

Estatística Básica e Introdução ao R

Prof^a. Dra. Natalia Giordani

Conteúdos da disciplina

- Súmula e Conteúdo programático
- Método de Avaliação
 - Presença e participação nas aulas
 - Entrega de atividades **em formato de relatório com interpretações relacionadas aos resultados**

Ciência de Dados e Estatística

- Ciência de Dados
 - Analisar grandes conjuntos de dados (megadados)

SERÁ??

- Estatística
 - Pequenos conjuntos de dados (microdados)

Ciência de Dados e Estatística

- Ciência de Dados
 - Estatística + Ciência da Computação
 - Estatística: guia a coleta e análise de dados complexos
 - Ciência da Computação: desenvolve algoritmos que, por exemplo, distribuem grandes conjuntos de dados por múltiplos processadores

Blei e Smyth, 2017

Ciência de Dados e Estatística

- “A ciência de dados contempla **modelos estatísticos** e **métodos computacionais** para resolver **problemas** específicos **de outras disciplinas**, entender o domínio desses problemas, decidir quais **dados** obter, como **processá-los**, **explorá-los** e **visualizá-los**, selecionar um **modelo estatístico e métodos computacionais** apropriados, além de **comunicar os resultados** da análise de forma inteligível para aqueles que propuseram os problemas.”

Blei e Smyth, 2017

Ciência de Dados e Estatística

- Ciência de dados é multidisciplinar
 - Problema a ser resolvido
 - Conjunto de dados, meios para sua obtenção e organização
 - Especificação do problema em termos de variáveis desse conjunto de dados
 - Descrição dos dados
 - Escolha das técnicas e algoritmos necessários para resolução do problema e implementação das técnicas
 - Apresentação dos resultados

Preparação dos Dados

- Para que coletar dados?
 - Obter informações
 - De quem?
 - População (Ex.: censo demográfico)
 - Amostra (Ex.: pesquisa nacional por amostras de domicílios - PNAD)
- Como?
 - Estudos observacionais (Ex.: registros de atendimento de SAC)
 - Estudos amostrais (Ex.: pesquisa de opinião)
 - Estudos experimentais (Ex.: teste de nova funcionalidade em app; ensaios clínicos)

Preparação dos Dados

- Dados

- Valores de um conjunto de **variáveis** obtidos pela observação de unidades de investigação (constituem uma amostra de uma população)
 - Unidades de investigação = onde as variáveis são observadas
- Exemplo: estudo em que se pretende avaliar relação entre motivo da reclamação e setor da empresa
 - Unidade de investigação = clientes
 - Variáveis a serem observadas = motivo da reclamação e setor da empresa relacionado

Preparação dos Dados

- Análise de dados de uma amostra -> Inferência
- Análise Exploratória de Dados
 - Organização e resumo dos dados (população ou amostra)

Preparação dos Dados

- Abordagem estatística para tratamento de dados
 - Planejamento da forma de coleta de dados considerando objetivos do estudo
 - Organização de tabela para coletar/armazenamento dos dados
 - Resumo dos dados através de tabelas e gráficos,
 - Identificação de possíveis erros de coleta e/ou digitação
 - Proposta de métodos de análise que respondam aos objetivos do estudo
 - Avaliação do ajuste dos métodos (técnicas de diagnóstico)
 - Tradução dos resultados em termos não técnicos

Adaptado de Morettin e Singer, 2022

Preparação dos Dados

- Tabela de dados
 - Matrizes onde se armazenam dados com o objetivo de permitir a realização de análises
 - Cada linha: uma unidade de investigação
 - Cada coluna: uma variável
 - Etapa importante é a construção do dicionário de dados: definição das variáveis; atribuição de rótulos; especificação de unidades de medida; especificação, quando pertinente, de limites

Preparação dos Dados

Exemplo de dados PNAD

Linha	ano	trimestre	id_uf	sigla_uf	capital	rm_ride	id_upa	id_estrato	id_domicilio
5	2016	1	12	AC	null	null	120005116	1250020	1200051160305
6	2016	1	52	GO	null	null	520068788	5252011	5200687881004
7	2016	1	52	GO	null	52	520048290	5220011	5200482900904
8	2016	1	21	MA	null	22	210057824	2140010	2100578240105
9	2016	1	32	ES	null	null	320058410	3252011	3200584101304
10	2016	1	17	TO	null	null	170008098	1752020	1700080980604
11	2016	1	13	AM	null	null	130013080	1352021	1300130801305
12	2016	1	11	RO	11	null	110000034	1110011	1100000341405
13	2016	1	42	SC	42	42	420057440	4210012	4200574400704
14	2016	1	42	SC	null	null	420049133	4253011	4200491330404
15	2016	1	42	SC	null	null	420081478	4253021	4200814780604
16	2016	1	21	MA	null	null	210008436	2153012	2100084360405
17	2016	1	25	PB	25	25	250030407	2510013	2500304071304
18	2016	1	26	PE	26	26	260035889	2610011	2600358890304
19	2016	1	51	MT	null	null	510022290	5153011	5100222901404
20	2016	1	43	RS	null	43	430094333	4321011	4300943330304
21	2016	1	24	RN	null	null	240017190	2452011	2400171900104
22	2016	1	50	MS	null	null	500008407	5052011	5000084070104

Resultados por página: 200 1 – 100 de 100



Software R

- Software livre
 - Disponível para download em <https://cran.r-project.org/>
- RStudio
 - Ambiente de desenvolvimento integrado
 - Disponível para download em <https://posit.co/download/rstudio-desktop/>
- Material apoio
 - [Curso R](#)
 - [R Markdown](#)

Software R

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

exemplos_hist.R

```
1 set.seed(35486) # Create random distributions
2 data <- data.frame(x1 = rbeta(1000, 10, 2),
3                   x2 = rbeta(1000, 5, 2),
4                   x3 = rnorm(1000),
5                   x4 = rbeta(1000, 2, 5),
6                   x5 = rbeta(1000, 2, 10),
7                   x6 = rnorm(1000, 0, 1))
8 head(data) # Print head of data
9
10 par(mfrow= c(2,2))
11 hist(data$x5, main = "Assimetria à direita")
12 hist(data$x1, main = "Assimetria à esquerda")
13 hist(data$x6, main = "Assimetria nula")
14 |
```

Environment History Connections Tutorial

R Global Environment

Data

data 1000 obs. of 6 variables

Files Plots Packages Help Viewer

Zoom Export Publish

Assimetria à direita

Assimetria à esquerda

Assimetria nula

Frequency

data\$x5

data\$x1

data\$x6

Editor de script

Console

R 4.1.0

```
+ x5 = rbeta(1000, 2, 10),
+ x6 = rnorm(1000, 0, 1))
> head(data) # Print head of data
  x1      x2      x3      x4      x5      x6
1 0.8219216 0.7678832 -0.4384896 0.25835433 0.12587011 0.8617332
2 0.9048126 0.7316123 -0.1695237 0.27571009 0.05149536 -0.8429724
3 0.8913348 0.6553369 0.1337007 0.21426531 0.25931552 -1.0046958
4 0.7111786 0.8620435 2.4944818 0.08016730 0.55864902 0.6278267
5 0.8204075 0.6466904 1.4550030 0.24385567 0.26146987 -0.6005240
6 0.9445704 0.2699652 1.4149316 0.08770332 0.07224937 -1.1551530
> par(mfrow= c(2,2))
```

Console

Software R

- Versões online
 - RStudio Cloud
- Atividade para próxima aula: ter acesso ao software

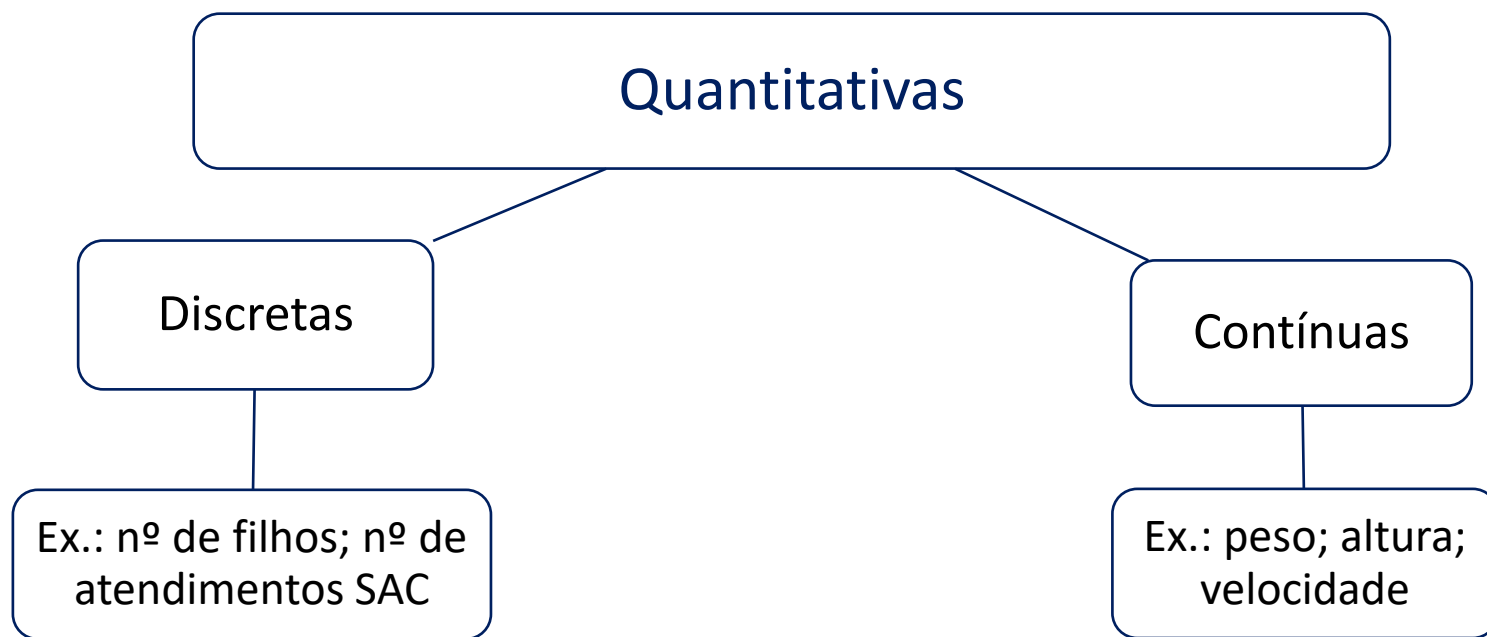
1. Análise exploratória de dados

1.1 Tipos de variáveis

- Numéricas
- Textuais (não numéricas)

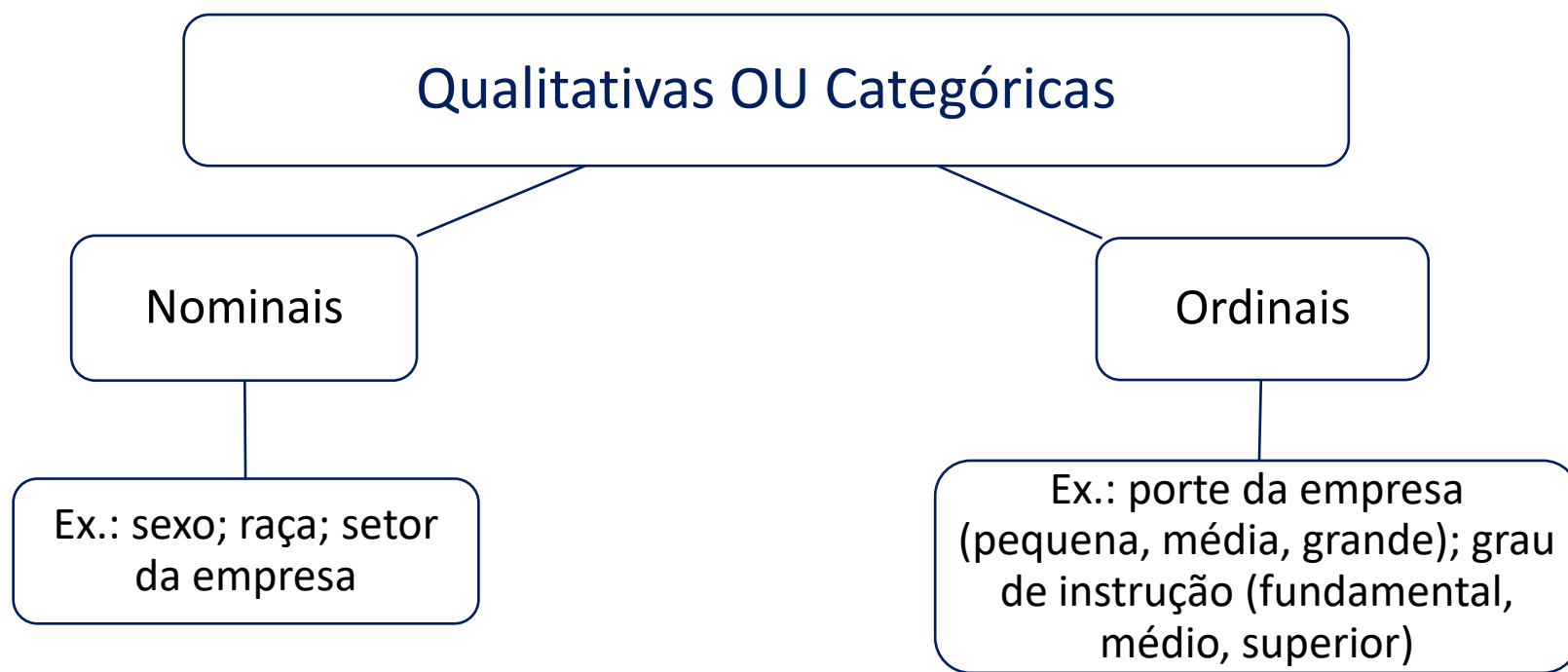
1. Análise exploratória de dados

1.1 Tipos de variáveis



1. Análise exploratória de dados

1.1 Tipos de variáveis



1. Análise exploratória de dados

1.1 Tipos de variáveis



<https://wordwall.net/play/68397/536/458>

1. Análise exploratória de dados

1.2 Análise de dados de uma variável

- Primeira etapa de uma análise de dados
 - Resumi-los
- Como?
 - Vai depender do tipo de variável



1.2.1 Análise exploratória de uma variável qualitativa

1. Distribuição de frequências

- Quantidade de cada categoria
 - Absoluta = número de unidades observadas
 - Relativa = porcentagem correspondente
- Gráficos
 - Barra
 - Pizza (será?)

1.2.1 Análise exploratória de uma variável qualitativa

■ Exemplo: Campeonato Brasileiro 2023

Temporada 2023 ▼					
Clube	Pts	PJ	VIT	E	DER
1  Palmeiras	70	38	20	10	8
2  Grêmio	68	38	21	5	12
3  Atlético-MG	66	38	19	9	10

1. Que informações essa tabela apresenta?
2. Como estão dispostos os dados que originaram essa tabela?

1.2.1 Análise exploratória de uma variável qualitativa

- Exemplo: Campeonato Brasileiro 2023

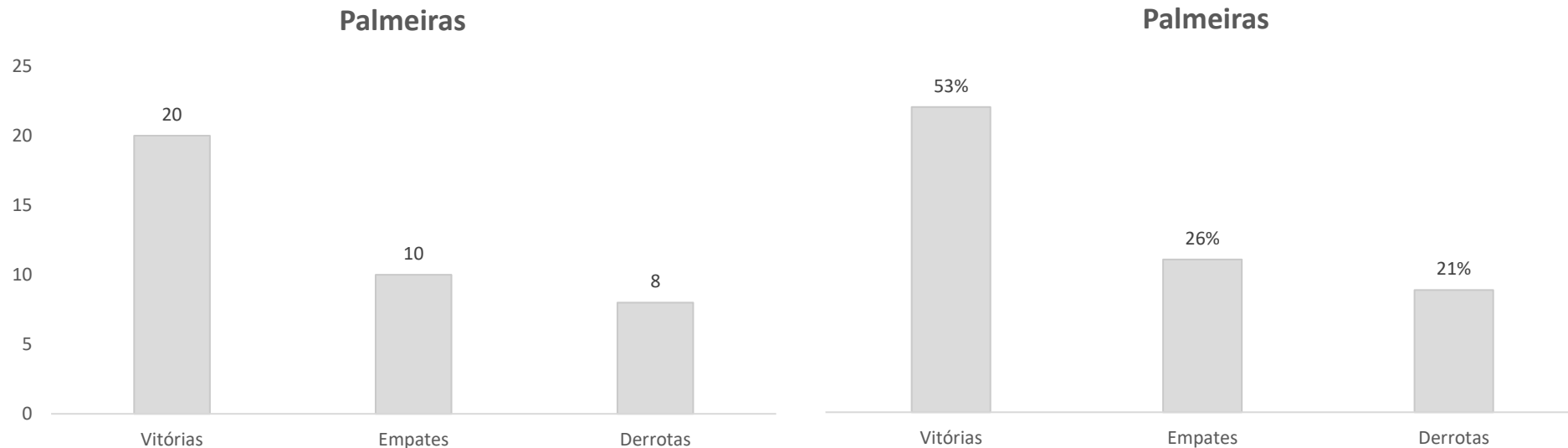
4. Quais seriam as frequências relativas?

Time	% Vitórias	% Empates	% Derrotas
Palmeiras	53%	26%	21%

1.2.1 Análise exploratória de uma variável qualitativa

- Exemplo: Campeonato Brasileiro 2023

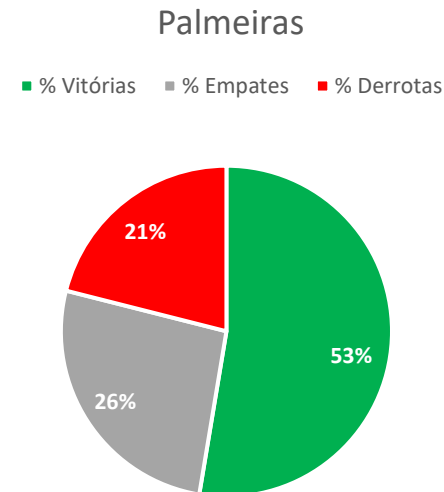
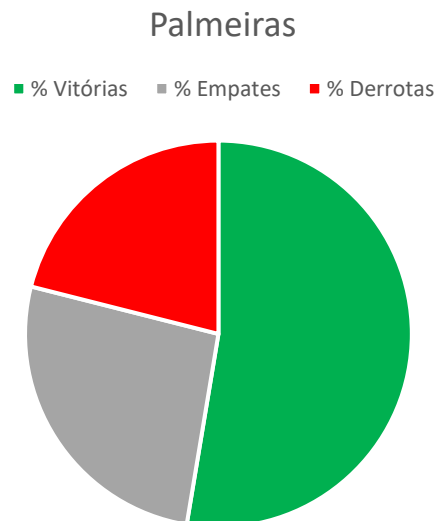
5. Quais as possibilidades de representação gráfica?



1.2.1 Análise exploratória de uma variável qualitativa

■ Exemplo: Campeonato Brasileiro 2023

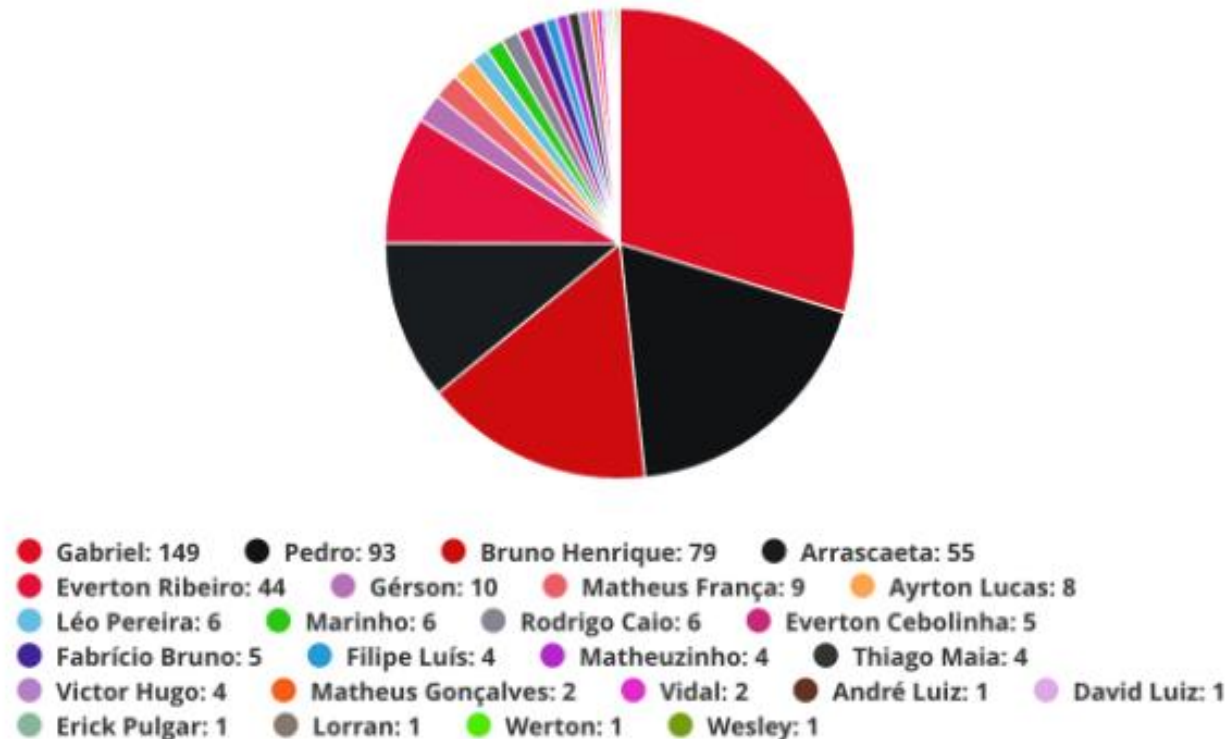
5. Quais as possibilidades de representação gráfica?



1.2.1 Análise exploratória de uma variável qualitativa

- Gráfico de pizza...

Os 501 gols do elenco atual pelo Flamengo



Leitura complementar sugerida

- [Tipos de dados no R](#)
- [Livro Curso R](#)

Para próxima aula...

- Providenciar acesso ao software R