Aprenda Estatística

Giseldo da Silva Neo

Contents

1 Introdução **5**

1.1 Estimativas de localização 5

1.2 Tipo do Dado 7

2 Análise exploratória dos dados **11**

2.1 Introdução 11

Chapter 1

Introdução

Acho importante começar pelos conceitos básicos para criar um vocabulário consistente e permitir uma comunicação clara entre os interessados em discutir e aprender estatística, análise de dados, ciência de dados e outras denominações dessa nova competência. A quantidade de dados disponível vem crescendo exponencialmente e analisá-los tem sido útil para diversas organizações. Logo, a habilidade em lidar com esses dados é crucial para o profissional que deseja extrair informação útil.

A Harvard Business Review, revista conceituada em administração e negócios, em uma matéria de opinião afirmou no título que este é um dos empregos mais sexy deste século 21 e exemplificou o conceito do que é um ser um cientista de dados [Review, 2012]. Porém, acredito que outras pessoas podem ter visões diferentes do que é um emprego sexy, como por exemplo Joel Grus (2016) [Grus, 2016], que acredita que a matéria da Harvard Business foi escrito por alguém que nunca visitou um quartel do corpo de bombeiros. Particularmente, a escolha do que é ser sexy foge do escopo desse livro e me parece uma questão de gosto. Mas concordo com Joel quando ele diz que o campo está em alta e evidência.

Nas análises preditivas são extraídas informações a partir dos dados. Entendemos os dados como fatos brutos, como por exemplo, o nome de um estudante e o número do CPF.

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.

R

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

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

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 enésimo item da lista. O termo média também pode ser representado pelo símbolo X

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

R

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

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

Python

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

```
## 15.2
```

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

R

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

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

Python

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

```
## 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 **mean**, da biblioteca statistics.

1.2 Tipo do Dado

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. Com o conhecimento do tipo do dado que estamos lidando poderemos realizar as conversões necessárias.

O tipo do dado pode ser **numérico** ou **categórico**, ou seja ele pode ser classificado como um dado númerico ou categórico, ou um ou outro.

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 de valores. A formação acadêmcia, por exemplo, é um dado do tipo **categórico**. Já o salário, é um dado to tipo **númerico**.

1.2.1 Numérico

O tipo do dado **numérico** ainda pode ser **numério contínuo** ou **númerico discreto**.

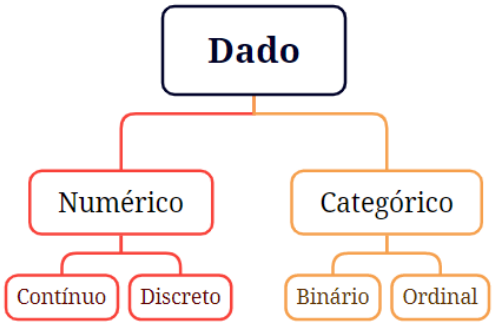


Figure 1.1: Tipo do dado

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

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

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

Um exemplo de **dado categórico**, 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 um tipo especial 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 do dado 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)	categório binário
Story Points (1, 3, 5, 7 …)	categorico 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

Chapter 2

Análise exploratória dos dados

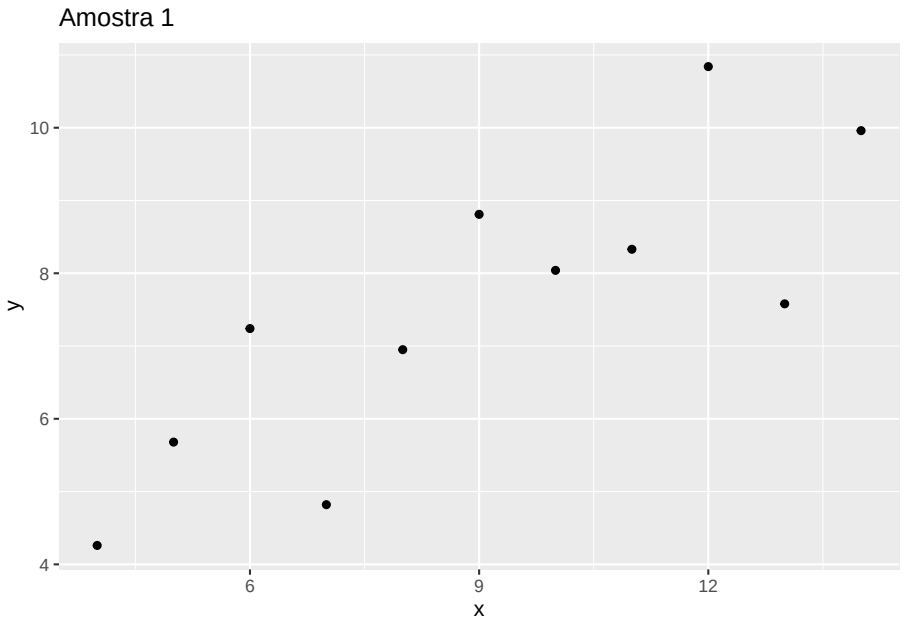
2.1 Introdução

Um dos pioneiros na definição da área de análise exploratória de dados (em inglês Estatistical Data Analisys, 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.

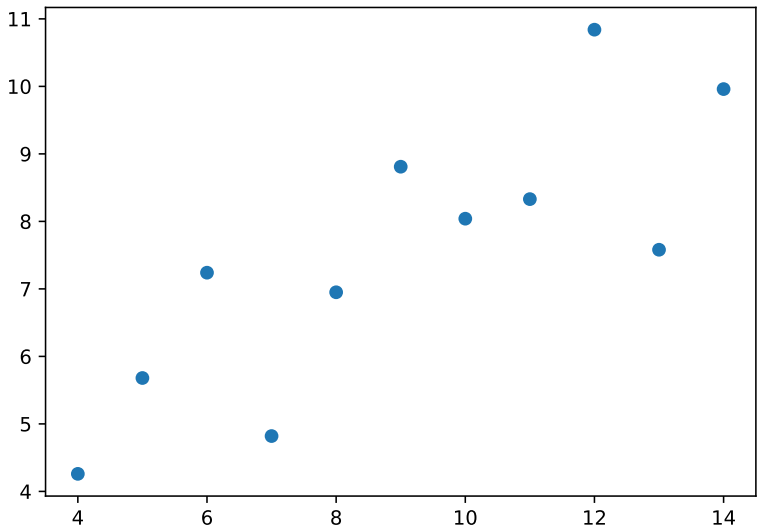
R

```
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))
```

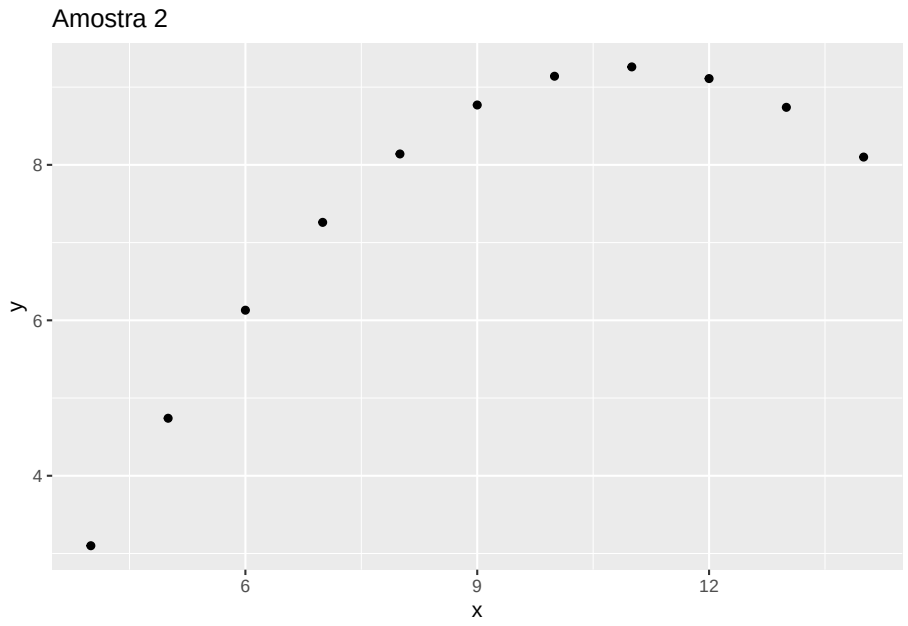
```
## 0.81642051634484
```

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

R

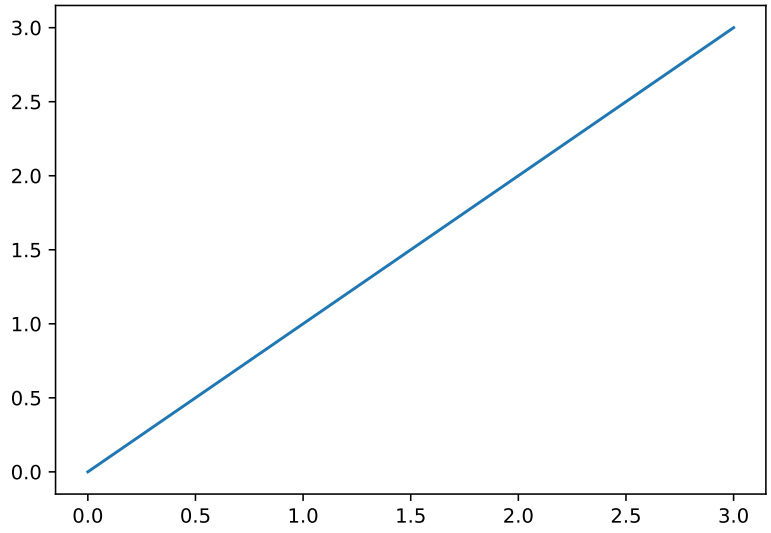
```
x <- c(10, 8, 13, 9, 11, 14, 6, 4 , 12, 7, 5)
y <- c(9.14, 8.14, 8.74,8.77,9.26,8.1,6.13,3.1,9.11,7.26,4.74)
DT = data.table(x, y)
```

```
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.

Joel Grus. *Data science do zero*. Alta books, 2016.

Harvard Business Review, 2012. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

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