Probabilidade e Estatística:

Correlação!

python™

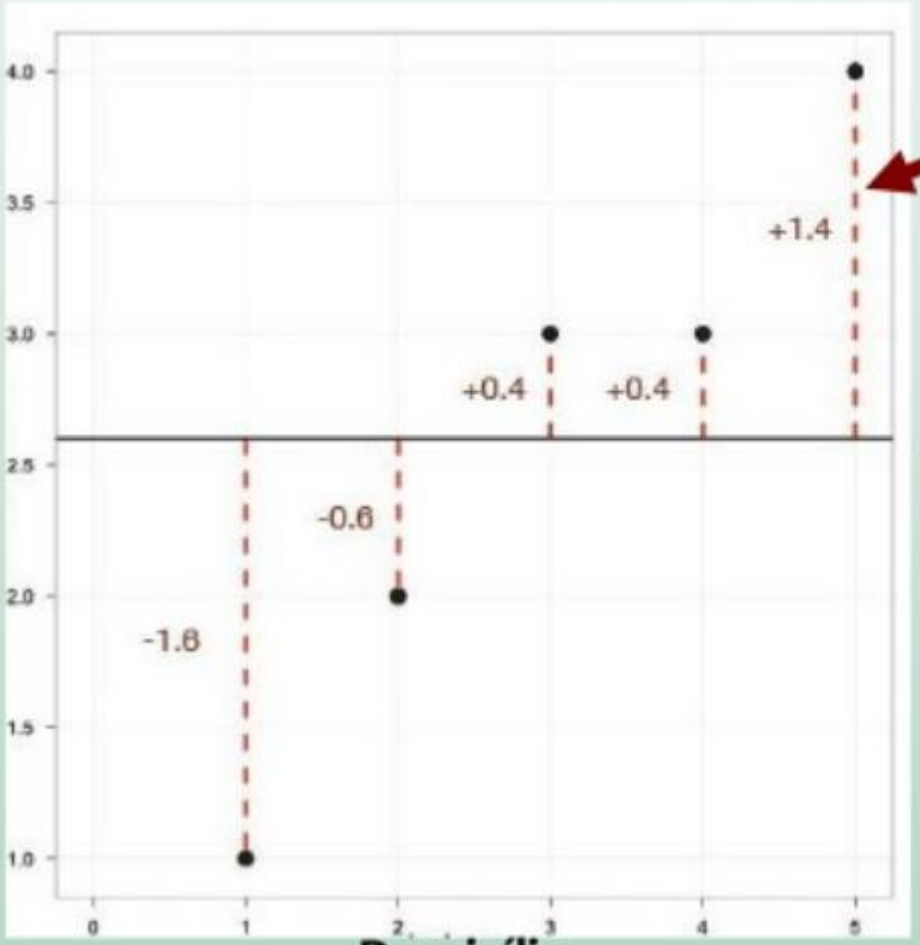
Média

Uma representação simplificada de uma característica do mundo real:

- *A média do consumo per capita de água na Região Sudeste*
- *A altura média dos edifícios em São Caetano*
- *O PIB médio dos municípios localizados no arco do desmatamento*

Modelo preciso?

O quão diferente nossos dados reais são do modelo criado?



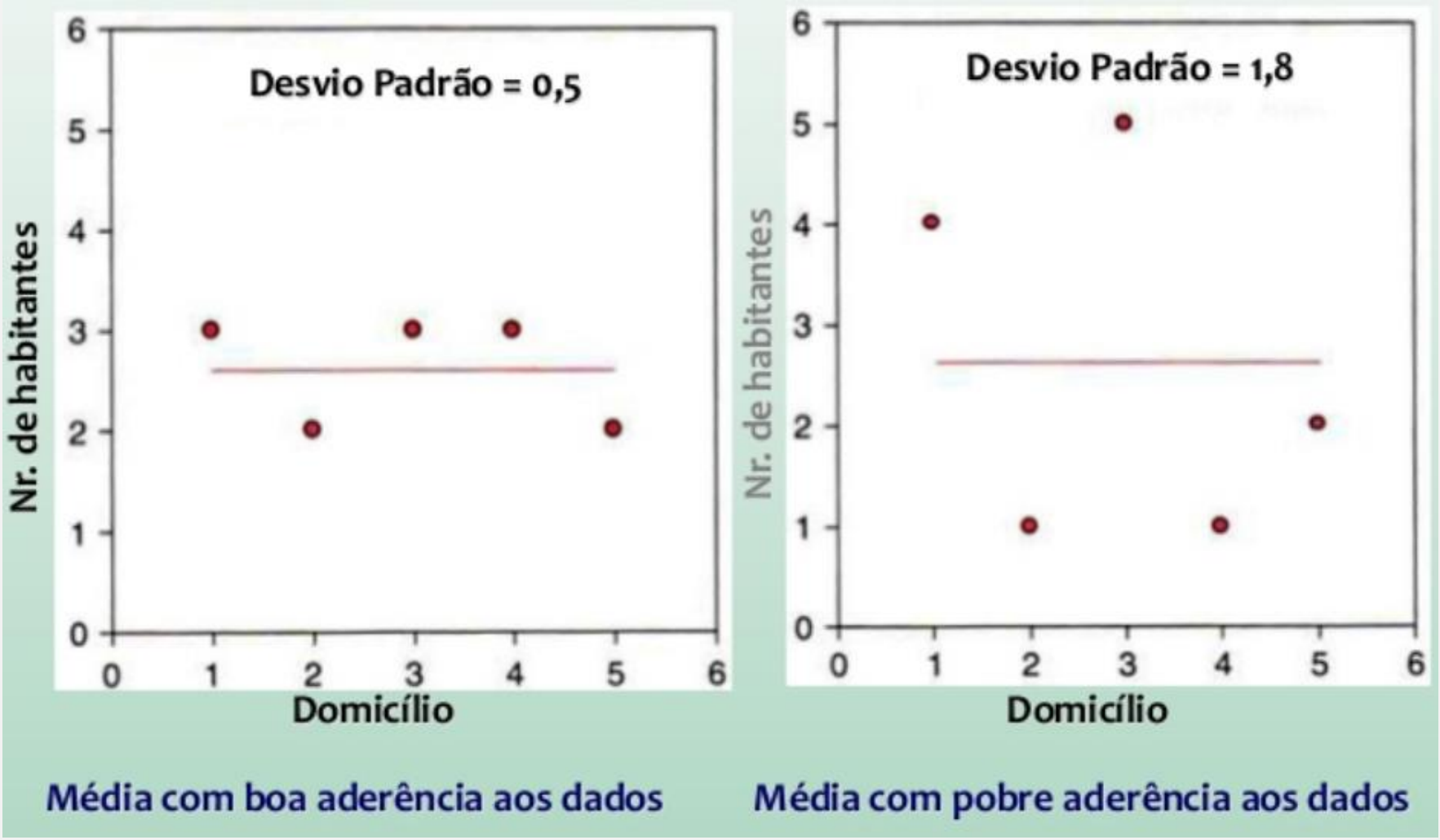
Desvios
(erro do modelo)

Média (2,6)

Conceitos:

- Variância
- Desvio Padrão

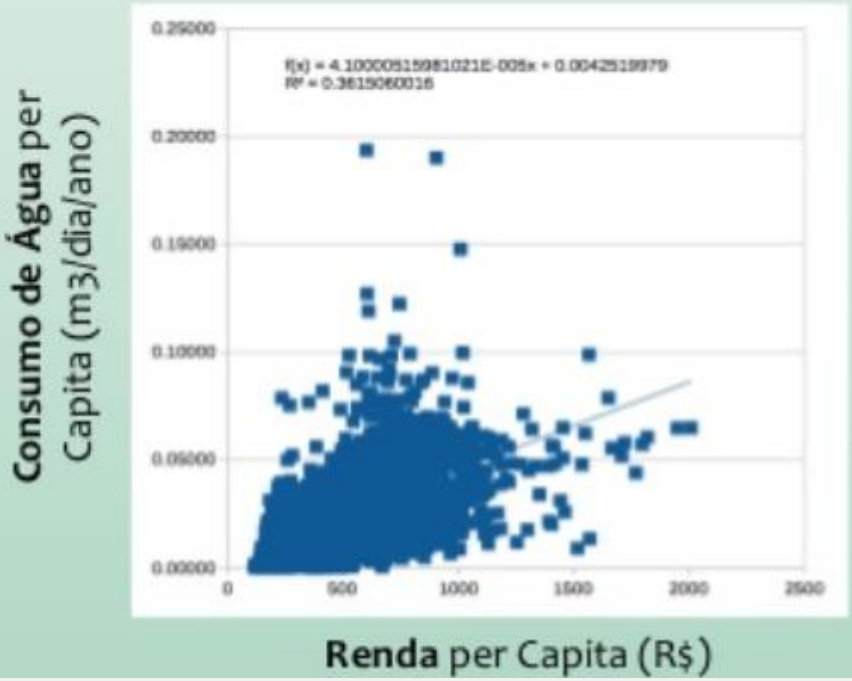
Médias iguais, desvio padrão diferente!



Modelos Lineares

- São modelos baseados sobre uma linha reta, utilizados para representar a relação entre variáveis
- Ou seja, geralmente estamos tentando resumir as **RELAÇÕES** observadas a partir de nossos dados observados em termos de uma linha reta.

RELAÇÃO ENTRE CONSUMO DE ÁGUA E RENDA



Correlação

É uma medida do relacionamento linear entre duas variáveis

Duas variáveis podem estar:

(a) Positivamente relacionadas → quando maior a renda, maior o consumo de água

(b) Negativamente relacionadas → quanto maior a renda, menor o consumo de água

(c) Não há relação entre as variáveis

Diagrama de Dispersão

Representando Relacionamentos Graficamente

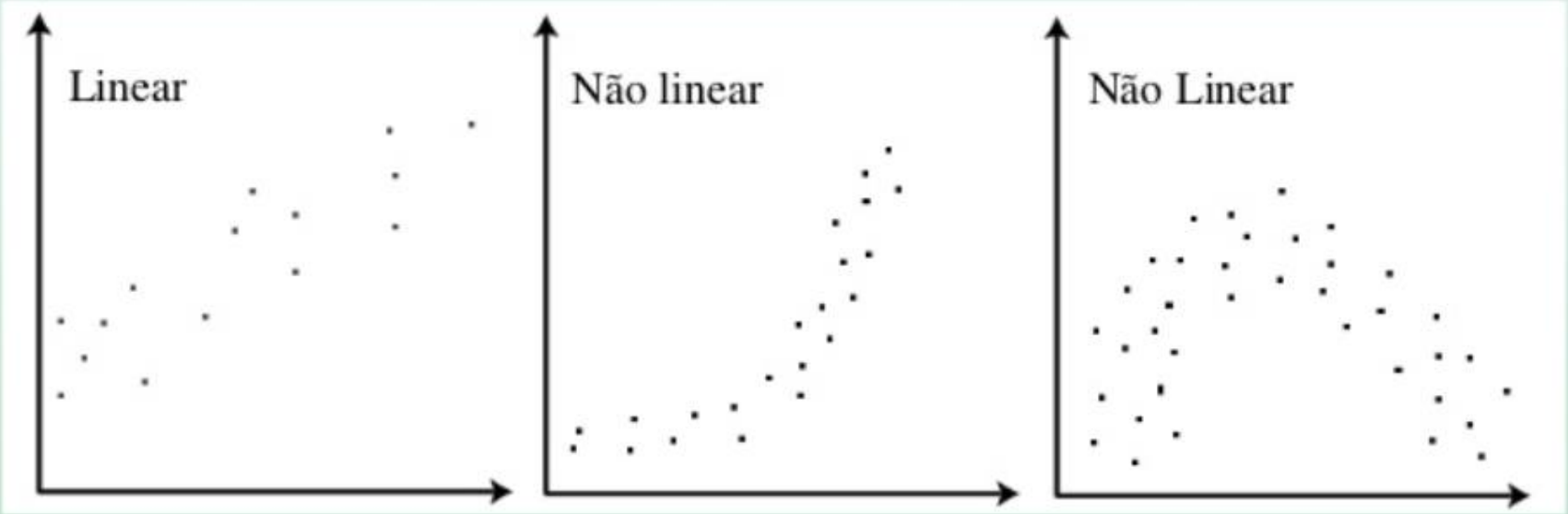
DIAGRAMA DE DISPERSÃO: Gráfico que coloca o escore de cada observação em uma variável contra seu escore em outra

Diagramas de dispersão que mostram correlação positiva entre as variaveis

Diagrama de Dispersão

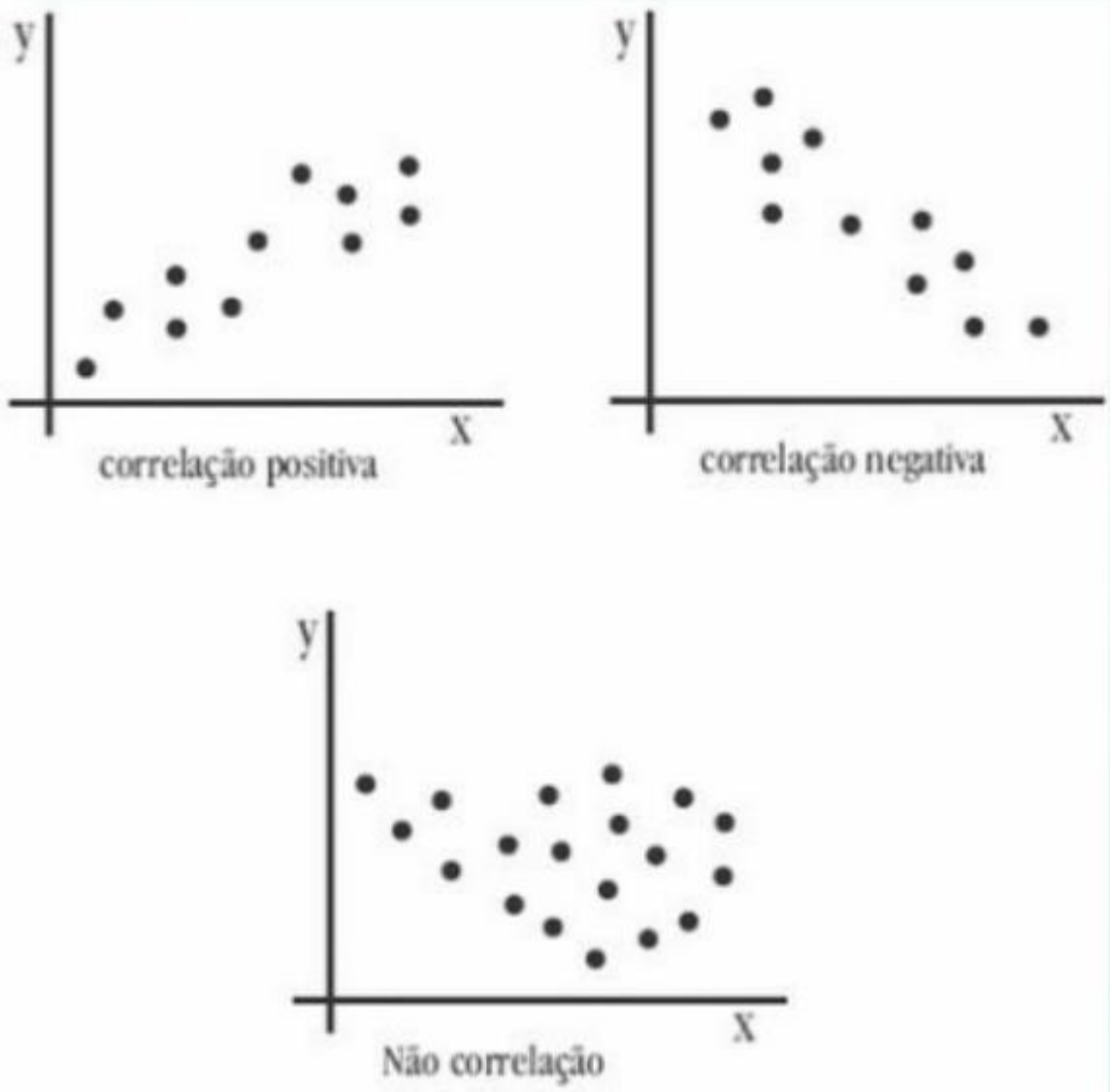
Importante começar por ele!

Nos diz se a relação entre variáveis é linear, se existem peculiaridades nos dados que valem a pena observar (outliers) e dá uma ideia da força do relacionamento entre as variáveis.



Exemplo de Correlação Não-Linear: Renda e proporção de domicílios próprios.

Diagrama de Dispersão



Como medimos relacionamentos?

Veremos duas medidas para expressar estatisticamente os relacionamentos entre variáveis:

1. Covariância

2. Coeficientes de correlação

Covariância

Uma maneira de verificar de duas variáveis estão associadas é ver se elas **variam** conjuntamente. Ou seja, ver se as mudanças em uma variável correspondem a mudanças similares na outra variável

RELEMBRANDO O CONCEITO DE VARIÂNCIA:

$$Variância = s^2 = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{\sum(x_i - \bar{x})\,(x_i - \bar{x})}{N-1}$$

Como medimos relacionamentos?

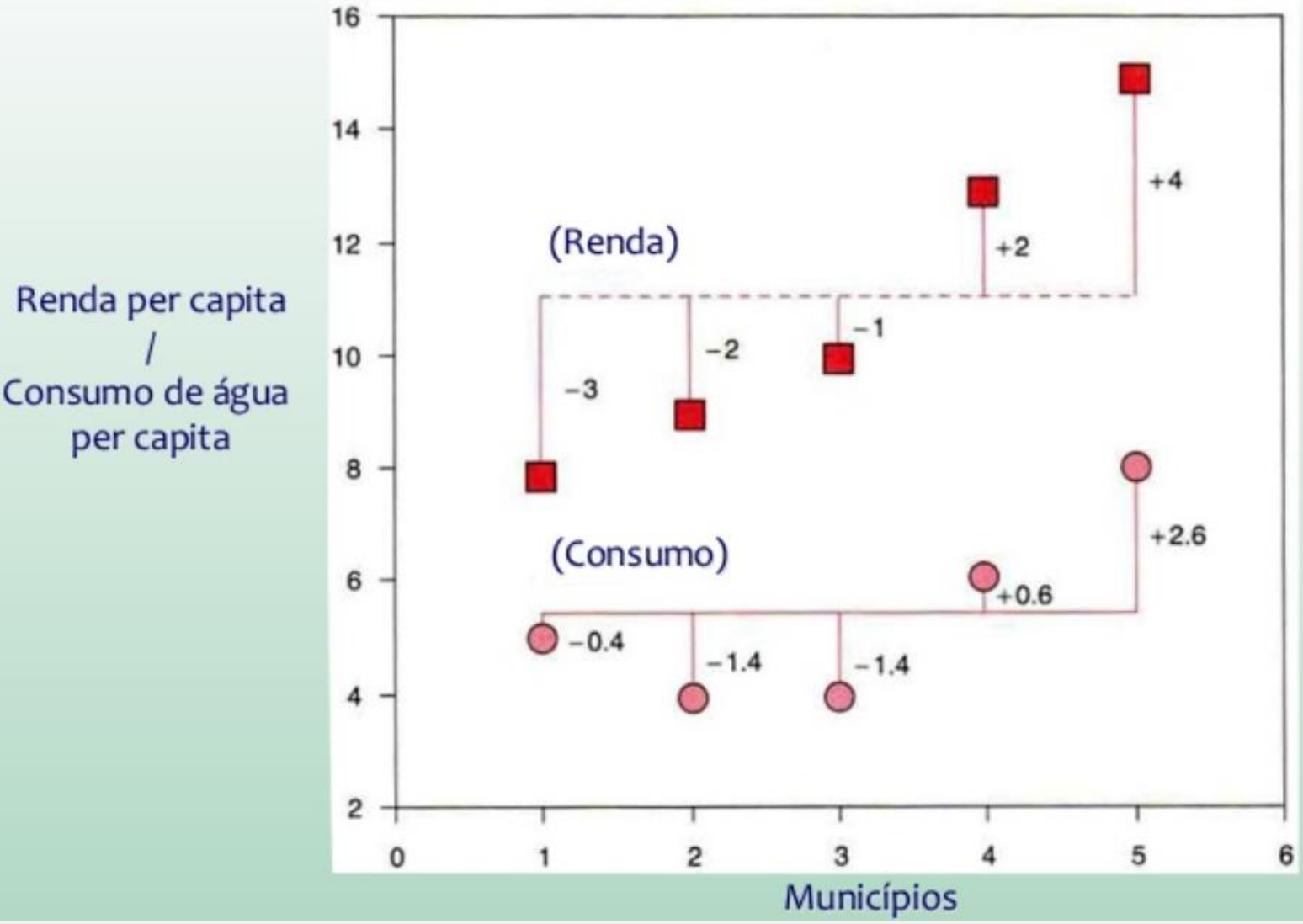
Em outras palavras:

Quando uma variável se desvia de sua média, esperamos que a outra variável se desvie da sua média de maneira similar (ou de maneira diretamente oposta).

RELEMBRANDO O CONCEITO DE VARIÂNCIA:

$$Variância = s^2 = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{\sum(x_i - \bar{x})\,(x_i - \bar{x})}{N-1}$$

Padrão similar nas diferenças de ambas as variáveis!



python™

Como calcular a semelhança entre o padrão das diferenças das 2 variáveis?

Multiplicando a diferença de uma variável pela diferença correspondente da segunda variável!

- Se ambos os erros são positivos ou negativos, isso nos dará um valor positivo (desvios na mesma direção)
- Se um erro for positivo e outro negativo, isso nos dará um valor negativo (desvios em direções opostas)

COVARIÂNCIA

$$cov(x,y) = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{N-1}$$

Covariância

Média das Diferenças Combinadas

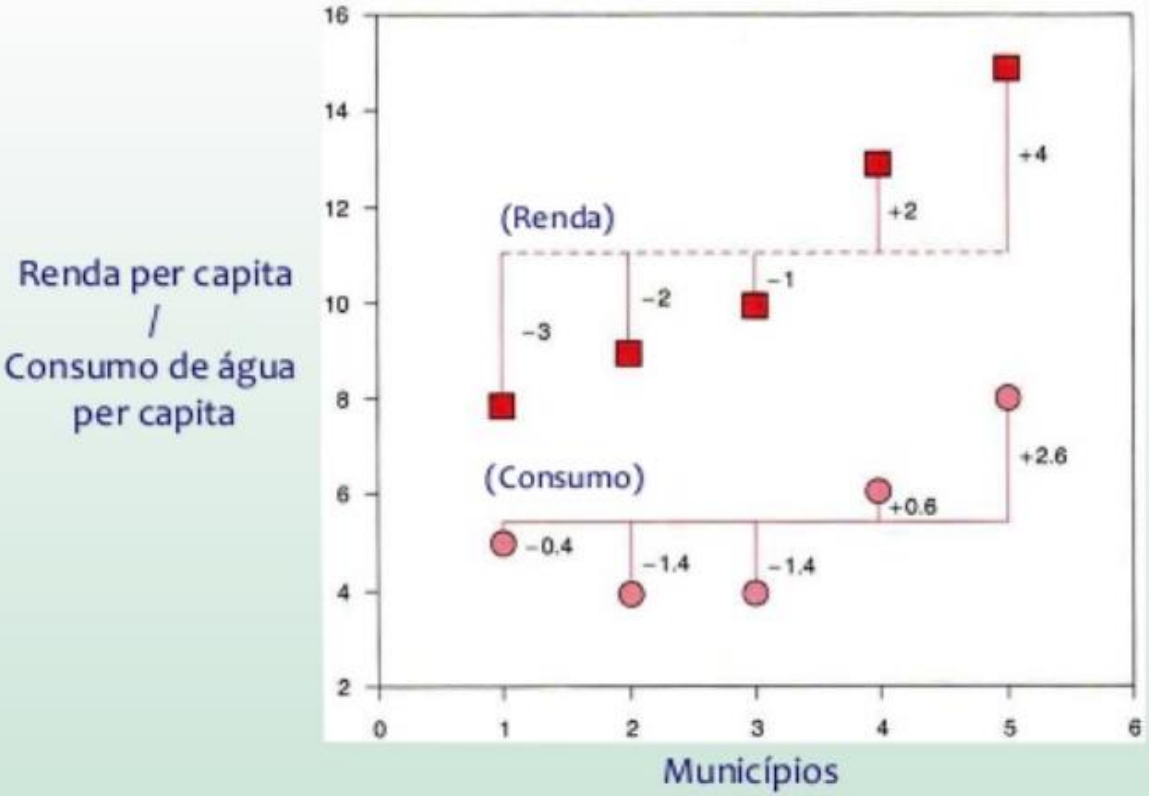
É uma medida de como duas variáveis variam conjuntamente.

Se a covariância entre duas variáveis é igual a zero, significa que elas são **independentes**.

COVARIÂNCIA

$$cov(x,y) = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{N-1}$$

Covariância



$$cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$= \frac{(-0{,}4)(-3) + (-1{,}4)(-2) + (-1{,}4)(-1) + (0{,}6)(2) + (2{,}6)(4)}{4}$$

$$= \frac{1{,}2 + 2{,}8 + 1{,}4 + 1{,}2 + 10{,}4}{4}$$

$$= \frac{17}{4} = 4{,}25$$

Covariância

Covariância Positiva: Quando uma variável se desvia da média, a outra variável se desvia na mesma direção.

Covariância Negativa: Quando uma variável se desvia da média, a outra variável se desvia na direção oposta.

COVARIÂNCIA

$$cov(x,y) = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{N-1}$$

Covariância

UM PROBLEMA!!!

A covariância depende das escalas de medida. Não é uma medida padronizada.

Ou seja, não podemos dizer se a covariância é particularmente grande ou pequena em relação a outro conjunto de dados a não ser que ambos os conjuntos fossem mensurados nas mesmas medidas.

Coeficiente de Correlação de PEARSON

O **COEFICIENTE DE CORRELAÇÃO**

é uma covariância padronizada

COEFICIENTE DE CORRELAÇÃO DE PEARSON

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{(N-1)s_x s_y}$$

Coeficiente de Correlação de PEARSON

Padronizando a covariância, encontramos um valor que deve estar entre -1 e +1

r = +1 → duas variáveis estão perfeitamente correlacionadas de forma positiva (se uma aumenta, a outra aumenta proporcionalmente)

r = -1 → relacionamento negativo perfeito (se uma aumenta, a outra diminui em valor proporcional

r = 0 → indica ausência de relacionamento linear

Interpretando r

- 0.9 para mais ou para menos indica uma correlação muito forte.
- 0.7 a 0.9 positivo ou negativo indica uma correlação forte.
- 0.5 a 0.7 positivo ou negativo indica uma correlação moderada.
- 0.3 a 0.5 positivo ou negativo indica uma correlação fraca.
- 0 a 0.3 positivo ou negativo indica uma correlação desprezível.

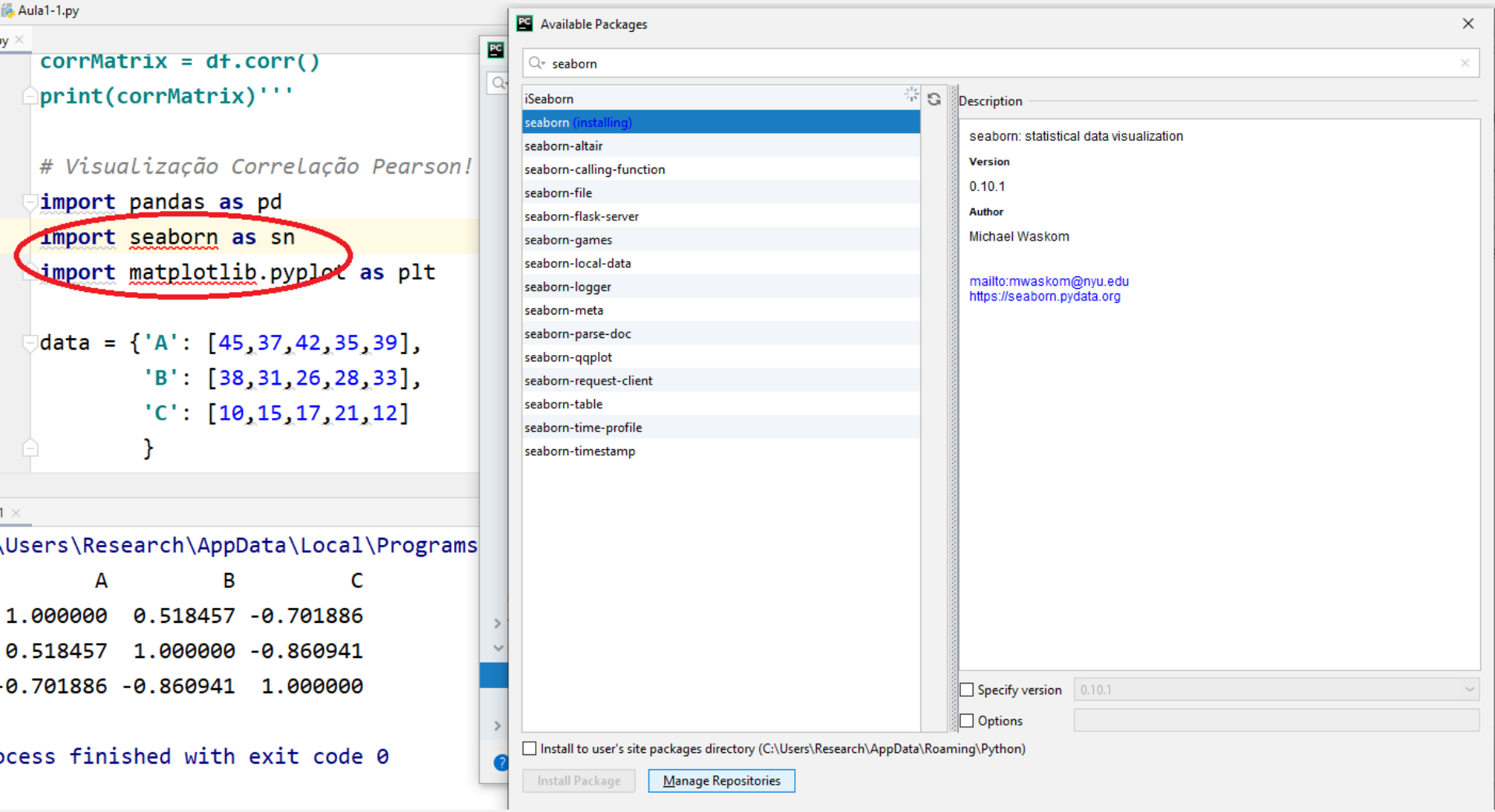
Coeficiente de Correlação de Pearson

```
9   # Correlação Pearson!
0   import pandas as pd
1   data = {'A': [45,37,42,35,39], 'B': [38,31,26,28,33], 'C': [10,15,17,21,12]}
2   df = pd.DataFrame(data,columns=['A','B','C'])
3   corrMatrix = df.corr()
4   print(corrMatrix)
```

Aula1-1

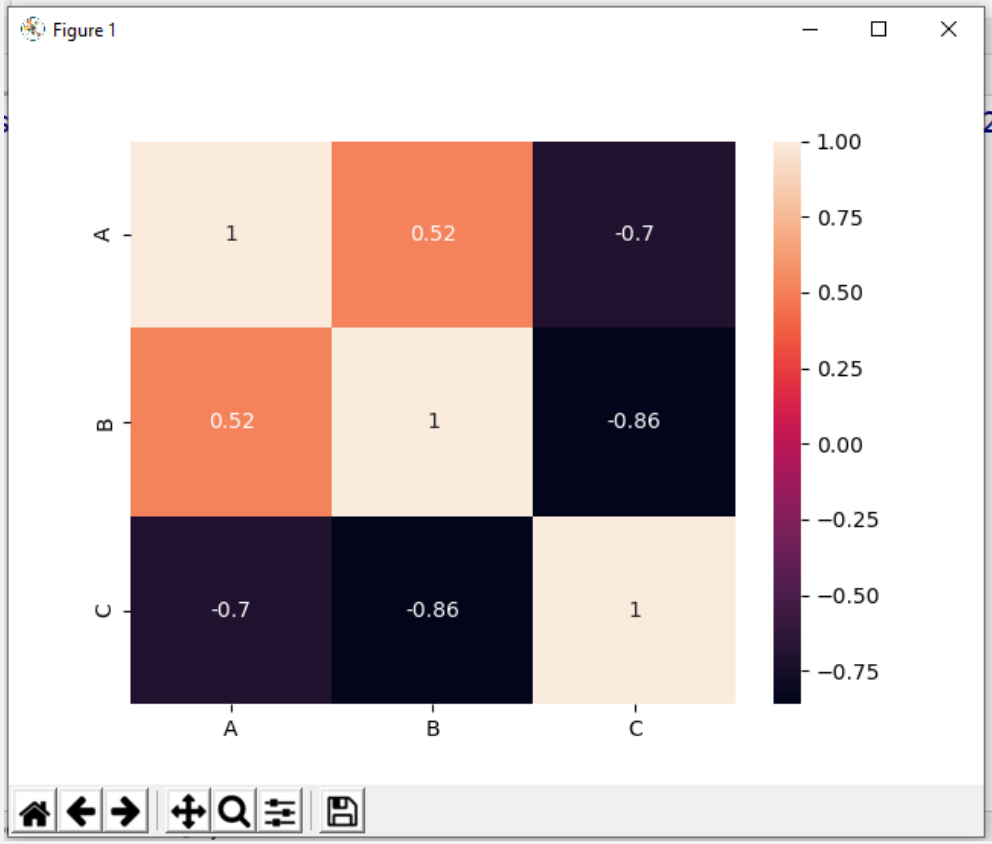
```
C:\Users\Research\AppData\Local\Programs\Python\Python38-32\python.exe "D:/IESB/Pr
          A         B         C
A  1.000000  0.518457 -0.701886
B  0.518457  1.000000 -0.860941
C -0.701886 -0.860941  1.000000
```

Bibliotecas para visualizar!



Coeficiente de Correlação de Pearson

```python
# Visualização Correlação Pearson!
import pandas as pd
import seaborn as sn
import matplotlib.pyplot as plt
data = {'A': [45,37,42,35,39], 'B': [38,31,26,28,33], 'C': [10,15,17,21,12] }
df = pd.DataFrame(data,columns=['A','B','C'])
corrMatrix = df.corr()
sn.heatmap(corrMatrix, annot=True)
plt.show()
```




```
2\python.exe "D:/IESB/Prol
```

python™

Lista de Exercícios!