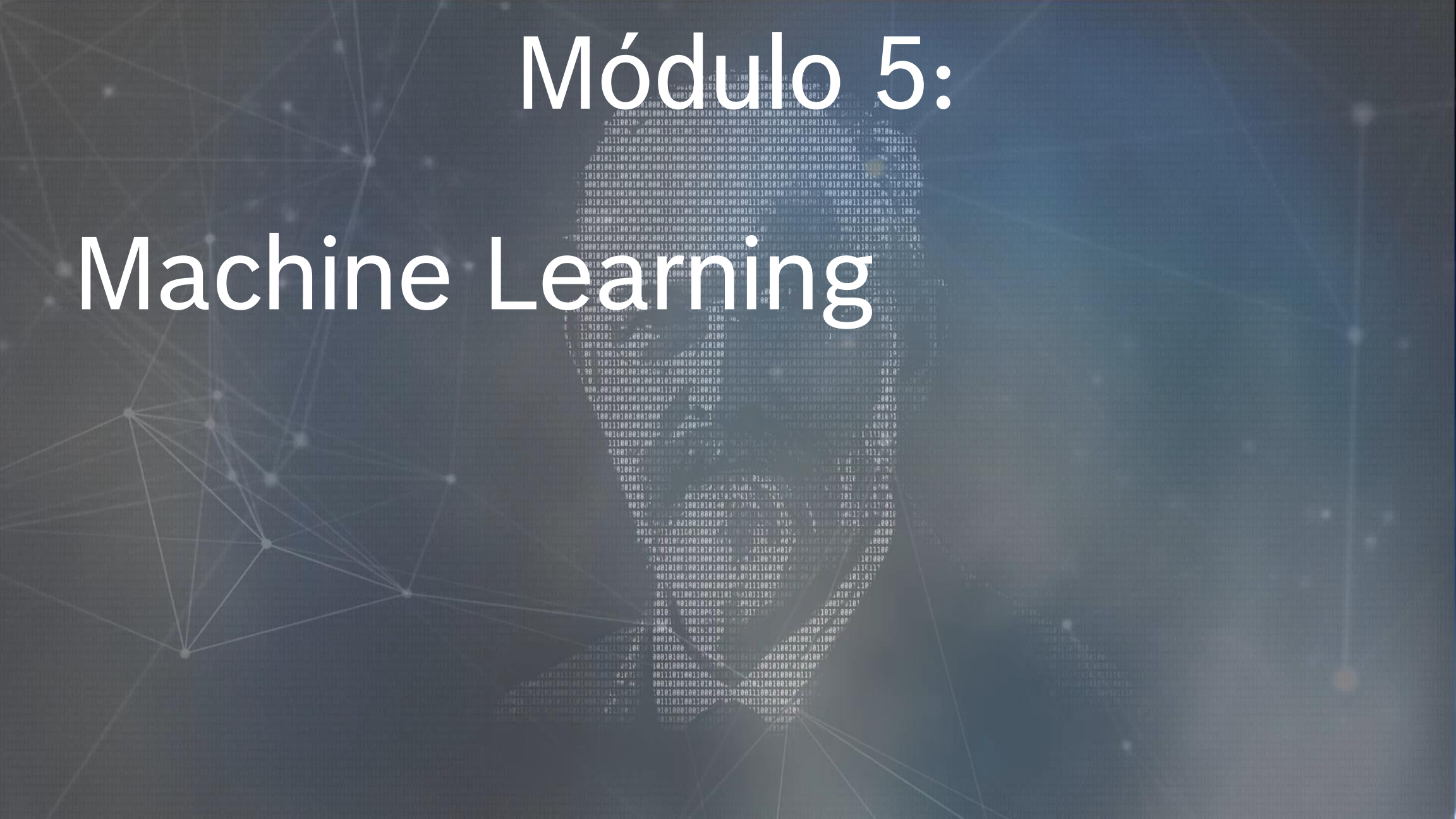
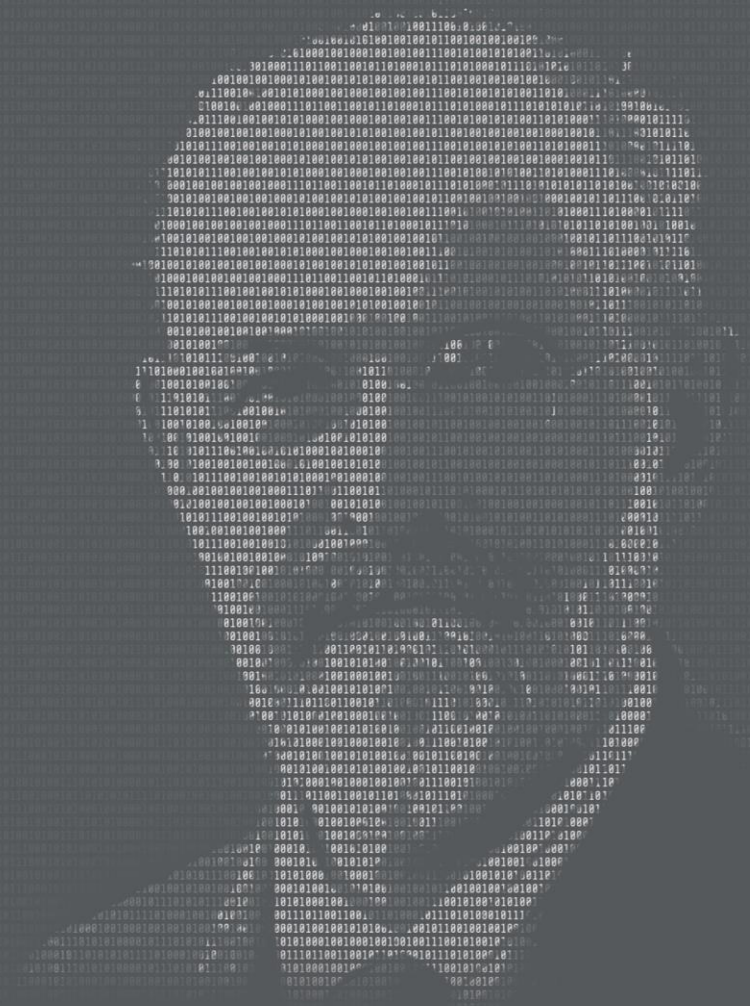
Módulo 5:

Machine Learning



Estatística Básica Para Inteligência Artificial

Estatística descritiva



Estatística Descritiva
O que é Estatística

Sala Digital

- Conjunto de métodos matemáticos usados para se analisar dados;
- A palavra estatística possui ao menos três significados:
  I. Coleção de informações numéricas, ou dados;
  II. Medidas resultantes de um conjunto de dados (média, mediana, etc.);
  III. Métodos usados na coleta e interpretação dos dados.
- Busca realizar pesquisas, colher dados e processá-los, analisar informações e apresentar situações através de gráficos de fácil compreensão;
- Em resumo:

É a parte da ciência responsável pela coleta, organização e interpretação de dados experimentais e pela extrapolação dos resultado da amostra para a população.

Estatística Descritiva
O que é Estatística

Sala Digital

- Estatística descritiva: organizar e representar os dados.
  - Frequências (n,%)
  - Medidas de posição e dispersão (média, mediana, desvio padrão)
- Estatística inferencial: realizar inferências, trabalhar com probabilidades.
  - Amostragens.
  - Medidas de precisão(IC).
  - Testes de hipótese.

Estatística Descritiva
População e Amostra

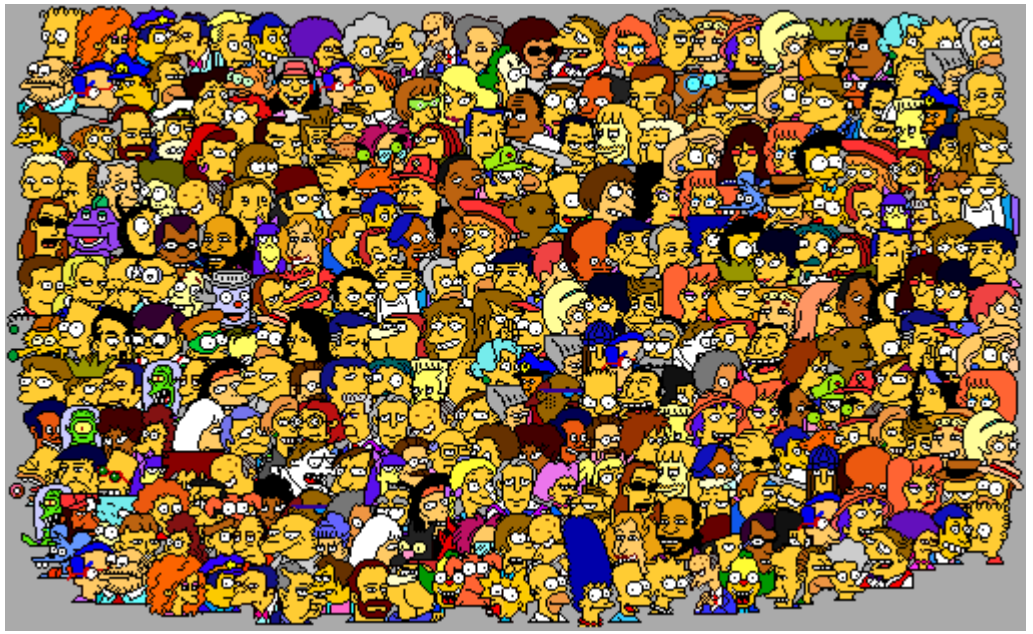
Sala Digital

- **População**: Público-alvo da pesquisa estatística. É o conjunto de pessoas, objetos ou informações, que os dados são coletados e analisados de acordo com o princípio da pesquisa.
  - ***População finita***: Número de elementos do grupo não é muito grande, a entrevista e análise deve envolver todo o grupo; Ex: número de colaboradores numa empresa.
  - ***População infinita***: Número de elementos nesse caso é muito elevado, sendo considerado infinito. Assim a entrevista e a análise devem ser feitas em uma amostra;
- **Amostra**: É um subconjunto da população, ou seja, é uma fração ou parte da população. É um grupo que representa a população. Ex: Número de homens colaboradores em uma empresa.
- Vejamos o exemplo a seguir nos próximos slides.

Estatística Descritiva
População e amostra: Exemplo

Sala Digital

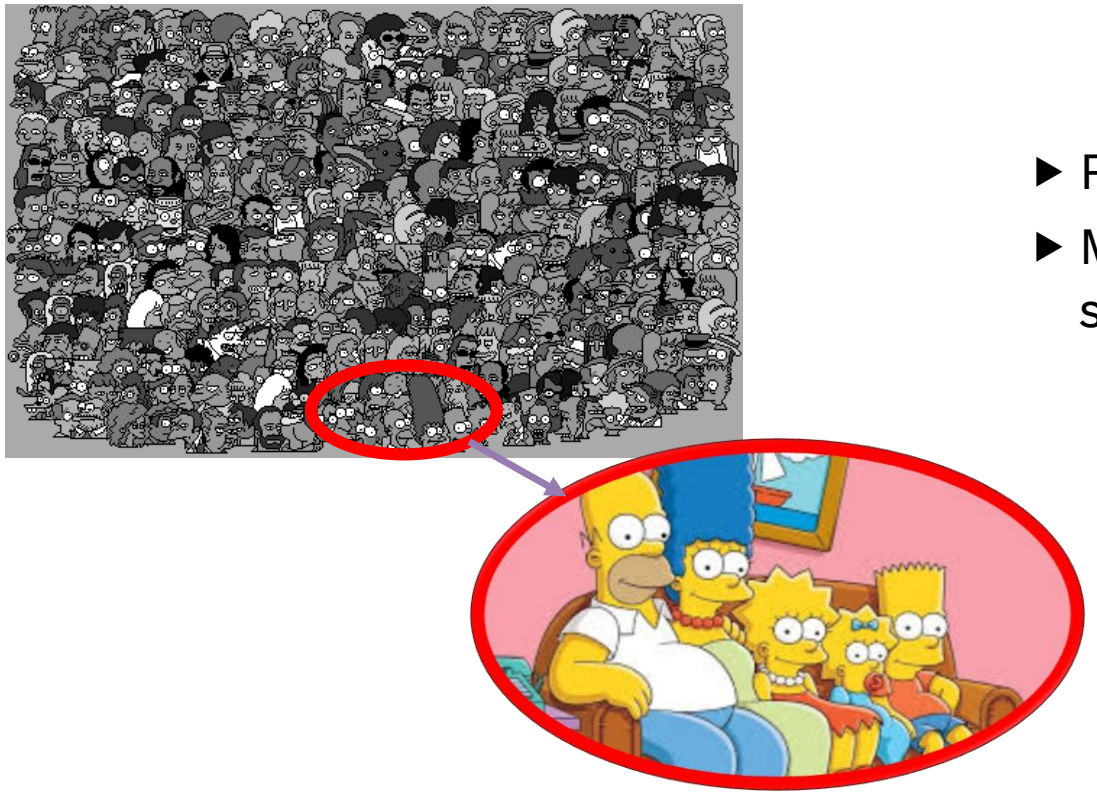
População



- Suponha que queremos estudar as idades de uma população de uma cidade, nesse caso Springfield, da série Os Simpsons;
- Nesse caso, a população é muito grande e demandaria muito tempo e esforço para coletar a idade de todos;
- Precisamos selecionar uma amostra.

Estatística Descritiva
População e amostra: Exemplo

Sala Digital

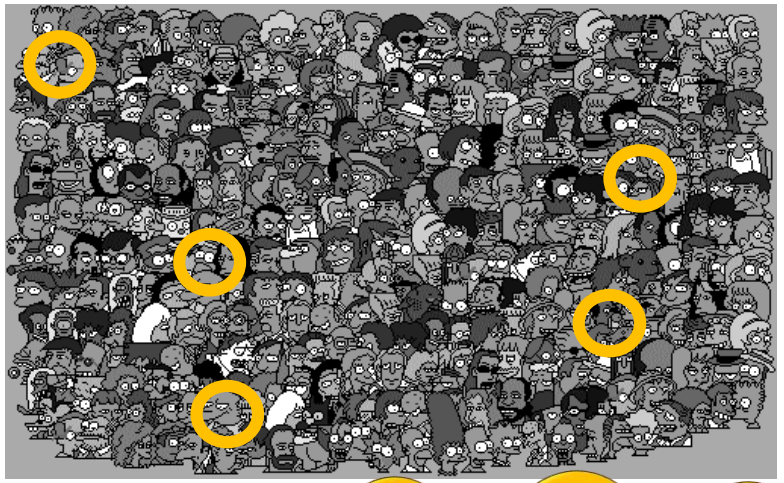


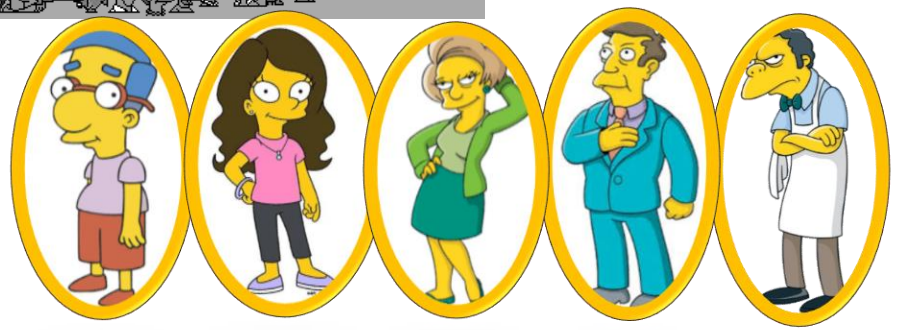
- Poderíamos selecionar apenas uma família;
- Mas talvez o resultado final não fosse o mais satisfatório.

BOSCH

Estatística Descritiva
População e amostra

Sala Digital





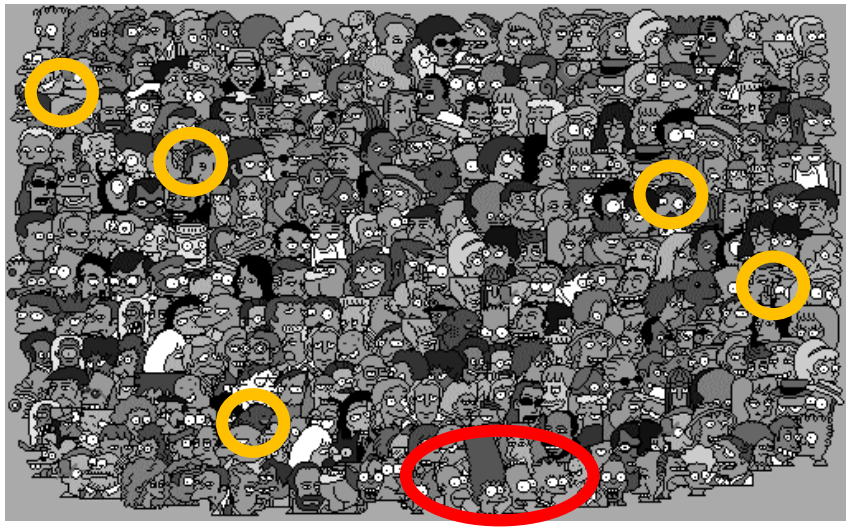
- Poderíamos selecionar pessoas aleatórias;

BOSCH

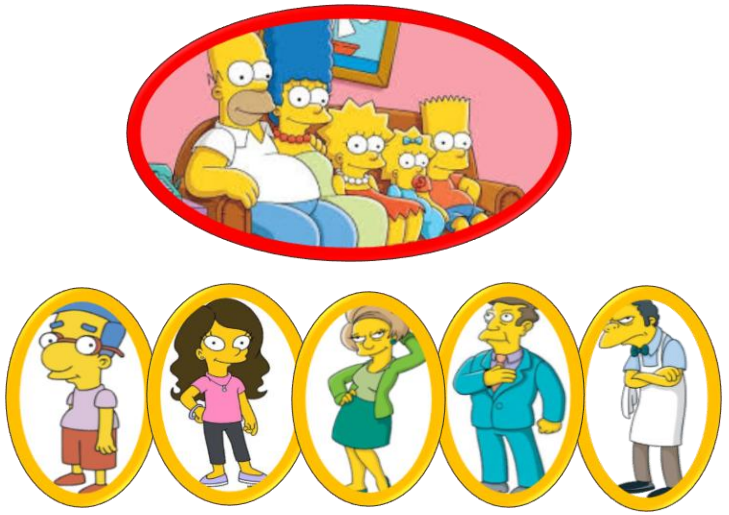
Estatística Descritiva
População e amostra

Sala Digital

População



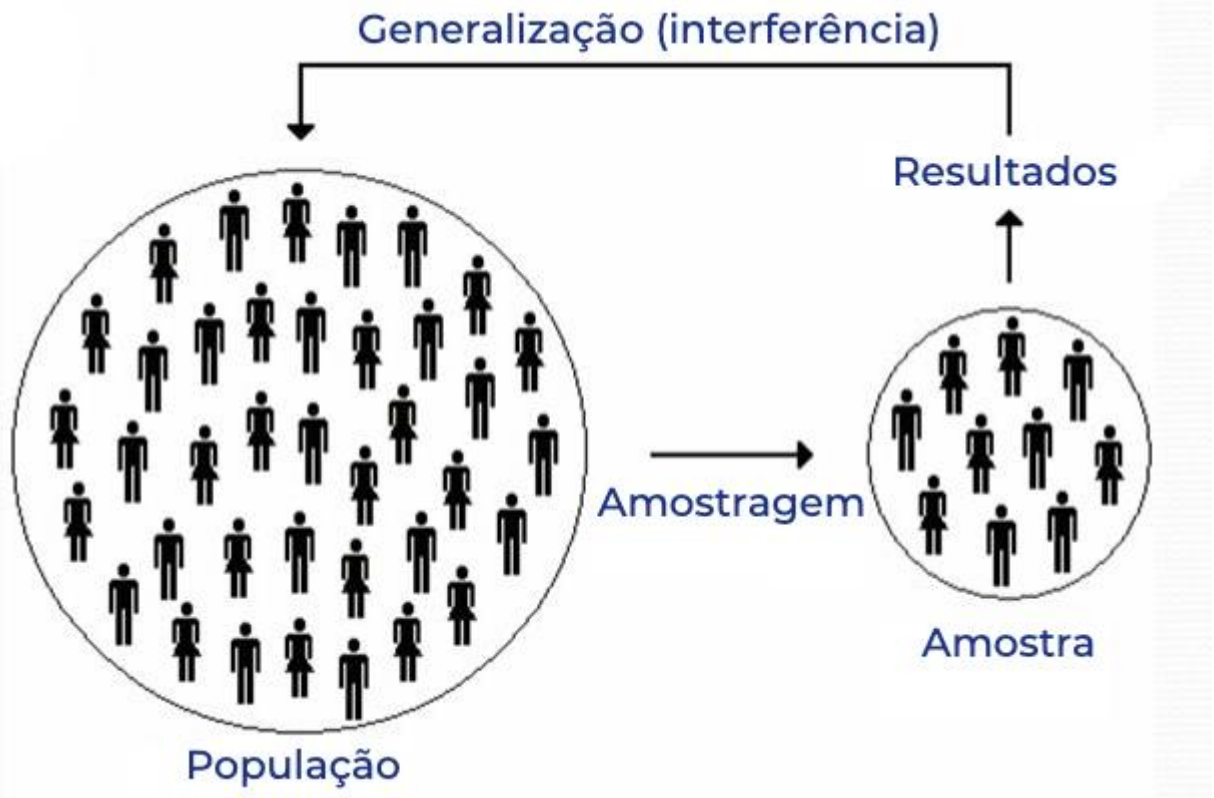
Amostras



- O ideal é que a amostra seja diversa e que o grupo tenha o número suficiente de indivíduos para poder generalizar.

Estatística Descritiva
População e amostra

Sala Digital



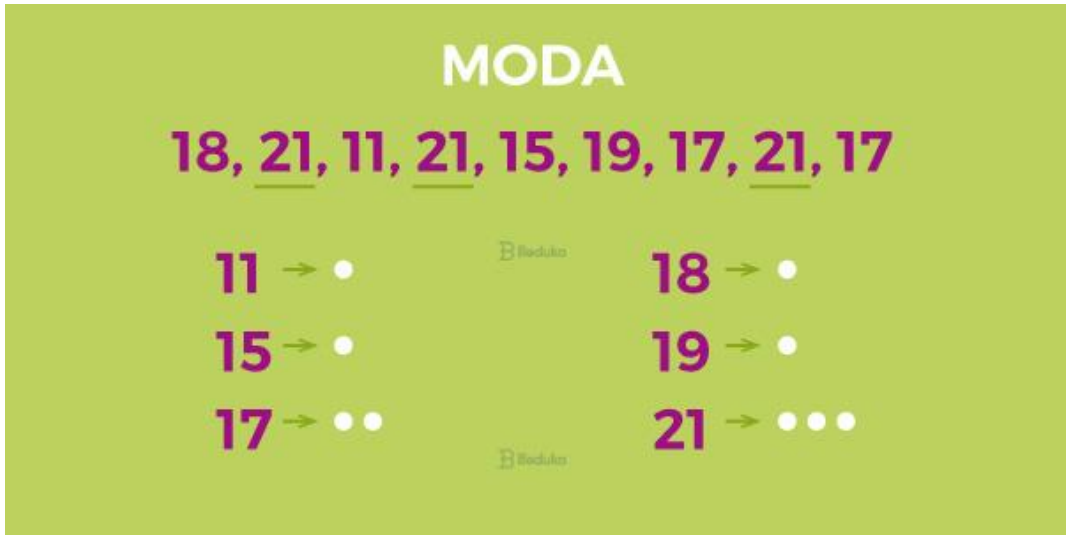
BOSCH

Estatística Descritiva

Medidas de Centralidade: Média, Moda e Mediana

Sala Digital

- São medidas obtidas de um conjunto de dados que podem ser usadas para *representar todo o conjunto*. A tendência dessas medidas é resultar em um *valor central*. Por isso elas são chamadas de *medidas de centralidade*;
- **Moda**: É o dado com o valor mais frequente de m conjunto. Podem existir conjuntos com duas modas (bimodal) ou três modas (trimodal, ou multimodal). A moda pode ser utilizada em dados não numéricos;



BOSCH

Estatística Descritiva

Medidas de Centralidade: Média, Moda e Mediana

Sala Digital

- **Mediana**: É o número que ocupa a posição central na lista de dados (organizada em ordem crescente ou decrescente). Se a lista for par, a mediana vai ser a média dos valores centrais;

$$2, 2, 3, \mathbf{7}, 8, 9, 9$$

$$\text{Mediana} = 7$$

$$1, 4, 4, \mathbf{5}, \mathbf{6}, 7, 7, 7$$

$$\text{Mediana} = (5+6) \div 2$$

$$= 5.5$$

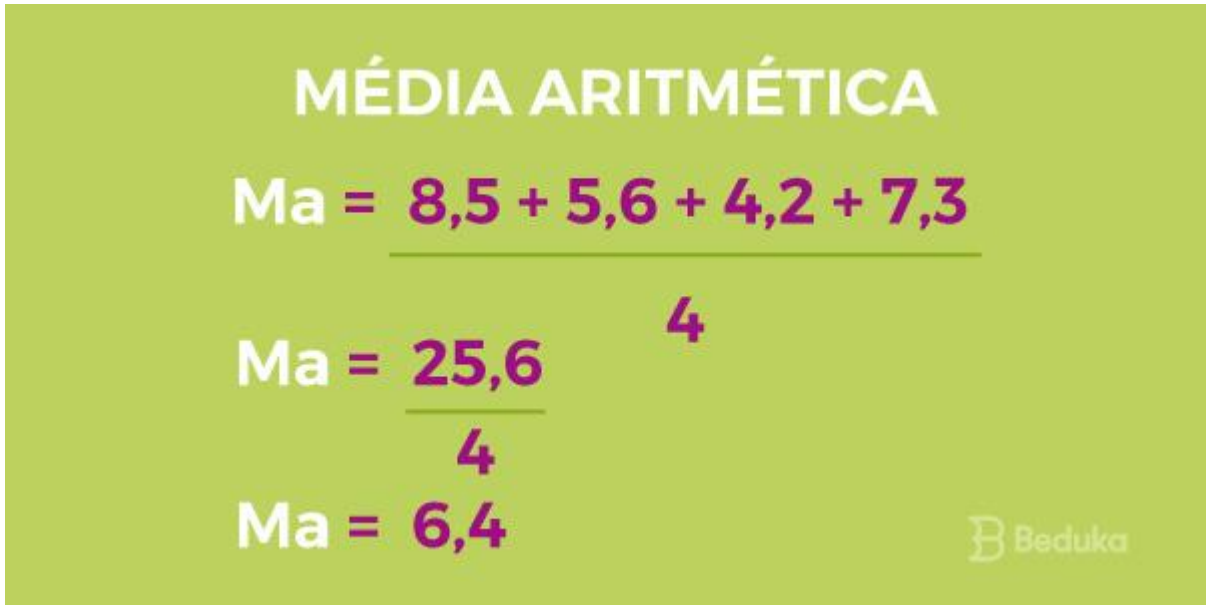
Estatística Descritiva

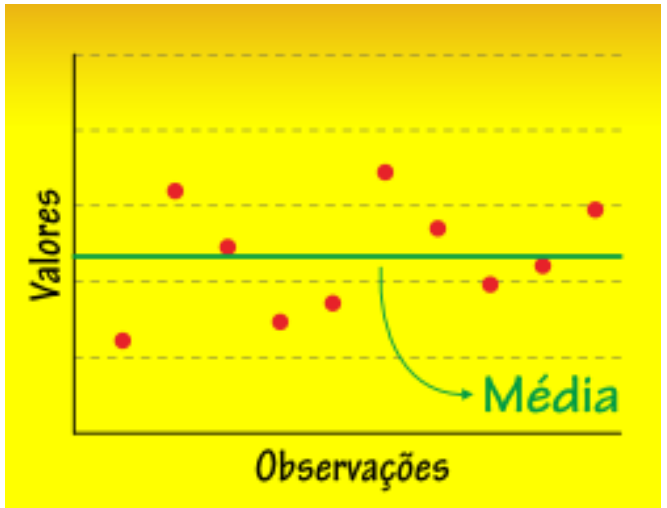
Medidas de Centralidade: Média, Moda e Mediana

Sala Digital

- **Média**: Mescla de maneira que mais uniforma os valores do conjunto de dados. Ela é o valor que aponta para onde mais se concentram os dados da distribuição, e pode ser considerada o pondo de equilíbrio das frequências em um histograma.

$$Média = \frac{\sum x}{N}$$

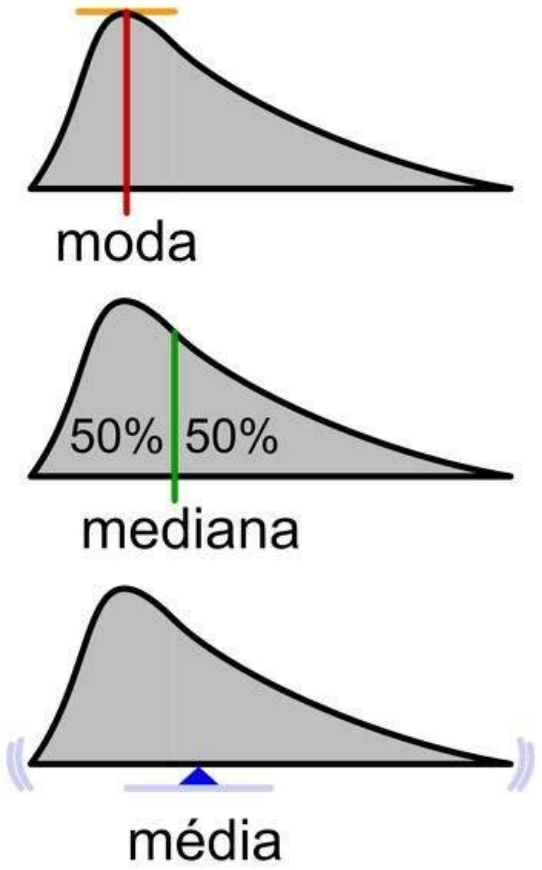




BOSCH

Estatística Descritiva

Medidas de Centralidade: Média, Moda e Mediana



Estatística Descritiva
Desvio Padrão

Sala Digital

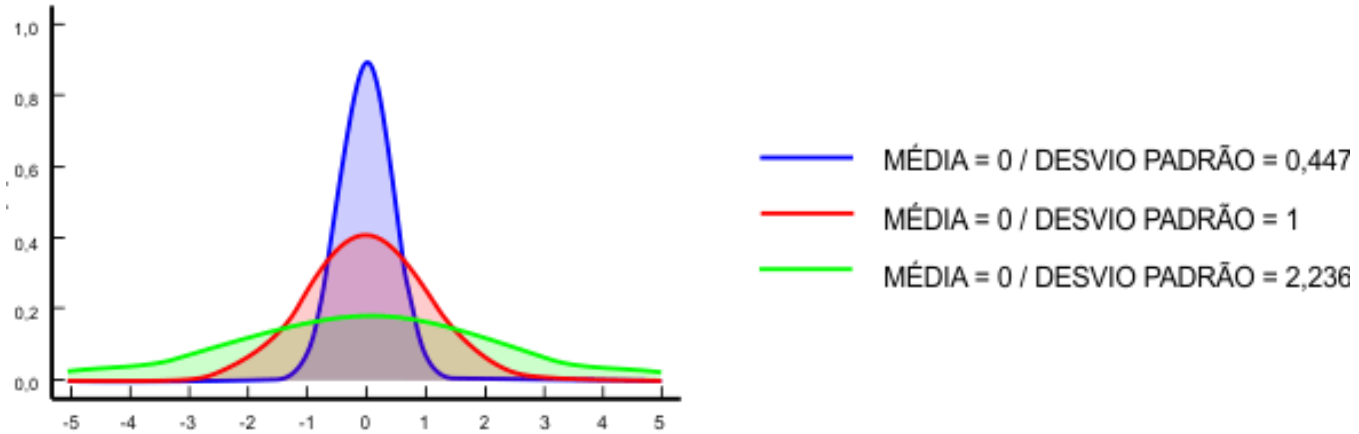
- É a maneira mais comum de dispersão estatística, e é representado pelo símbolo σ (sigma);
- O desvio padrão mostra quanto de variação existe em relação a média;
- Um valor baixo de desvio padrão significa que os dados tendem a estar próximos da média. Já um valor alto de desvio padrão indica que os dados estão espalhados e muito diferentes entre si.

Desvio padrão (σ) = $\sqrt{\frac{\sum(x - media)^2}{N}}$

Onde:

X é o valor individual

N é o tamanho do conjunto de dados



BOSCH

Estatística Descritiva
Desvio Padrão

Sala Digital

- Exemplo: Temos uma turma de 17 alunos que fizeram a peça, escolher 3 pessoas aleatoriamente para fazerem parte da minha amostra. Os diâmetros das peças foram 1,55 mm; 1,7mm; 1,8mm. Quero saber quanto esses diâmetros foram diferenciados da média.

$$DP = \sqrt{\frac{(1{,}55-1{,}68)^2+(1{,}70-1{,}68)^2+(1{,}80-1{,}68)^2}{3}}$$

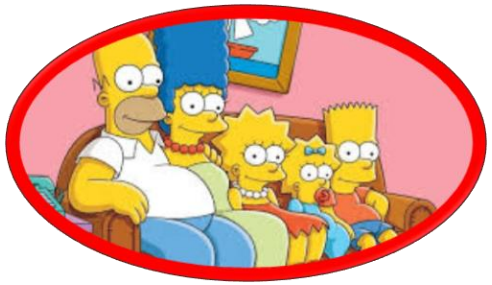
$$DP = \sqrt{\frac{(0{,}13)^2+(0{,}02)^2+(0{,}12)^2}{3}} = \sqrt{\frac{0{,}0317}{3}}$$

$$DP = \sqrt{0{,}01005} = 0{,}1027$$

Estatística Descritiva
Exemplo

Sala Digital

Grupo 1



Mag = 1

Lisa = 8

Bart = 10

Marge = 38

Homer = 39

Grupo 2

Isabela = 8

Millhouse = 10

Edna = 39

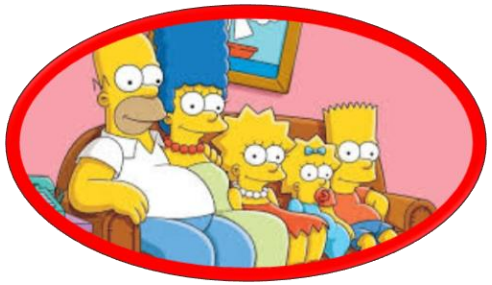
Moe = 45

Skinner = 49

- Vamos primeiro utilizar os dois grupos separados, utilizaremos as bibliotecas numpy (para fazer a média, mediana e desvio padrão) e pandas (para fazer o média, mediana, desvio padrão e moda) da linguagem Python.
- Calcular:
  - Média;
  - Mediana;
  - Desvio Padrão.

Estatística Descritiva
Exemplo: Grupo 1

Sala Digital



Mag = 1

Lisa = 8

Bart = 10

Marge = 38

Homer = 39

Grupo 1: (1, 8, 10, 38, 39)

$$\boldsymbol{Média} = \frac{\sum x}{N}$$

Média = (1 + 8 + 10 + 38 + 39)/5 = **19,5**

Mediana = **10**

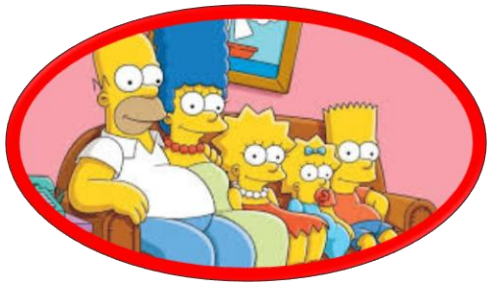
Desvio padrão (σ) = $\sqrt{\frac{\sum(x - media)^2}{N}}$

std = $\sqrt{\frac{(1-19{,}5)^2+(8-19{,}5)^2+(10-19{,}5)^2+(38-19{,}5)^2+(39-19{,}5)^2}{5}} = \mathbf{17,93}$

BOSCH

Estatística Descritiva
Exemplo: Grupo 1

Sala Digital



Mag = 1

Lisa = 8

Bart = 10

Marge = 38

Homer = 39

```
import pandas as pd
import numpy as np
```

```
grupo_1 = (1, 8, 10, 38, 39)

print("Media: ", np.mean(grupo_1))
print("Mediana: ", np.median(grupo_1))
print("Desvio Padrão: ", np.std(grupo_1))
```

```
Media:  19.2
Mediana:  10.0
Desvio Padrão:  17.93599732381782
```


BOSCH

Estatística Descritiva
Exemplo: Grupo 2

Sala Digital

Isabela = 8

Millhouse = 10

Edna = 39

Moe = 45

Skinner = 49

```
grupo_2 = (8, 10, 39, 45, 49)

print("Media: ", np.mean(grupo_2))
print("Mediana: ", np.median(grupo_2))
print("Desvio Padrão: ", np.std(grupo_2))
```

```
Media:  30.2
Mediana:  39.0
Desvio Padrão:  19.690099034794112
```


BOSCH

Estatística Descritiva
Exemplo: Utilizando o Pandas

Sala Digital

```
dados = {'Grupo 1': [1, 8, 10, 38, 39],
         'Grupo 2':[8, 10, 39, 45, 49]}
dataframe = pd.DataFrame(data=dados)
dataframe
```

	Grupo 1	Grupo 2
0	1	8
1	8	10
2	10	39
3	38	45
4	39	49

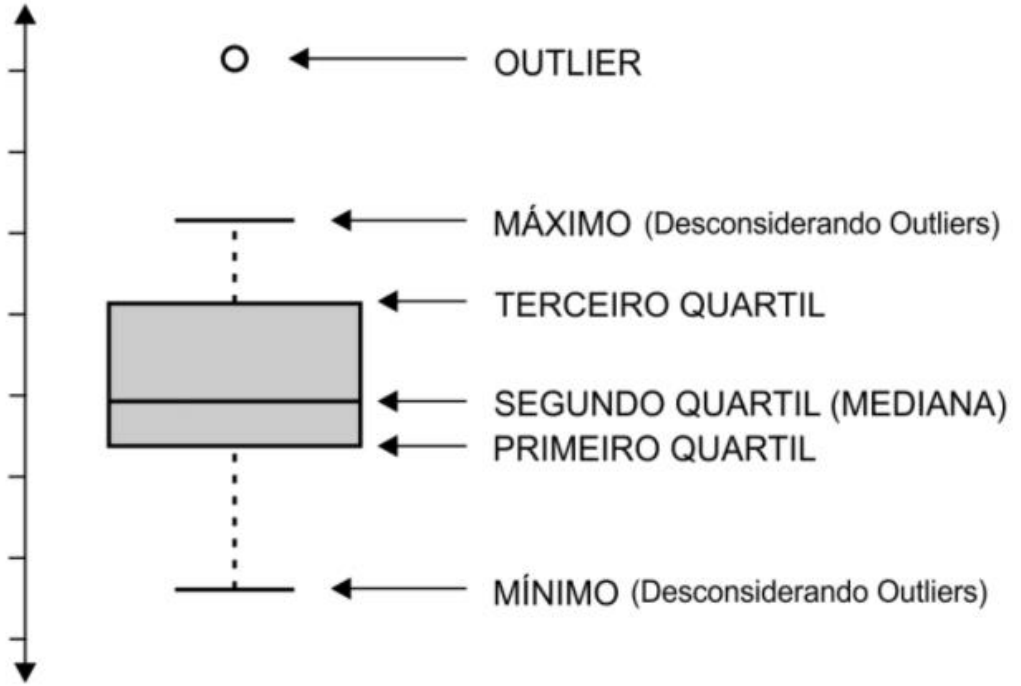
```
print("Media: ", dataframe['Grupo 1'].mean())
print("Mediana: ", dataframe['Grupo 1'].median())
print("Moda: ", dataframe['Grupo 1'].mode())
print("Desvio Padrão: ", dataframe['Grupo 1'].std())
```

```
Media:  19.2
Mediana:  10.0
Moda:  0     1
1     8
2    10
3    38
4    39
dtype: int64
Desvio Padrão:  17.93599732381782
```

Estatística Descritiva
Boxplot (Diagrama de Caixa)

Sala Digital

- O boxplot ou diagrama de caixa é uma ferramenta gráfica que permite visualizar a distribuição e valores discrepantes (outliers) dos dados, fornecendo assim um meio complementar para desenvolver uma perspectiva sobre o caráter dos dados. Além disso, o boxplot também é uma disposição gráfica comparativa.

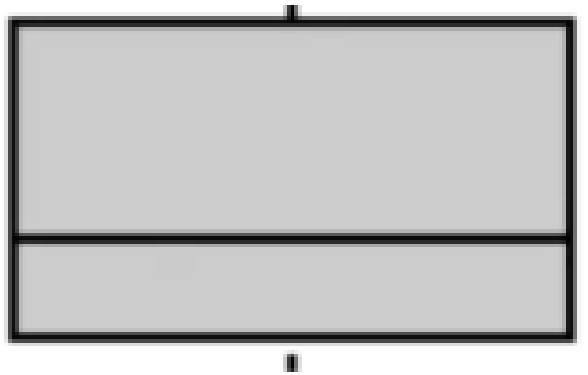


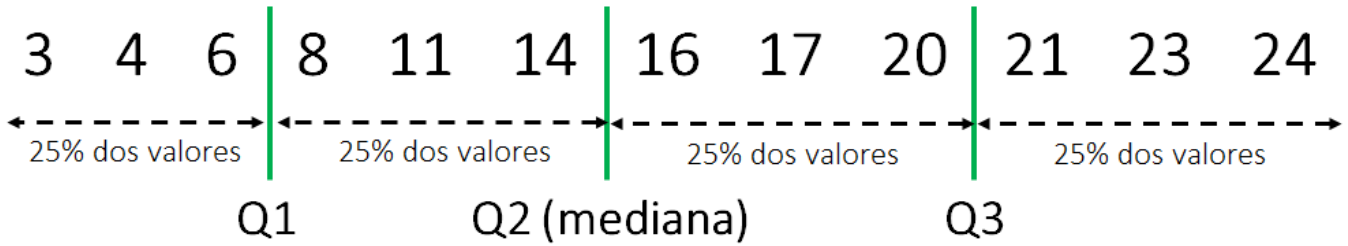
BOSCH

Estatística Descritiva
Quartil

Sala Digital

- Os quartis nada mais são que os percentis 25, 50 e 75, representando, respectivamente, o primeiro, segundo e terceiro quartil. Veja que o segundo quartil equivale ao percentil 50, valor em que pelo menos 50% da amostra está acima dele e pelo menos 50% está abaixo.
- Não é isso a definição de mediana?
- Sim! O percentil 50 ou segundo quartil equivalem à mediana!



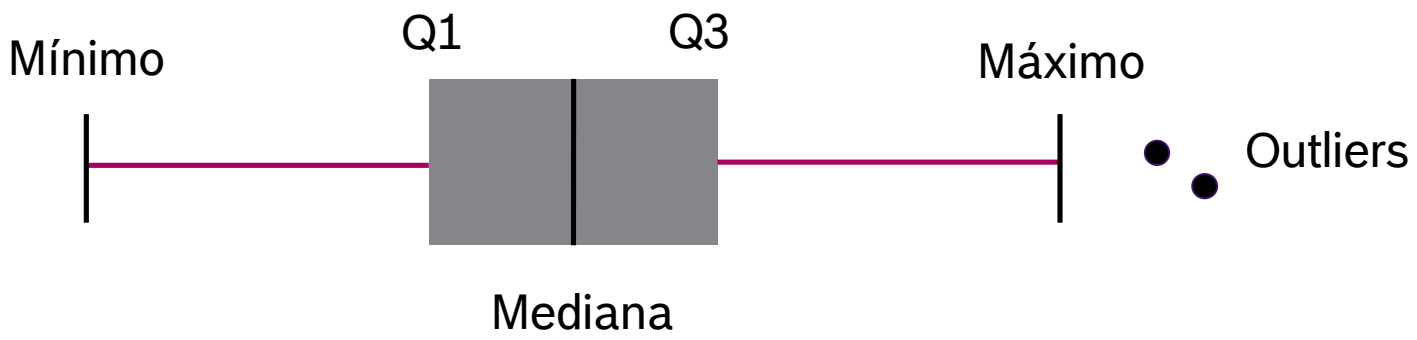


BOSCH

Estatística Descritiva
Boxplot (Diagrama de Caixa)

Sala Digital

- Para alguns conjuntos de dados é necessário mais do que as medidas de centralidade, podendo ser necessário informações como variabilidade e dispersão dos dados;
- Boxplot é uma forma padronizada de exibir a distribuição dos dados com base em cinco itens:
  - **Mínimo**: Q1 – 1.5 * (Q3-Q1);
  - **Primeiro Quartil (Q1)**: O número do meio entre o menor número dos dados (não o *mínimo*) e a *mediana*;
  - **Mediana**: O valor do meio dos dados;
  - **Terceiro Quartil (Q3)**: O número do meio entre a *mediana* e o maior número dos dados (não o *máximo*);
  - **Máximo**: Q3 – 1.5 * (Q3-Q1);
  - **Outlier**: Valores que estão fora do range entre o mínimo e o máximo.



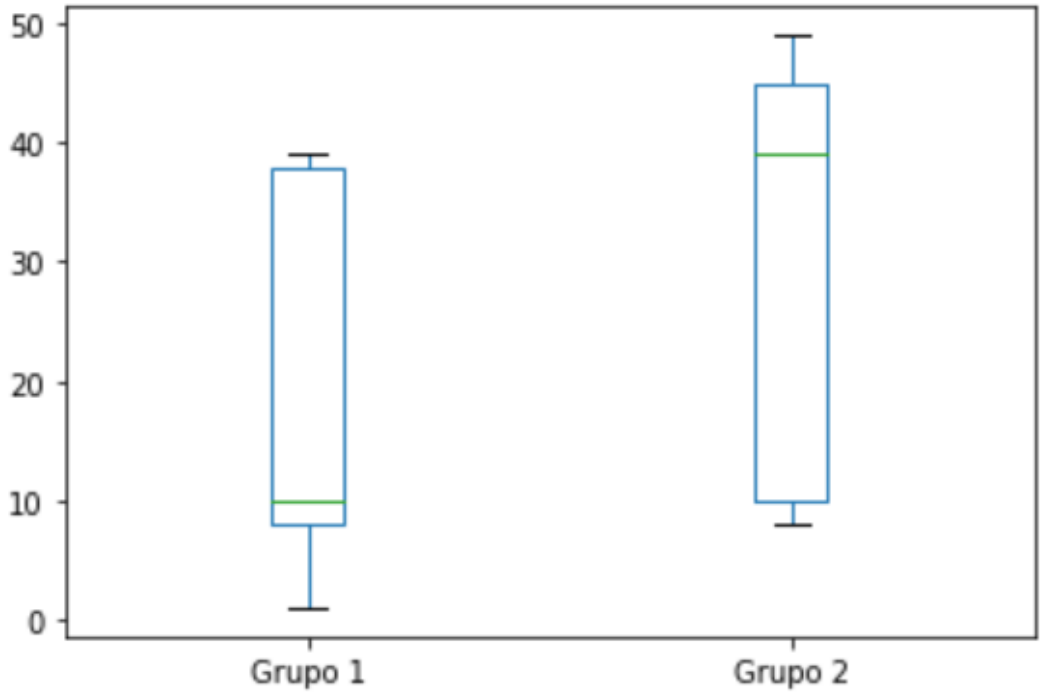
BOSCH

Estatística Descritiva
Exemplo: Boxplot

Sala Digital

- Vamos utilizar o dataframe do exemplo anterior.

```
boxplot = dataframe.boxplot(column=['Grupo 1', 'Grupo 2'], grid=False)
```



BOSCH

Estatística Descritiva
Exercício 1

Sala Digital

- Uma fábrica produz massa pronta para pasteis. Diariamente é medida a densidade da mistura das massas em máquinas diferentes. Dadas as medições a seguir, calcule a média, mediana, desvio padrão e crie o gráfico boxplot utilizando a biblioteca pandas.
- Tempo estimado: 30 min

Mistura 1	Mistura 2	Mistura 3
22.02	21.49	20.33
23.83	22.67	21.67
26.67	24.62	24.67
25.38	24.18	22.45
25.49	22.78	22.29
23.50	22.56	21.95
25.90	24.46	20.49
24.89	23.79	21.81

Estatística Descritiva
Exercício 1 - Resposta

Sala Digital

```
#Criando o dataframe
dados = {'Mistura 1': [22.02, 23.83, 26.67, 25.38, 25.49, 23.50, 25.90, 24.89],
        'Mistura 2': [21.49, 22.67, 24.62, 24.18, 22.78, 22.56, 24.46, 23.79],
        'Mistura 3': [20.33, 21.67, 24.67, 22.45, 22.29, 21.95, 20.49, 21.81]}

df_misturas = pd.DataFrame(data=dados)
```

```
#Calculos da Mistura 1
print("Media: ", df_misturas['Mistura 1'].mean())
print("Mediana: ", df_misturas['Mistura 1'].median())
print("Desvio Padrão: ", df_misturas['Mistura 1'].std())
```

```
Media:  24.71
Mediana:  25.134999999999998
Desvio Padrão:  1.5034246619919853
```

```
#Calculos da Mistura 2
print("Media: ", df_misturas['Mistura 2'].mean())
print("Mediana: ", df_misturas['Mistura 2'].median())
print("Desvio Padrão: ", df_misturas['Mistura 2'].std())
```

```
Media:  23.31875
Mediana:  23.285
Desvio Padrão:  1.1078091313166598
```

```
#Calculos da Mistura 3
print("Media: ", df_misturas['Mistura 3'].mean())
print("Mediana: ", df_misturas['Mistura 3'].median())
print("Desvio Padrão: ", df_misturas['Mistura 3'].std())
```

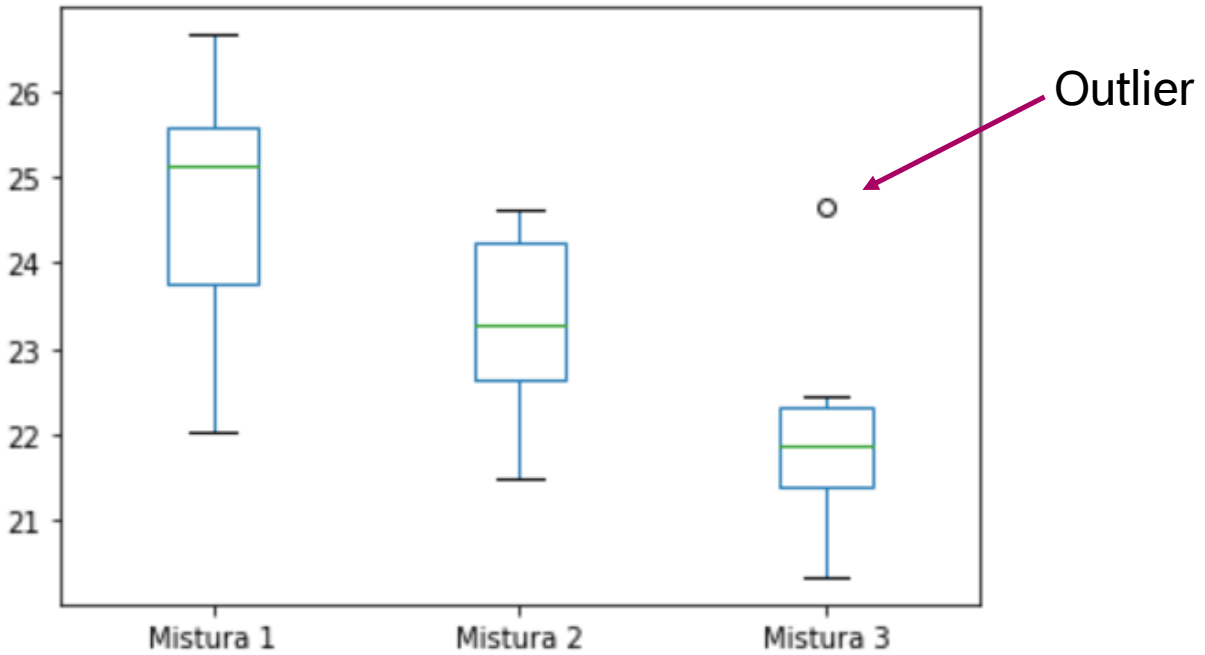
```
Media:  21.9575
Mediana:  21.88
Desvio Padrão:  1.3425748183461306
```


BOSCH

Estatística Descritiva
Exercício 1 - Resposta

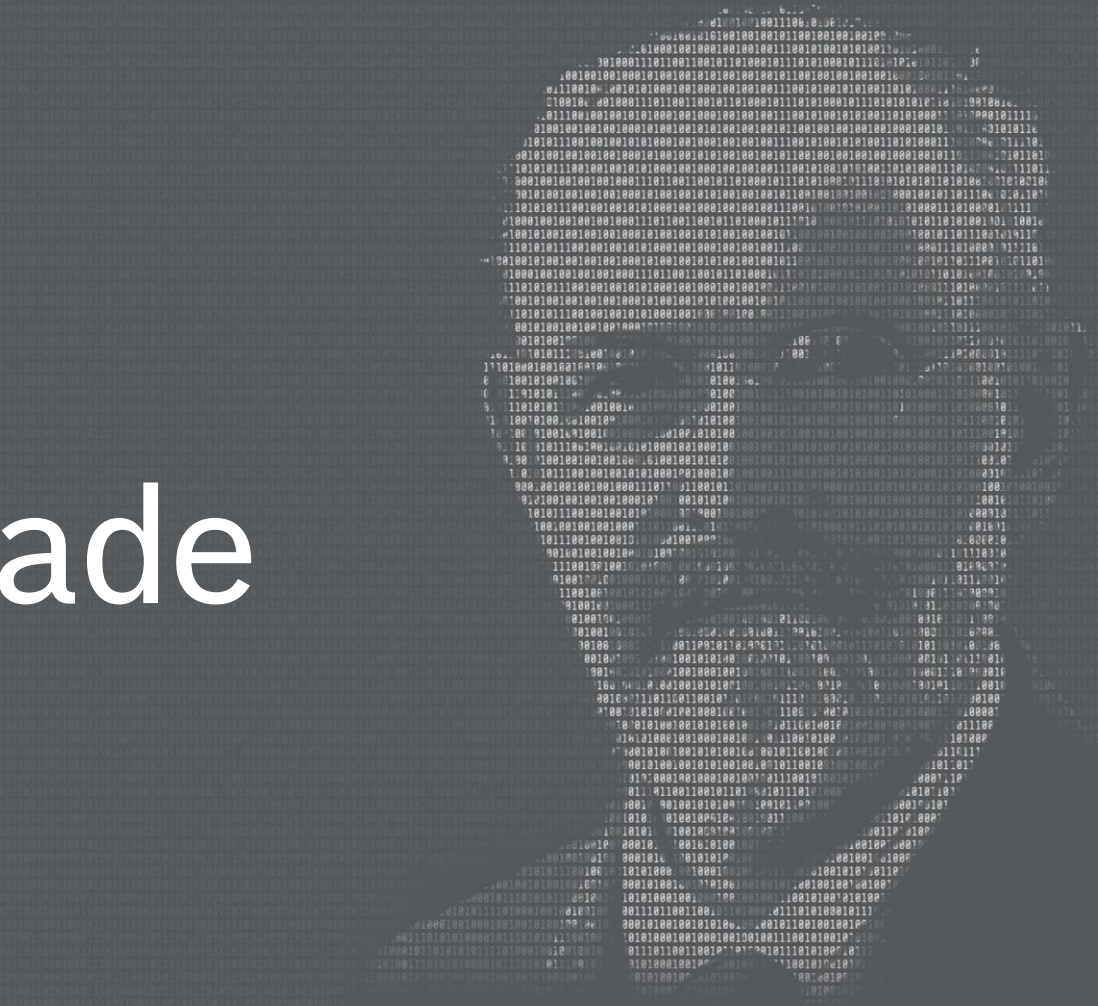
Sala Digital

```
#Gerando o boxplot dos dados
plot = df_misturas.boxplot(column=['Mistura 1', 'Mistura 2', 'Mistura 3'], grid=False)
```



BOSCH

Probabilidade



BOSCH

Probabilidade
O que é Probabilidade

Sala Digital

- Área da matemática que estuda a chance de determinado evento acontecer;
- É necessário entender alguns conceitos importantes sobre probabilidade:
  - Experimento aleatório;
  - Evento;
  - Espaço Amostral;
  - Eventos equiprováveis;
- O valor da probabilidade sempre é um número entre 0 e 1 ou uma porcentagem entre 0% e 100%.

Probabilidade
Conceitos

Sala Digital

Experimento Aleatório

- É o experimento que, mesmo sendo realizado várias vezes sempre nas mesmas condições, possui um resultado imprevisível.
- Um bom exemplo de experimento aleatório é o lançamento de um dado. Ainda que seja possível calcular a chance de cada um dos resultados ocorrer, é impossível sabermos qual será o resultado do lançamento.

Espaço Amostral

- É o conjunto de todos os resultados possíveis de um experimento aleatório. Pode ser conhecido também como universo amostral. O espaço amostral é representado por Ω (ômega);
- No exemplo do dado comum, o espaço amostral seria:
  - Ω: {1, 2, 3, 4, 5, 6}

Probabilidade
Conceitos

Sala Digital

Eventos

- É qualquer subconjunto do espaço amostral. Geralmente, é um conjunto com resultados satisfatórios, ou seja, um subconjunto do espaço amostral que contém elementos com os quais se calcula probabilidade;
  - **Evento certo**: Possui 100% de chance de ocorrer. Ex.: Em um dado de 6 lados e um evento E{1, 2, 3, 4, 5, 6};
  - **Evento impossível**: Possui 0% de chance de ocorrer. Ex.: Em um dado de 6 lados e um evento M{7, 9};

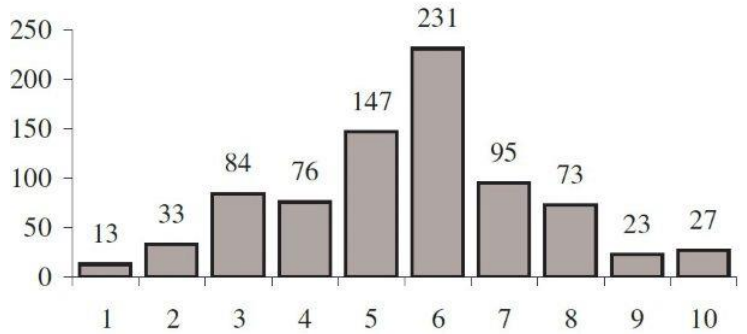
Evento equiprovável

- São eventos que possuem a mesma chance de ocorrer. Ex.: Em um dado de 6 lados, e dois eventos A{1, 3, 5} e B{2, 4, 6}. Os eventos A e B são equiprováveis pois possuem a mesma chance de acontecer.

Probabilidade
Histograma

Sala Digital

- O histograma é um gráfico de barras que demonstra a distribuição de frequências, onde cada coluna representa uma classe (valor do dado), e a altura representa a quantidade ou frequência absoluta em que ocorre.
- Tem como objetivo ilustrar como determinada amostra, ou população, está distribuída, com intuito de facilitar a visualização dessa distribuição;
- Utilidades de um histograma:
  - Resumir grande volume de dados;
  - Comunicar as informações por meio de gráfico.
  - Ex: satisfação em aplicativos.



BOSCH

Probabilidade
Histograma: Exemplo

Sala Digital

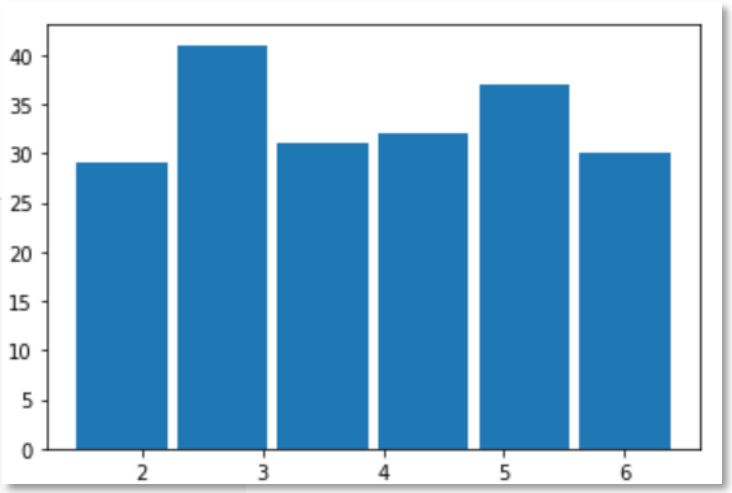
- Seguindo com o exemplo de dados, vamos montar um histograma para verificar se um dado é normal ou viciado.

```
#Importando as bibliotecas
import random
import pandas as pd
```

```
#Criando uma lista com números de 1 a 6 com tamanho 200
lista_lados = [random.randint(1,6) for i in range(200)]

#Passando a lista para um dataframe
lista_lados = pd.DataFrame(data=lista_dados, columns=['Lado'])

#Colocando os dados em um histograma
lista_lados.plot.hist(align='right', rwidth=0.9, bins=6, legend=False)
```

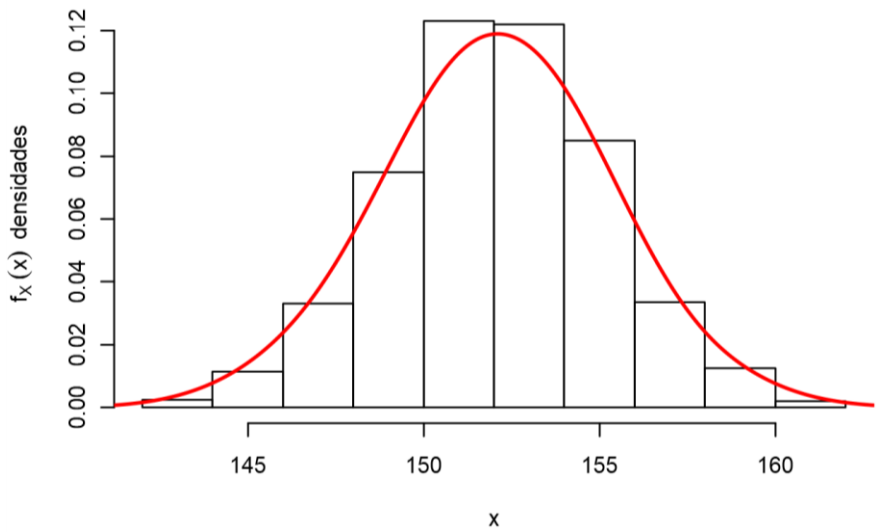


BOSCH

Probabilidade
Distribuição Normal

Sala Digital

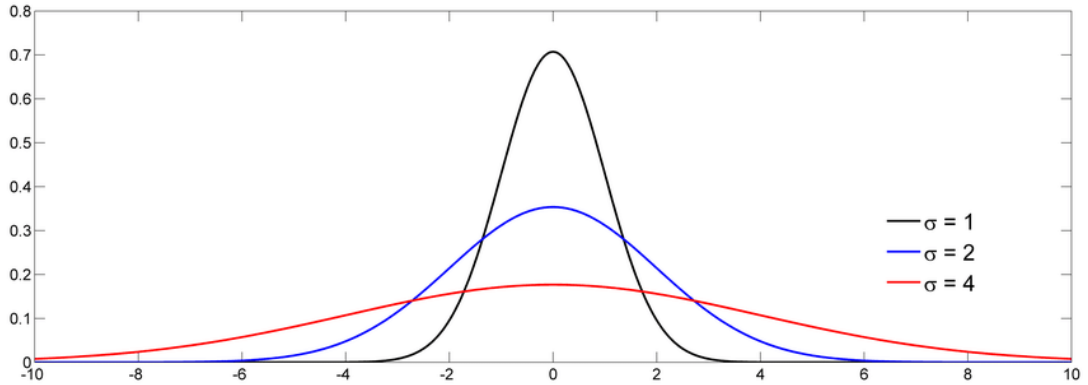
- Uma distribuição estatística é uma função que define uma curva, e uma área sob essa curva que determina a probabilidade de um determinado evento acontecer;
- Muitos fenômenos aleatórios se comportam próximos a essa distribuição. Como por exemplo altura das pessoas, peso, pressão sanguínea, etc.
- Características:
  - O ponto central é a média. (Simétrica, e valor mais provável)
  - Curva em formato de sino;
  - Unimodal;
  - A probabilidade tende a zero quando se afasta da média;
- A área total da curva normal representa 100%, ou 1, de probabilidade;



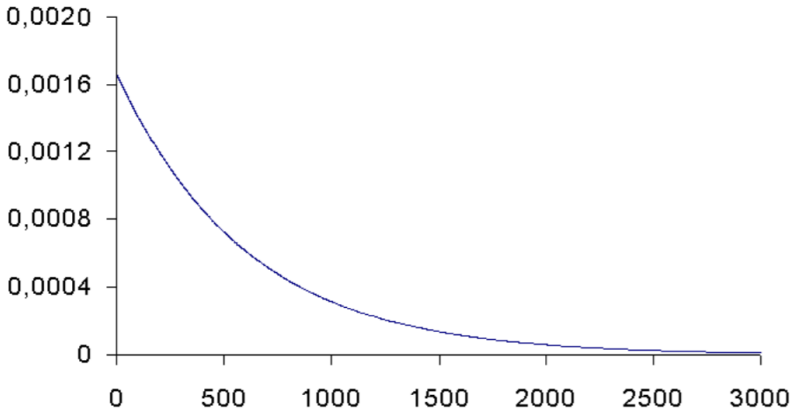
Probabilidade
Distribuição Normal

Sala Digital

- Diferentes valores de média e desvio padrão, nos darão distribuições com formas diferentes;
- A média indica o centro da distribuição, e o desvio padrão indica a variabilidade dos dados;
- Nem todos os fenômenos podem ser representados pela distribuição normal. Por exemplo a duração de uma lâmpada de certa marca selecionada ao acaso



Diferentes formas de distribuição normal.



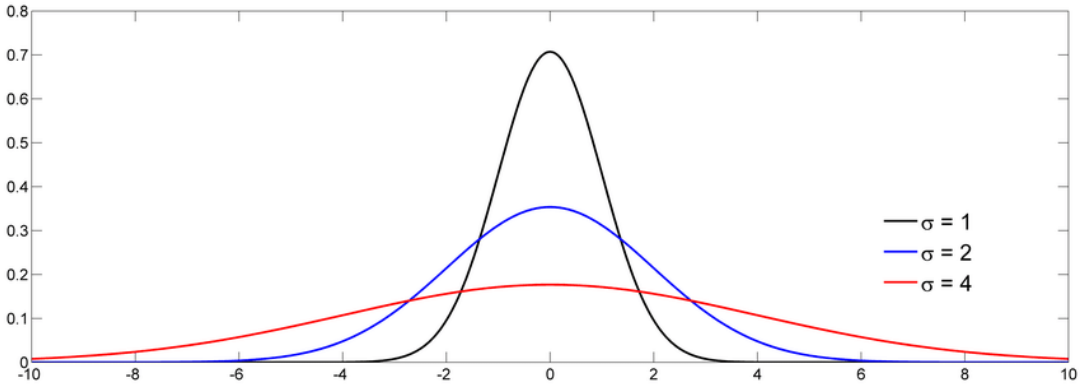
Distribuição assimétrica:
Duração de uma lâmpada.

BOSCH

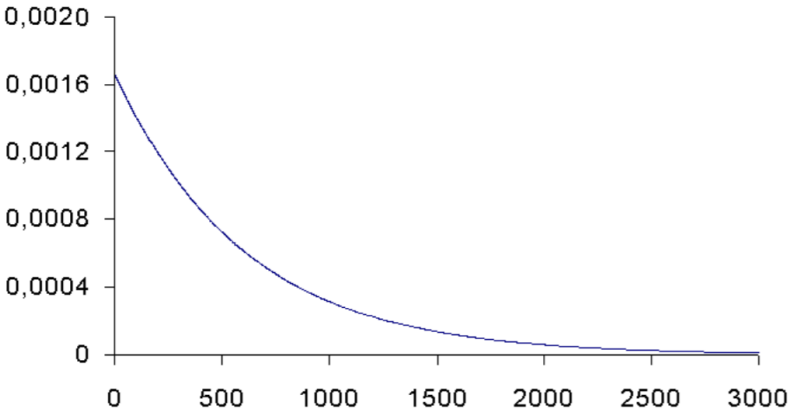
Probabilidade
Distribuição Normal

Sala Digital

- Se o sino é menor, teremos uma menor variabilidade, os valores vão estar mais perto da média.
- Se a curva tende a ser uma linha, os eventos tem probabilidade de ocorrer mais próximos.
- Detalhe: nem todos os fenômenos tendem a ter um comportamento de distribuição normal.



Diferentes formas de distribuição normal.



Distribuição assimétrica:
Duração de uma lâmpada.

BOSCH

Probabilidade
Distribuição Normal: Exemplo

Sala Digital

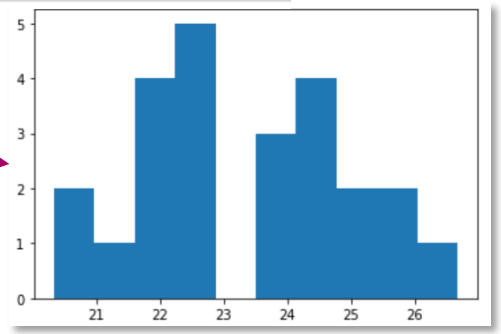
- Vamos utilizar os dados de densidade das misturas do exercício anterior para calcular a distribuição normal.

```
#Importando as bibliotecas
import numpy as np
from matplotlib import pyplot as plt
from scipy.stats import norm
```

```
#Criando um dataset (conjunto de dados)
dados = {'Densidade': [22.02, 23.83, 26.67, 25.38, 25.49, 23.50, 25.90, 24.89,
                       21.49, 22.67, 24.62, 24.18, 22.78, 22.56, 24.46, 23.79,
                       20.33, 21.67, 24.67, 22.45, 22.29, 21.95, 20.49, 21.81]}
```

```
#Visualizando a distribuição dos dados
plt.hist(dados['Densidade'])
plt.show()
```

Histograma dos dados



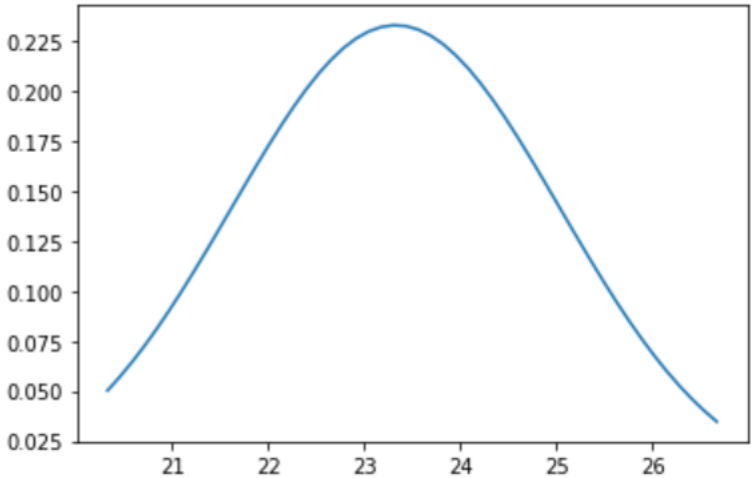
BOSCH

Probabilidade
Distribuição Normal: Exemplo

Sala Digital

```
#Calculando média e desvio padrão
desvio_padrao = np.std(dados['Densidade'], ddof=1)
media = np.mean(dados['Densidade'])
```

```
#Visualizando a curva da distribuição normal
dominio = np.linspace(np.min(dados['Densidade']), np.max(dados['Densidade']))
plt.plot(dominio, norm.pdf(dominio, media, desvio_padrao))
```

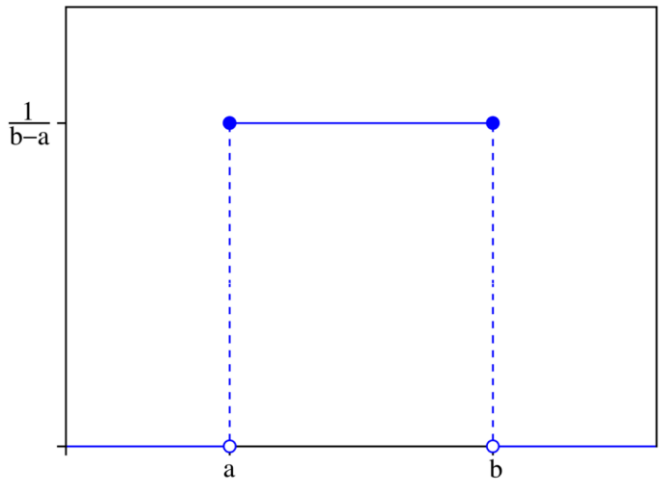


BOSCH

Probabilidade
Distribuição Uniforme

Sala Digital

- É uma distribuição de probabilidade contínua, ou seja, possui um número finito de resultados com chances iguais de acontecer;
- É usada quando assumimos intervalos iguais da variável que a mesma probabilidade;
- Um bom exemplo de distribuição uniforme seria um lançamento de um dado não viciado. Como todos os lados tem chances iguais, a distribuição das probabilidades seria uma reta num intervalo entre 1 e 6.
- A área abaixo da curva equivale a 1.



Probabilidade
Exercício 2

Sala Digital

- Em uma determinada escola de ensino médio, alguns alunos foram selecionados aleatoriamente para saber qual idade eles tem. Abaixo temos a amostra das idades desse alunos. Calcule a média e o desvio padrão, e elabore os gráficos de boxplot, histograma e a curva da distribuição normal.

Idades
14, 17 ,18, 15, 15, 16, 17, 15, 16, 16, 15, 17, 15, 16, 16, 18, 18, 19, 17, 16, 17, 15, 16, 17, 17, 19, 20, 18, 17, 16, 15, 16, 16, 17, 18, 18, 17, 17, 15, 16, 16, 15

Probabilidade
Exercício 2 - Resposta

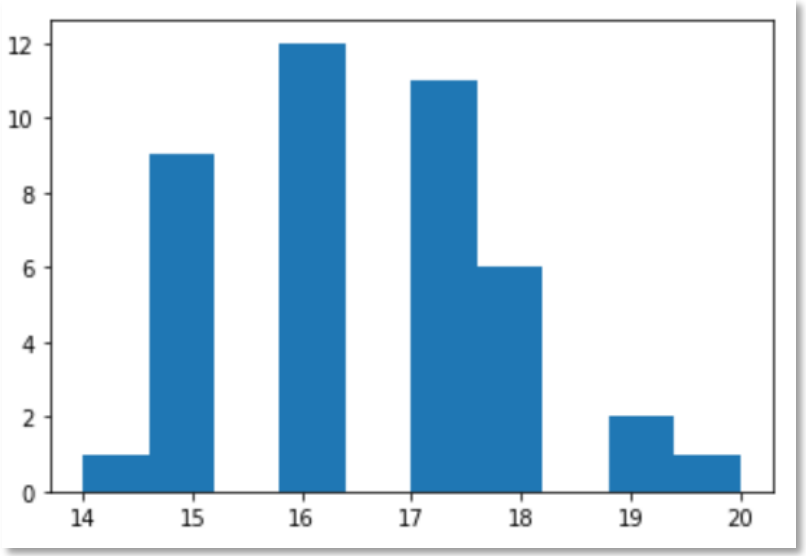
Sala Digital

```
#Importando as bibliotecas
import numpy as np
from matplotlib import pyplot as plt
from scipy.stats import norm
```

```
#Criando o dataset
alunos = {'Idades': [14, 17 ,18, 15, 15, 16, 17, 15, 16, 16, 15, 17, 15, 16,
                     16, 18, 18, 19, 17, 16, 17, 15, 16, 17, 17, 19, 20, 18,
                     17, 16, 15, 16, 16, 17, 18, 18, 17, 17, 15, 16, 16, 15]}
```

Probabilidade
Exercício 2 - Resposta

```
#Mostrando a distribuição dos dados
#Através de um histograma
plt.hist(alunos['Idades'])
```

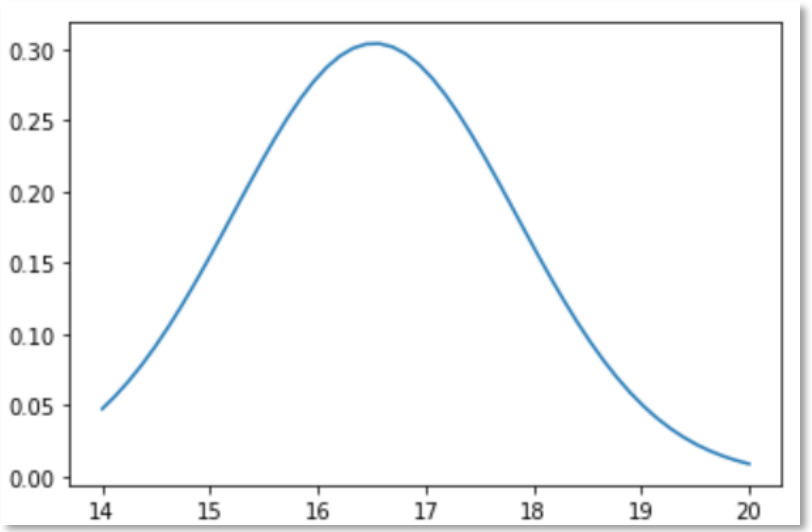



```
#Plotando o boxplot
plt.boxplot(alunos['Idades'])
```

Probabilidade
Exercício 2 - Resposta

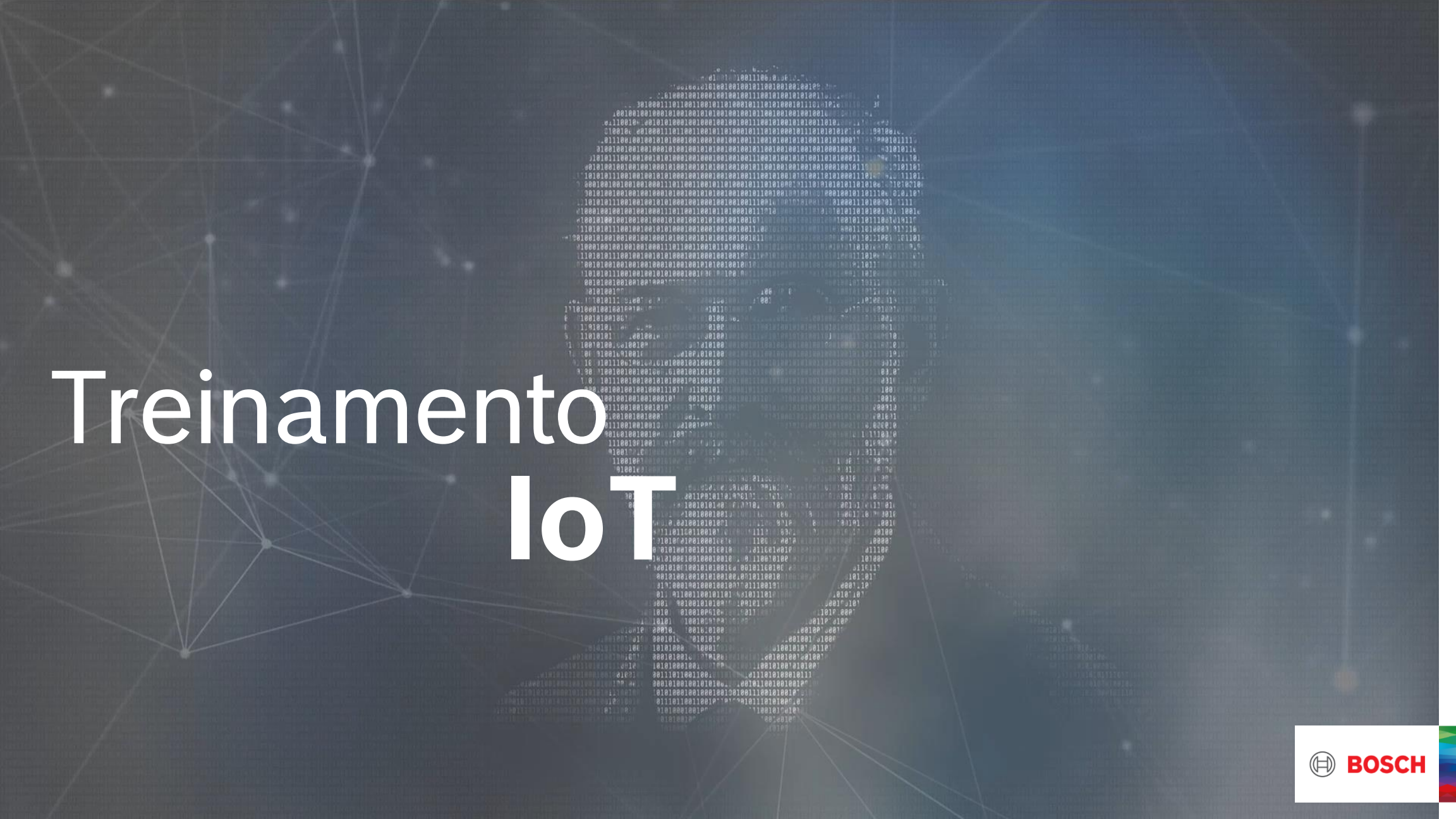
Sala Digital

```
#Calculando média e desvio padrão
desvio = np.std(alunos['Idades'], ddof=1)
media = np.mean(alunos['Idades'])
```




```
#Contruindo a curva de distribuição normal
dominio = np.linspace(np.min(alunos['Idades']), np.max(alunos['Idades']))
plt.plot(dominio, norm.pdf(dominio, media, desvio))
```


BOSCH



Treinamento
IoT

BOSCH