SENAI

Ciência de Dados (CIDA)

Aula – Resumo

Prof.: Hugo S. Idagawa

Resumo

CIDA

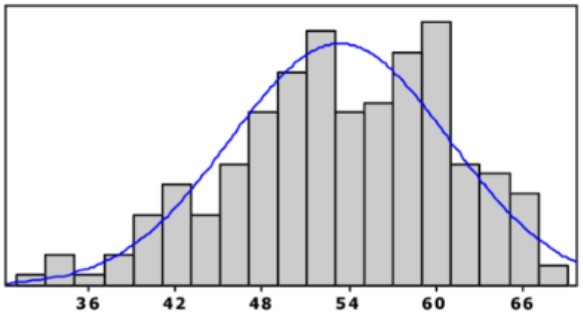
Visualização e exploração dos dados

Extração de dados, criação de modelos e tomada de decisão

Estatística

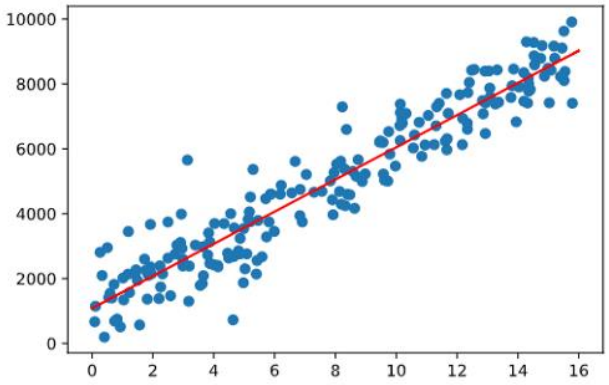
Descritiva

- Medidas resumo e tendências:
  - Média
  - Mediana
  - Moda
  - Desvio-padrão
  - Quantis
- Visualizações e gráficos:
  - Boxplot
  - Histogramas
  - Dispersões



Inferencial

- Correlação de Pearson
- Regressão linear



Resumo

CIDA

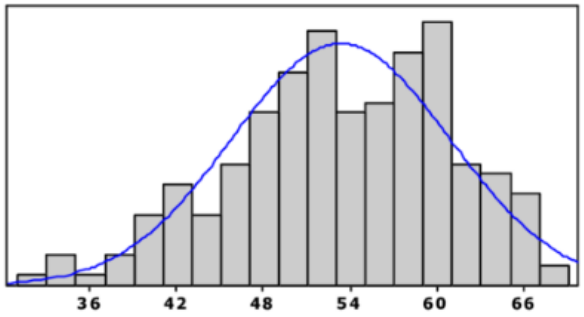
Visualização e exploração dos dados

Estatística

Extração de dados, criação de modelos e tomada de decisão

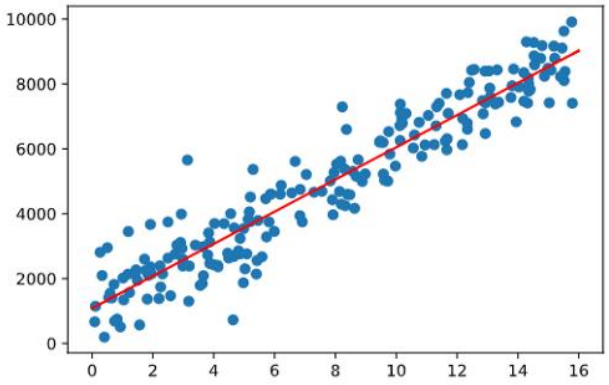
Descritiva

- Medidas resumo e tendências:
  - Média
  - Mediana
  - Moda
  - Desvio-padrão
  - Quantis
- Visualizações e gráficos:
  - Boxplot
  - Histogramas
  - Dispersões



Inferencial

- Correlação de Pearson
- Regressão linear



Resumo – Estatística Descritiva

- Medidas resumo e tendências:
  - **Média:** ponto de equilíbrio dos dados:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

  - **Mediana:** valor que divide os dados ordenados no meio:

$$md(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & se\ n\ é\ ímpar, \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & se\ n\ é\ par. \end{cases}$$

  - **Moda:** valor que aparece mais vezes na amostra.
  - **Desvio-padrão:** medida de de dispersão dos dados em relação à média:

$$var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

$$dp(x) = \sqrt{var(x)}$$

Resumo – Estatística Descritiva

- Medidas resumo e tendências:
  - **<u>Quantis:</u>** medida que permite representar os dados em posições não centrais.
    - <u>Ex:</u> Q(0,60) = 10 significa que 60% dos dados estão abaixo do valor 10 (inclusive).

$$Q(p) = \begin{cases} x_{(i)}, & \text{se } p = p_i = (i - 0{,}5)/n,\ i = 1, \ldots, n \\ (1 - f_i)Q(p_i) + f_i Q(p_{i+1}), & \text{se } p_i < p < p_{i+1} \\ x_{(1)}, & \text{se } 0 < p < p_1 \\ x_{(n)}, & \text{se } p_n < p < 1, \end{cases}$$

  - Quando queremos dividir os dados em 4 partes iguais os quantis são chamados de quartis:
    - ✓ Q_1: primeiro quartil (25% dos dados)
    - ✓ Q_2: segundo quartil (50% dos dados)
    - ✓ Q_3: terceiro quartil (75% dos dados)
    - ✓ Q_4: quarto quartil (100% dos dados)

Resumo – Estatística Descritiva

- **Exemplo:** um conjunto de corpos de prova foram fabricados utilizando-se a impressão 3D e, em seguida, algumas propriedades mecânicas desses corpos de prova foram avaliadas. Os corpos foram impressos em dois diferentes tipos de materiais (ABS e PLA) e a resistência mecânica, o alongamento total e a rugosidade dos corpos foram analisados. As tabelas abaixo apresentam os resultados da coleta de dados desse experimento.

Amostra	Material	Resistência Mecânica	Alongamento	Rugosidade
1	PLA	24	1,4	24
2	PLA	27	2,2	126
3	PLA	23	1,9	145
4	PLA	33	2,1	92
5	PLA	14	1,5	121
6	PLA	4	0,7	163

Resumo – Estatística Descritiva

Amostra	Material	Resistência Mecânica	Alongamento	Rugosidade
1	ABS	35	3,3	212
2	ABS	34	3,1	276
3	ABS	28	2,2	298
4	ABS	28	1,6	360
5	ABS	21	1,1	357
6	ABS	27	2,4	168

- A partir dos dados das duas tabelas, levante as medidas de resumo (média, mediana e desvio-padrão) das propriedades mecânicas de cada material e responda os itens a seguir:

a) Usando apenas a média como informação, qual dos materiais apresenta a melhor resistência mecânica, alongamento e rugosidade?

Resumo – Estatística Descritiva

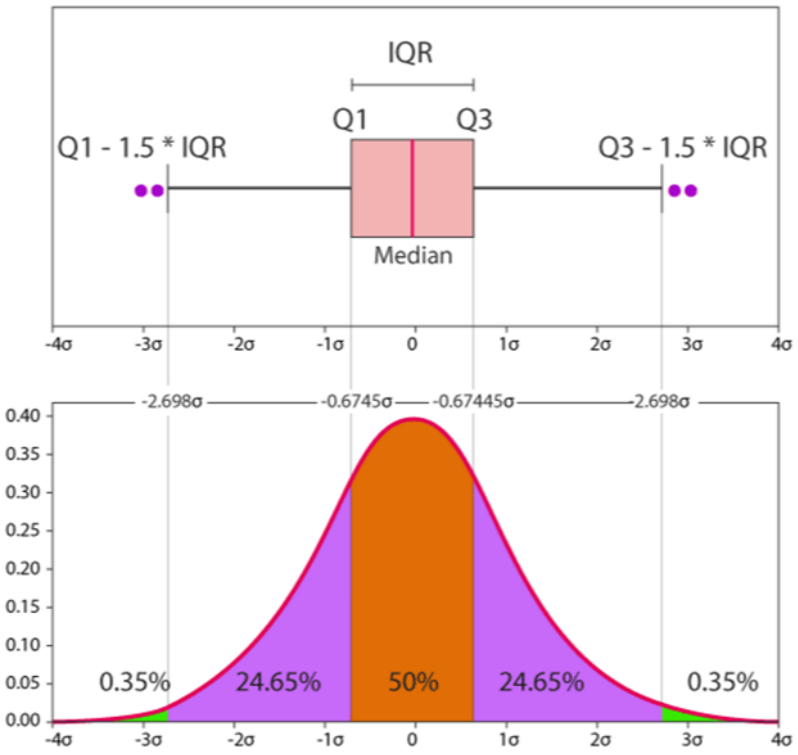
b) Em relação à dispersão dos resultados, qual dos materiais apresenta um resultado mais consistente em relação a cada uma das propriedades mecânicas?

c) Utilizando-se a média e o desvio-padrão como informação de decisão, qual dos materiais apresenta a maior resistência mecânica, alongamento e rugosidade?

d) Para cada um dos materiais e propriedade mecânica, qual seria a faixa de valores onde encontraríamos 50% dos valores?

Resumo – Estatística Descritiva

- Visualizações e gráficos:
  - **Boxplot:** visualização da distribuição dos dados que utiliza os quartis como pontos importantes:
    - ✓ **Limite inferior:** max(x_1; Q1 -1,5*dq)
    - ✓ **Limite superior:** min(x_n; Q3 + 1,5*dq)



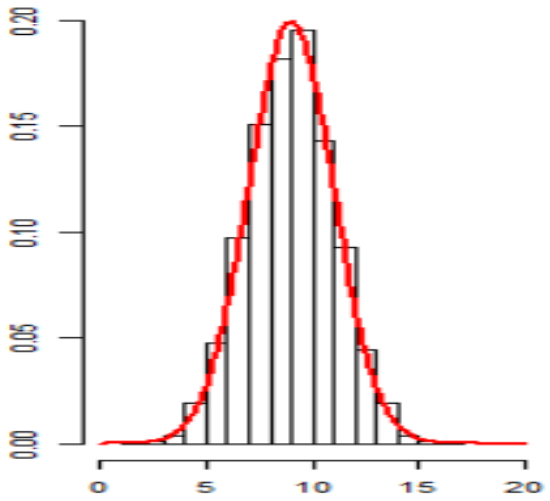
Resumo – Estatística Descritiva

- Visualizações e gráficos:

- **<u>Histogramas:</u>** distribuição em frequência que serve tanto para variáveis qualitativas, quanto quantitativas. Servem para identificar o tipo de distribuição que os dados apresentam e, em seguida, aplicar os testes estatísticos adequados. A quantidade de classes pode ser definida pela equação abaixo:

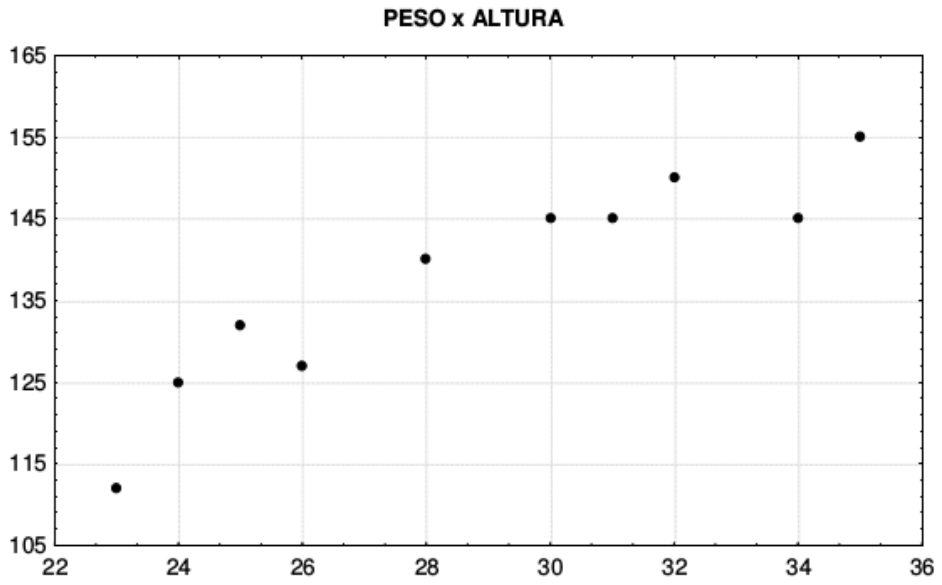
  ✓ **Fórmula de Sturges:**

$$k = 1 + 3{,}32 * \log_{10}(N)$$



Resumo – Estatística Descritiva

- Visualizações e gráficos:
  - **Dispersões:** as dispersões são gráficos de pontos (ou linhas) que permitem visualizar a relação existente entre duas ou mais variáveis. São úteis para posteriormente aplicar algum método de regressão e, assim, obter um modelo de previsão dos dados.



Resumo – Estatística Descritiva

- **Exemplo:** Utilizando os dados da tabela do experimento sobre impressão 3D, construa as visualizações abaixo:

  a) Para cada propriedade mecânica estudada, apresente uma comparação entre os boxplots de cada material (um boxplot para o PLA e outro para o ABS).
  b) Utilizando os resultados anteriormente, podemos identificar algum “outlier” nos dados?
  c) Com essa representação podemos supor que existe alguma diferença entre as propriedades mecânicas para cada material?

Resumo – Estatística Descritiva

- **Exemplo:** mais informações sobre o processo de impressão 3D foram coletadas e apresentadas na tabela abaixo, agora temos também os dados sobre preenchimento e espessura da camada. Utilizando esses novos dados, responda os itens a seguir:

Amostra	Material	Preenchimento	Espessura da camada	Resistência Mecânica	Alongamento
1	PLA	40	0,02	24	1,4
2	PLA	90	0,06	27	2,2
3	PLA	40	0,06	23	1,9
4	PLA	80	0,06	33	2,1
5	PLA	30	0,10	14	1,5
6	ABS	80	0,20	35	3,3
7	ABS	90	0,20	34	3,1
8	ABS	30	0,20	28	2,2
9	ABS	90	0,20	28	1,6

Resumo – Estatística Descritiva

a) Monte um histograma utilizando todas as amostras da tabela para a resistência mecânica e para o alongamento.

b) Faça um rascunho do gráfico de dispersão para a resistência mecânica em função do preenchimento para cada material. Verifique se existe alguma relação entre os dados.

c) Faça o mesmo para o alongamento em função do preenchimento.

Implementação – Estatística Descritiva

- A seguir estão apresentadas as funções de python utilizadas para se obter as medidas de resumo apresentadas anteriormente. Em vermelho temos os parâmetros obrigatórios e o parâmetro "dados" deve ser uma lista.

- Medidas resumo e tendências:
  - **Média:** `np.mean(dados)`
  - **Mediana:** `np.median(dados)`
  - **Moda:** `stats.mode(dados)`
  - **Desvio-padrão:** `np.std(dados)`
  - **Quantis:**
```
np.percentile(dados,
    percentual,
    method="averaged_inverted_cdf")
```

OBS: as bibliotecas utilizadas acima são: o *numpy (np)* e o *statistics (stats)*.

Implementação – Estatística Descritiva

- A seguir estão apresentadas as funções de python utilizadas para criar os gráficos de visualização. Em vermelho temos os parâmetros obrigatórios e os parâmetros "dados", "x" e "y" devem ser uma lista. Os parâmetros em azul são opcionais e são utilizados para obter maior flexibilidade no uso das funções.

- Visualizações e gráficos:
  - **Boxplot:** `plt.boxplot(dados)`
  - **Histograma:** `plt.hist(dados, bins=10)`
    `np.histogram(dados, bins=10)`
  - **Dispersões:**

```
plt.scatter(x, y, s=20, c='r', marker='o')
plt.plot(x, y, color='r', marker='o')
```

OBS: as bibliotecas utilizadas acima são: o *numpy (np)* e o *matplotlib (plt)*.

Resumo

CIDA

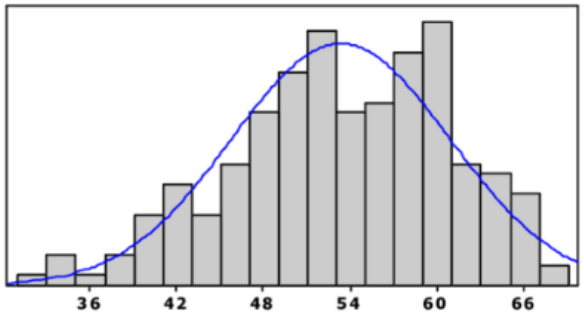
Visualização e exploração dos dados

Estatística

Extração de dados, criação de modelos e tomada de decisão

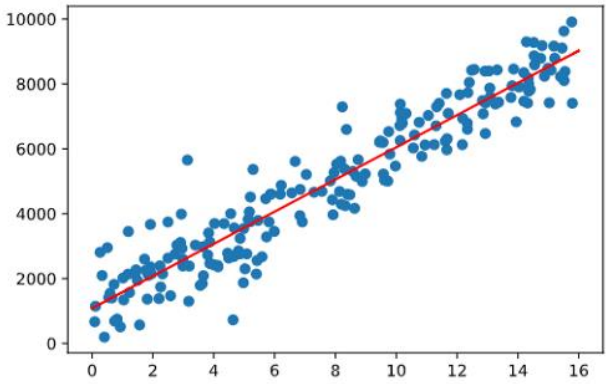
Descritiva

- Medidas resumo e tendências:
  - Média
  - Mediana
  - Moda
  - Desvio-padrão
  - Quantis
- Visualizações e gráficos:
  - Boxplot
  - Histogramas
  - Dispersões



Inferencial

- Correlação de Pearson
- Regressão linear



Estatística Inferencial

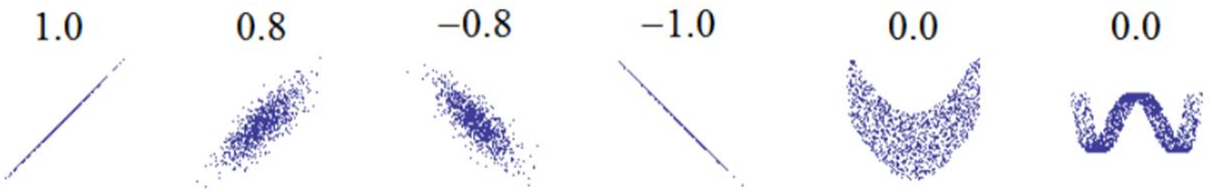
- Estatística Inferecial

- **<u>Correlação:</u>** medida estatística que quantifica a “força” da relação existente entre duas variáveis. Serve para identificar se algum parâmetro tem influência na variação de outro parâmetro de saída em estudo.

  Aqui vamos utilizar a correlação linear de Pearson, que é calculada segundo a fórmula abaixo e deve ser um valor entre -1 (correlação negativa perfeita) e 1 (correlação positiva perfeita).

$$\rho = \frac{Cov(X,Y)}{S_X S_Y}$$

  A seguir temos uma comparação entre o valor da correlação e o gráfico de dispersão:



Estatística Inferencial

- Estatística Inferecial
  - **Correlação:** o quadro abaixo apresenta algumas regras que auxiliam na interpretação do valor do coeficiente de correlação de Pearson:

Interpretação	Coeficiente de correlação linear de Pearson
Forte associação positiva	(0,9; 1]
Alta associação positiva	(0,7; 0,9]
Moderada associação positiva	(0,5; 0,7]
Baixa associação positiva	(0,3; 0,5]
Associação nula	[0; 0,3]

Interpretação	Coeficiente de correlação linear de Pearson
Forte associação negativa	[-1; -0,9)
Alta associação positiva	[-0,9; -0,7)
Moderada associação negativa	[-0,7; -0,5)
Baixa associação negativa	[-0,5; -0,3)
Associação nula	[-0,3; 0]

Estatística Inferencial

- **Exemplo:** ao avaliar os dados do processo de impressão 3D apresentados no exemplo anterior, foi obtida a seguinte matriz de correlação de Pearson:

	Preenchimento	Espessura de Camada	Resistência Mecânica	Alongamento
Preenchimento	1,0	0,2826	0,7030	0,5025
Espessura de Camada	0,2826	1,0	0,4472	0,5647
Resistência Mecânica	0,7030	0,4472	1,0	0,7625
Alongamento	0,5025	0,5647	0,7625	1,0

A partir dos resultados dessa tabela, responda os itens a seguir:

a) Existem algum parâmetro de impressão que afeta de maneira positiva a resistência mecânica?
b) Existem algum parâmetro de impressão que afeta de maneira positiva o alongamento?

Estatística Inferencial

c) Se aumentarmos o valor de preenchimento das peças impressas, o que podemos esperar que vai acontecer com a resistência mecânica e com o alongamento?