lipídico (contínua)

**Esta relação pode ser expressa como**

$$\text{insulina} \sim \text{conteúdo lipídico}$$

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

| % C20–22 Polyunsaturated Fatty Acids | Insulin Sensitivity ($mg/m^2/min$) |
|---|---|
| 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 |




Obs: na verdade, $r = 0.77$.

- Tamanho da amostra: $n = 13$
- **Premissa:** ambas variáveis tem o mesmo $n$
- **Premissa:** mensurações vem da mesma população
- **Premissa:** população Normal

$H_0$: Não há relação entre as variáveis na população:

$H_0 : \rho = 0$

**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 |

## Resultados brutos do exemplo

```
Pearson's product-moment correlation

data:  Gordura and Insulina
t = 4.0026, df = 11, p-value = 0.002077
alternative hypothesis: true correlation
 is not equal to 0
95 percent confidence interval:
 0.3804100 0.9274906
sample estimates:
      cor
0.7700025
```

## Resultados

- $r = 0.77,\ p = 0.0021.$
- Interpretação: se não houver relação entre as variáveis ($H_0$), existe apenas 0.21% de chance de observamos uma correlação tão (ou mais) forte com um estudo deste tamanho
- $IC = [0.38, 0.93]$
- Interpretação: (...) temos 95% de confiança que a correlação real está entre 0.38 e 0.93.
- (...) e que ela é positiva!

Por que estas variáveis parecem 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

Por que estas variáveis parecem 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

## Exemplo 17.1



Por que estas variáveis parecem 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

Por que estas variáveis parecem 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

# Mantra



Repita várias vezes mentalmente

## Correlação não implica causalidade

- 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

# Sumário

- Se $r \approx 0$, então X e Y não variam juntos

  (independentes, i.e., não têm associação linear)

- Se $r > 0$, então quando uma aumenta, a outra aumenta em proporção direta

  (associação linear positiva)

- Se $r < 0$, então quando uma aumenta, a outra diminui em proporção inversa

  (associação linear negativa)

Porém...

## Exemplo

- Em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita

## Exemplo – correlação sem causalidade

- 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:**

## Exemplo – correlação sem causalidade

- 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

## Cuidado!

Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável

# Sumário

- 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.

- 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).

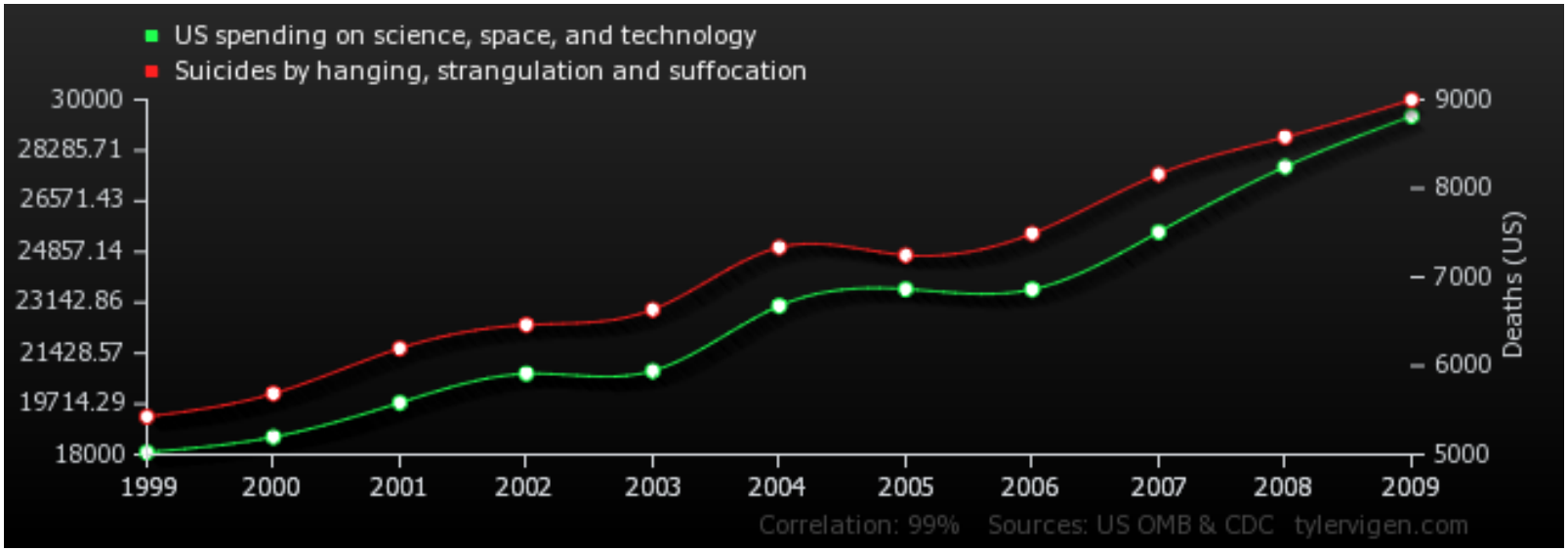
# Mantra



**Repita várias vezes mentalmente**

## Correlação não implica causalidade

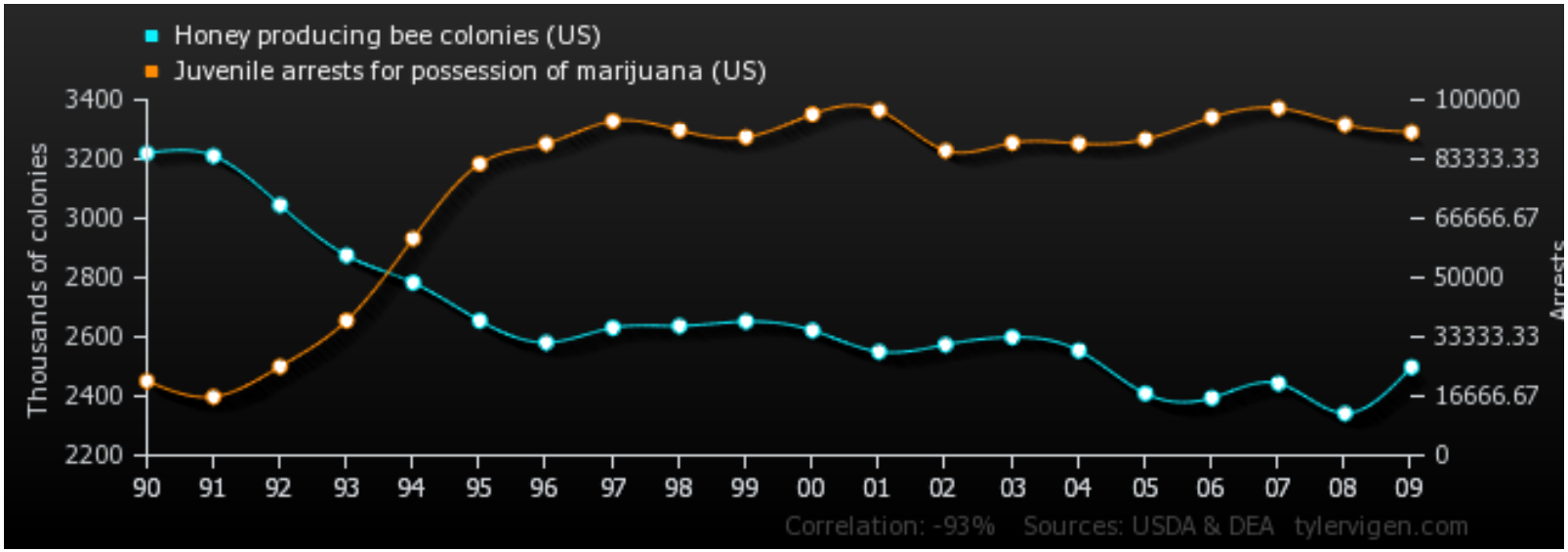
Gasto com C&T (EUA) x Suicídios por enforcamento




Correlação: 0.992082

Fonte: Spurious correlations

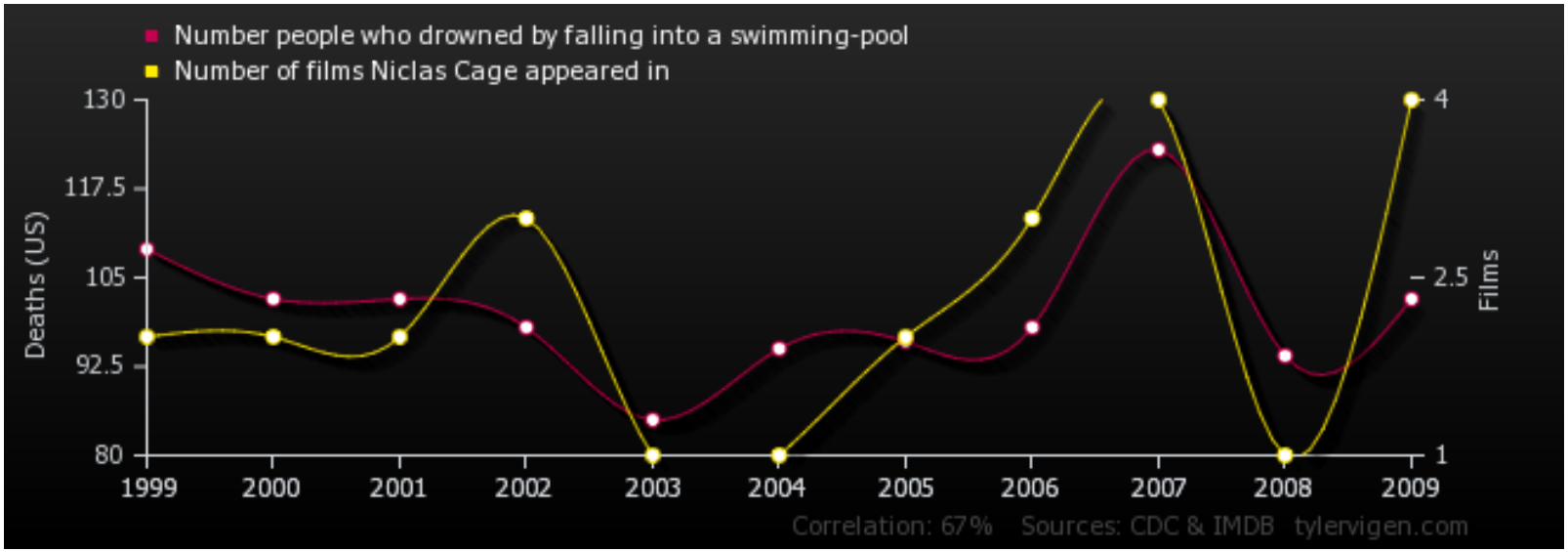
Produção de mel x Prisões por posse de maconha




Correlação: -0.933389

Fonte: Spurious correlations

Afogamentos em piscina x Filmes com Nicholas Cage




Correlação: 0.666004

Fonte: Spurious correlations

# Mantra



**Repita várias vezes mentalmente**

## Correlação não implica causalidade

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?
   (variável de confundimento)
4. É possível que a relação entre duas variáveis seja uma coincidência?
   (erro tipo I)

Estas perguntas estão fora do escopo da Bioestatística!

Só o especialista pode investigar (e discutir) estas possibilidades.

# Sumário

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
  (pense nas quatro perguntas anteriores)

# Sumário

## Leitura obrigatória

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

## Leitura recomendada

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