Estatística Descritiva

Ma. Pétala Tuy
Cientista de Dados – ATOS Brasil
Pesquisadora associada do Centro de Excelência em Pesquisa Aplicada
em Inteligência Artificial para a Industria do SENAI CIMATEC/ATOS

Conteúdo

- Introdução
- Tipos de Variáveis
- Análise Univariada
- Distribuições de Frequências
- Gráficos para Variáveis Qualitativas
- Construindo gráficos no R com ggplot2
- Gráficos para Variáveis Quantitativas
- Medidas Resumo:
  - Medidas de posição
  - Medidas de dispersão
  - Quantis empíricos

Introdução

- **Estatística descritiva** é um ramo da estatística que utiliza diversas técnicas para **descrever** e **sumarizar** conjuntos de dados.

- **Conjunto de dados** é uma coleção de dados normalmente organizados de forma matricial, onde cada **linha** é uma **observação** e cada **coluna** é uma **variável**.

- **Variáveis** são valores que assumem determinadas características.

- **Observação** é o valor obtido para cada variável.

Introdução

- Inspeção dos dados

Observações

Variáveis

```
> df <- as.data.table(df)
> head(df)
   DATA DA NOTIFICACAO DOR DE GARGANTA DISPNEIA FEBRE TOSSE OUTROS E PROFISSIONAL DE SAUDE?
1:          07/06/2020             SIM      NAO   NAO   NAO                           NAO                                SIM
2:          05/06/2020             NAO      NAO   SIM   NAO                           SIM                                NAO
3:          16/06/2020             SIM      NAO   NAO   SIM                           SIM                                SIM
4:          30/06/2020             NAO      NAO   NAO   NAO                           SIM                                NAO
5:          27/05/2020             NAO      NAO   NAO   NAO                           SIM                                NAO
6:          17/04/2020             NAO      NAO   NAO   SIM                           NAO                                NAO
                    DESCRICAO DO SINTOMA DATA DO INICIO DOS SINTOMAS
1:                                  <NA>                  03/06/2020
2:                           DOR TORACICA                 01/06/2020
3:                               CEFALEIA                 15/06/2020
4:                               CEFALEIA                 30/06/2020
5: ASSINTOMATICA CONTATO COM CASO POSITIVO                 25/05/2020
6:                                  <NA>                  13/04/2020
```

Conjunto de Dados

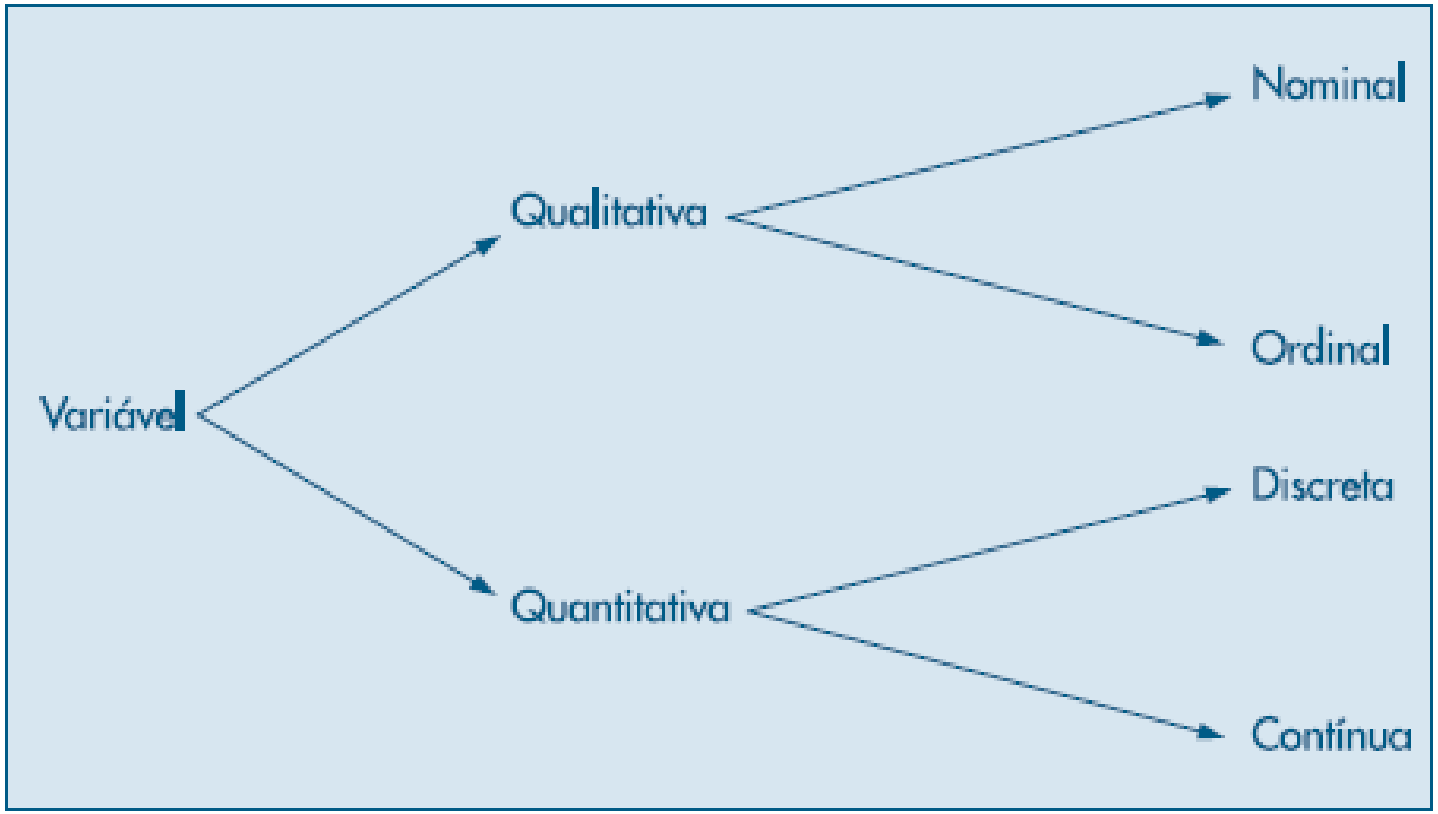
Introdução

- A **análise descritiva** do conjunto de dados é a **fase mais importante** no processo de análise de dados.

- Muitas vezes, uma boa análise descritiva é **suficiente** para o **responder** às perguntas de pesquisa.

- Cerca de 80% do tempo do Cientista de Dados é gasto com a fase de pré-processamento e análise descritiva dos dados.

Tipos de Variáveis



Sexo, Estado, etc..

Nível de escolaridade, Classe social

Número de filhos, Idade

Altura, Peso

Tipos de Variáveis

Nome da variável

Tipo da variável

```
> str(df)
Classes 'data.table' and 'data.frame':  5837 obs. of  29 variables:
 $ DATA DA NOTIFICACAO                                  chr  "07/06/2020" "05/06/2020" "16/06/2020" "30/06/2020
 ...
 $ DOR DE GARGANTA                                      chr  "SIM" "NAO" "SIM" "NAO" ...
 $ DISPNEIA                                             chr  "NAO" "NAO" "NAO" "NAO" ...
 $ FEBRE                                                chr  "NAO" "SIM" "NAO" "NAO" ...
 $ TOSSE                                                chr  "NAO" "NAO" "SIM" "NAO" ...
 $ OUTROS                                               chr  "NAO" "SIM" "SIM" "SIM" ...
 $ E PROFISSIONAL DE SAUDE?                             chr  "SIM" "NAO" "SIM" "NAO" ...
 $ DESCRICAO DO SINTOMA                                 chr  NA "DOR TORACICA" "CEFALEIA" "CEFALEIA" ...
 $ DATA DO INICIO DOS SINTOMAS                          chr  "03/06/2020" "01/06/2020" "15/06/2020" "30/06/2020
 ...
 $ DOENCAS RESPIRATORIAS CRONICAS DESCOMPENSADAS        chr  "NAO" "NAO" "NAO" "NAO" ...
 $ DOENCAS CARDIACAS CRONICAS                           chr  "NAO" "NAO" "NAO" "NAO" ...
 $ DIABETES                                             chr  "NAO" "NAO" "NAO" "NAO" ...
 $ DOENCAS RENAIS CRONICAS EM ESTAGIO AVANCADO (GRAUS 3, 4 OU 5)  chr  "NAO" "NAO" "NAO" "NAO" ...
 $ IMUNOSSUPRESSAO                                      chr  "NAO" "NAO" "NAO" "NAO" ...
 $ GESTANTE DE ALTO RISCO                               chr  "NAO" "NAO" "NAO" "NAO" ...
```

Tipos de Variáveis

- Cada **tipo** de **variável** exige uma **técnica** de análise **diferente**.
- **Relações entre** as **variáveis** também precisam ser avaliadas.
- Os tipos de análises dividem-se em três:
  - **Análise univariada:** cada variável é avaliada individualmente.
  - **Análise bivariada**: relações existentes entre 2 variáveis são avaliadas.
  - **Análise multivariada:** a complexidade resultante da multiplicidade das variáveis é avaliada.

Análise Univariada

- **Variável qualitativa nominal**:
  - Tabela de frequências (absolutas e/ou relativas)
  - Gráfico de setores
  - A moda, i.e., o valor que ocorre com maior frequência

- **Variável qualitativa ordinal**:
  - Tabela de frequências (absolutas e/ou relativas)
  - Gráfico de barras

Análise Univariada

- **Variável quantitativa discreta**:
  - Frequencias absolutas e relativas
  - Gráfico de barras
  - Medidas de posição: moda, mediana, média, etc.
  - Medidas de dispersão: desvio padrão, quartis, máximo, mínimo, amplitude, coeficiente de variação, etc.

- **Variável quantitativa contínua**:
  - Histograma
  - Boxplot
  - Medidas de posição
  - Medidas de dispersão
  - Variável pode ser categorizada e analisada com técnicas para variáveis qualitativas

Distribuições de Frequências

- **Distribuições** de **frequência** possibilitam o entendimento do **comportamento** das **variáveis**.

- Distribuição de frequência para a variável “Classificação Final”:

```
> kable(cbind(freq,prop),col.names = c("Freq","Prop"))
```

	Freq	Prop
CONFIRMADO CLINICO-EPIDEMIOLOGICO	14	0.24
CONFIRMADO LABORATORIALMENTE	960	16.45
CONFIRMADO SOROLOGIA	2	0.03
CONFIRMADO TESTE RAPIDO	557	9.54
DESCARTADO	2358	40.40
DESCARTADO LABORATORIALMENTE	1027	17.59
SUSPEITO	919	15.74

Distribuições de Frequências

- **Variáveis quantitativas** devem ser **agrupadas** antes da construção de distribuições de frequências.

- Distribuição de frequência para a variável “Idade em Anos”:

```
> kable(cbind(freq,prop),col.names = c("Freq","Prop"))
```

	Freq	Prop
18-24	270	4.81
25-34	1167	20.79
35-44	2022	36.02
45-54	1260	22.45
55 +	894	15.93

Gráficos para Variáveis Qualitativas

- Existem diversas variações de **gráficos** para representar **variáveis qualitativas**.

- Os dois tipos de gráficos principais são:

  - **Gráficos em barras**

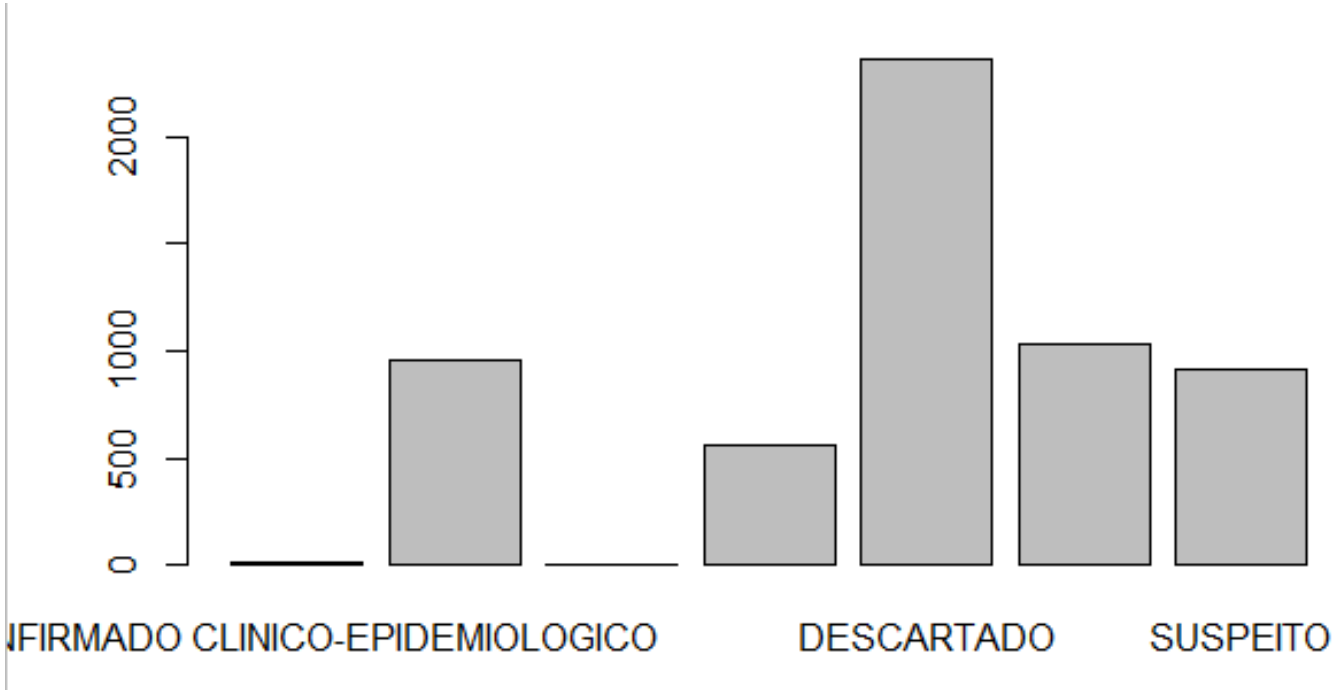
  - **Gráficos em setores (“pizza”)**

Gráficos para Variáveis Qualitativas

Gráficos em Barras

- Gráfico em barras para a variável “Classificação Final”

```
barplot(table(df$`CLASSIFICACAO FINAL`))
```



Construindo gráficos no R com ggplot2

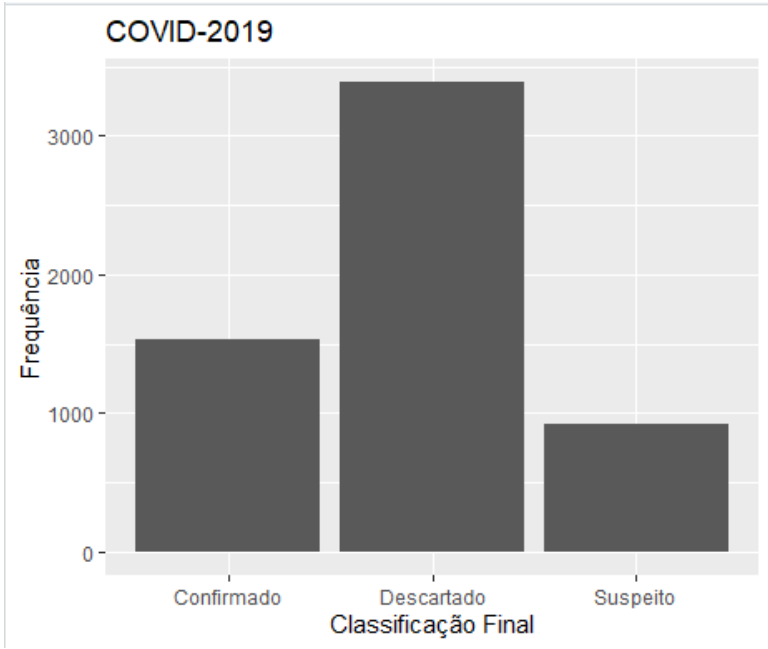
- Gráficos são construídos por camadas
- A camada base do gráfico é criada com a função ggplot()
- Camadas são adicionadas com um ‘+’
- A função aes(), responsável por descrever como as variáveis serão mapeadas nos aspectos visuais do gráfico

Gráficos para Variáveis Qualitativas

Gráficos em Barras

- Gráficos mais elegantes com *ggplot2:*

```
ggplot(df, aes(x=classificacao_final)) +
  geom_bar() +
  labs(x="Classificação Final", y="Frequência", title="COVID-2019")
```

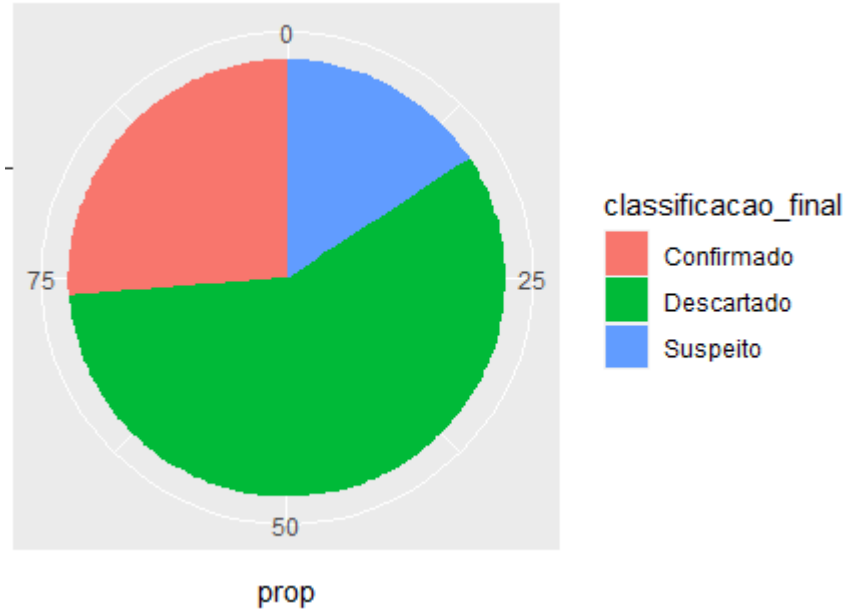


Gráficos para Variáveis Qualitativas

Gráficos em Setores

```
# Preparando o conjunto de dados
freq = table(df$classificacao_final)
prop = round(prop.table(table(df$classificacao_final))*100,2)
df_pie = data.table(cbind(freq,prop))
df_pie[,classificacao_final := levels(df$classificacao_final)]

# Construindo o gráfico
ggplot(df_pie, aes(x='',y=prop,fill = classificacao_final)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0)
```



Exercício

- Para o dataset de óbitos, escolha uma variável quantitativa e faça a análise com as técnias aprendidas até o momento.

Gráficos para Variáveis Quantitativas

- Existe uma variedade maior de **gráficos** para análises de **variáveis quantitativas**.

- Podem-se **categorizar** as variáveis quantitativas e utilizar as **mésmas técnicas** de análises utilizadas nas **variáveis qualitativas**.

- Exemplos:

  - **Histograma**

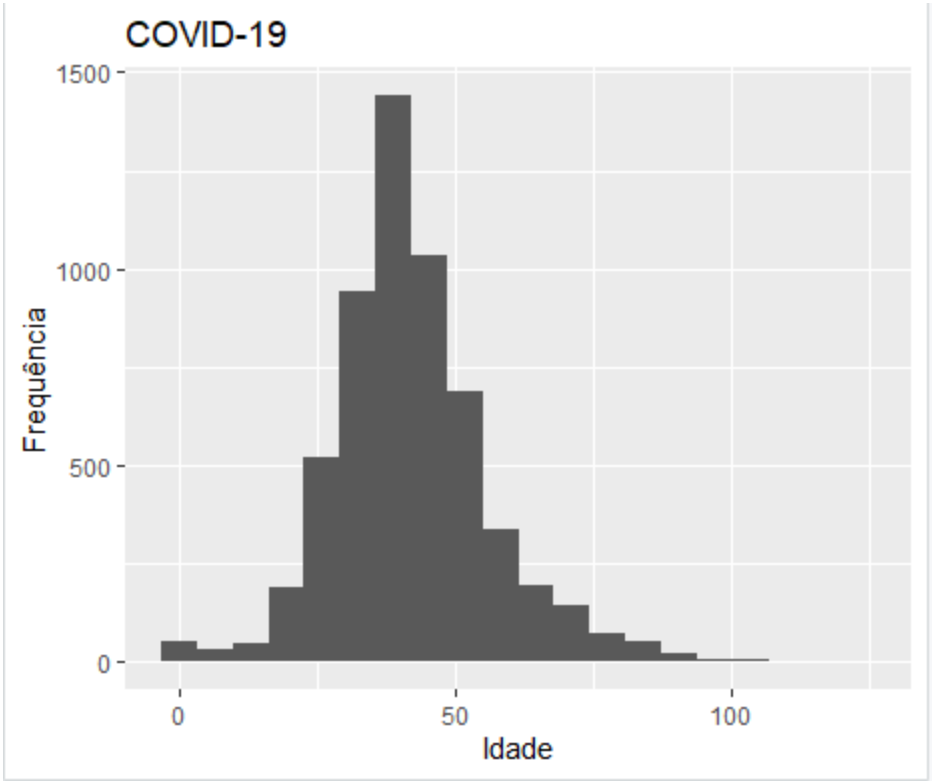
  - **Gráfico de dispersão**

  - **Boxplot (veremos após estudo das medidas resumo)**

Gráficos para Variáveis Quantitativas

Histograma

```
ggplot(df[!is.na(idade_em_anos)], aes(x=idade_em_anos)) +
  geom_histogram(bins=20) +
  labs(x="Idade", y="Frequência", title="COVID-19")
```

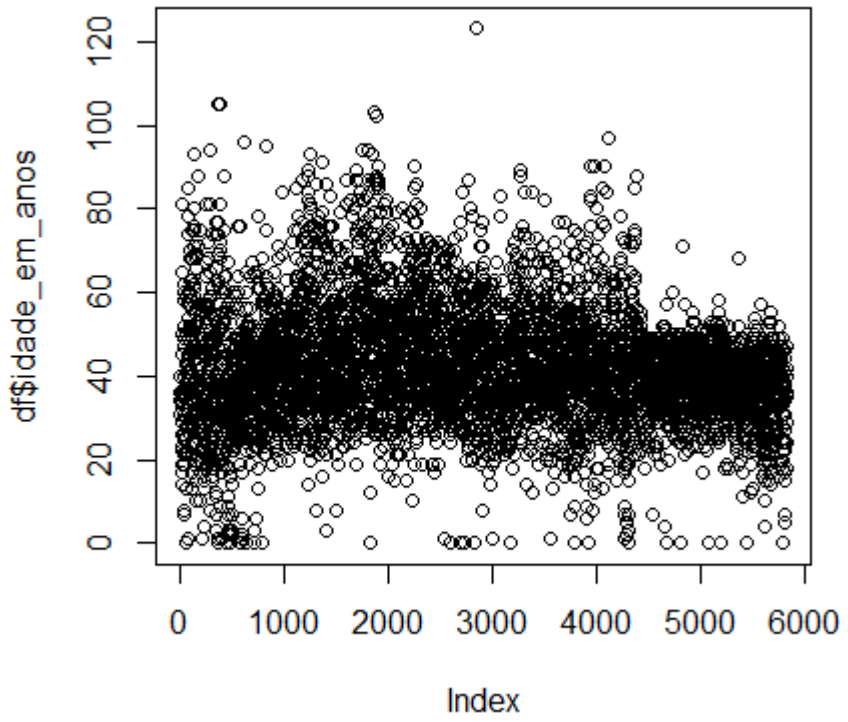


Gráficos para Variáveis Quantitativas

Gráfico de dispersão

- Gráficos de dispersão são mais informativos quando avaliando relações entre variáveis.

```
plot(df$idade_em_anos)
```



Medidas Resumo

Medidas de Posição

- **Resumir** os dados graficamente fornece uma maneira mais **fácil** de **interpretar** o **comportamento** de uma variável.

- É possível resumir ainda mais estes dados, apresentando um ou alguns **valores** que sejam *representativos* da **série toda**.

- As seguintes medidas de posição central são normalmente utilizadas:
  - **Média**
  - **Mediana**
  - **Moda**

Medidas Resumo

Medidas de Posição - Moda

- A ***moda*** é definida como a realização **mais freqüente** do conjunto de valores observados.

- Em alguns casos, pode haver mais de uma moda, ou seja, a distribuição dos valores pode ser **bimodal**, **trimodal** etc.

Medidas Resumo

Medidas de Posição - Mediana

- A ***mediana*** é a realização que ocupa a **posição central** da série de observações, quando estão ordenadas em ordem crescente. Ou seja, a mediana deixa **metade** dos dados **abaixo** dela e **metade acima**.

- Exemplo:

x<- c(3, 4, 7, 8, 8) **Número Ímpar**

median(x) = 7

x<- c(3, 4, 7, 8, 8, 9) **Número par**

median(x) = (7+8)/2 = 7,5

Medidas Resumo

Medidas de Posição - Média

- A ***média aritmética***, é a soma das observações dividida pelo número delas.

```
x<- c(3, 4, 7, 8, 8)
```

mean(x) = (3 + 4 + 7 + 8 + 8)/5 = 6

Medidas Resumo

- Qual a idade mediana dos individuos testados para o covid-19?
- E qual a idade média?
- Quais as idades média e mediana entre os testados positivos e negativos para o covid-19?

Medidas Resumo

Medidas de Posição - Dispersão

- O resumo de um conjunto de dados por uma única medida representativa de **posição central** esconde toda a informação sobre a **variabilidade** do conjunto de observações.

- Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

  - grupo A (variável *X*): 3, 4, 5, 6, 7
  - grupo B (variável *Y*): 1, 3, 5, 7, 9
  - grupo C (variável *Z*): 5, 5, 5, 5, 5
  - grupo D (variável W): 3, 5, 5, 7
  - grupo E (variável V): 3, 5, 5, 6, 6

- A **média** para todos os grupos é **igual** a 5,0.

Medias Resumo

Medidas de Dispersão

- A média dos grupos não informa suas diferentes variabilidades.
- O princípio básico de variabilidade é analisar os **desvios** das observações em **relação** à **média** dessas observações.
- Para o grupo A, os desvios em relação a média são: –2, –1, 0, 1, 2

grupo A : 3, 4, 5, 6, 7

Média = (3+4+5+6+7)/5 = 5

- É fácil ver que, para *qualquer* conjunto de dados, a soma dos desvios é igual a zero.

(3-5) + (4-5) + (5-5) + (6-5) + (7-5) = 0

Medias Resumo

Medidas de Dispersão

- Nestas condições, a soma dos desvios não é uma boa medida de dispersão para o conjunto A.

- Duas medidas são as mais usadas:

  - **Desvio padrão**

$$s = \sqrt{\frac{\sum_{i}^{N}(X - \overline{X})^2}{N-1}}$$

  - **Variância**

$$s^2 = \frac{\sum_{i}^{N}(X - \overline{X})^2}{N-1}$$

Medias Resumo

Medidas de Dispersão

FÓRMULAS

- **Amplitude Total** = max(n) – min(n)
- **Variância (var)** = $\frac{\sum(x_i - \overline{x})^2}{n}$
- **Desvio Padrão (σ)** = $\sqrt{\frac{\sum(x_i - \overline{x})^2}{n}}$
- **Coeficiente de Variação** = $\frac{\sigma}{\overline{x}}$

O coeficiente de variação (CV) é ajustado de modo que os valores estão em uma escala sem unidade.

CV pode ser utilizado para comparar a variação nos dados que tem unidades diferentes ou que tem médias muito diferentes.

Medias Resumo

Quantis Empíricos

- **Média** e **desvio padrão** podem não ser medidas adequadas para representar um conjunto de dados, pois:

- São afetados, de forma exagerada, por **valores extremos**;

- Apenas com estes dois valores não temos idéia da **simetria** ou **assimetria** da distribuição dos dados.

Medias Resumo

Quantis Empíricos

- **q(p):** Quantil de ordem ***p*** ou ***p-quantil***
  - ***p*** é uma proporção qualquer, 0 < ***p*** < 1, tal que 100p% das observações sejam menores do que **q(p)**.

- Indicamos, abaixo, alguns quantis e seus nomes particulares.

  - ***q*(0,25)** = ***q*1**: 1° Quartil = 25° Percentil
  - ***q*(0,50)** = ***q*2**: Mediana = 2° Quartil = 50° Percentil
  - ***q*(0,75)** = ***q*3**: 3° Quartil = 75° Percentil
  - ***q*(0,40)**: 4° Decil
  - ***q*(0,95)**: 95° Percentil

Medias Resumo
Quantis Empíricos

- Os cinco valores, ***mínimo***, ***q*1**, ***q*2**, ***q*3** e ***máximo*** são importantes para se ter uma boa idéia da **assimetria** da distribuição dos dados.

  - **q2** – **mínimo** é chamada ***dispersão inferior***
  - **máximo** – **q2** é a ***dispersão superior***

- Estas duas dispersões devem ser **aproximadamente iguais**, para uma distribuição **aproximadamente simétrica**.

Medias Resumo
Quantis Empíricos

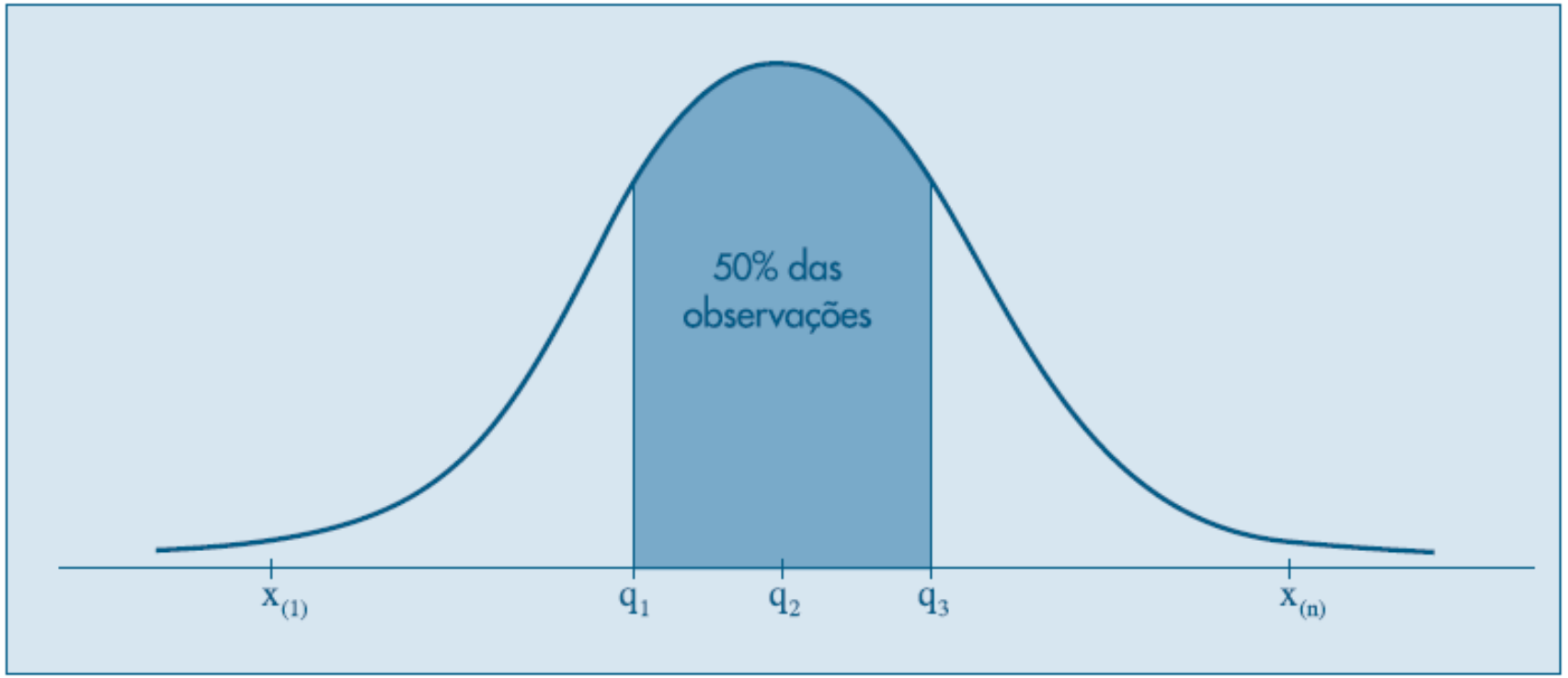
- *Uma distribuição simétrica deve apresentar as seguintes características*

  - $q2 - x(1) \approx x(n) - q2$
  - $q2 - q1 \approx q3 - q2$
  - $q1 - x(1) \approx x(n) - q3$
  - Distâncias entre **mediana** e $q1, q3$ menores do que distâncias entre os extremos e $q1, q3$.

Medias Resumo

Quantis Empíricos

- A chamada *distribuição normal* ou *gaussiana* possui as características de simetria



Gráficos para Variáveis Quantitativas

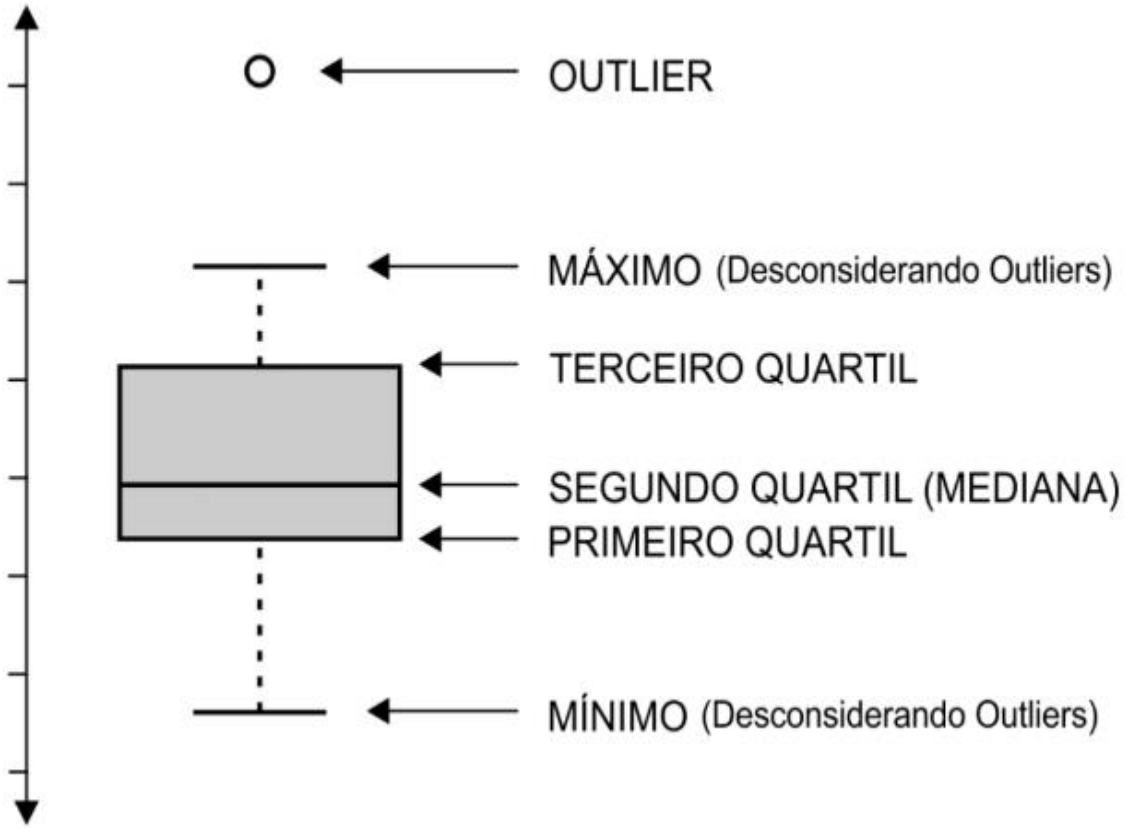
Boxplot

- Os **principais quantis** empíricos podem ser traduzidos graficamente pelo *box plot*.

- O boxplot fornece evidência acerca da:
  - **Posição**
  - **Dispersão**
  - **Assimetria**
  - **Valores extremos (atípicos)**

MÁXIMO = Q1 + 1,5(Q3 – Q1)

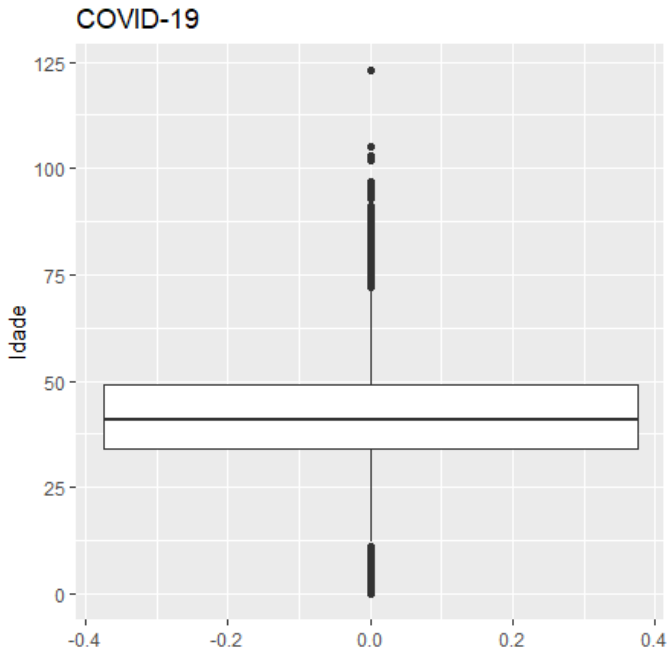
MÍNIMO = Q1 – 1,5(Q3 – Q1)



Gráficos para Variáveis Quantitativas

Boxplot

```
ggplot(df[!is.na(idade_em_anos)], aes(y=idade_em_anos)) +
  geom_boxplot() +
  labs(x="", y="Idade", title="COVID-19")
```



Exercícios

- Para os dados de óbitos, faça a análise da variável “Idade” com as técnicas para variável quantitativa aprendidas até o momento.