Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 - Bioestatística

Estatística Descritiva

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

- Para aqueles que já tem o repositório local:

```
> cd /path/to/IMD0601
```

```
> git pull
```

Estatística

Ciência que tem por objetivo **planejar**, **coletar**, **organizar**, **resumir**, **analisar** e **interpretar** dados.

- Estatística Descritiva: descreve um conjunto de dados a partir de medidas de centralidade e dispersão.
- Estatística Inferencial ou Indutiva: faz afirmações sobre uma população a partir da análise de uma amostra.
- Probabilidade: ramo da matemática que estuda eventos aleatórios e analisa as chances de um determinado evento ocorrer.

Estatística

- **População**: conjunto de elementos que possuem pelo menos uma característica em comum de interesse a ser analisado.

- **Amostra**: subconjunto finito de elementos em uma população, que são representativos para o estudo de uma determinada característica de interesse na população.

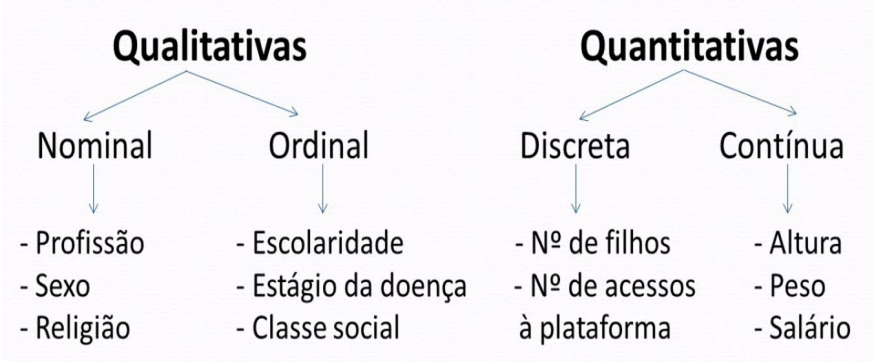
Estatística

Variáveis Quantitativas

- **Contínuas**: assumem qualquer valor dentro de um intervalo de interesse.
- **Discretas**: assumem somente valores.

Variáveis Qualitativas ou Categóricas

- **Nominais**: assumem estados ou categorias, que não implicam em precedência.
- **Ordinais**: assumem categorias que são ordenadas ou avaliadas segundo algum critério.

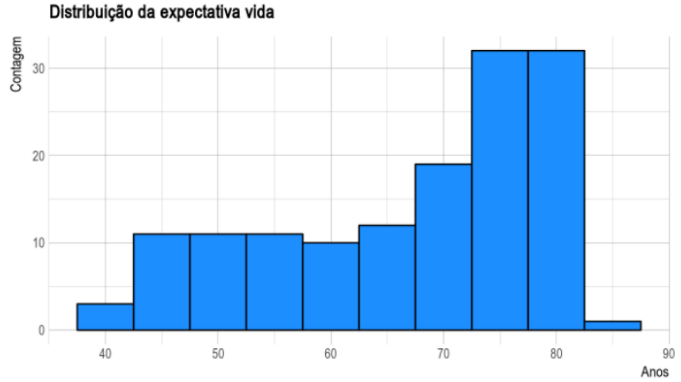


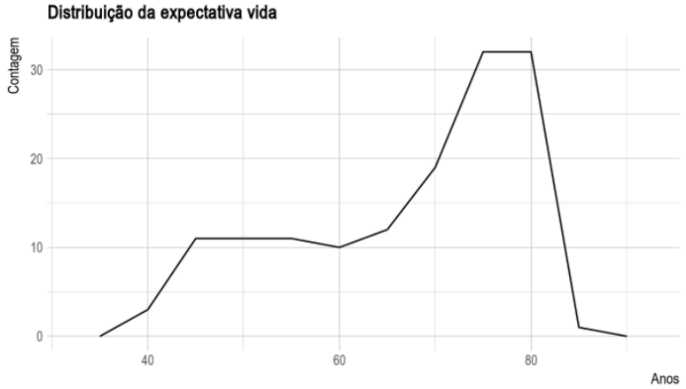
Estatística Descritiva

- Gráficos são uma boa forma de mostrar como uma amostra se comporta, mas os descrições numéricas são bem informativas e precisas!

- Podemos começar perguntando qual é a média da expectativa de vida medida em 2007? Há algum valor típico?

- Quais são os limites inferior e superior desta distribuição? Qual é a dispersão?





Medidas de Tendência Central (MTC)

- Representam um conjunto de dados com valores centrais pelos quais os dados tendem a concentrar-se.
- Também chamadas de estatísticas de centralidade ou de localização.
- Indicam onde localiza-se o centro (o ponto médio) de um conjunto de dados.
- São medidas de resumo dos dados.
- As mais comuns: **Média**, **Mediana** e **Moda**.

Média Aritmética

- É a soma dos valores observados dividido pelo número de observações.
- Indica o “centro de gravidade” dos dados e funciona bem com valores que se ajustam a distribuição normal.
- É altamente sensível a valores discrepantes ou atípicos (“outliers”).

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mediana

- É a observação que indica exatamente a posição central de um conjunto de dados quando estes estão ordenados (crescente ou decrescente).

- Divide o conjunto de dados em duas partes com o mesmo número de elementos:
  - Quantidade ímpar de valores: termo central do conjunto de dados.
  - Quantidade par de valores: média dos dois termos centrais.

- Útil para lidar com distribuições altamente distorcidas.

- Útil quando é impraticável medir todos os valores, como em ‘tempo para evento’.

Método para determinação da Mediana

Amostra obtida

$\{ x_1 , x_2, \ldots , x_k , \ldots, x_n \}$ 

Amostra ordenada

$\{ x_{(1)} , x_{(2)}, \ldots , x_{(k)} , \ldots, x_{(n)} \}$

min — k-ésima — max

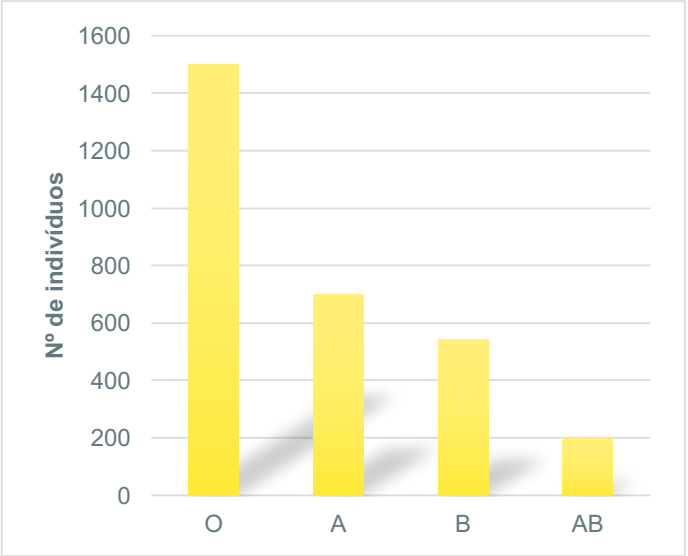
$$n \text{ ímpar}: \quad Med(x) = x_{\left(\frac{n+1}{2}\right)}$$

$$n \quad par: \; Med(x) = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}}{2}$$

Moda

- É a observação mais comum (frequente) em um conjunto de dados.
- Requer que uma variável contínua seja agrupada em um número relativamente pequeno de classes.
- É também usada com variáveis categóricas.
- Útil para distinguir distribuições:
  - Amodal: não apresenta uma moda.
  - Unimodal: 1 único valor aparece mais.
  - Bimodal: 2 valores aparecem mais.
  - Multimodal: 3 ou mais valores se destacam.

Tipo Sanguíneo	Número de indivíduos
O	1500
A	700
B	543
AB	200



Média, Moda e Mediana

Considerando os conjuntos A e B, abaixo:

A = (1, 3, 5, 7, 9)

Média = 5

Mediana = 5

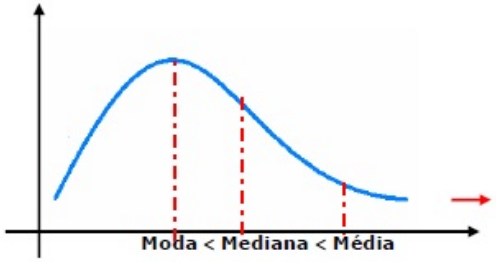
B = (2, 3, 5, 7, 58)

Média = 15

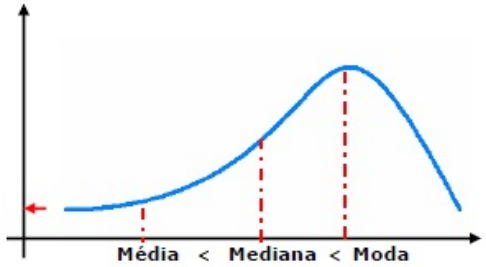
Mediana = 5

Enquanto a média é afetada por valores extremos, a mediana é mais “robusta”, ou seja, não sofre influência de valores extremos.

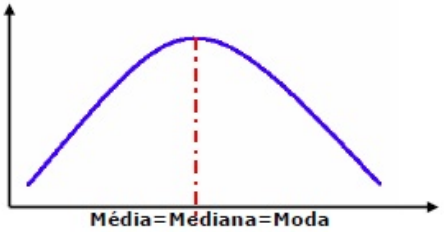
→ Distribuição Assimétrica à Direita (ou de Assimetria Positiva):



→ Distribuição Assimétrica à Esquerda (ou de Assimetria Negativa):



→ Distribuição Simétrica:



Estatística descritiva em R

The blacknose dace, *Rhinichthys atratulus.*

- O Maryland Biological Stream Survey usou a pesca elétrica para contar o número de indivíduos de cada espécie de peixe em segmentos de riachos de 75 m de comprimento selecionados aleatoriamente em Maryland.

- Calcule a média, mediana e a moda da amostra.

Stream	fish/75m
Mill_Creek_1	76
Mill_Creek_2	102
North_Branch_Rock_Creek_1	12
North_Branch_Rock_Creek_2	39
Rock_Creek_1	55
Rock_Creek_2	93
Rock_Creek_3	98
Rock_Creek_4	53
Turkey_Branch	102

http://rcompanion.org/rcompanion/c_01.html

Estatística descritiva em R

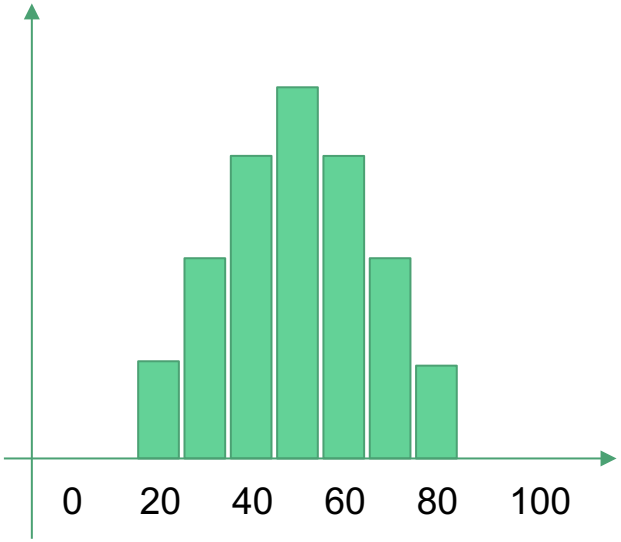
```r
# Média Aritimética
# Use na.rm=TRUE para não retornar erro. Algumas funções excluem NA por padrão.
mean(Data$Fish, na.rm=TRUE)

# Mediana
median(Data$Fish, na.rm=TRUE)

# Moda
Mode(Data$Fish)

# Resumos de estatisticas descritivas e gráficos
# funcionam com todo o dataframe ou com variáveis individuais
summary(Data$Fish)            # quartis
describe(Data$Fish, type=2)   # outras estatisticas

# Histograma
hist(Data$Fish,
     col="gray",
     main="Maryland Biological Stream Survey",
     xlab="Fish count")

# Adicione uma variável numérica com os mesmos valores de Fish
Data$Fish.num = as.numeric(Data$Fish)

# Descfunction produz informações resumidas para cada tipo de variável e gráficos
Desc(Data, plotit=TRUE)
```

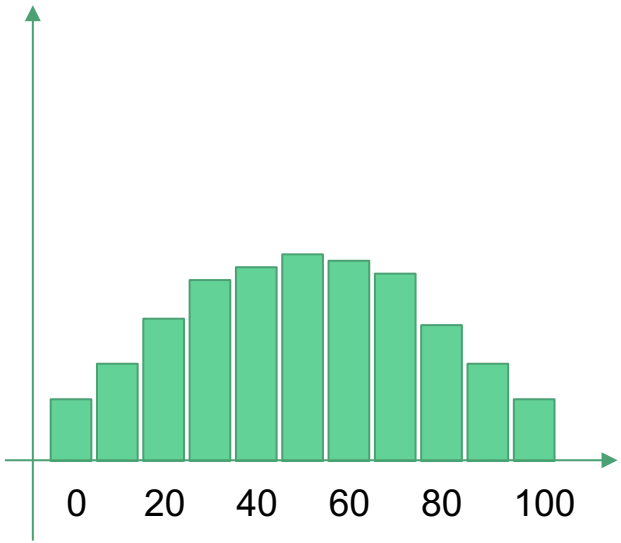
The blacknose dace, *Rhinichthys atratulus.*

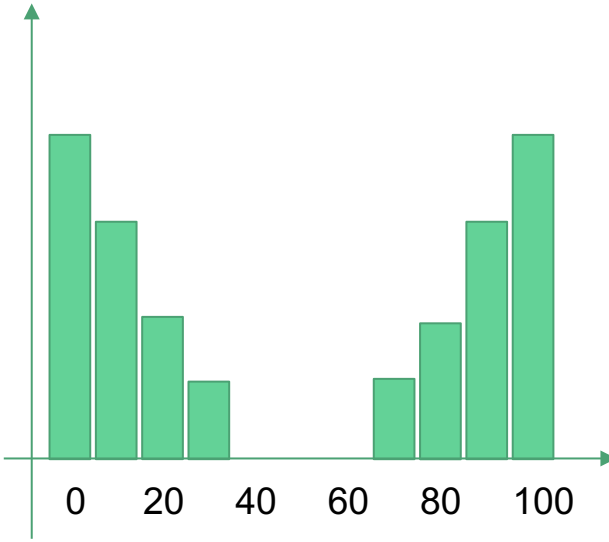
Stream	fish/75m
Mill_Creek_1	76
Mill_Creek_2	102
North_Branch_Rock_Creek_1	12
North_Branch_Rock_Creek_2	39
Rock_Creek_1	55
Rock_Creek_2	93
Rock_Creek_3	98
Rock_Creek_4	53
Turkey_Branch	102

http://rcompanion.org/rcompanion/c_01.html

Como são os dados que possuem média = 50?





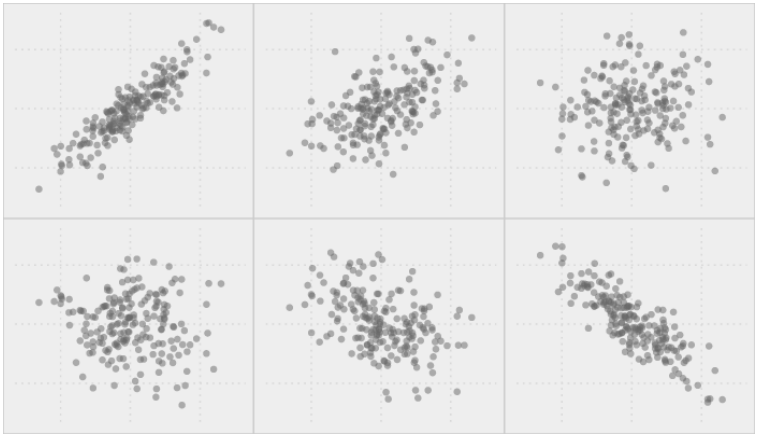


Medidas de dispersão

São medidas que descrevem o espalhamento dos dados e junto com as MTCs descrevem a distribuição.

Medida de dispersão ideal:

- Definição clara e rígida;
- Fácil cálculo e entendimento;
- Não deve ser muito afetada por flutuações;
- Baseado em todas as observações.



Fonte: https://uc-r.github.io/correlations

Medidas de dispersão

Medidas de dispersão absoluta

- Quantificam a variação em termos da unidade de medida dos dados
    - Amplitude;
    - Desvio entre quartis;
    - Desvio absoluto da média;
    - Desvio padrão.

Medidas de dispersão relativa

- Não possuem unidade de medida, comparação entre as distribuições.
    - Coeficiente de amplitude;
    - Coeficiente de desvio entre quartis;
    - Coeficiente de desvio da média;
    - Coeficiente de variação;

Amplitude

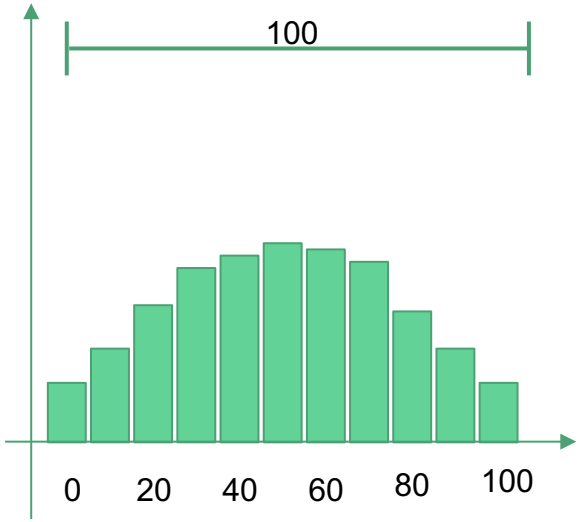
- Diferença entre o valor máximo e mínimo.
- Pode não ser muito informativo.
- Depende do número de observações.

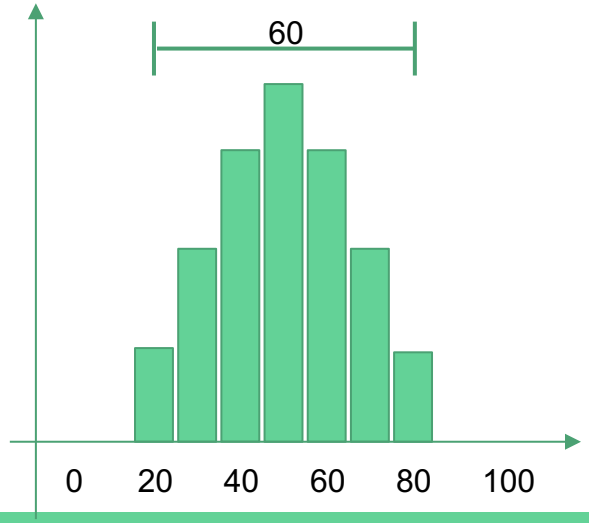
Vantagens:

- Medidas de dispersão mais simples;
- Fácil cálculo e entendimento.

Desvantagens:

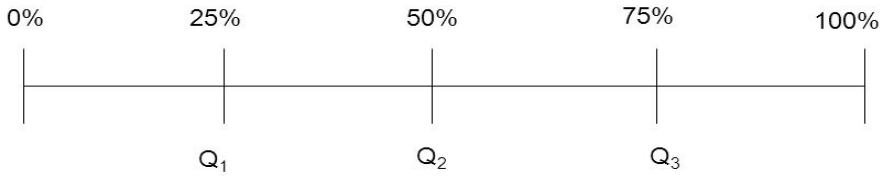
- Baseado em apenas duas observações extremas;
- Não é uma medida de dispersão confiável.





Quantis

- São as observações em um conjunto de dados ordenados, que estabelecem divisões do conjunto em partes iguais.
- Denota-se Pk o percentil de ordem k, ao valor que deixa k% dos dados abaixo de si.
- Neste grupo uma conotação de interesse são os "**quartis**", que dividem o conjunto de dados em quatro partes iguais, cada uma contendo 25% das observações.

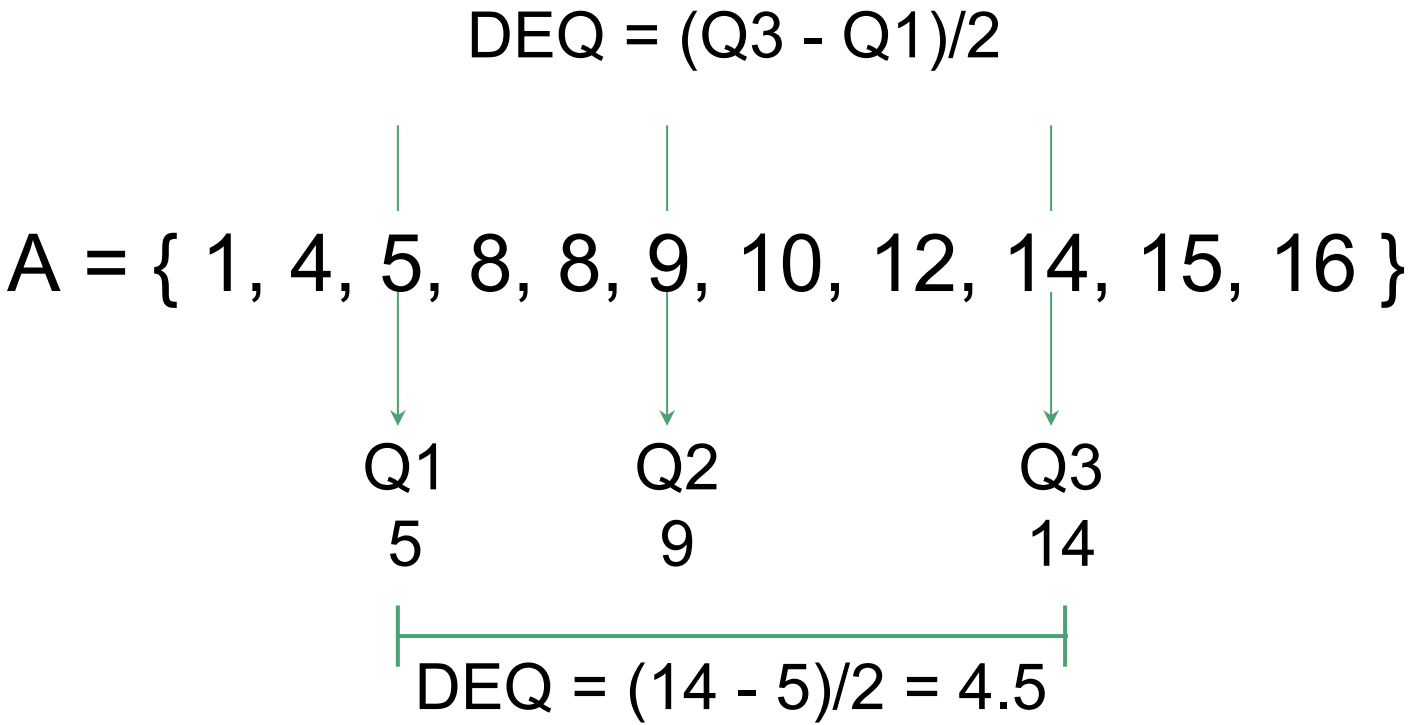


Q_1= 1º QUARTIL, DEIXA 25% DOS ELEMENTOS.

Q2 = 2º QUARTIL, COINCIDE COM A MEDIANA, DEIXA 50% DOS ELEMENTOS.

Q3 = 3º QUARTIL, DEIXA 75% DOS ELEMENTOS.

Desvio entre Quartis (IQR)



Desvio entre quartis

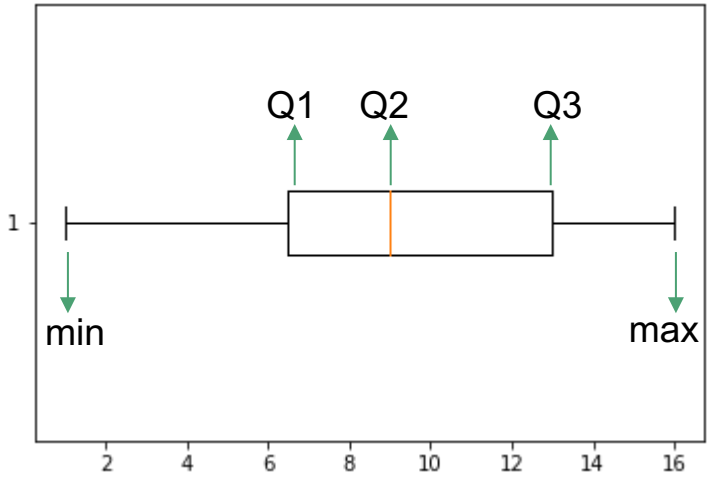
Vantagens:

- Fácil de calcular;
- O cálculo envolve apenas o Q1 e o Q3;
- Não é afetado por valores extremos;

Desvantagens:

- Utiliza apenas 50% dos dados para o seu cálculo;

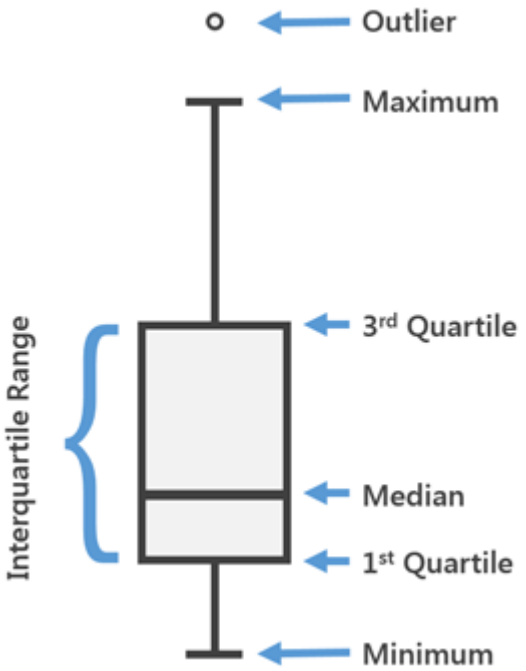
Representação gráfica em Boxplot



A = { 1, 4, 5, 8, 8, 9, 10, 12, 14, 15, 16 }

Outliers

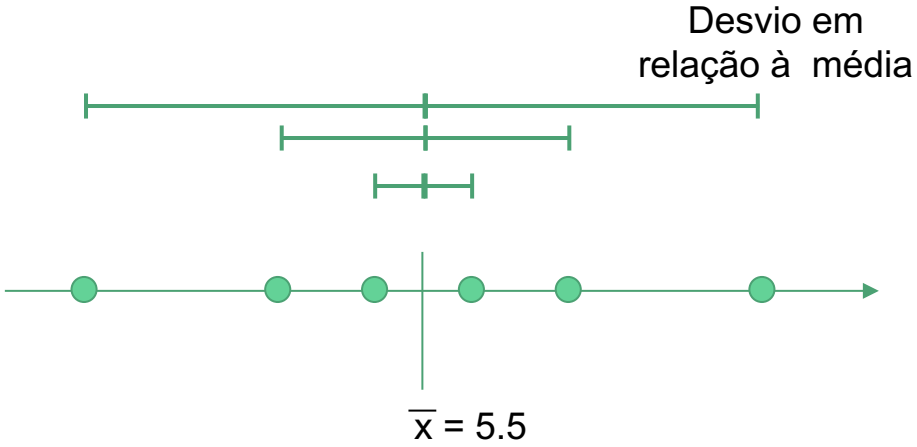
- Determinação dos *outliers* utilizando a separação dos dados em quartis:
- Limite inferior: **Q1 - 1.5 * IQR**
- Limite superior: **Q3 + 1.5 * IQR**
- Onde **IQR = Q3 - Q1**
- Os dados que estiverem fora deste limite são considerados *outliers*.



Fonte: https://sudar.me/pgdmlai-notes/eda/assets/images/box-plot.png

Desvio envolvendo a média

xi	xi - $\overline{x}$
2	2 - 5.5 = -3.5
4	4 - 5.5 = -1.5
5	5 - 5.5 = -0.5
6	6 - 5.5 = 0.5
7	7 - 5.5 = 1.5
9	9 - 5.5 = 3.5



Média do desvio = $\sum(xi - x)/n = 0$

Desvio envolvendo a média

Desvio absoluto

xi	$xi - \overline{x}$	$\lvert xi - \overline{x} \rvert$
2	2 - 5.5 = -3.5	|2 - 5.5| = 3.5
4	4 - 5.5 = -1.5	|4 - 5.5| = 1.5
5	5 - 5.5 = -0.5	|5 - 5.5| = 0.5
6	6 - 5.5 = 0.5	|6 - 5.5| = 0.5
7	7 - 5.5 = 1.5	|7 - 5.5| = 1.5
9	9 - 5.5 = 3.5	|9 - 5.5| = 3.5

Média do desvio absoluto $= \sum(|xi - x|)/n = 5.5$

Desvio envolvendo a média

Quadrado do Desvio

xi	xi - $\overline{x}$	| xi - $\overline{x}$ |	(xi - $\overline{x}$)**2
2	2 - 5.5 = -3.5	|2 - 5.5| = 3.5	(2 - 5.5)**2 = 12.25
4	4 - 5.5 = -1.5	|4 - 5.5| = 1.5	(4 - 5.5)**2 = 2.25
5	5 - 5.5 = -0.5	|5 - 5.5| = 0.5	(5 - 5.5)**2 = 0.25
6	6 - 5.5 = 0.5	|6 - 5.5| = 0.5	(6 - 5.5)**2 = 0.25
7	7 - 5.5 = 1.5	|7 - 5.5| = 1.5	(7 - 5.5)**2 = 2.25
9	9 - 5.5 = 3.5	|9 - 5.5| = 3.5	(9 - 5.5)**2 = 12.25

Média do quadrado dos desvios

=

∑((xi -x)**2)/n

=

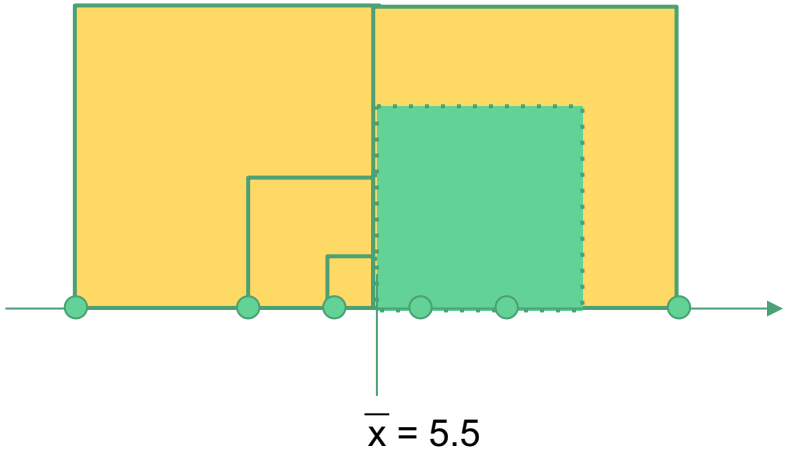
4.916

Variância

Variância

1. Subtraia a média de cada observação, eleve ao quadrado, some => soma dos quadrados
2. Divida pelo número de observações => desvio quadrático médio (da população)!

- O resultado é um número elevado ao quadrado => quadrado da unidade de medida!
- Quanto maior a variância, maior o espalhamento dos dados.



Desvio-padrão

- É uma medida mais fácil de entender uma vez que não é elevada ao quadrado!

Vantagens:

- Envolve todas as observações para o seu cálculo;
- É pouco afetado por flutuações dos valores;
- Bem definido;

Desvantagens:

- Seu cálculo pode ser laborioso, especificamente se o tamanho dos dados é grande o suficiente;
- Pode ser afetado por valores extremos;

Variância

$$\sigma^2 = \frac{\sum(xi-x)^2}{(n)}$$

Desvio padrão

$$\sigma = \sqrt{\frac{\sum(xi-x)^2}{n}}$$

Muitas vezes...

- Em relação às medidas de variância e desvio padrão, usamos o divisor como n-1.
- Ele indica os graus de liberdade da amostra, ou seja, o número de desvios que estão livres para variar.
- Lembrando que a soma dos desvios é sempre zero, em uma amostra com **n** elementos, eu posso variar **n-1** destes elementos.
- Depois de determinados eles **n-1** valores, o último só pode assumir um único valor.

Desvio padrão da população

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Desvio padrão da amostra

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}}$$

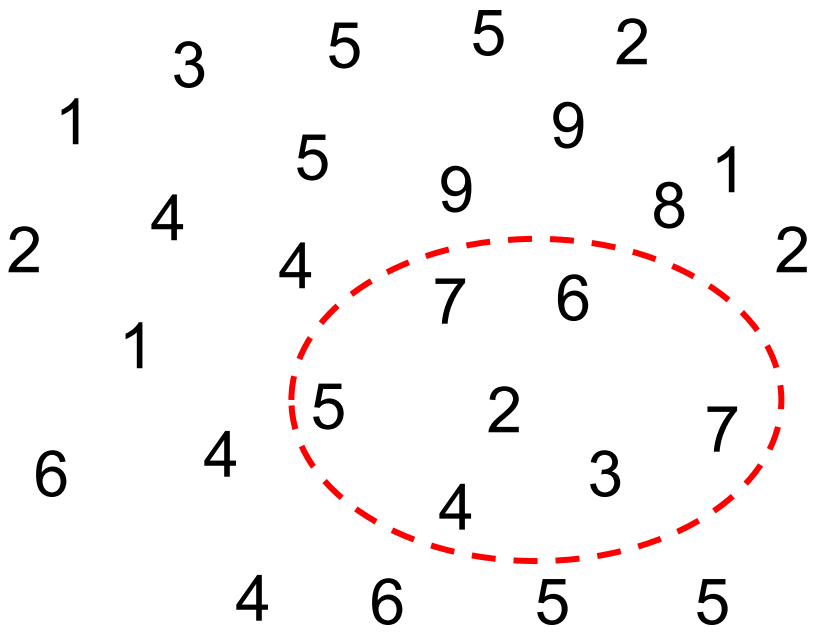
Desvio padrão amostral

Uma amostragem não consegue representar toda a variabilidade da população, por isso utilizamos a **correção de Bessel**, para corrigir esta limitação da amostragem.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

$\sigma = 2.243756$

$s = 1.9518$ $(\sigma = 1.807016)$



Medidas de dispersão relativa

- Não possuem unidade de medida.
- Permitem a comparação entre distribuições.
  - Coeficiente de Amplitude: (H - L)/(H + L)
  - Coeficiente de desvio entre quartis: (Q3 - Q1)/(Q3 + Q1)
  - Coeficiente de desvio da média: (desvio da média)/(média ou mediana)
  - Coeficiente de variação: (desvio padrão)/(média)

Medidas de dispersão relativa

- Resume a quantidade de variação como uma porcentagem ou proporção do total.
- É útil ao comparar a quantidade de variação de uma variável entre grupos com médias diferentes ou entre variáveis de medição diferentes.

Coef. de variação = (desvio padrão)/(média)

Idade grupo 1	Idade grupo 2
1	53
3	55
5	57

$S^2 = 4$

CV(1) = 2/3 * 100 = 66,67 %

CV(2) = 2/55 * 100 = 3,64 %

Estatística descritiva em R

The blacknose dace, *Rhinichthys atratulus.*

- O Maryland Biological Stream Survey usou a pesca elétrica para contar o número de indivíduos de cada espécie de peixe em segmentos de riachos de 75 m de comprimento selecionados aleatoriamente em Maryland.

- Calcule a amplitude, variância, desvio padrão e coeficiente de variação da amostra.

Stream	fish/75m
Mill_Creek_1	76
Mill_Creek_2	102
North_Branch_Rock_Creek_1	12
North_Branch_Rock_Creek_2	39
Rock_Creek_1	55
Rock_Creek_2	93
Rock_Creek_3	98
Rock_Creek_4	53
Turkey_Branch	102

http://rcompanion.org/rcompanion/c_02.html

Estatística descritiva em R

The blacknose dace, *Rhinichthys atratulus.*

```
## Estatísticas de Dispersão ------------

# Intervalo
range(Data$Fish, na.rm=TRUE)
max(Data$Fish, na.rm=TRUE) - min(Data$Fish, na.rm=TRUE)

# Variância (amostra)
var(Data$Fish, na.rm=TRUE)

# Desvio Padrão (amostra)
sd(Data$Fish, na.rm=TRUE)
round(sd(Data$Fish, na.rm=TRUE), 2)

# Coeficiente de variaçao, como porcentagem
sd(Data$Fish, na.rm=TRUE)/mean(Data$Fish, na.rm=TRUE)*100
```

Stream	fish/75m
Mill_Creek_1	76
Mill_Creek_2	102
North_Branch_Rock_Creek_1	12
North_Branch_Rock_Creek_2	39
Rock_Creek_1	55
Rock_Creek_2	93
Rock_Creek_3	98
Rock_Creek_4	53
Turkey_Branch	102

http://rcompanion.org/rcompanion/c_02.html

Referências

- Vieira, S. *Introdução à Bioestatística*. 4ª ed. Rio de Janeiro: Elsevier, 2008.

- McDonald, J.H. 2014. *Handbook of Biological Statistics* (3rd ed.). Sparky House Publishing, Baltimore, Maryland. http://www.biostathandbook.com/index.html

- Mangiafico, S.S. 2015. *An R Companion for the Handbook of Biological Statistics*, version 1.3.2. rcompanion.org/rcompanion/.