Probabilidade e Estatística

Aula 3

Medidas Numéricas Descritivas

Leitura: Levine et al. Capítulo 3

Objetivos

Nesta parte, aprenderemos:

- a descrever as propriedades de tendência central, variação e formato em dados numéricos
- a calcular medidas resumo para a população
- a construir e interpretar um Box- plot
- a descrever a covariância e o coeficiente de correlação

Exemplo

- Uma pesquisa em uma certa cidade perguntou a 15 pessoas, escolhidas aleatoriamente, o tempo de viagem de casa para o trabalho em minutos:

  30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

- Em rol:

  5 10 10 10 10 12 15 20 20 25 30 30 40 40 60

Exemplo

- Diagrama ramo-e-folha:

- Distribuição assimétrica
- Maior tempo de viagem = 60 min

0	5
1	000025
2	005
3	00
4	00
5	
6	0

- Vamos aprender a descrever, com números, o centro e a dispersão das distribuções de dados!!

Definições

- A **tendência central** corresponde à extensão na qual todos os valores de dados se agrupam em torno de um valor central típico.
- A **variação** corresponde ao montante de dispersão, ou espalhamento, de valores em relação a um valor central.
- O **formato** corresponde ao padrão da distribuição de valores do valor mais baixo para o mais alto.

Medidas de Tendência Central

- **tendência central:** valores no centro da distribuição, em torno dos quais os dados se agrupam.

- Medidas tipicamente usadas:
  - Média aritmética
  - Mediana
  - Moda

Média

- A média aritmética (média) é a mais comum das medidas de tendência central.

Para uma amostra de tamanho n:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Tamanho da amostra

Valores observados

Exemplo: Média

- A tabela abaixo lista o tempo de viagem de casa para o trabalho de 15 pessoas em minutos:

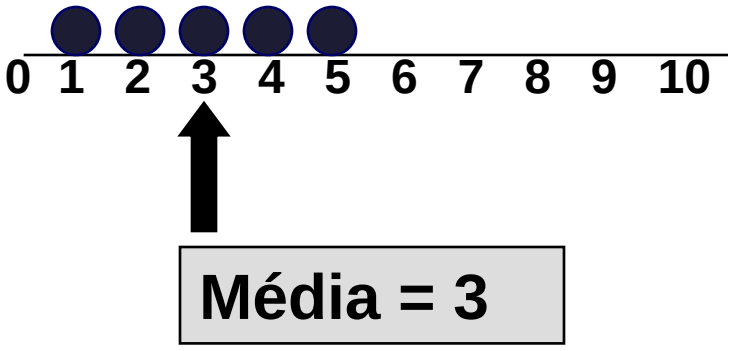
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
30	20	10	40	25	20	10	60	15	40	5	30	12	10	10

- O tempo médio de viagem das pessoas é:

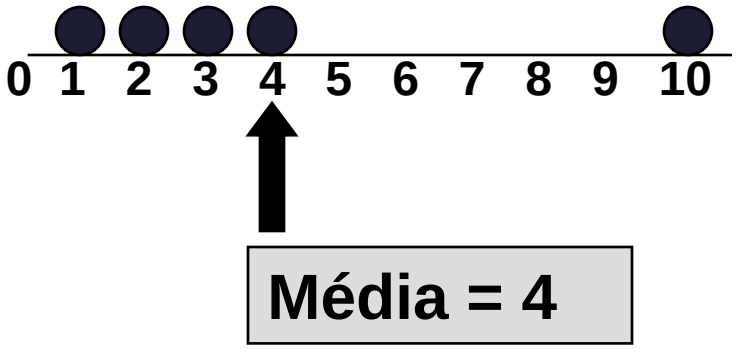
$$\bar{x} = \frac{\sum_{i=1}^{15} x_i}{n} = \frac{30 + 20 + \cdots + 10}{15} = \frac{337}{15} = 22.5\ min$$

Média

- Média = soma dos valores dividido pelo número de valores
- Afetada por valores atípicos, também chamados de valores extremos ou outliers.



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

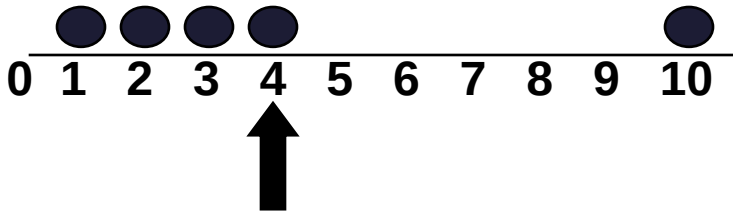


$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Média:o ponto de equilíbrio

- A média é « Ponto de equilíbrio » em um conjunto de dados (gangorra),

  onde todos os valores desempenham um papel igual (mesma massa).



Média = 4

$$(1-4)+(2-4)+(3-4)+(4-4)+(10-4)=0$$

Média:o ponto de equilíbrio

- **Propriedade**: A soma dos desvios em relação a média é nula.

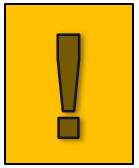
$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X}$$

$$= \sum_{i=1}^{n} X_i - n\bar{X} = 0$$

Desvio de X_i em relação a média $\bar{X}$:

o desvio mede a "distância" entre o valor e a média: tem sinal "-" para valores abaixo da média e "+" para valores acima da média.

Média



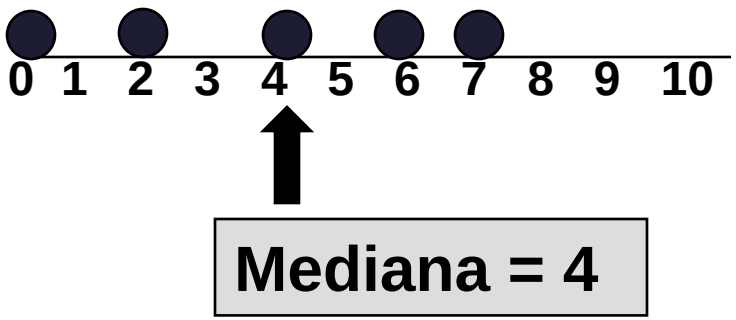
- **Propriedade**: é o valor que minimiza a soma do quadrado dos desvios:

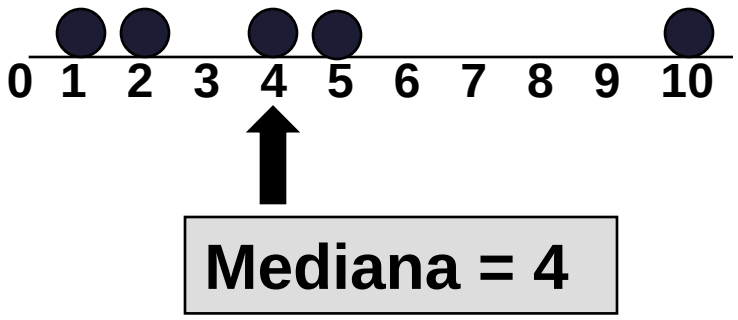
$$\bar{X} = arg \min_{c} \sum_{i=1}^{n} (X_i - c)^2$$

Ou seja, imagine que estamos buscando o valor c (que vamos considerar como "centro" dos dados), tais que a "distância" (medida pela soma dos quadrados dos desvios) dos outros valores em relação a c seja a menor possível. Este valor c sempre é a média!

Mediana

- Em **um rol** (lista dos dados em ordem crescente), a mediana é o "número" do meio, (50% acima, 50% abaixo)





- Não é afetada por valores atípicos (extremos)

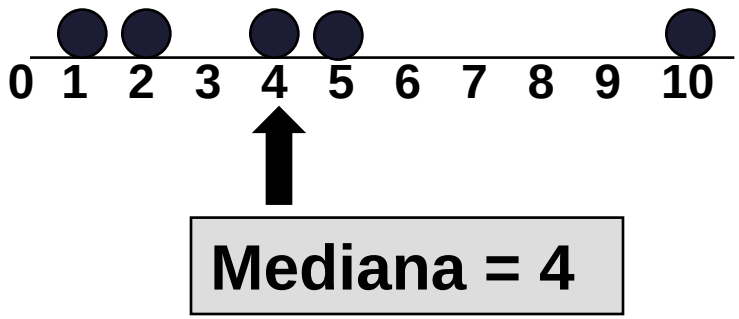
Localizando a Mediana

- A mediana de um **conjunto de dados ordenados** é *localizada* na posição: $\frac{n+1}{2}$.

  - Se o número de valores é *ímpar*, $\frac{n+1}{2}$ é inteiro. Então, a mediana é o número do meio.
  - Se o número de valores é *par*, $\frac{n+1}{2}$ não é inteiro. Então, adotamos a convenção de que mediana é a média dos dois valores do meio.

A Mediana

- Em um rol, a mediana é o “número” do meio, (50% acima, 50% abaixo)

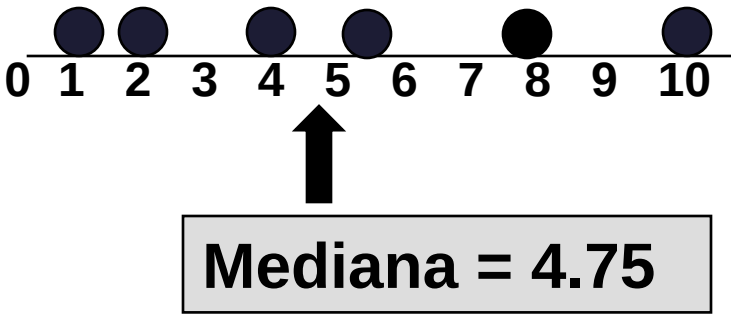


- Exemplo com n=5 (número ímpar).
- Posição da mediana = (5+1)/2=3.
- Então, a mediana é o 3º valor no rol, ou seja,

  mediana=4

A Mediana

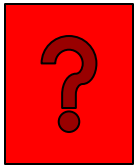
- Em um rol, a mediana é o "número" do meio, (50% acima, 50% abaixo)
- Exemplo: valores dos dados são: 1.1, 2.1, 4, 5.5, 7.9, 10



- Exemplo com $n = 6$ (número par)
- Posição da mediana = (6+1)/2=3.5, entre o 3º e o 4º valor no rol:

$$mediana = \frac{3º\ valor\ + 4º\ valor}{2} = \frac{4 + 5.5}{2} = 4.75$$

Exercício: Mediana



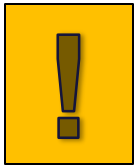
Exercício: Determine o tempo **mediano** de viagem de casa para o trabalho para as pessoas da cidade.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
30	20	10	40	25	20	10	60	15	40	5	30	12	10	10

Solução:

- O tamanho da amostra é: $n = 15$ (ímpar)
- Posição da mediana: $\frac{n+1}{2} = \frac{15+1}{2} = 8$
- A mediana é o 8º valor no rol!
- Para estes dados o rol é: 5,10,10,10,10,12,15,20,20,25,30,30,40,40,60
- Então a mediana é: 20 minutos.

Mediana



- **Propriedade**: A mediana é o valor que minimiza a soma do valor das distâncias (valor absoluto dos desvios):

$$\boldsymbol{Mediana} = arg \min_{c} \sum_{i=1}^{n} |X_i - c|$$

Ou seja, imagine que estamos buscando o valor c (que vamos considerar como "centro" dos dados), tais que a distância dos outros valores em relação a c seja a menor possível. Este valor c sempre é a mediana!

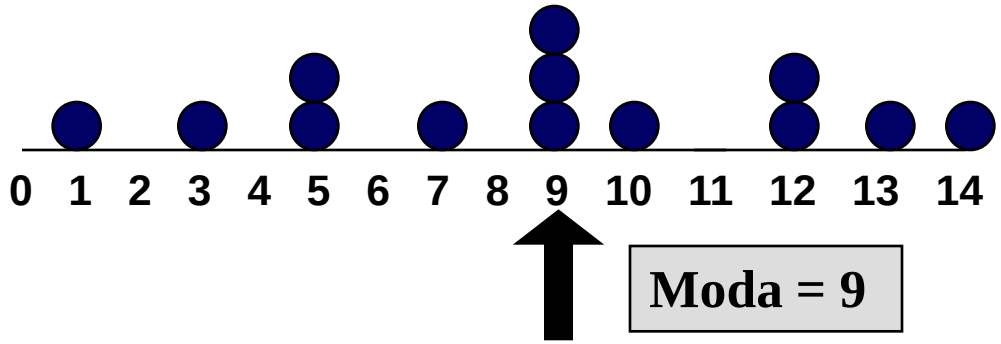
Média x Mediana

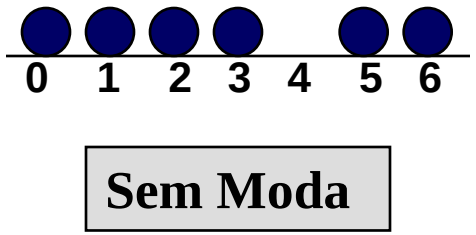
- Vimos que a média é afetada por valores extremos, enquanto a mediana é robusta a valores extremos.

- Para visualizar melhor esta diferença de comportamento entre média e mediana, vejam o applet « Mean and Median ».
  - Exercício: Escolher 9 pontos ao acaso no aplicativo. Agora adicione um 10º ponto bem afastado dos demais. O que vc espera que aconteça com a média? E com a mediana?
  - Exercício: Escolher 5 pontos ao acaso no aplicativo. Agora tente acrescentar pontos de forma que a média e a mediana coincidam.

Medidas de Tendência Central: a moda

- A moda é o valor que ocorre com maior frequência.
- Usada tanto para dados numéricos quanto para dados categóricos (cuidado: afetada pela escolha de classes de agrupamento)
- Pode não haver moda e pode haver várias modas
- Não é afetada por valores extremos





Exemplo

- Uma pesquisa em uma certa cidade perguntou a 15 pessoas, escolhidas aleatoriamente, o tempo de viagem de casa para o trabalho em minutos:

  30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

- Em rol:

  5 10 10 10 10 12 15 20 20 25 30 30 40 40 60

Qual é a moda?

Medidas de Tendência Central: Exemplo

Preços das casas:

$2,000,000
500,000
300,000
100,000
100,000

Soma **3,000,000**

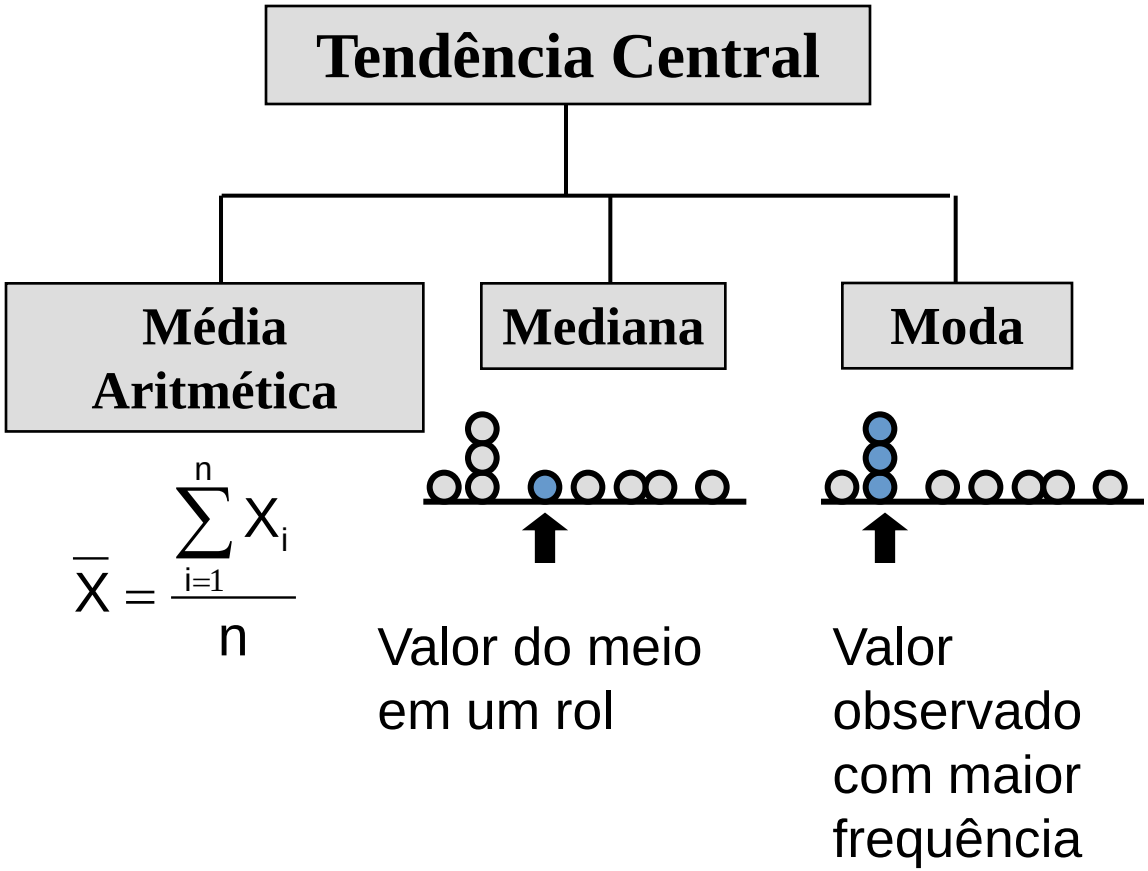
- **Média:** ($3,000,000/5)
  = **$600,000**
- **Mediana:** valor do meio dos dados ordenados
  = **$300,000**
- **Moda:** valor mais frequente
  = **$100,000**

Medidas de Tendência Central: Qual medida escolher?

- A média geralmente é usada, a menos que existam valores extremos e com distribuição muito assimétricas.

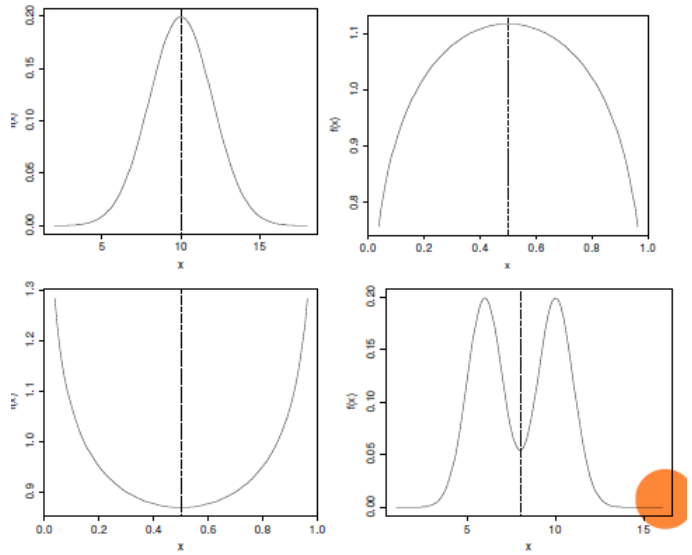
- Nesse caso, a mediana é a mais usada, uma vez que não é sensível a valores extremos. Por exemplo, o preço mediano de casas pode ser registrado para uma região por ser menos sensível a valores extremos.

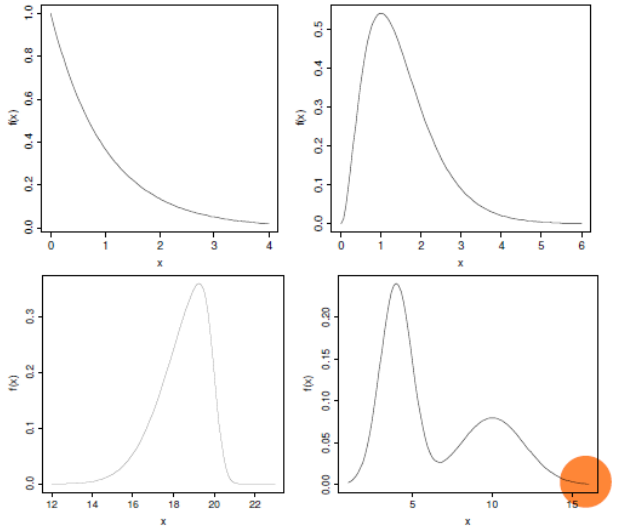
Medidas de Tendência Central: Resumo



Formato de uma Distribuição

- Medidas de formato tentam captar, em um número, características da distribuição dos dados como assimetria e "achatamento".





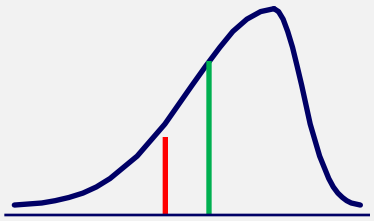
- Não vamos ver medidas numéricas de formato. As medidas mais usadas são: assimetria e curtose.

Formato de uma Distribuição

- Para dados com uma única moda, a relação entre moda, mediana e média nos fornecem uma ideia sobre a simetria de uma distribuição:

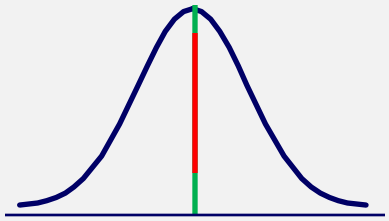
Assimétrica à esquerda

Média < Mediana < Mo



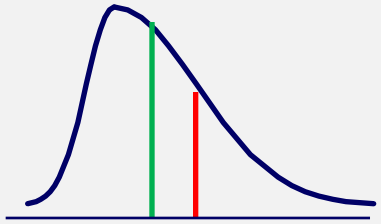
Simétrica

Média = Mediana = Mo



Assimétrica à direita

Mo < Mediana < Média



Obs: a assimetria segue a direção da cauda longa da distribuição.

Medidas de Variação

- Medidas de **variação** medem a dispersão de valores em um conjunto de dados, i. e., o grau de afastamento dos dados em torno de um valor central.

- Medidas absolutas: (Amplitude, Amplitude interquartil, Variância e Desvio-padrão)

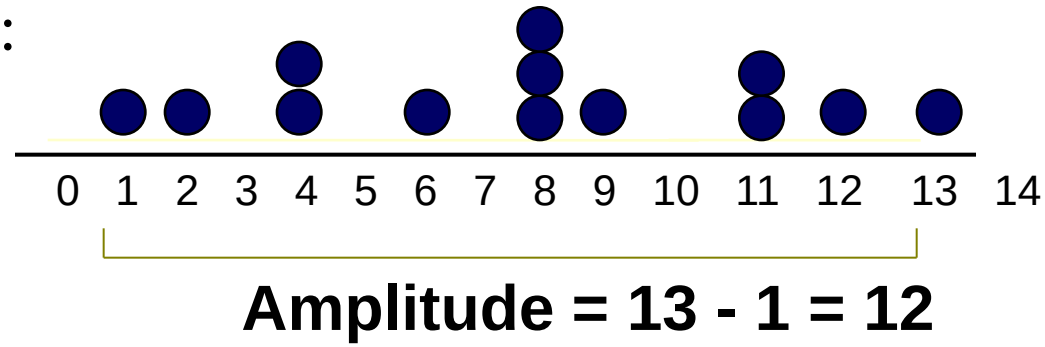
- Indicam se um conjunto de dados é *homogêneo* ou *heterogêneo*.

Amplitude

- Medida de variação mais simples
- Amplitude é definida como a diferença entre o maior e o menor dos valores:

$$\text{Amplitude} = X_{\text{maior}} - X_{\text{menor}}$$

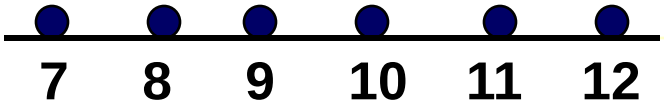
Exemplo:



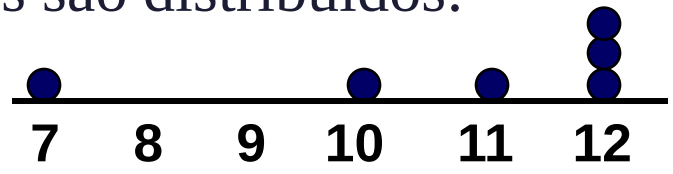
Amplitude = 13 - 1 = 12

Desvantagens da Amplitude

- Ignora a forma na qual os dados são distribuídos:



Amplitude = 12 - 7 = 5



Amplitude = 12 - 7 = 5

- Sensível a outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Amplitude = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Amplitude = 120 - 1 = 119

Exemplo

- Uma pesquisa em uma certa cidade perguntou a 15 pessoas, escolhidas aleatoriamente, o tempo de viagem de casa para o trabalho em minutos:

  30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

- Em ordem crescente:

  5 10 10 10 10 12 15 20 20 25 30 30 40 40 60

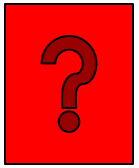
- A amplitude é: 60 – 5 = 55 min
- Afetada pelo valor atípico…
- **Como podemos ter uma ideia da variação que não seja sensível a valores atípicos?**

Medidas Separatrizes

- **Medidas separatrizes** são valores que dividem o rol em partes iguais.

- Medidas separatrizes tipicamente usadas:
  - Quartis (4 partes)
  - Decis (10 partes)
  - Centis (100 partes)
  - A nomenclatura geral é: quantil ou pertencil

Exercício: Quartis

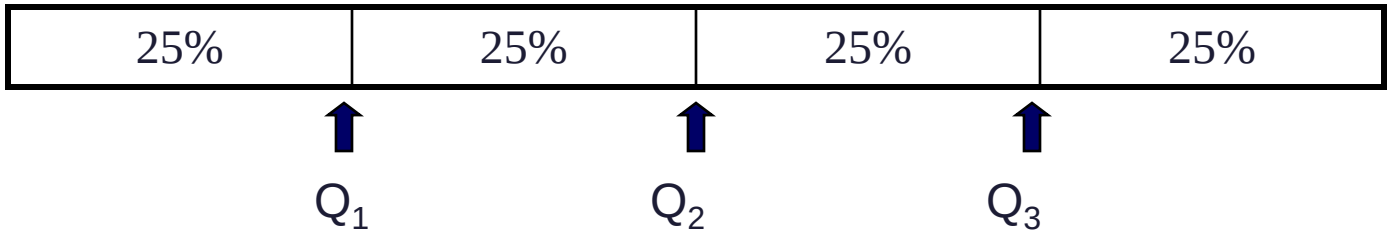


- **Exercício:** Você tem uma corda de um metro e deseja separá-la em 4 pedaços de 25 cm.

- Você deve cortar a corda em quantos pontos?
- Quais são estes pontos?

Quartis

- Quartis dividem os dados ordenados em 4 segmentos com o mesmo número de valores por segmento.



- O primeiro quartil, Q_1, é o valor para o qual 25% das observações são menores e 75% são maiores do que ele.
- Q_2 é o mesmo que a mediana (50% são menores, 50% são maiores)
- Apenas 25% dos valores são maiores do que o terceiro quartil, Q_3.

Localizando Quartis

Encontre os quartis ao determinar o valor correspondente a posição apropriada nos dados ordenados, onde

Posição do primeiro quartil: $\mathbf{Q_1}$ **= (n+1)/4° valor ordenado**

Posição do segundo quartil: $\mathbf{Q_2}$ **= (n+1)/2° valor ordenado**

Posição do terceiro quartil: $\mathbf{Q_3}$ **= 3(n+1)/4° valor ordenado**

em que **n** é o número observado de valores

ESTA É A **POSIÇÃO** DOS QUARTIS NOS **DADOS ORDENADOS**!!

Localizando Quartis

Posição dos quartis:

$$P_{Q_1} = \frac{1}{4}(n+1)$$

$$P_{Q_2} = \frac{1}{2}(n+1)$$

$$P_{Q_3} = \frac{3}{4}(n+1)$$

- **Regra 1:** se a posição de um quartil é um número inteiro, então o quartil corresponde ao valor ordenado nesta posição.
- **Regra 2:** se a posição é uma fração com 0.5 (2.5, 3.5, etc), então o quartil é igual a **média dos valores** correspondendo as posições adjacentes (2 e 3, 3 e 4, etc).
- **Regra 3:** se a posição não é um nº inteiro, nem uma fração com 0.5, então **arredonda-se** a posição para o inteiro mais próximo e determina-se o valor correspondente.

Localizando o Primeiro Quartil

- Exemplo: Encontre o primeiro quartil para os dados a seguir:

11 12 13 16 16 17 18 21 22

Primeiro, note que n = 9.

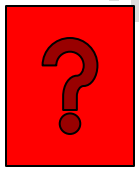
Q_1 esta na posição **(9+1)/4 = 2.5 dos dados ordenados**,

então é o valor médio entre os 2° e 3° valores ordenados,

Q_1 = 12.5

Q_1 e Q_3 são medidas de locação não centrais
Q_2 = mediana, é uma medida de tendência central

Exercício: Quartis



- Uma pesquisa em uma certa cidade perguntou a 15 pessoas, escolhidas aleatoriamente, o tempo de viagem de casa para o trabalho em minutos:

  30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

- Em rol:

  5 10 10 10 10 12 15 20 20 25 30 30 40 40 60

- **Quais são os quartis da distribuição do tempo de viagem??**

Medidas de Variação: Amplitude Interquartil

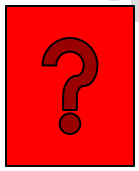
- Uma boa medida de dispersão dos dados, que não é sensível a valores atípicos, é a **Amplitude Interquartil (AIQ)**.

- A Amplitude Interquartil elimina alguns dos maiores e menores valores e calcula a amplitude apenas com os valores restantes.

- **Amplitude Interquartil =** 3° quartil – 1° quartil

$$= Q_3 - Q_1$$

Exercício: Quartis



- Uma pesquisa em uma certa cidade perguntou a 15 pessoas, escolhidas aleatoriamente, o tempo de viagem de casa para o trabalho em minutos:

  30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

- Em rol:

  5 10 10 10 10 12 15 20 20 25 30 30 40 40 60

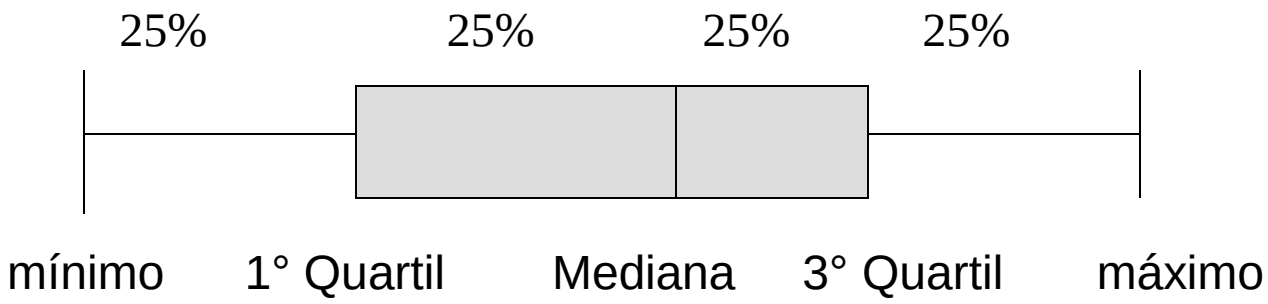
- **Qual é a Amplitude Interquartil do tempo de viagem?**

Resumo de Cinco Números

- Um Resumo de Cinco números consiste de:
  - Mínimo (X_{menor})
  - Primeiro Quartil (Q_1)
  - Mediana (Q_2)
  - Terceiro Quartil (Q_3)
  - Máximo (X_{maior})

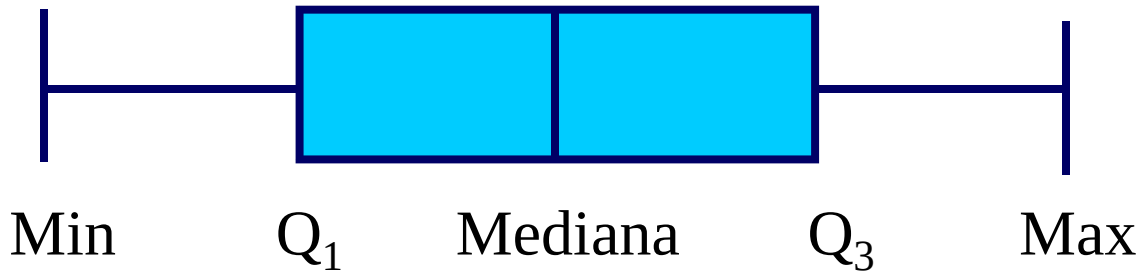
Box-Plot (diagrama de caixa)

- O Box-Plot é uma apresentação gráfica dos resumo de 5 números.



Box-Plot

- O quadro e a linha central estão localizados no meio dos pontos extremos se os dados forem simétricos em torno da média.

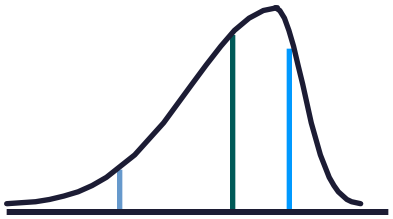


- Um gráfico Box-Plot pode ser apresentado tanto na vertical quanto na horizontal.

Box-Plot

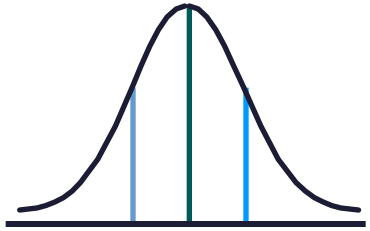
Quando os dados tem uma única moda, o box-plot nos dá uma ideia da direção da assimetria nos dados (sem precisar olhar a distribuição).

Assim. à Esq

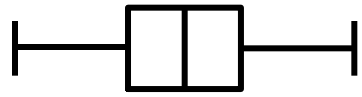


Q1 Q2Q3

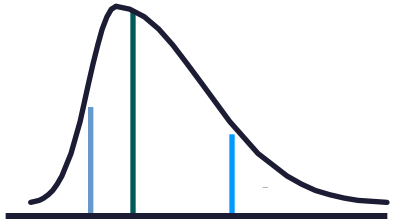
Simétrica



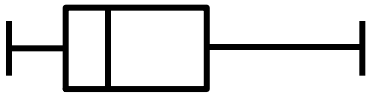
Q1Q2Q3



Assim. à Dir



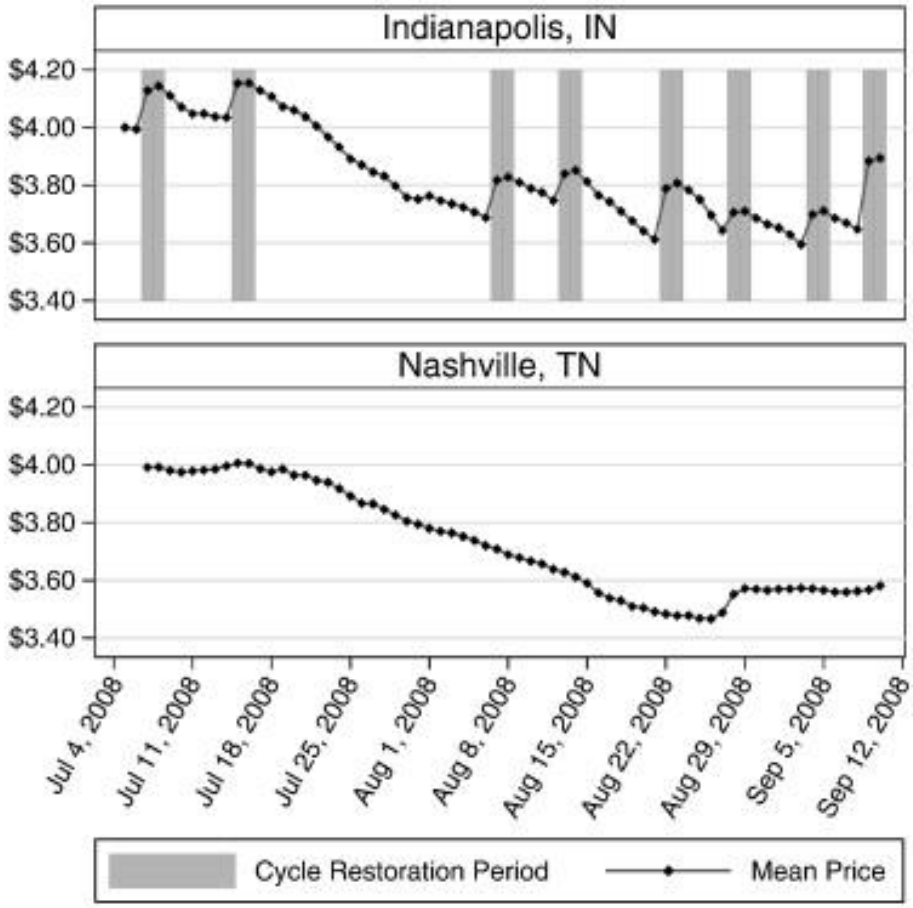
Q1 Q2 Q3



Análise Exploratória de Dados: Box-Plot

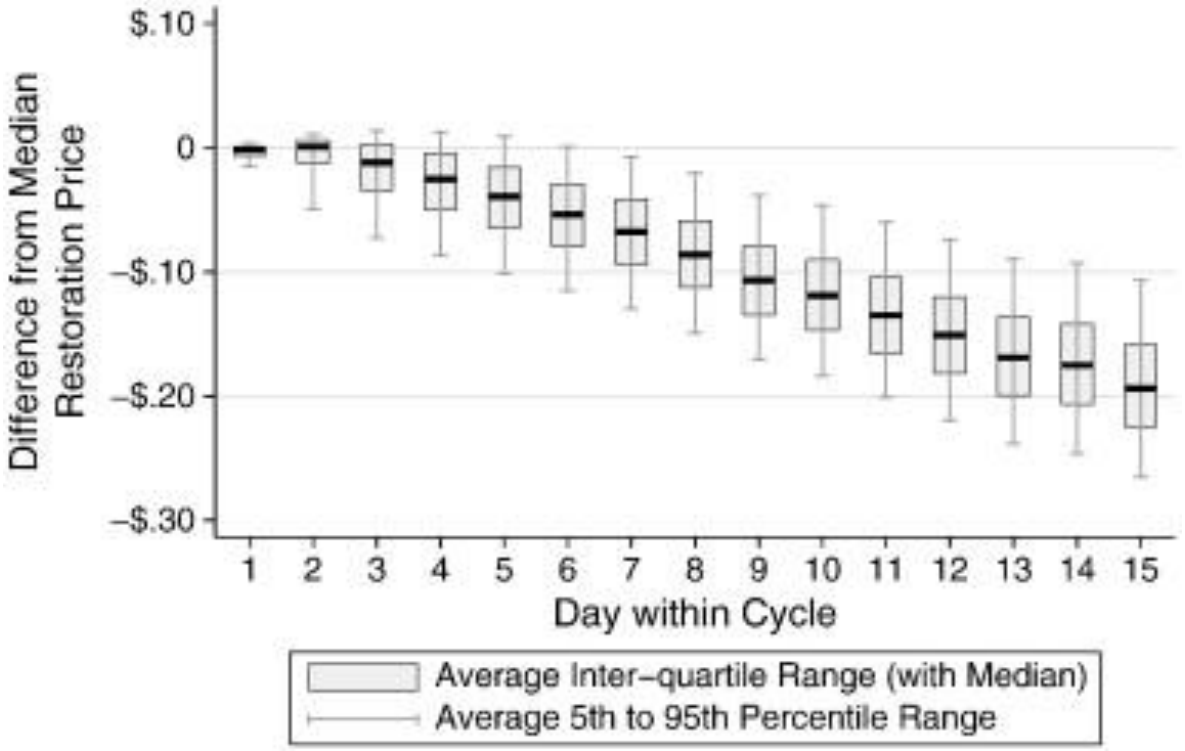
- Formas alternativas para o « bigode » do box-plot:
  - Mínimo e máximo
  - 2° e 98° percentis.
  - 1° e 99° percentis.
  - Um desvio padrão abaixo e acima da média.
  - O menor dado dentro de 1.5 AIQ (Amplitude interquartil) de Q_1 e o maior dado dentro de 1.5 AIQ de Q_3.

Análise Exploratória de Dados: Box-Plot



Fonte: Lewis (2012)
http://www.sciencedirect.com/science/article/pii/S0167718711001081

Análise Exploratória de Dados: Box-Plot



Fonte: Lewis (2012)
http://www.sciencedirect.com/science/article/pii/S0167718711001081

Variância e Desvio-padrão

- O resumo dos 5 números **não** é a descrição numérica mais comum de uma distribuição de dados.

- As medidas mais usadas para descrever os dados são:
  - Média (tendência central)
  - **Variância ou desvio-padrão** (variação)!!

Medem o quanto as observações se afastam da média…

Medidas de Variação: Variância

- A variância é a média (aproximadamente*) do **quadrado dos desvios** dos valores em relação a média.

Variância Amostral:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Em que $\overline{X}$ = média aritmética

n = tamanho da amostra

X_i = i^{esimo} valor da variável X

* (n-1: graus de liberdade)

Medidas de Variação: Desvio-padrão

- Medida de variação mais utilizada: "desvio *médio** dos dados em relação a média".
- Mostra variações em relação a média
- Raiz quadrada da variância
- Tem a mesma unidade dos dados originais

Desvio-padrão amostral:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

Medidas de Variação: Desvio-padrão

Passos para computar o **desvio-padrão amostral**

1. Compute a diferença entre cada valor e a média.
2. Eleve esta diferença ao quadrado.
3. Some os quadrados das diferenças.
4. Divida o total por n-1 para obter a variância amostral.
5. Tire a raiz quadrada da variância amostral para obter o desvio padrão amostral.

Medidas de Variação: Desvio-padrão

Dados

Amostrais (X_i) : 10 12 14 15 17 18 18 24

$n = 8$ Média = $\overline{X} = 16$

$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

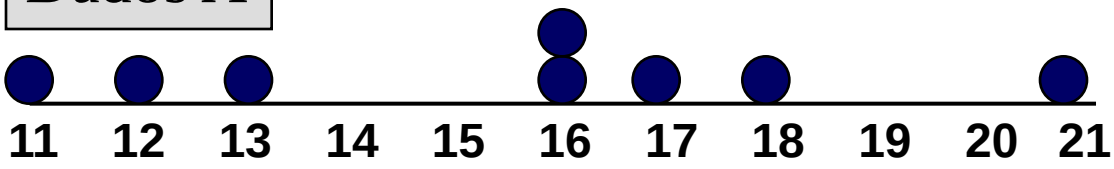
$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = 4.31$$

Uma medida de afastamento "médio" dos dados em relação à média.

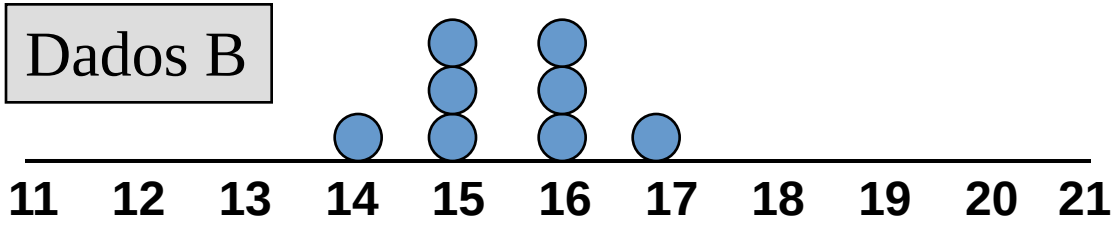
Medidas de Variação: Comparando Desvios-padrão

Dados A



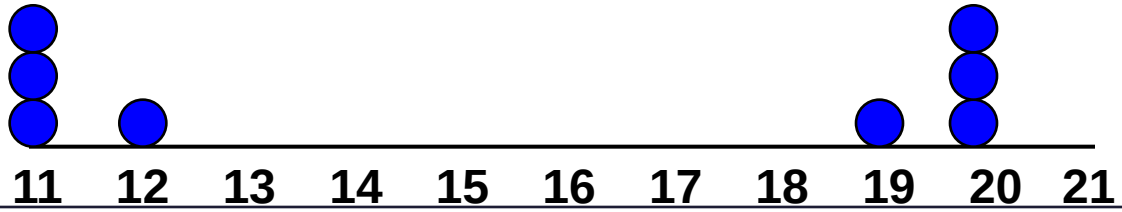
Média = 15.5
S = 3.338

Dados B



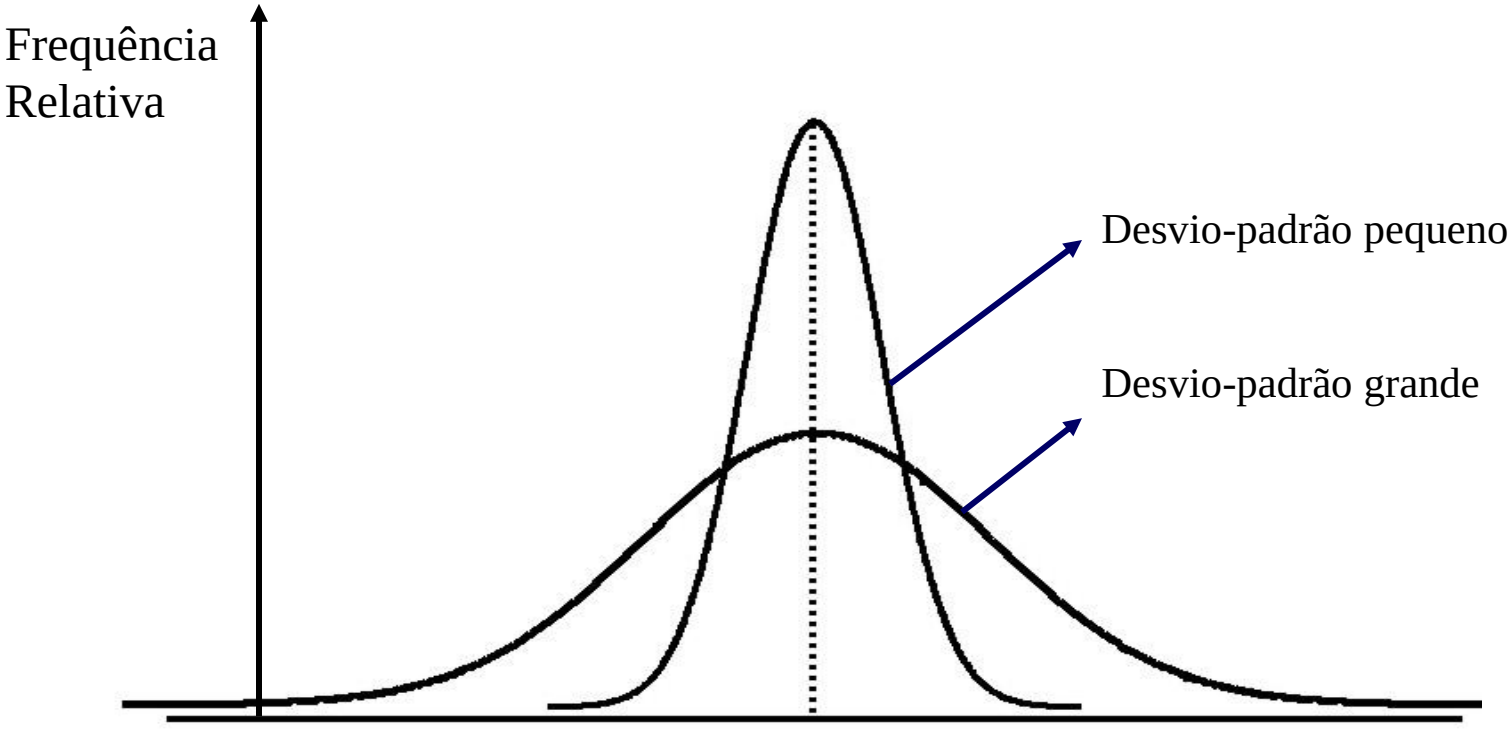
Média = 15.5
S = 0.926

Dados C

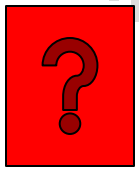


Média = 15.5
S = 4.570

Medidas de Variação: Comparando Desvios-padrão



Exercício: Desvio-padrão



- A taxa metabólica de uma pessoa é a taxa segundo a qual o corpo consume energia. Veja abaixo a taxa metabólica (calorias/dia) de 3 homens que participaram de uma dieta.

1792 1666 1362

a) Determine a taxa metabólica média.

b) Determine o desvio-padrão.

Exercício: Desvio-padrão

- Solução: n=3

1792 1666 1362

a) Determine a taxa metabólica média.

	X_i
	1792
	1666
	1362
Soma:	4820
Média:	1606.67

Exercício: Desvio-padrão

- Solução: n=3

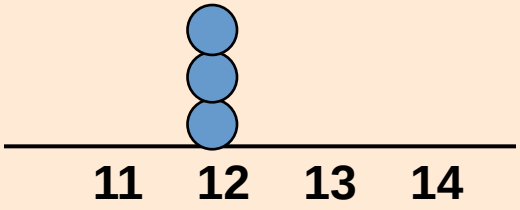
b) Determine o desvio-padrão.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

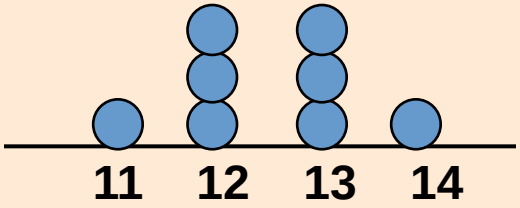
	X_i	(X_i-Média)	(X_i-Média)^2	
	1792	1792-1606.67= 185.33	185.33^2= 34347.21	
	1666	1666-1606.67 59.33	59.33^2= 3520.05	
	1362	1362-1606.67 -244.67	$(-244.67)^2$= 59863.40	
Soma:	**4820**	**-0.01**	**97730.67**	
	Média= 1606.67		**97730.67/2=48865.33**	**S=raiz(48865.33) = 221.05**

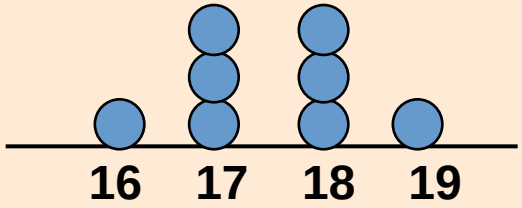
Medidas de Variação: Propriedades da Variância

Propriedade 1: A variância de uma constante é nula;



Propriedade 2: A variância da soma ou diferença de uma constante k com uma variável é igual a variância da variável;





Medidas de Variação: Propriedades da Variância

Propriedade 1: A variância de uma constante é nula;

Para uma base de dados com n dados: $x_1 = x_2 = \cdots = x_n = k$

$$S^2(k) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(k-k)^2}{n-1} = 0$$

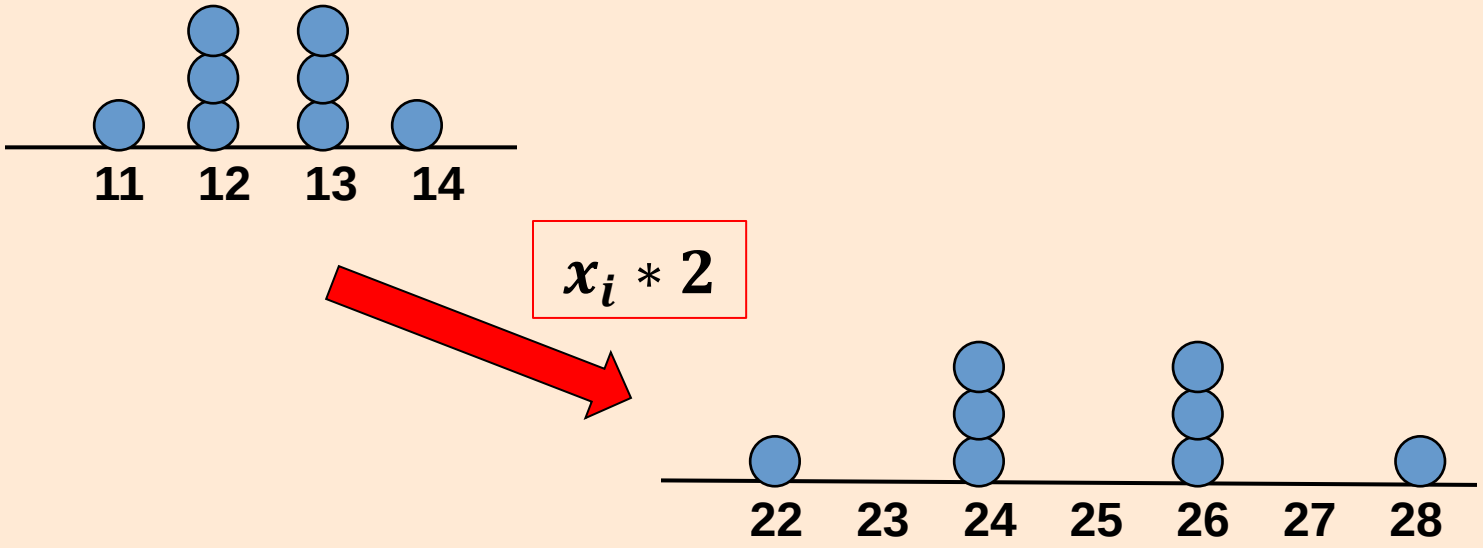
Propriedade 2: A variância da soma ou diferença de uma constante k com uma variável é igual a variância da variável;

Para uma base de dados com n dados: $x_1 = x_2 = \cdots = x_n$. Some k unidades para cada valor. A variância é:

$$S^2(x+k) = \frac{\sum_{i=1}^{n}(x_i + k - (\bar{X} + k))^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = S^2$$

Medidas de Variação: Propriedades da Variância

Propriedade 3: A variância do produto de uma constante por uma variável é igual ao produto do quadrado da constante pela variância da variável.



Medidas de Variação: Propriedades da Variância

Propriedade 3: A variância do produto de uma constante por uma variável é igual ao produto do quadrado da constante pela variância da variável.

Para uma base de dados com n dados: $x_1 = x_2 = \cdots = x_n$.

$$S^2(kx) = \frac{\sum_{i=1}^{n}\left(kx_i - \overline{kX}\right)^2}{n-1} = \frac{\sum_{i=1}^{n} k^2(x_i - \overline{X})^2}{n-1}$$
$$= k^2 \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} = k^2 S^2(x)$$

Medidas Numéricas Descritivas para a População

- As estatísticas descritivas discutidas descrevem uma *amostra* e não a *população*.
- Medidas descritivas para a população são chamadas de **parâmetros** e geralmente denotadas por letras gregas.
- Parâmetros de população importantes são a média populacional, a variância populacional e desvio-padrão populacional.

Média Populacional

- A **média populacional** é a soma dos valores na população dividida pelo tamanho da população, N.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Em que μ = média populacional

N = tamanho da população

X_i = i$^{\text{ésimo}}$ valor da variável X

Variância Populacional

- A **variância populacional** é a média do quadrado dos desvios dos valores em relação a média populacional.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Em que μ = média populacional

N = tamanho da população

X_i = i^{ésimo} valor da variável X

Desvio-Padrão Populacional

- O **desvio-padrão populacional** é a medida de variação populacional mais usada.
- A raiz da variância.
- Ele tem a mesma unidade que os dados originais.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Em que μ = média populacional

N = tamanho da população

X_i = i^{ésimo} valor da variável X

Estatísticas Amostrais Versus Parâmetros Populacionais

Medida	Parâmetro Populacional	Estatística Amostral
Média	μ	$\bar{X}$
Variância	σ^2	S^2
Desvio - Padrão	σ	S

Localizando Valores Extremos

- Duas alternativas diferentes são usadas para localizar valores atípicos (extremos) dependendo das medidas usadas para variação:
  - **Regra 1:** Usando Amplitude Interquartil
  - **Regra 2:** Usando o Desvio-padrão (Escore-Z)

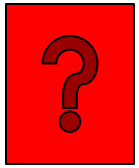
Localizando Valores Extremos 1.5 AIQ

- Vimos que a Amplitude Interquartil compreende 50% dos dados.
- Uma regra para localizar valores extremos é identificar dados que são:
  - **Menores do que Q1 – 1.5AIQ**
  - **Maiores do que Q3 + 1.5AIQ**

Um valor, X_i, é considerado **extremo** se:

$$X_i \leq Q_1 - 1.5(Q_3 - Q_1) ou\ X_i \geq Q_3 + 1.5(Q_3 - Q_1)$$

Localizando Valores Extremos 1.5 AIQ



Exercício: Abaixo estão descritos os tempos de viagem para 20 cidadãos de Nova Yorque, já arranjados em ordem crescente.

5 10 10 15 15 15 15 20 20 20 | 25 30 30 40 40 45 60 60 65 85

Existe algum valor extremo?

Localizando Valores Extremos 1.5 AIQ

- **Solução:**

  Os quartis destes dados são: Q_1=15, Q_2=22.5 e Q_3=45

  Amplitude Interquartil: AIQ = 45 – 15 = 30

  1.5*AIQ = 1.5*30 = 45

- Os valores extremos caem:
  - Abaixo de Q1-1.5*AIQ = 15 – 45 = -30
  - Acima de Q3+1.5*AIQ = 45 + 45 = 90
  - **Portanto, o tempo de viagem de 85 min não é extremo (ou atípico).**

Localizando Valores Extremos: Escore-Z

- O Escore-Z, Z_i, de um valor X_i é a "distância" que este valor está da média medida em unidades de desvio-padrão.
- Para computar o escore-Z de um dado, diminua a média e divida pelo desvio-padrão.
- Quanto maior o valor absoluto do escore-Z, mais longe o valor está da média.

Um valor X_i é considerado **extremo** se e somente se:

$$Z_i \leq -3 \; ou \; Z_i \geq 3$$

Localizando Valores Extremos: Escore-Z

$$Z_i = \frac{X_i - \overline{X}}{S}$$

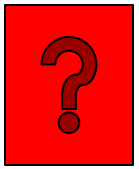
Em que X_i representa o valor do dado observado

$\overline{X}$ é a média amostral

S é o desvio-padrão amostral

Mede a distância em desvio-padrões de um certo valor X_i em relação a média.

Localizando Valores Extremos: Escore-Z



Exercício: Suponha que a nota média de um teste seja de 490 e desvio-padrão de 100.

Calcule o Escore-Z de um aluno com nota 620.

$$Z_i = \frac{X_i - \overline{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

Um escore de 620 equivale a 1.3 desvios-padrão acima da média e portanto não seria considerado um valor extremo.

Medidas numéricas para 2 variáveis

- Até agora trabalhamos com medidas para a descrição de apenas uma variável.
- Geralmente temos diversas variáveis que se relacionam entre si…
- Veremos agora medidas para a força da relação entre 2 variáveis!!

Covariância Amostral

- A **covariância amostral** mede a força da relação *linear* entre duas variáveis.

- A covariância mede se as duas variáveis se movem juntas!

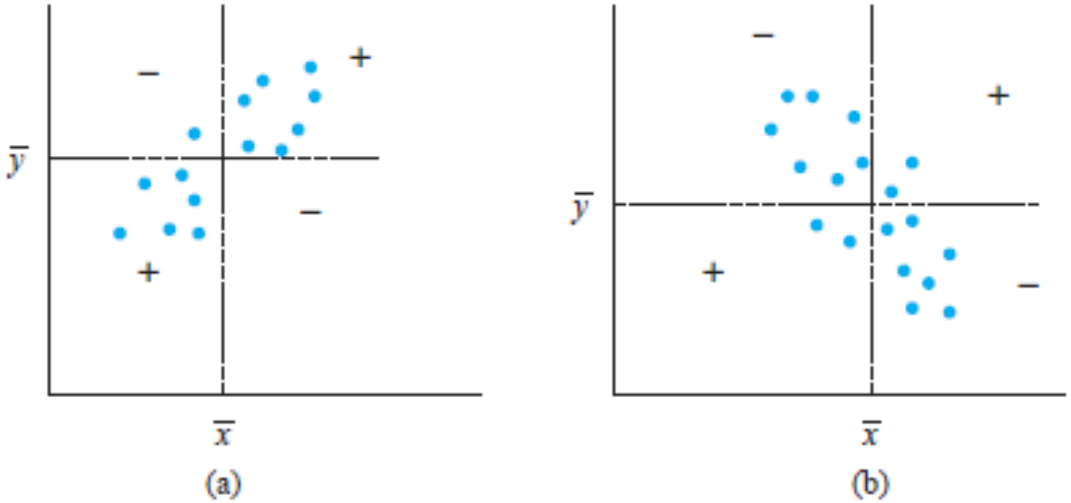
- Covariância amostral:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}\{(X_i - \bar{X}) * (Y_i - \bar{Y})\}}{n-1}$$

Covariância Amostral

- A **covariância amostral**

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}\{(X_i - \bar{X}) * (Y_i - \bar{Y})\}}{n-1}$$



Covariância Amostral

- **Covariância,** cov(X,Y), entre duas variáveis:

  Positiva: *X* e *Y* tendem a se mover na mesma direção.

  - X_i's grandes observados ao mesmo tempo que Y_i's grandes
  - X_i's pequenos observados ao mesmo tempo que Y_i's pequenos

  Negativa: *X* e *Y* tendem a se mover em direções opostas.

  - X_i's grandes observados ao mesmo tempo que Y_i's pequenos
  - X_i's pequenos observados ao mesmo tempo que Y_i's grandes

  Nula: *X* e *Y* são linearmente independentes.

Covariância Amostral

- A covariância depende das dimensões usadas…

- Assim, ao olharmos o valor calculado podemos apenas analisar o seu sinal, a magnitude não contém informação alguma sobre a força da relação entre variáveis.
  - Ex: Cov(X,Y) = 25 kg*m quando X é medido em m e Y em kg.
    → Cov(X,Y) = 2500 kg*cm quando X em cm e Y em kg.

- Por isso, usamos a **correlação**!

Coeficiente de Correlação

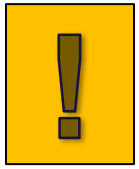
- O **coeficiente de correlação** mede a força *relativa* da relação *linear* entre duas variáveis.

- Coeficiente de correlação amostral:

$$r = \frac{\sum_{i=1}^{n}\{(X_i - \bar{X}) * (Y_i - \bar{Y})\}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} * \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left\{\frac{(X_i - \bar{X})}{S_X} * \frac{(Y_i - \bar{Y})}{S_Y}\right\} = \frac{Cov(X,Y)}{S_X * S_Y}$$

Coeficiente de Correlação: Propriedades

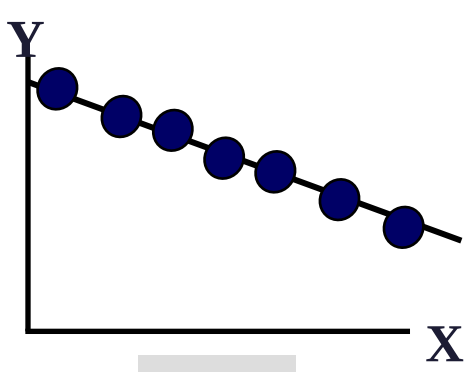


Propriedades do coeficiente de correlação:

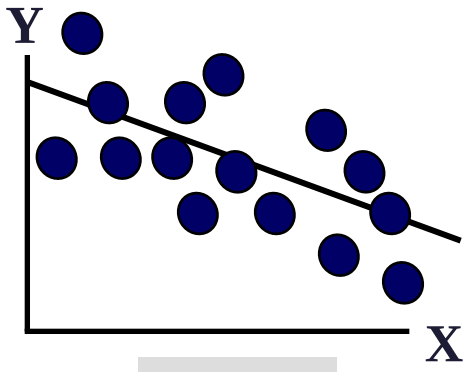
- Adimensional
- Varia entre -1 e 1
- Quanto mais próximo de -1 mais forte é a relação linear negativa entre as variavies
- Quanto mais próximo de 1, mais forte é a relação linear positiva entre as variáveis.
- Quanto mais próximo de 0, mais fraca é a relação linear entre as variáveis.

Ver applet « regression by eye »

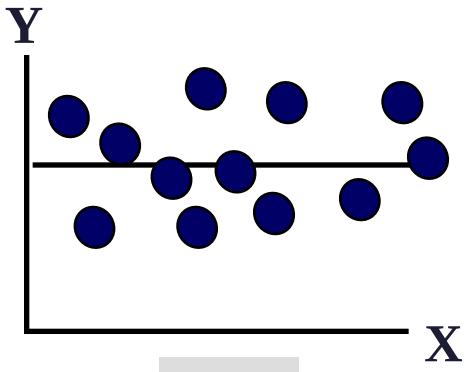
Coeficiente de Correlação



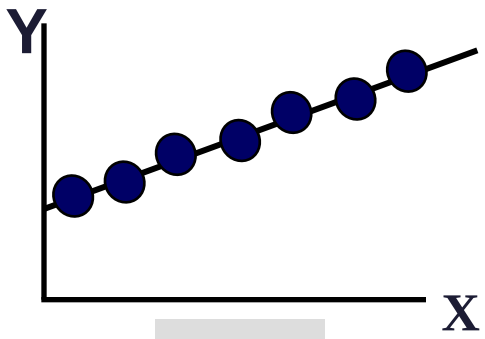
r = -1



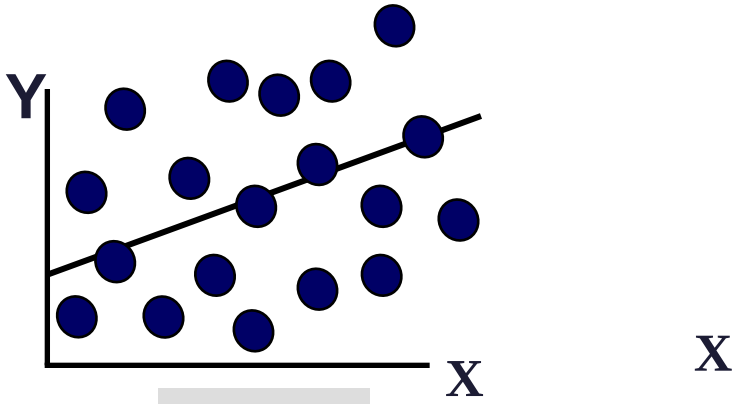
r = -0.6



r = 0

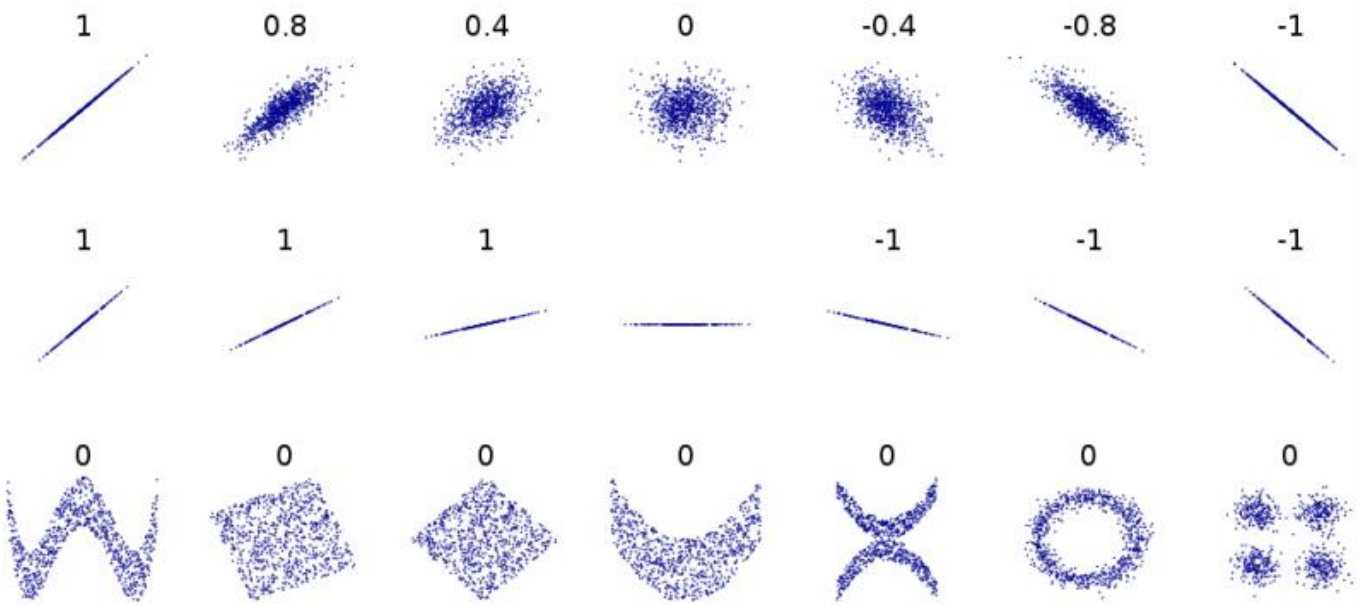


r = +1



r = +0.3

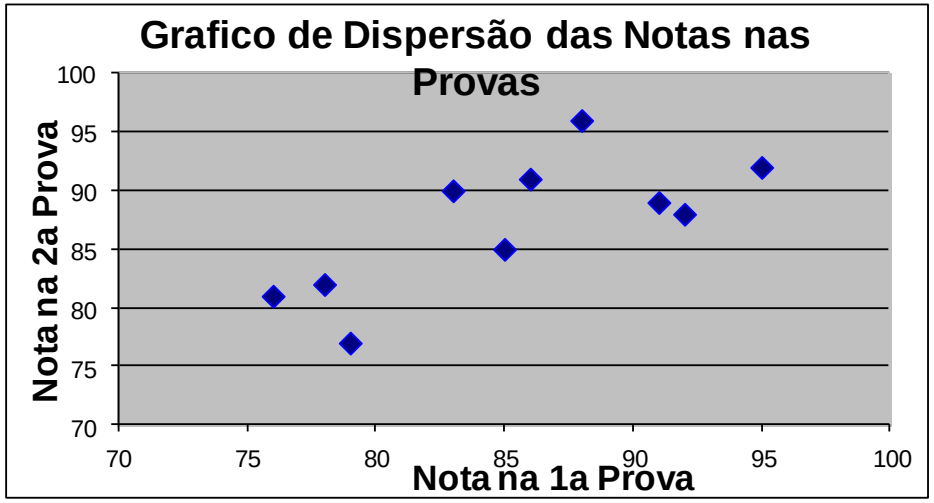
Coeficiente de Correlação



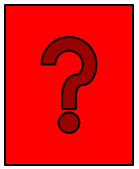
A correlação mede apenas a grau em que uma reta aproxima a relação entre duas variáveis e a direção da relação linear entre elas.
A correlação não mede a inclinação da relação ou relações não lineares entre 2 variáveis.

Coeficiente de Correlação: Exemplo

- r = 0.733
- Claramente existe uma relação linear positiva entre a nota na 1ª prova e a nota na 2ª prova.
- Alunos que tiraram notas boas na 1ª prova tendem a tirar notas boas na 2ª prova.



Coeficiente de Correlação



Exercício: Supõe-se que o conteúdo de hidrogênio (X) seja um fator importante na porosidade (Y) de fundições de liga de alumínio. Utilize os dados abaixo para calcular a correlação entre conteúdo de hidrogénio e porosidade:

X	0.18	0.20	0.21	0.22	0.30
Y	0.46	0.70	0.41	0.44	0.72

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{\mathrm{cov}(X,Y)}{S_X S_Y}$$

Coeficiente de Correlação

- **Solução:**
- Para calcularmos a correlação:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

X	Y
0.18	0.46
0.20	0.70
0.21	0.41
0.22	0.44
0.30	0.72
Soma=1.11	Soma=2.73

- Primeiro, calculamos as médias:

$$\bar{X} = \frac{1.11}{5} = 0.22$$
$$\bar{Y} = \frac{2.73}{5} = 0.55$$

Coeficiente de Correlação

- **Solução:**
- Em seguida, calculamos os desvios em relação à média:

	X_i	$(X_i-\bar{X})$	$(X_i-\bar{X})^2$	Y_i	$(Yi-\bar{Y})$	$(Yi-\bar{Y})^2$	$(X_i-\bar{X})(Yi-\bar{Y})$
	0.18	-0.04	0.0016	0.46	-0.09	0.0081	0.0036
	0.20	-0.02	0.0004	0.70	0.15	0.0225	-0.003
	0.21	-0.01	0.0001	0.41	-0.14	0.0196	0.0014
	0.22	0	0	0.44	-0.11	0.0121	0
	0.30	0.08	0.0064	0.72	0.17	0.0289	0.0136
Soma:	**1.11**	**0.01**	**0.0085**	**2.73**	**-0.02**	**0.0912**	**0.0156**

- Então:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{0.0156}{\sqrt{0.0085} * \sqrt{0.0912}} = \frac{0.0156}{0.092 * 0.31} = 0.55$$

Correlação x Causalidade

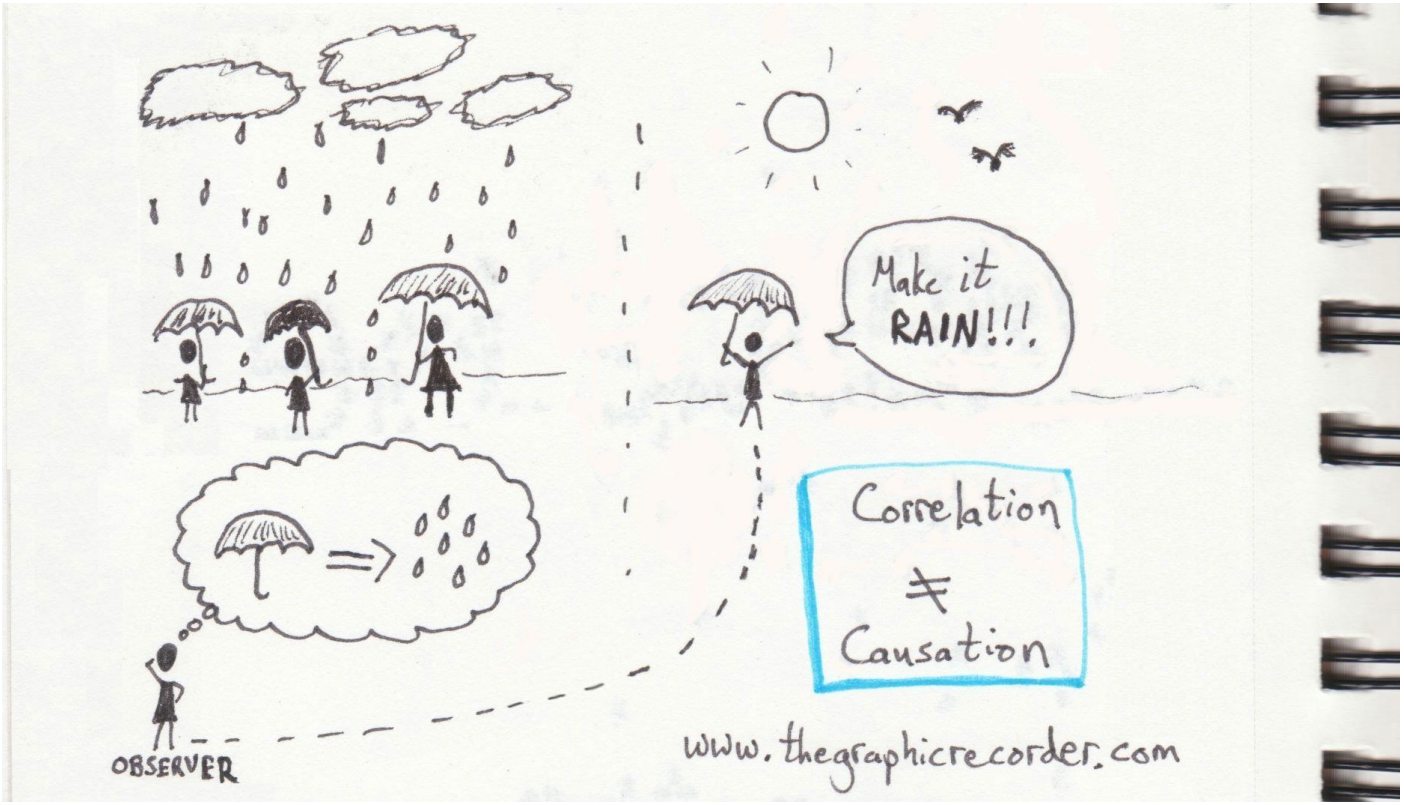
Correlação não é a mesma coisa que causalidade!!

- Na causalidade, uma variável apenas acontece por causa da outra.
- Quando há correlação positiva observamos que duas variáveis costumam andar juntas.
  - Pode ser que Y cause X, ou que X cause Y ou que exista outra variável (omitida) Z que cause as duas coisas...

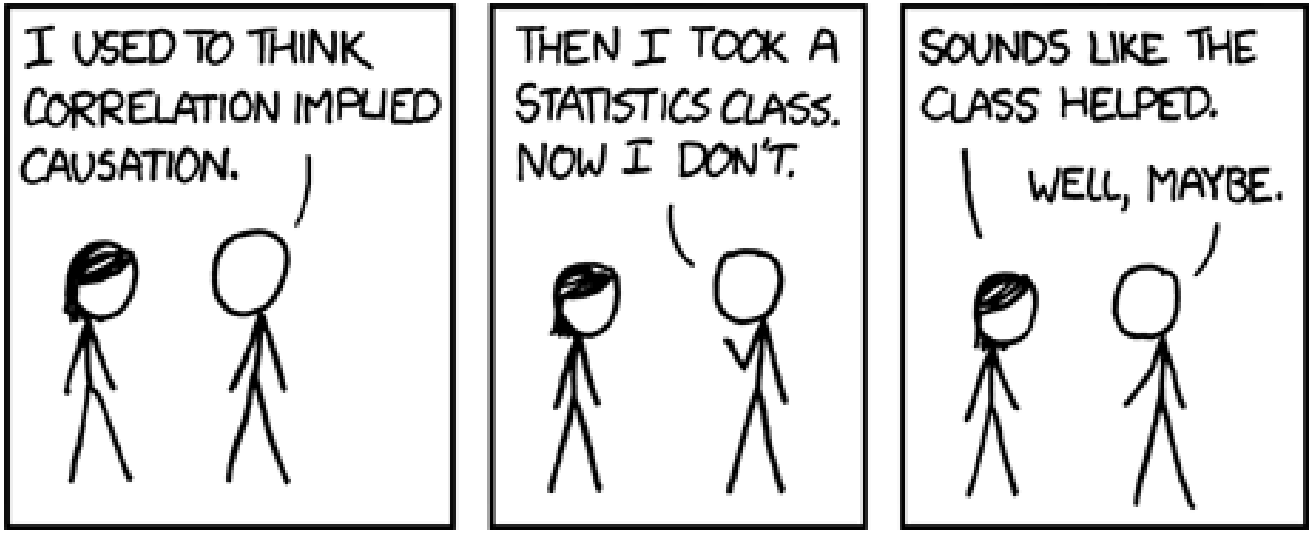
- Exemplos de correlação e não causalidade:
  - Com o passar do tempo, observamos primeiro o cantar do galo e uns minutos depois o nascer do sol – mas isso não quer dizer que é o cantar do galo que causa o nascer do sol;
  - Pessoas que dormem de sapato acordam com dor de cabeça. Dormir de sapato causa dor de cabeça?
  - Pessoas que dormem tarde tem salários mais elevados. Vou dormir mais tarde hoje para ver se acordo amanhã com o salário mais alto.. ☺

Correlação x Causalidade

Correlação não é a mesma coisa que causalidade!!



Correlação x Causalidade



- Eu achava que correlação implica causalidade

- Aí eu fiz um curso de estatística e agora não acho mais.

- Parece que o curso ajudou.

- Pode ser.

Resumo

Nesta parte da estatística descritiva, vimos:

- Medidas de tendência central: média, mediana e moda;
- Medidas de variação: amplitude, amplitude interquartil, desvio-padrão e variância, o resumo de 5 números e o box-plot.
- Como identificar valores extremos: usando a aplitude interquartil ou o escore-Z.
- Medidas de relação linear entre duas variáveis: a covariância e o coeficiente de correlação.