Big Data

Wellington Franco
Universidade Federal do Ceará – UFC
Campus de Crateús
wellington@crateus.ufc.br

Big Data

Módulo 1: Análise Estatística de Dados

Aula 3: Análise Exploratória de Dados

Roteiro

Análise Estatística dos Dados

- Medidas de Tendência Central
- Tabela de Frequência e Histograma
- Distribuição dos Dados

Medidas de Tendência Central e Medidas de Dispersão

Introdução

- **Medidas de Tendência Central ou Estimativa de Parâmetro**
  - Média Aritmética
  - Mediana
  - Valor máximo e mínimo
- **Medida de Dispersão ou Estimativa de Variabilidade**
  - Desvio em relação à média
  - Variância
  - Desvio Padrão

Estimativa de Parâmetros (*Estimates of Location*)

Identificar a localização dos dados (ou seja, a sua tendência central). A seguir, os principais conceitos-chave:

- **Média**: a soma de todos os valores dividido pelo número de valores.
  - Sinônimo: *average*, *mean*
  - Ex: Seja A = {3, 5, 1, 2}, média = (3+5+1+2)/4 = 11/4 = 2.75
  - x-barra é a média amostral da população

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

Estimativa de Parâmetros (*Estimates of Location*)

- **Média ponderada**: cada item da amostra estará associado a um peso.
  - Sinônimo: *weighted mean*
  - Ex: Seja A = {(10,1), (7,2), (8,3)},
    - Média ponderada = ((10*1) + (7*2) + (8*3)) / (1+2+3) = (10 +14 + 24) / 6 = 8

Em *data science*, na prática, algumas *features* (atributos) são mais relevantes do que outras e, portanto, precisamos destacar a relevância desses atributos;

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$

Estimativa de Parâmetros (*Estimates of Location*)

- **Mediana**: é o número central de uma lista ordenada de valores.
    - Sinônimo: *Median*, 50*th percentile*.
  - A mediana é mais interessante que a média pois seu resultado não é influenciado por outlier.
  - Numa lista que contém n valores:
    - se n é ímpar: obtém-se o valor do meio;
    - se n é par: obtém-se a média dos dois valores centrais.

Exemplo: Seja uma lista contendo os seguintes valores: {2, 2, 3, 3, 5, 7, 8, 130}.

Média: (2+2+3+3+5+7+8+130)/8= **20**

Mediana: (3 +5)/2 = **4**

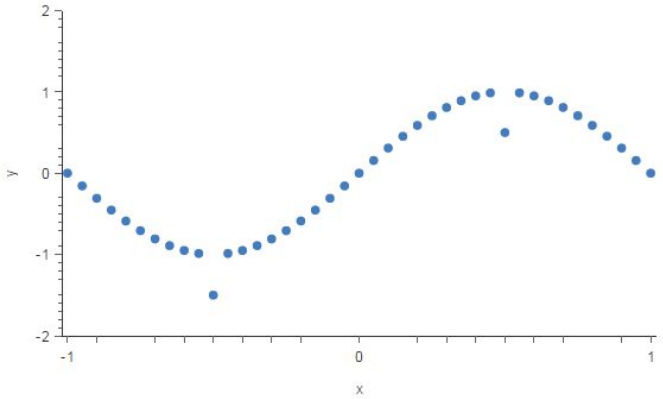
A mediana representa muito melhor o comportamento dos dados do que a média.

Estimativa de Parâmetros (*Estimates of Location*)

- ***Outlier***: o valor de um item que é muito diferente das outras amostras.
  - Na prática, um outlier PODE ser um erro na entrada ou coleta dos dados:
    - Ex: erro de digitação, erro no funcionamento de um sensor, etc.

No exemplo anterior, a lista {2, 2, 3, 3, 5, 7, 8, 130} contém um *outlier*.

Um exemplo de outlier no $\mathbb{R}^2$:



Estimativa de Variabilidade

Estimativa de parâmetro é apenas uma das dimensões para resumir um atributo;

Uma segunda dimensão, **variabilidade**, também referida como **dispersão**, serve para identificar o quão agrupados ou distantes estão os dados.

Estimativa de Variabilidade

As estimativas de variabilidade mais usadas são baseadas nas diferenças, ou *desvios*, entre a estimativa de parâmetro e o dado observado.

- **Desvio**: a diferença entre o valor observado e a estimativa de parâmetro selecionada.
    - Sinônimo: erro, *deviation, residuals*. $(x - \bar{x})$
- **Desvio Absoluto Médio** (ou Média do Desvio Absoluto):

$$\text{Mean Absolution Deviation} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

Estimativa de Variabilidade

Exemplo: seja o conjunto de dados {1, 4, 4}.

- Suponha que cada valor representa o número de curtidas de uma foto;
- A média desse conjunto é 3 e a mediana é 4;
- Os desvios são dados por 1-3 = -2, 4-3 = 1 e 4-3 = 1.
- Estes resultados, (-2,1,1), nos dizem o quão dispersos estão os dados do valor central.
- Podemos também gerar o módulo desse resultado, (2, 1, 1), e calcular a **desvio absoluto médio**:
  - (2 + 1 + 1) / 3 = 1.33.
- Como interpretar esse valor?
  - Na média, cada foto ficou em torno de 1.33 "curtidas" distante da média.

Estimativa de Variabilidade

- **Variância**: a soma dos erros (*deviations*) da média dividida por N-1 onde N é o número de dados.
    - Sinônimo: Erro Quadrado Médio

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

- **Desvio padrão**: raiz quadrada da variância.
    - Sinônimo: Norma-l2, norma Euclidiana

$$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

Estimativa de Variabilidade

Exemplo: seja o conjunto de dados {1, 4, 4}.

- Suponha que cada valor representa o número de curtidas de uma foto;
- A média desse conjunto é 3 e a mediana é 4;
- Os desvios são dados por:
  - $(1-3)^2 = 4$
  - $(4-3)^2 = 1$
  - $(4-3)^2 = 1$
- Variância = (4 + 1 + 1) / (3-1) = 6/2 = 3
- Desvio Padrão = $\sqrt{3}$ = 1.73

Estimativa Baseada em Percentis

Em um conjunto de dados, o percentil P é um valor tal que pelo menos P por cento dos valores assumem esse valor ou menos, e pelo menos (100-P) por cento dos valores assumem esse valor ou mais.

- Estimando Percentil P:

$$100 * \frac{j}{n} \leq P < 100 * \frac{j+1}{n}$$

Estimativa Baseada em Percentis

Em um conjunto de dados, o percentil P é um valor tal que pelo menos P por cento dos valores assumem esse valor ou menos, e pelo menos (100-P) por cento dos valores assumem esse valor ou mais.

- Exemplo: sejam os valores (3, 1, 5, 3, 6, 7, 2, 9)
  - Ordenando esses conjunto, temos: (1, 2, 3, 3 ,5 ,6, 7, 9)
  - Iremos calcular o 25th e 75th percentil, ou seja:
    - (25*n)/100 = (25*8)/100 = 2, ou seja, posição 2 e (2+1) 3 para tirar a média (2.5)
    - (75*n)/100 = (75*8)/100 = 6, ou seja, posição 6 e (6+1) 7 para tirar a média (6.5)
  - O 25th percentil é 2.5, pois (1, **2, 3**, 3 ,5 ,6, 7, 9) implica que (2+3)/2 = 2.5
  - O 75th percentil é 6.5, pois (1, 2, 3, 3 ,5 ,**6, 7**, 9), implica que (6+7)/2 = 6.5
  - A diferença entre o 25th e o 75th é chamada de **IQR** = 6.5 - 2.5 = **4** (range)
    - **IQR = 75th - 25th**

Considerações

- Cada uma das estimativas descritas anteriormente **resume os dados em um único número** para descrever a localização ou variabilidade dos dados;
- Também é útil explorar como os dados são distribuídos em geral.
- Caso haja remoção de outlier, verificar novamente as estimativas, pois a maioria são sensíveis a outliers:
    - Em muitos casos recomenda-se não remover os outliers (principalmente se estes não frem comprovadamente um erro na entrada de dados).

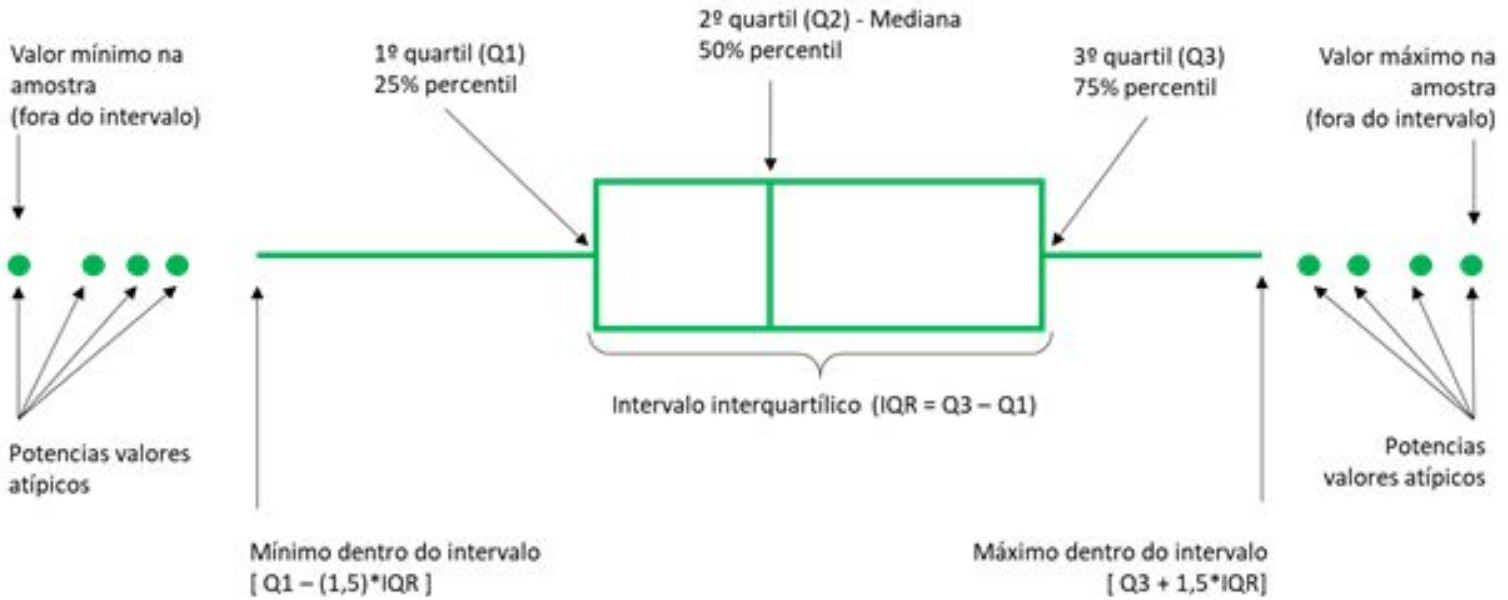
Explorando a Distribuição dos Dados

Explorar a distribuição dos dados é explorar como são os dados em sua totalidade.

- **Boxplot**: maneira rápida de visualização da distribuição dos dados.
    - Sinônimo: *Whiskers plot*
  - Informa percentis (25th, 50th, 75th) e quartis;
  - Informa mediana;
  - Informa máximos e mínimos;
  - Informa *outliers*.

Explorando a Distribuição dos Dados

- *Boxplot*



Explorando a Distribuição dos Dados

- *Boxplot:* Exemplo

Na Tabela a seguir temos as medidas da altura de 20 hastes. Faça o boxplot correspondente.

Dados da usinagem			
903,88	1036,92	1098,04	1011,26
1020,70	915,38	1014,53	1097,79
934,52	1214,08	993,45	1120,19
860,41	1039,19	950,38	941,83
936,78	1086,98	1144,94	1066,12

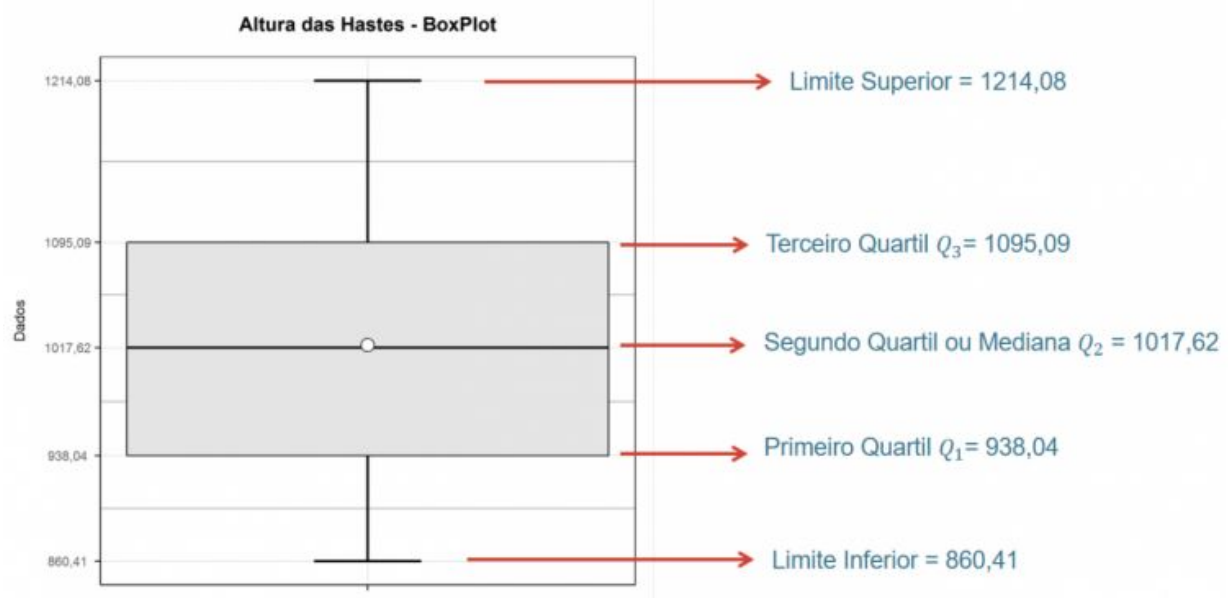
Exemplo retirado de: http://www.portalaction.com.br/estatistica-basica/31-boxplot

Explorando a Distribuição dos Dados

- *Boxplot:* Exemplo

Agora um resumo estatístico no *boxplot*:

Dados da usinagem			
903,88	1036,92	1098,04	1011,26
1020,70	915,38	1014,53	1097,79
934,52	1214,08	993,45	1120,19
860,41	1039,19	950,38	941,83
936,78	1086,98	1144,94	1066,12



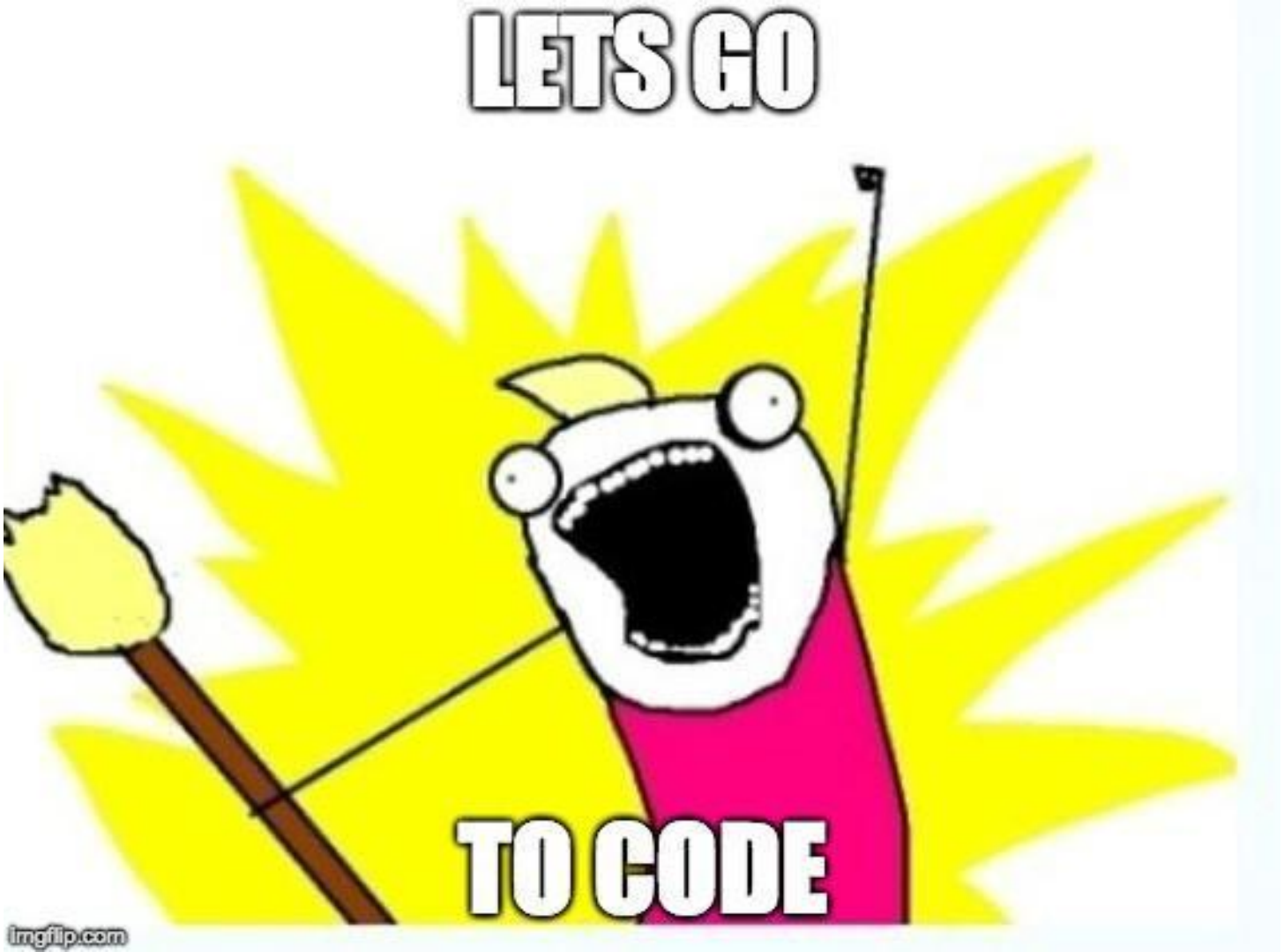

LETS GO
TO CODE
imgflip.com

Tabela de Frequência e Histograma

Introdução

- Para dados **quantitativos contínuos**, geralmente resultante de medições de características de qualidade (por exemplo: de peças ou produtos), dividimos a faixa de variação dos dados em intervalos de classes;
- Abaixo, uma ilustração do processo de geração de tabela de frequência e histograma:

Diametro do Eixo de 100 motores									
4,8	4,2	5,1	5,2	4,8	4,7	4,9	4,5	4,9	4,5
4,9	5,1	4,8	4,9	4,8	5	5,3	4,9	5,5	5,2
5,1	4,6	4,9	4,8	5,1	4,6	4,3	4,9	4,7	5,2
4,8	4,4	5,6	5	5	5	4,8	5,2	4,5	5,1
5,1	4,9	4,8	4,8	5	4,8	5,1	5,4	4,2	5,1
4,9	4,6	5,4	4,9	4,3	4,6	4,7	4,7	5,3	4,4
4,7	4,8	5,2	4,5	5,1	4,6	5,8	4,9	5,2	4,8
4,9	4,9	4,4	4,7	4,8	5,1	5,4	5	4,4	5,1
4,9	4,9	5,1	5,2	4,7	4,8	4,6	5,2	5,5	5,2
4,2	4,9	4,9	4,8	4,2	5,2	4,7	4,8	4,6	5,2

→

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

→

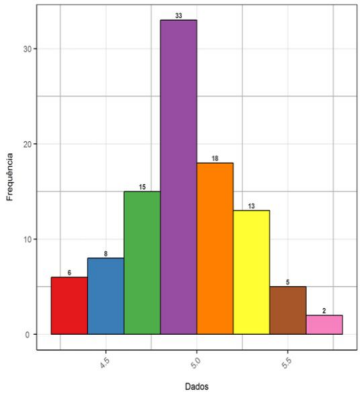


Tabela de Frequência

- Primeiro, com os dados que desejamos gerar a tabela de frequência, precisamos calcular o número de classes (**grupos**) para agrupar alguns valores.
- Para calcular o **número de grupos** necessários, existem duas opções:
  - ***k* = sqrt(n)** , onde *sqrt()* é a raiz quadrada e “n” o tamanho do dataset;
  - Fórmula de Sturges: $k = \lceil 1 + 3{,}3 \log_{10}(n) \rceil$, onde “n” é o tamanho do dataset.

- A partir daí precisamos gerar os grupos:
  - O menor valor do grupo é denominado limite inferior (l_i)
  - O maior valor do grupo é denominado limite superior (L_i)

Tabela de Frequência: exemplo

- Seja a seguinte tabela:

A seguir, iremos construir sua tabela de frequência. Atente-se a alguns detalhes que serão discutidos.

Diametro do Eixo de 100 motores									
4,8	4,2	5,1	5,2	4,8	4,7	4,9	4,5	4,9	4,5
4,9	5,1	4,8	4,9	4,8	5	5,3	4,9	5,5	5,2
5,1	4,6	4,9	4,8	5,1	4,6	4,3	4,9	4,7	5,2
4,8	4,4	5,6	5	5	5	4,8	5,2	4,5	5,1
5,1	4,9	4,8	4,8	5	4,8	5,1	5,4	4,2	5,1
4,9	4,6	5,4	4,9	4,3	4,6	4,7	4,7	5,3	4,4
4,7	4,8	5,2	4,5	5,1	4,6	5,8	4,9	5,2	4,8
4,9	4,9	4,4	4,7	4,8	5,1	5,4	5	4,4	5,1
4,9	4,9	5,1	5,2	4,7	4,8	4,6	5,2	5,5	5,2
4,2	4,9	4,9	4,8	4,2	5,2	4,7	4,8	4,6	5,2

Tabela de Frequência: exemplo

- O intervalo das classes pode ser representado das seguintes maneiras:
  - **(l_i) |- (L_i):** o limite inferior da classe é incluído na contagem da frequência absoluta, mas o superior não;
  - **(l_i) -| (L_i):** o limite superior da classe é incluído na contagem da frequência absoluta, mas o inferior não.
- Na tabela de distribuição de frequência, acrescentamos uma coluna com os pontos médios de cada intervalo de classe, denominado por x_i (média dos pontos), onde $x_i = (l_i + L_i)/2$

Tabela de Frequência: exemplo

- Na medida do possível, as classes deverão ter **amplitudes** iguais;
  - Amplitude = A = $(L_i - l_i)/k$
- Escolher os limites dos intervalos entre duas possíveis observações;
- Escolher os limites dos intervalos entre duas possíveis observações;
- O número de intervalos não deve ultrapassar 20;
- Escolher limites que facilitem o agrupamento;
- Marcar os pontos médios dos intervalos.

Diametro do Eixo de 100 motores									
4,8	4,2	5,1	5,2	4,8	4,7	4,9	4,5	4,9	4,5
4,9	5,1	4,8	4,9	4,8	5	5,3	4,9	5,5	5,2
5,1	4,6	4,9	4,8	5,1	4,6	4,3	4,9	4,7	5,2
4,8	4,4	5,6	5	5	5	4,8	5,2	4,5	5,1
5,1	4,9	4,8	4,8	5	4,8	5,1	5,4	4,2	5,1
4,9	4,6	5,4	4,9	4,3	4,6	4,7	4,7	5,3	4,4
4,7	4,8	5,2	4,5	5,1	4,6	5,8	4,9	5,2	4,8
4,9	4,9	4,4	4,7	4,8	5,1	5,4	5	4,4	5,1
4,9	4,9	5,1	5,2	4,7	4,8	4,6	5,2	5,5	5,2
4,2	4,9	4,9	4,8	4,2	5,2	4,7	4,8	4,6	5,2

→

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

Tabela de Frequência: exemplo

- Gerando tabela de distribuição de frequência:
  - **1º Passo:** calculando o número de classes (grupos):

- $k = \lceil 1 + 3{,}3 \log_{10}(100) \rceil = \lceil 7{,}6 \rceil =$ **8**
- Menor valor do dataset (l_i): = 4,2
- Maior valor do dataset (L_i) = 5,8
- Amplitude (A) = $(L_i - l_i)/k = (5{,}8 - 4{,}2)/8 = 0{,}2$
- Intervalo de classe definido como fechado em l_i e aberto em L_i --> $[l_i ; L_i)$ com range (amplitude) de 0,2

Tabela de Frequências

Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

Tabela de Frequência: exemplo

- Gerando tabela de distribuição de frequência:
  - **2º Passo:** Calcular frequência (número de ocorrências dos valores naquele intervalo)

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

Tabela de Frequência: exemplo

- Gerando tabela de distribuição de frequência:
  - **3º Passo:** Calcular frequência relativa (número de ocorrências dos valores naquele intervalo dividido pelo tamanho do dataset)

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

Tabela de Frequência: exemplo

- Gerando tabela de distribuição de frequência:
  - **4º Passo:** Calcular a frequência percentual: Frequência Relativa * 100

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

Tabela de Frequência: exemplo

- Gerando tabela de distribuição de frequência:
  - **5º Passo:** Calcular a porcentagem acumulada

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

Tabela de Frequência: exemplo

- Gerando tabela de distribuição de frequência:
  - **6º Passo:** Calcular a densidade: Frequência relativa dividida pela amplitude
    - Na prática isso corresponde à altura do retângulo no gráfico

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

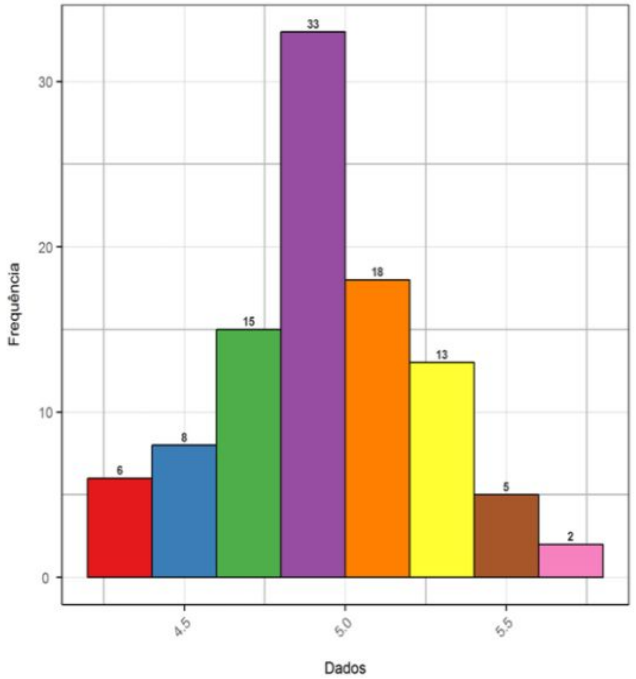
Tabela de Frequência: exemplo

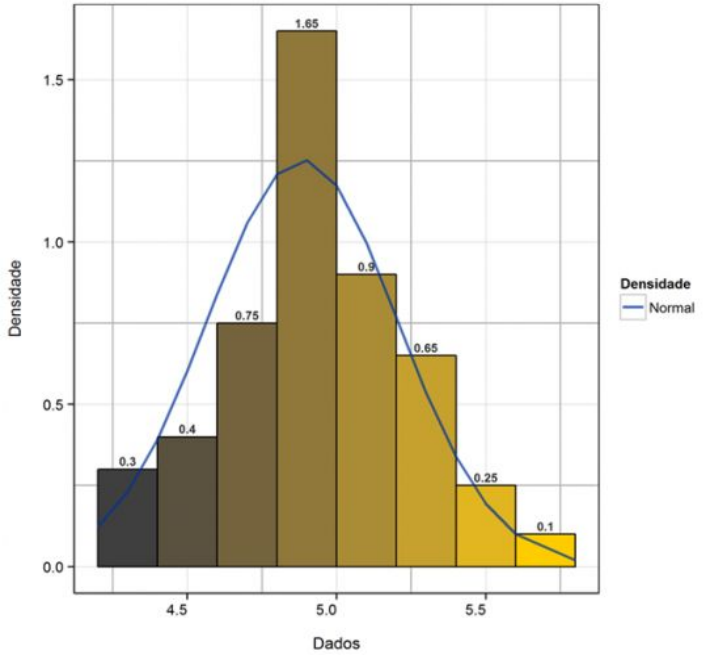
- Gerando tabela de distribuição de frequência:
  - **7º Passo:** Calcular o ponto médio (x_i) do grupo

Tabela de Frequências						
Classe	Frequência	Freq. Rel.	Freq. Perc.	Freq. Acum.	Densidades	Ponto Médio
[4,2 ; 4,4)	6	0,06	6	6	0,3	4,3
[4,4 ; 4,6)	8	0,08	8	14	0,4	4,5
[4,6 ; 4,8)	15	0,15	15	29	0,75	4,7
[4,8 ; 5)	33	0,33	33	62	1,65	4,9
[5 ; 5,2)	18	0,18	18	80	0,9	5,1
[5,2 ; 5,4)	13	0,13	13	93	0,65	5,3
[5,4 ; 5,6)	5	0,05	5	98	0,25	5,5
[5,6 ; 5,8)	2	0,02	2	100	0,1	5,7

Histograma

- Por fim, o histograma (representada pela frequência ou pela densidade):





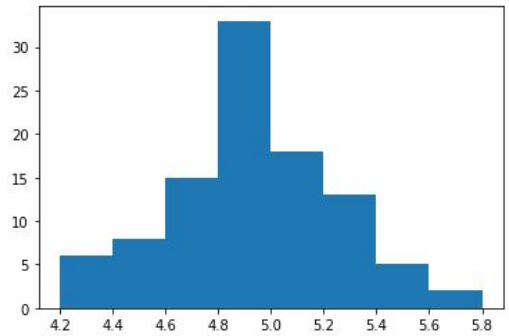
Histograma

- Usando Matplotlib

Calculando o número ideal de bins k = ⌈1 + 3,3 log10(tamanho_do_dataset)⌉

```
In [4]: k = math.ceil(1 + 3.3 * math.log10( Data_df.size ))
        k
Out[4]: 8

In [5]: import matplotlib.pyplot as plt
        plt.hist(Data_df[0], bins=k)
        plt.show()
```

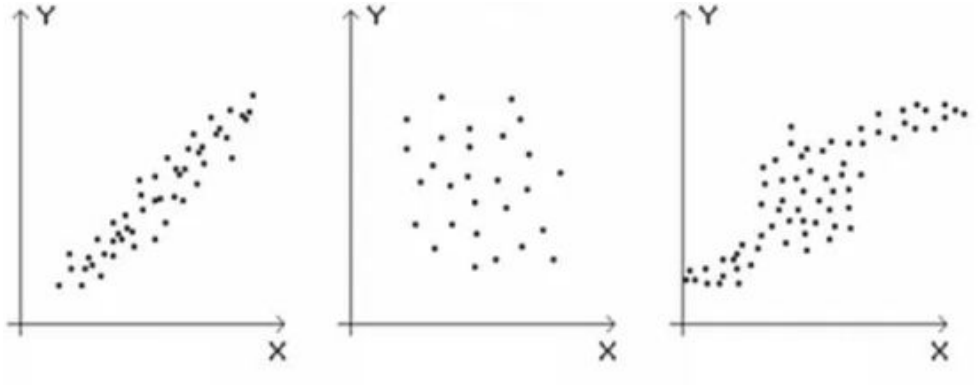


Para gerar a distribuição de frequência, basta descobrir o número de classes/ grupos (*bins*) e enviar como parâmetro para função hist() junto com os dados.

Correlações

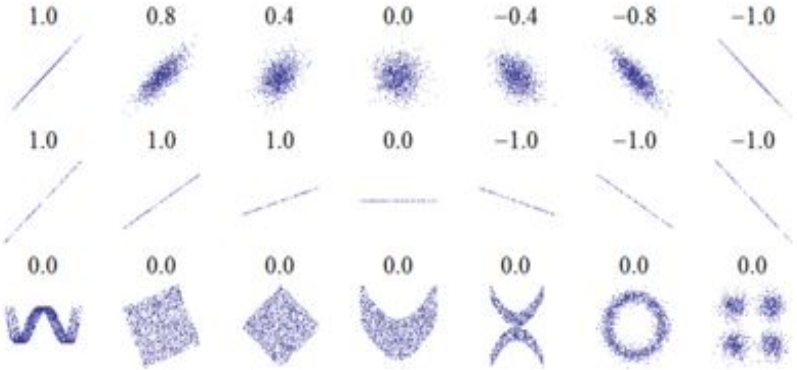
Dada duas variáveis X e Y (atributos, por exemplo), dizemos que:

- Se X e Y são **fortemente e positivamente correlacionadas**, então se X cresce, Y cresce;
- Se X e Y são **fortemente e negativamente correlacionadas**, então se X cresce, Y diminui;
- Se X e Y **não possuem correlação**, o comportamento de X não possui associação com Y.



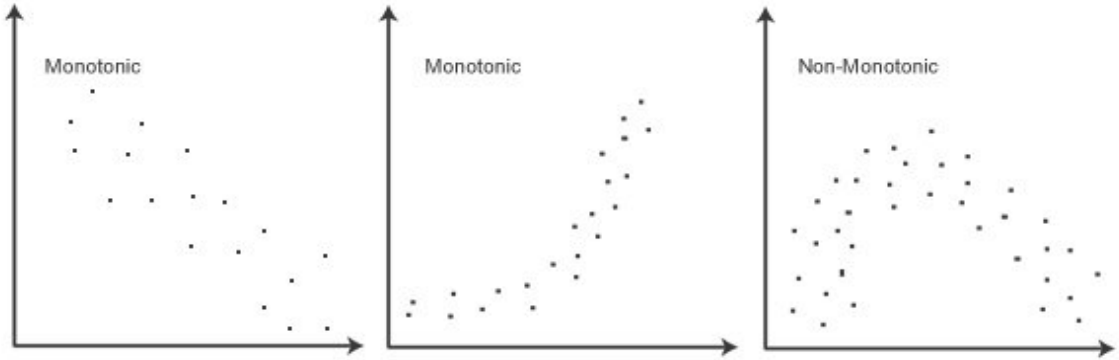
Correlações

- Coeficiente de correlação de ***Pearson***:
  - **Técnica Paramétrica**: Usado para variáveis que possuem **distribuição Normal** (estudaremos distribuição dos dados mais adiante);
  - Serve para medir a **correlação linear** entre duas variáveis;
  - O coeficiente varia de **-1 a 1:**
    - -1 é correlação forte e negativa;
    - 0 sem nenhuma correlação e
    - 1 correlação forte e positiva.



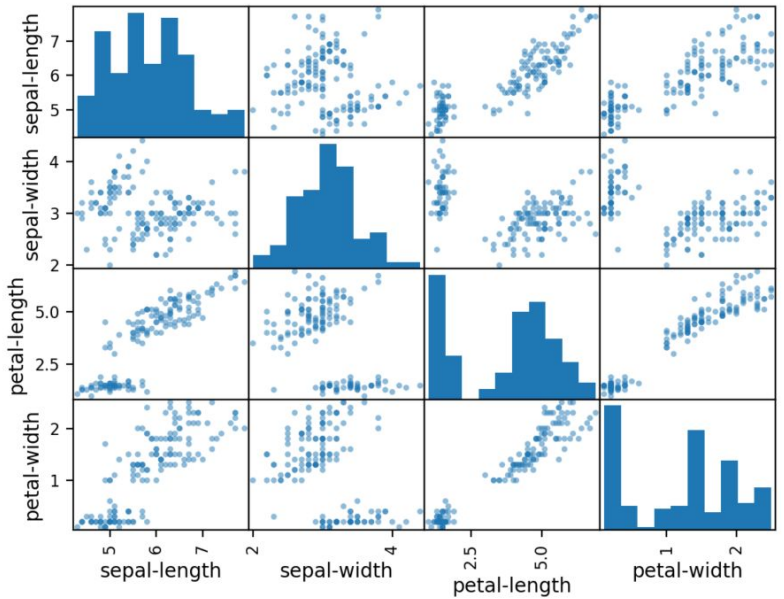
Correlações

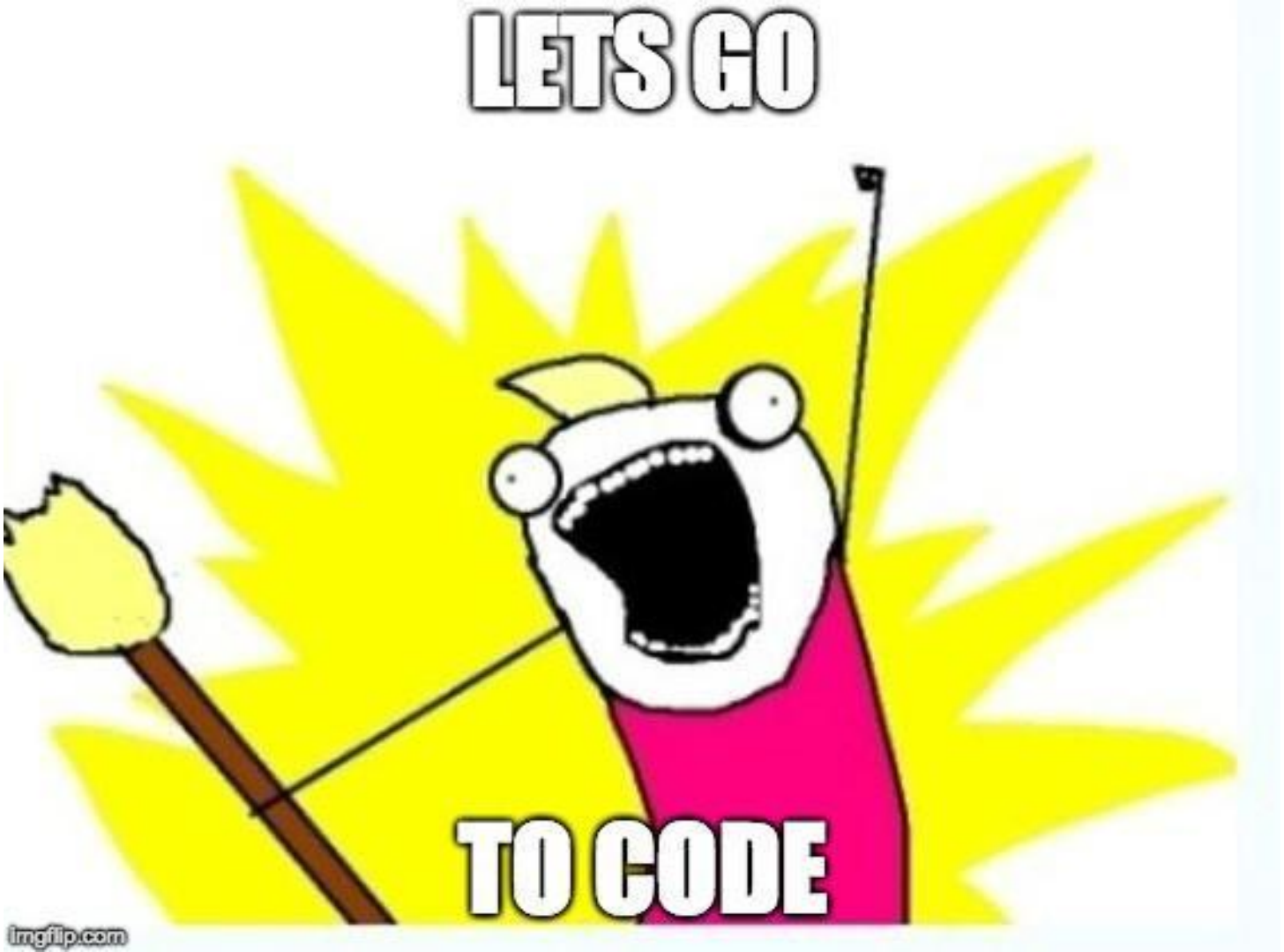
- Coeficiente de correlação de ***Spearman***:
  - Técnica Não-Paramétrica: Não requer que as variáveis sigam uma distribuição específica para ser aplicado;
  - Serve para medir a correlação **não-linear** entre duas variáveis;
  - O coeficiente varia de -1 a 1, onde:
    - -1 é correlação forte e negativa,
    - 0 sem nenhuma correlação e
    - 1 correlação forte e positiva.



Explorando duas ou mais variáveis

1. Em caso de múltiplas variáveis independentes (atributos, *features* ou colunas) no qual contém **distribuições Normal e não-Normal**, pode-se **aplicar *Spearman*** em todas;
2. Pode também aplicar **Pearson somente nas colunas que apresentarem distribuição Normal** e analisá-las isoladamente das demais.




LETS GO
TO CODE
imgflip.com