

# Aprenda Estatística

Giseldo da Silva Neo



# Contents

<b>1</b>	<b>Introdução</b>	<b>5</b>
1.1	Estimativas de localização . . . . .	5
1.2	Tipo do Dado (ou caracterização dos dados - numérico, categórico)	7
1.3	Tipo do Atributo (Preditor, Alvo) . . . . .	10
<b>2</b>	<b>Análise exploratória dos dados</b>	<b>11</b>
2.1	Introdução . . . . .	11



# Chapter 1

## Introdução

É importante começarmos pelos conceitos básicos para criar um vocabulário consistente e permitir uma comunicação clara entre os interessados em discutir e aprender estatística e análise de dados. Estatística e análise de dados é uma habilidade crucial no processo de extração de informação para tomada de decisão. A quantidade de dados disponível para consulta vem crescendo exponencialmente e analisar estes dados tem sido útil para diversas organizações.

### 1.1 Estimativas de localização

Muitas vezes é conveniente representar um conjunto de números de uma forma mais simples. Nem sempre temos a possibilidade de lidar com vários números, por limitação ou por falta de praticidade. Por exemplo, imagine uma sala de aula com 5 estudantes, vamos montar uma lista da idade de todos os estudantes nessa sala no R e no Python, duas linguagens de programação comumente utilizadas em análise de dados.

Código R

```
idades <-c(14,15,16,14,17)
idades
```

```
## [1] 14 15 16 14 17
```

Código Python

```
idades = [14, 15, 16, 14, 17]
print(idades)
```

```
## [14, 15, 16, 14, 17]
```

Podemos representar essa lista com um número mais simples, que pode resumir ou representar aquela lista original. Para isso, utilizamos as **estimativas de localização** [Bruce et al., 2020]. As mais comuns são **média** e **mediana**.

### 1.1.1 Média

A média é calculada dividindo a soma de todos os números da lista pela quantidade de itens. Sua fórmula matemática é apresentada em FIGURA XXX. Onde  $i$  é a quantidade de itens da lista e  $x_i$  é o  $i$ -ésimo item da lista. O termo média também pode ser representado pelo símbolo  $\bar{X}$

No nosso exemplo se fossemos calcular manualmente a média da lista **idade**, o cálculo seria:

Código R

```
( 14 + 15 + 16 + 14 + 17 ) / 5
```

```
## [1] 15.2
```

Código Python

```
print(( 14 + 15 + 16 + 14 + 17 ) / 5)
```

```
## 15.2
```

Porém, podemos utilizar algumas funções que já disponibilizam esse recurso de calcular a média. O código para criar uma lista e verificar a média dessa lista, utilizando as funções, no R e no Python, seria o seguinte:

Código R

```
idades <- c(14, 15, 16, 14, 17)
mean(idades)
```

```
## [1] 15.2
```

Código Python

```
from statistics import mean
idades = [14, 15, 16, 14, 17]
print(mean(idades))
```

## 1.2. TIPO DO DADO (OU CARACTERIZAÇÃO DOS DADOS - NÚMÉRICO, CATEGÓRICO)7

## 15.2

A função **mean**, no R, recebe como parâmetro uma lista de itens e retorna a média dessa lista, no python utilizei a função mesmo nome, porém disponível na biblioteca statistics do python.

## 1.2 Tipo do Dado (ou caracterização dos dados - numérico, categórico)

É necessário classificar o tipo do dado (também chamado de variável) pois os algoritmos de aprendizagem de máquina, ou os modelos estatísticos, irão funcionar com determinados tipos e de determinadas formas. Com o conhecimento do tipo do dado que estamos lidando poderemos realizar as conversões ou tratamentos adicionais que forem necessários.

O tipo do dado pode ser **numérico**, **categórico**, texto ou data. Vide figura abaixo. Dados do tipo numérico e categórico, que são os comumente utilizados em estatística e aprendizagem de máquina serão melhor detalhados nas seções seguintes.

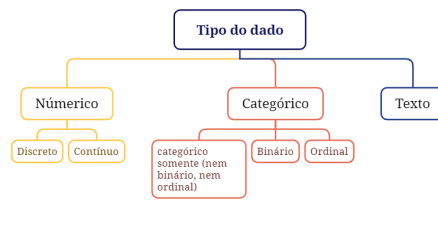


Figure 1.1: Tipo do dado

Dado do tipo **numérico** é expresso geralmente expresso como um número inteiro ou real. Porém, existem casos em que números inteiros também expressam dados **categóricos**. Já o dado do tipo **categórico** está relacionado a um valor dentro de um lista (geralmente finita) de valores. A formação acadêmica de uma determinada pessoa (Ensino Fundamental, Médio ou Superior), por exemplo, é um dado do tipo **categórico**. Já o salário, é um dado do tipo **numérico**.

### 1.2.1 Numérico (ou quantitativo)

O tipo do dado **numérico** ainda pode ser sub classificado como **numérico contínuo** ou **numérico discreto**.

Um dado **numérico contínuo** é quando o dado pode ser qualquer número em um intervalo de números reais - lembrando que o conjunto de número reais engloba os números inteiros -. Geralmente é o resultado de uma medida, por exemplo, a altura dos estudantes é um dado do tipo **numérico contínuo**.

O dado **numérico discreto** geralmente é resultado de uma contagem - um número inteiro -, por exemplo, a idade é uma contagem de anos do estudante, é um dado **numérico discreto**.

Na nossa lista de idades, a idade é um dado do estudante do tipo **numérico discreto**.

### 1.2.2 Categórico (ou qualitativo)

Um dado é do tipo **categórico** representa um valor dentro de um conjunto ou de uma categoria.

O dado **categórico** pode ser **categórico binário** ou **categórico ordinal**, ou nenhuma das duas subcategorias, ou seja **categórico somente**.

Um exemplo de **dado categórico somente**, é uma lista com as cores preferidas dos estudantes (azul, verde, vermelho), ou o estado civil de uma pessoa (casado, solteiro).

O dado do tipo **categórico binário** é quando ele somente pode assumir dois valores no universo de valores possíveis. Por exemplo 0 ou 1, existente ou ausente, true ou false, sim e não.

O dado do tipo **categórico ordinal** é quando o valor é um elemento de um conjunto que pode ser ordenado, por exemplo, imagine a classificação de altura de estudantes onde somente os valores alto, médio e baixo são permitidos. Nesse exemplo, existe uma ordem, o aluno com altura classificado como baixo tem uma altura menor do que o aluno com altura média.

Variável	Tipo do dado
Idade (Ex: 14, 17, 23)	numérico discreto
Doença (Ausente, Presente)	categórico binário
Story Points (1, 3, 5, 7 ... )	categórico ordinal
Ano (2021, 2022, ...)	numérico discreto
Altura (1,79 - 2,05 - ...)	numérico contínuo
Estado Civil (Casado, Solteiro)	categórico binário
Cores preferidas (Azul, verde, vermelho)	categórico somente (nem binário, nem ordinal)

Os dados geralmente são organizados em formato de tabelas. Onde as linhas representam as observações (ou instâncias) e as colunas representam as variáveis.



## 1.2. TIPO DO DADO (OU CARACTERIZAÇÃO DOS DADOS - NÚMÉRICO, CATEGÓRICO)9

Vamos utilizar o exemplo de uma empresa que desenvolve software e registra os dados relacionados a seus projetos. Essa empresa mantém o registro de determinada funcionalidade e do tamanho dessa funcionalidade. Cada linha da tabela representa uma funcionalidade (chamada de User Story em projetos que utilizam SCRUM). Cada coluna representa uma informação dessa User Story. As informações que a empresa mantém registro são as variáveis, as colunas da tabela. Uma dessas variáveis é a descrição, outra é uma estimativa que o desenvolvedor atribui do tamanho funcional, chamado Story Point. Essas informações estão dispostas em um arquivo no formato CSV. O código abaixo, carrega esse arquivo e exibe parte de seu conteúdo. Iremos então classificar cada uma das colunas de acordo com o tipo do dado.

Código R

```
df <- read_csv('data/neodataset/7764.csv')
```

```
## Rows: 355 Columns: 5
## -- Column specification -----
## Delimiter: ","
## chr  (2): title, description
## dbl  (2): issuekey, storypoints
## dtm  (1): created
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(df)
```

```
## # A tibble: 6 x 5
##   issuekey created          title description storypoints
##   <dbl> <dtm>          <chr>      <chr>          <dbl>
## 1 29688087 2020-01-17 00:50:48 Update templates for web- "Relates t~          1
## 2 29682716 2020-01-16 19:21:38 Make sure that we Captur~ "This was ~          1
## 3 29644971 2020-01-15 21:17:03 Propose new IA for Brand~ "## Goals\~          1
## 4 29494181 2020-01-10 19:20:50 Cache `node_modules` for~ "# UPDATE ~          1
## 5 29437529 2020-01-09 10:26:51 Disable all remaining un~ "Similar t~          1
## 6 29358963 2020-01-07 08:35:44 Disable unnecessary jobs~ "As discus~          1
```

Código Python

```
import pandas as pd
#pd.set_option('max_columns', None)
df = pd.read_csv('data/neodataset/7764.csv')
df.head()
```

```
##      issuekey  ... storypoints
## 0  29688087  ...             1
## 1  29682716  ...             1
## 2  29644971  ...             1
## 3  29494181  ...             1
## 4  29437529  ...             1
##
## [5 rows x 5 columns]
```

Nome da Coluna	Tipo do dado	Observação
Issuekey	categorico somente	Apesar de ser um número, não são realizadas operações no número, ele é um identificador único da User Story
stroypoints	numérico discreto	É um número geralmente de 1 á 100
created	data	Data em que a User Story Foi criada
title	texto	Título da User Story
description	texto	Desrição da User Story

A tabela acima apresenta a caracterização dos dados do conjunto de dados neo-dataset (esse conjunto de dados pode ser baixado em ...). Nessa tabela foram tipificados os dados. É interessante apresentar essa tipificação em estudos científicos e trabalhos de conclusão de curso, quando estamos lidando com conjuntos de dados. Cabe ressaltar que essa tipificação independe da linguagem. Internamente cada linguagem de programação tem seus tipos específicos e que podem ter pequenas diferenças entre as linguagens.

### 1.3 Tipo do Atributo (Preditor, Alvo)

Nos modelos de aprendizagem de máquina (quando lidamos com algoritmos classificados como supervisionados) e de inferência estatística o dado também pode ser classificado entre atributo preditor ou atributo alvo. Atributo preditor, são os atributos que serão utilizados para realizar a previsão, geralmente um ou mais atributos. Atributo alvo é o atributo que queremos ‘advinhar (ou dar o melhor chute técnico)’ com os modelos preditivos. Atributo preditor muitas vezes é chamado de vairável independente, e atributo alvo de variável dependente.

## Chapter 2

# Análise exploratória dos dados

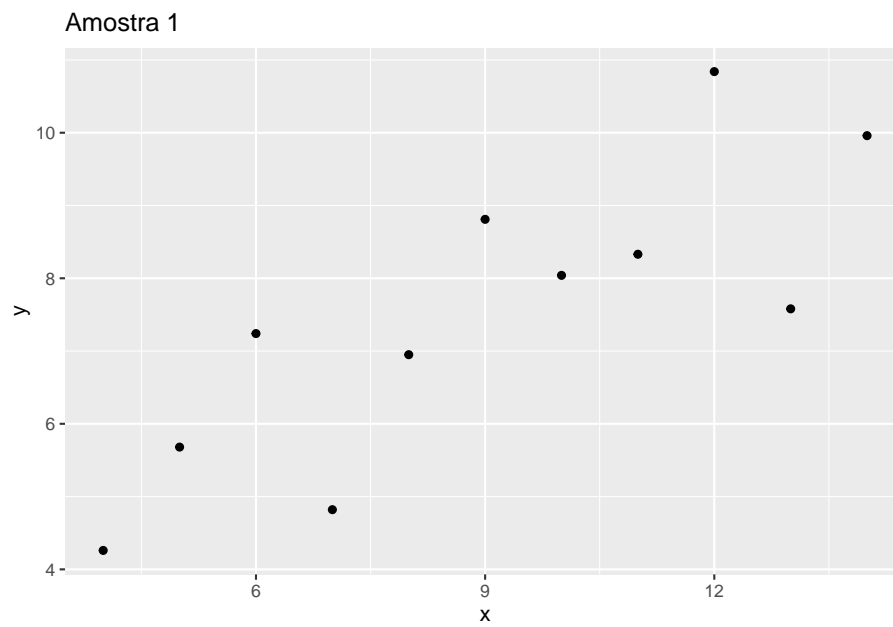
### 2.1 Introdução

Um dos pioneiros na definição da área de análise exploratória de dados (em inglês *Statistical Data Analysis*, ou EDA) foi Tukey (1997) [Tukey et al., 1977]. Tukey (1997) argumenta que seu foco, até aquele momento, estava em desenvolver novas técnicas para inferência. Porém, depois de reflexão, ele chega a conclusão de que o foco dele, e de outros estatísticos, seria melhor aplicado no desenvolvimento de técnicas para a etapa de preparação desses dados. Era nos procedimentos de estruturar os dados que estava o verdadeiro desafio. Problemas, tais como, lidar com dados faltantes ou *outliers*, traziam impactos negativos na inferência e novas técnicas nessa etapa precisavam ser estudadas. Sua recomendação era uma mudança de paradigma e novos estudos, voltando mais para a preparação dos dados. Sua visão é de que isso iria trazer enorme avanços como um todo. O que de fato aconteceu.

Podemos considerar essa necessidade de estudos anterior ao processo de inferência analisando o exemplo criado por Ancobe.

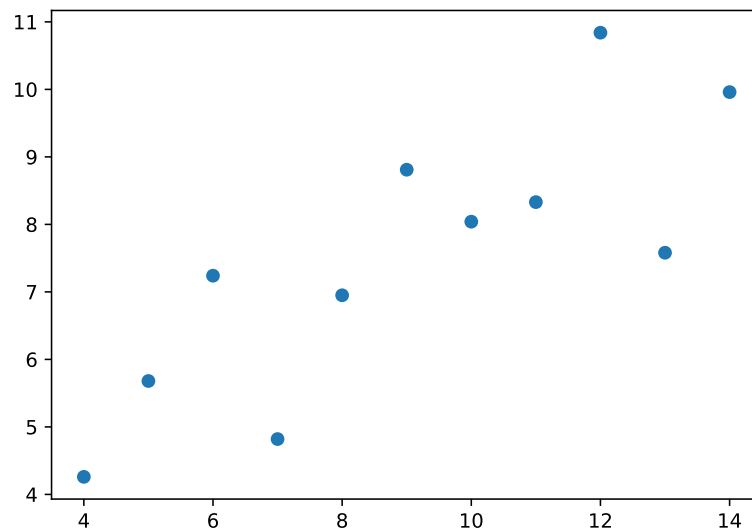
R

```
x <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
y <- c(8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68)
DT = data.table(x, y)
ggplot(DT, mapping = aes(x = x, y = y)) +
  geom_point() +
  labs(title = "Amostra 1")
```



Python

```
import matplotlib.pyplot as plt
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
plt.scatter(x, y)
plt.show()
```



Veja a imagem “Amostra 1” acima. Nela visualmente percebemos uma relação linear direta entre as duas variáveis, podemos confirmar isso analisando o gráfico de pontos e o valor da correlação, abaixo.

R

```
cor(x, y)
```

```
## [1] 0.8164205
```

Python

```
from statistics import correlation
print(correlation(x, y))
```

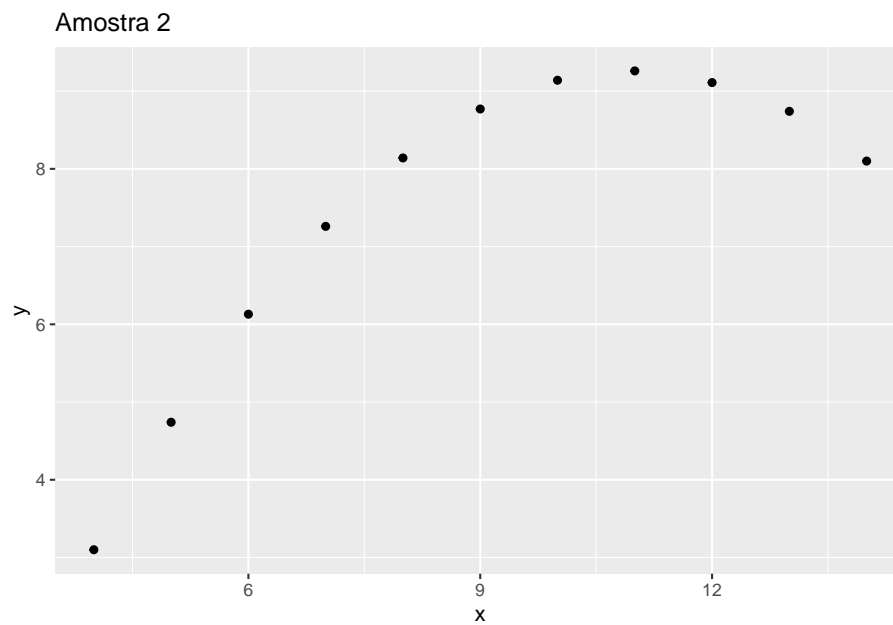
```
## 0.81642051634484
```

Para os dados da Amostra 1, acima temos uma correlação de 0.81.

R

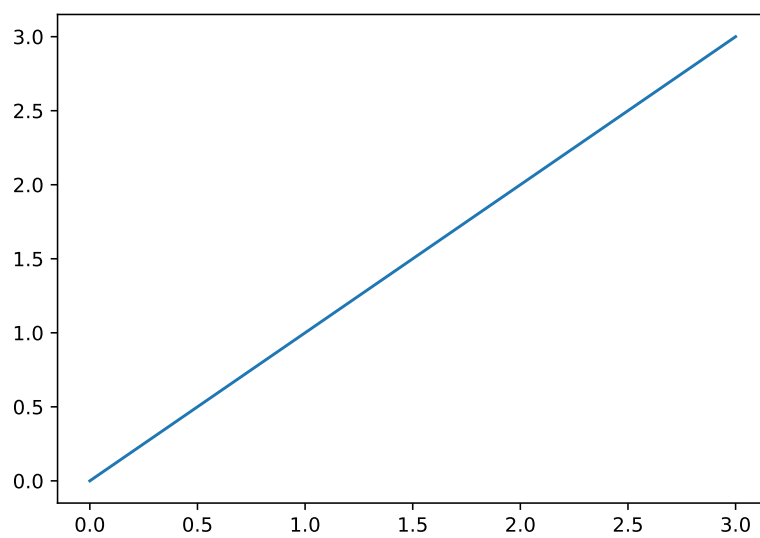
```
x <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
y <- c(9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.11, 7.26, 4.74)
DT = data.table(x, y)
```

```
ggplot(DT, mapping = aes(x = x, y =y)) +  
  geom_point() +  
  labs(title = "Amostra 2")
```



Python

```
import matplotlib.pyplot as plt  
plt.plot([0, 1, 2, 3])  
plt.show()
```



Na amostra 2 percebemos uma relação em forma de curva, quando verificamos a correlação verificamos o mesmo valor de 0.81 dos dados da amostra 1.





# Bibliography

Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media, 2020.

John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.