Formação Livre em Ciência de Dados

Aula 3 | Módulo Básico II

João Pedro Passos Pereira

Roteiro Aula 3

1. Parte 1:
    a. Conceitos básicos em Ciência de Dados II;
    b. Lógica II;
2. Parte 2:
    a. Estatística Descritiva I.

Material por: João Pedro Passos Pereira

Parte 1

Material por: João Pedro Passos Pereira

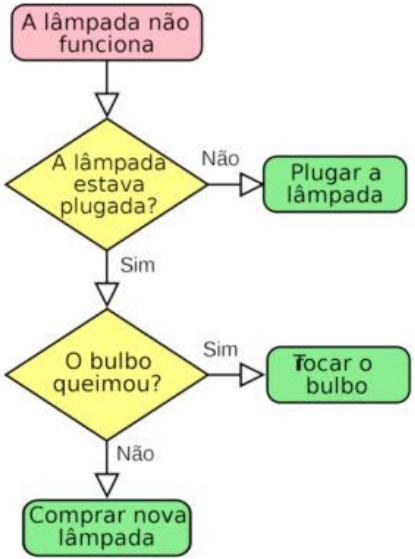
(1) Conceitos básicos em CIência de Dados II

Material por: João Pedro Passos Pereira

Conceitos básicos (1)

Algoritmo

"Um algoritmo é uma série de **instruções** que se dá a um computador para que ele **solucione algum problema proposto (...)**. Essas instruções podem ser simples, como retirar todos os caracteres especiais de um número de telefone, ou mais complexas, como prever o número de vendas que um representante irá fazer em determinado período no futuro."

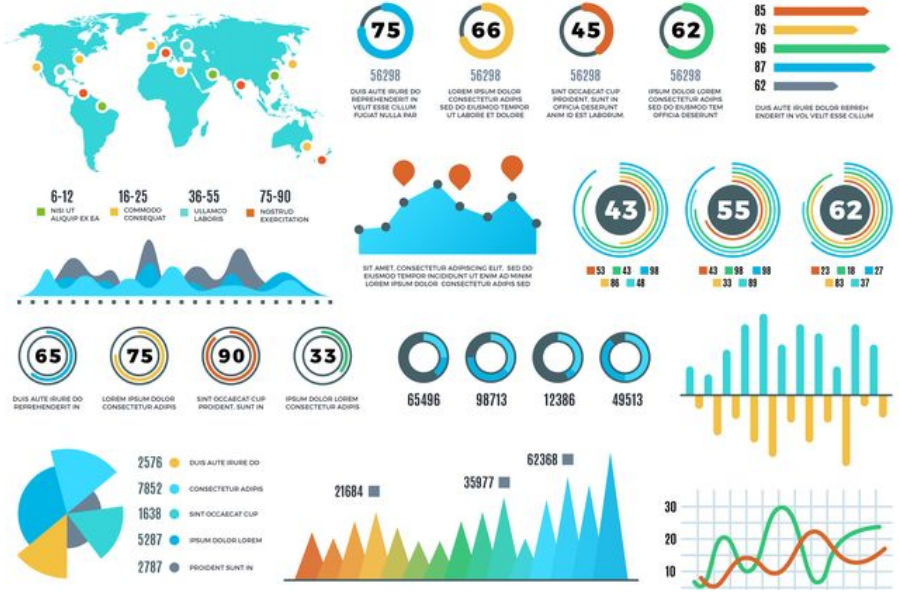


Material por: João Pedro Passos Pereira

Conceitos básicos (2)

Data Visualization

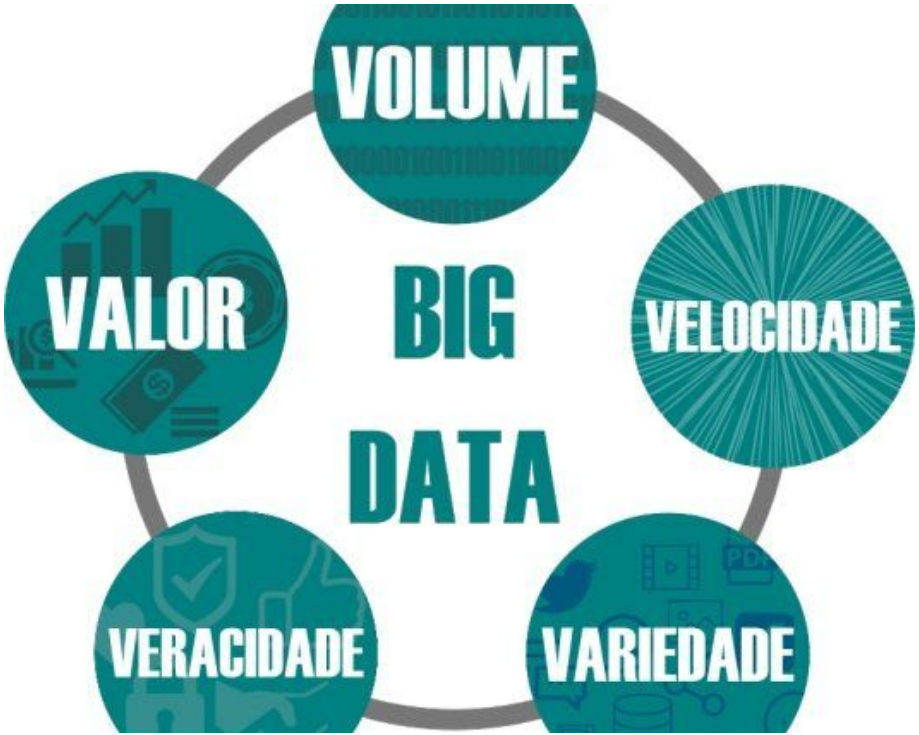
"A visualização de dados é o processo de **traduzir grandes conjuntos de dados e métricas em tabelas, gráficos e outros recursos visuais** . A representação visual de dados resultante torna mais fácil identificar e compartilhar tendências em tempo real, outliers e novos insights sobre as informações representadas nos dados."



Material por: João Pedro Passos Pereira

Conceitos básicos (3)

Big Data
"Big data é a área do conhecimento que estuda como tratar, analisar e obter informações a partir de **conjuntos de dados grandes demais para serem analisados por sistemas tradicionais.**"



Conceitos básicos (4)

Business Intelligence (BI)

"Inteligência de negócios refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem **suporte a gestão de negócios.**"

Material por: João Pedro Passos Pereira

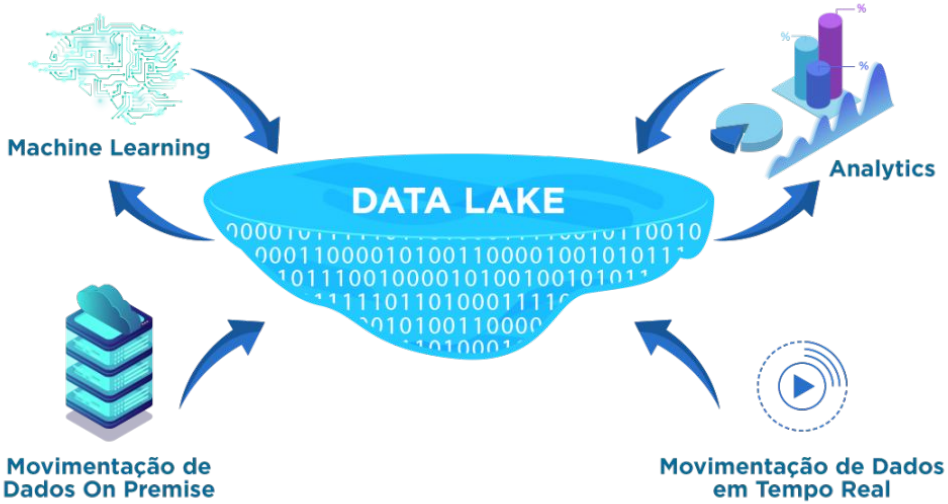
Conceitos básicos (5)

Cloud Computing

"Computação em nuvem é um termo coloquial para a **disponibilidade sob demanda de recursos do sistema de computador**, especialmente armazenamento de dados e capacidade de computação, **sem o gerenciamento ativo direto do utilizador.**"

Conceitos básicos (6)

Data Lake
“Um data lake (“lago de dados” em português) é um **repositório** para armazenamento de dados, que na maioria das vezes **não são estruturados ou estão em seu estado natural (não tratados).**”



Material por: João Pedro Passos Pereira

Conceitos básicos (7)

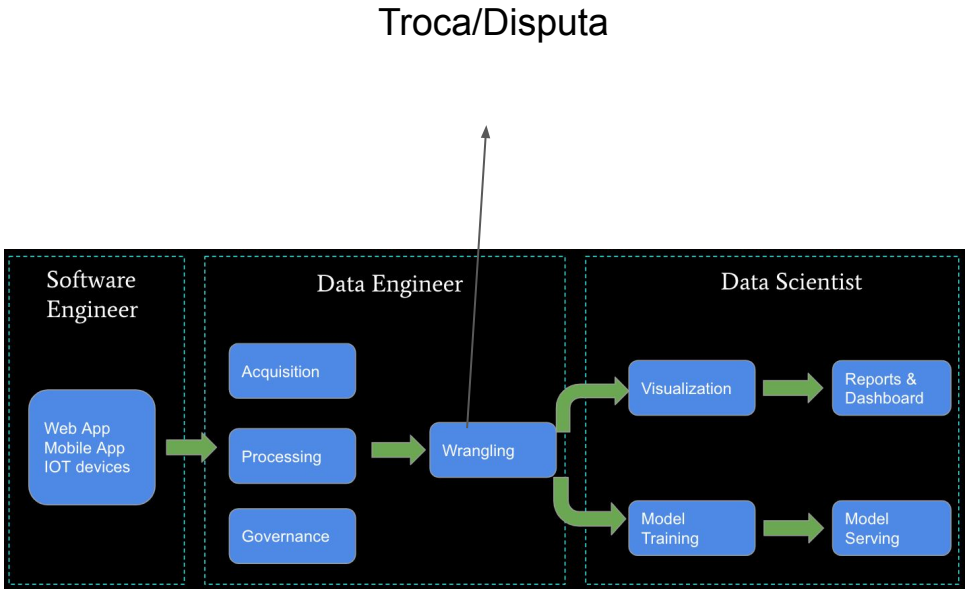
Data Transformation
“Transformação de dados é o processo de **conversão de formato, estrutura ou valores de dados**. Esse processo pode ser realizado antes ou depois da chegada dos dados em um base de dados”

Material por: João Pedro Passos Pereira

Conceitos básicos (8)

Engenheiro de Dados

"O engenheiro de dados é responsável por implementar mecanismos que façam a **coleta, armazenamento e transformação de dados**, para que estes, ao serem disponibilizados ao usuário final, sejam usáveis."



Material por: João Pedro Passos Pereira

Conceitos básicos (9)

Data Warehouse
“O data warehouse é um **local usado para armazenar grandes quantidades de dados do negócio e suportar atividades de análise e business intelligence**. Ele se diferencia do data lake nos seguintes aspectos:

Data Warehouse

Material por: João Pedro Passos Pereira

Conceitos iniciais (10)

ELT	**ETL**
“A sigla ELT significa “Extract, Load, Transform”, ou em português “Extrair, Carregar, Transformar”, e refere-se ao processo usado para replicar dados de uma fonte em uma base de dados. No ELT, **os dados são extraídos da fonte, carregados no seu destino e só então transformados.”**	“A sigla ETL significa “Extract, Transform, Load”, ou em português “Extrair, Transformar, Carregar”, e refere-se também ao processo usado para replicar dados de uma fonte em um destino. Em contrapartida ao ELT, **os dados são extraídos e carregados no destino somente após realizada a transformação neles.**”

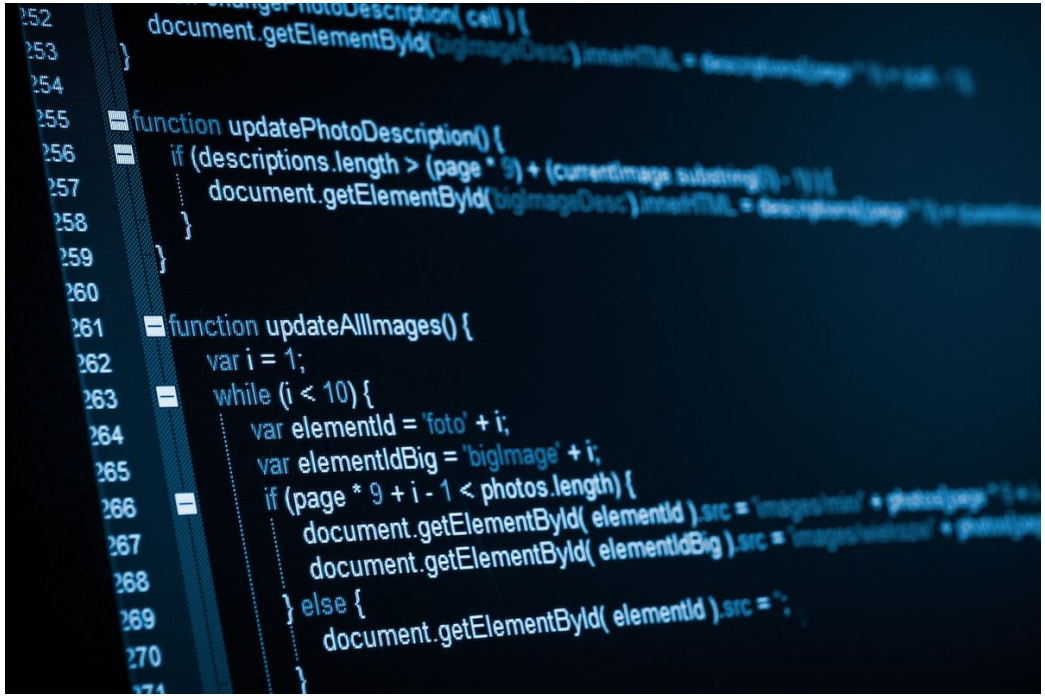
ETL vs ELT

Material por: João Pedro Passos Pereira

Conceitos básicos (11)

Machine Learning

“Machine Learning, ou aprendizagem de máquina, é um processo em que **um computador usa um algoritmo para entender um conjunto de dados**, e com base nesse entendimento **fazer previsões ou tomar decisões.**”

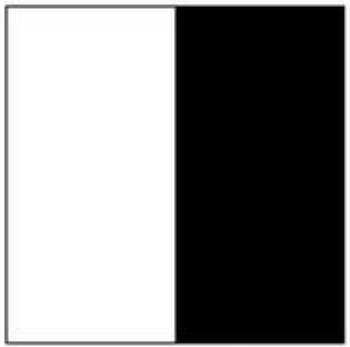


Material por: João Pedro Passos Pereira

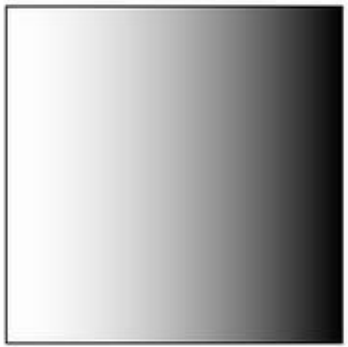
(2) Lógica II

Material por: João Pedro Passos Pereira

Observação inicial - Alternativa à Lógica Clássica



Lógica Clássica

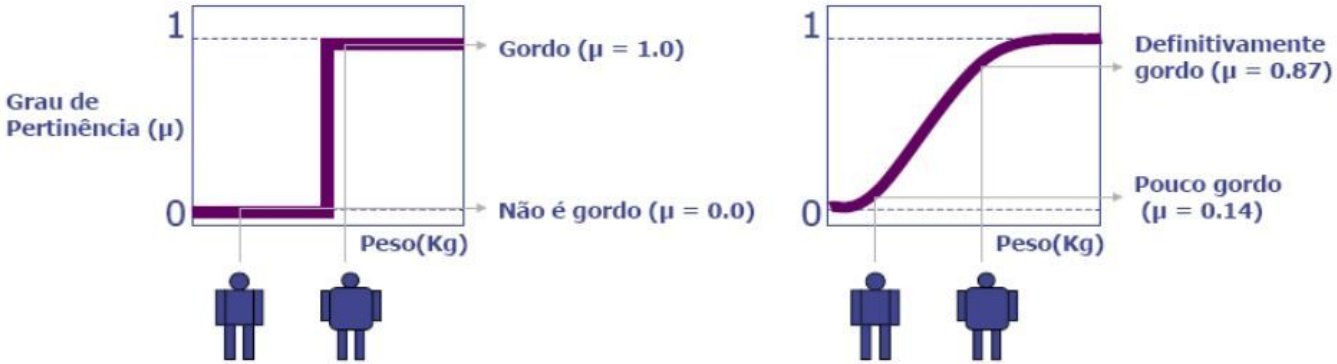


Lógica Fuzzy

Material por: João Pedro Passos Pereira

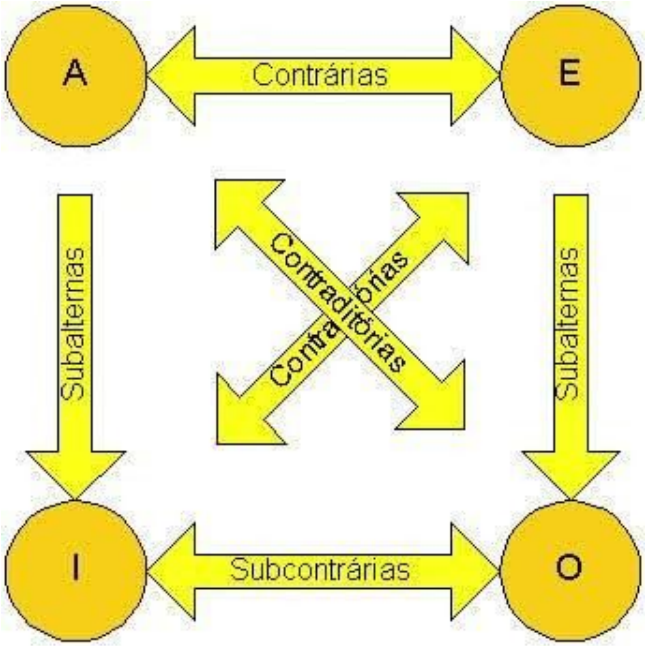
O que é Lógica Fuzzy?

Exemplo:
Quando uma pessoa é considerada gorda?



Material por: João Pedro Passos Pereira

O quadrado lógico



Material por: João Pedro Passos Pereira

Exemplo

Se todos Wargs são Twerps e nenhum dos Twerps são Gollums, então nenhum dos Gollums são definitivamente Wargs?

Wargs = w
Twerps = t
Gollums = g

w C t
t /C g
logo
g /C W (verdade)

-
C = CONTIDO
/C = NÃO CONTIDO

Parte 2

Material por: João Pedro Passos Pereira

(3) Estatística Descritiva I

Material por: João Pedro Passos Pereira

Estatística descritiva

"A estatística descritiva é um ramo da estatística que aplica várias técnicas para **descrever** e **sumarizar** um conjunto de dados. Diferencia-se da estatística inferencial, ou estatística indutiva, pelo objetivo: **organizar, sumarizar dados ao invés de usar os dados em aprendizado sobre a população.**"

Estatística descritiva - Métodos

Os dois principais:

- Tabelas de frequência;
- Gráficos.

Métodos acessórios:

- Medidas de tendência central;
- Medidas de dispersão;
- Medidas de distribuição dos dados (assimetria, curtose, boxplot);
- Estatísticas de resumo;
- Estatísticas de tendência.

Material por: João Pedro Passos Pereira

Tabelas de frequência

- É uma das formas de sumarizar um conjunto de dados. E a mais utilizada;
- Consiste em obter uma representação tabular ou gráfica baseada nas frequências com que os dados aparecem dentro do conjunto de dados utilizado, de modo a observar tendências básicas e possíveis caminhos iniciais de análise;
- Para cada tipo de variável, um jeito diferente de realizar uma distribuição de frequências.
  - Para variáveis quantitativas discretas, utiliza-se a **tabela de frequências simples**.
  - Para variáveis quantitativas contínuas, utiliza-se a **tabela de classes de frequência**.
- Na tabela de frequências simples é realizada uma simples contagem de cada valor individual que aparece nos dados e a contabilidade de quantas vezes tal valor aparece.
- Já na tabela de classes de frequência você separa os dados em intervalos e agrupa os dados dentro de cada intervalo.

Material por: João Pedro Passos Pereira

Tabelas de frequência - Casos

Variável discreta - Tabela de Frequência simples	Variável contínua - Tabela de Classes de Frequência
xi = variável na posição i **fi = quantas vezes (frequência/ contagem) o valor x na posição i aparece** **ex: idades de crianças (primeira coluna) X quantas vezes as idades aparecem (segunas coluna)**	

x_i	f_i
0	8
1	5
2	5
3	2

Classe	Intervalo de classe	f_i
1	2 |——— 4	4
2	4 |——— 6	12
3	6 |——— 8	10
4	8 |——— 10	4

Material por: João Pedro Passos Pereira

Tabelas de frequência - Pontos importantes

- No caso do cálculo da distribuição de frequência para variáveis discretas, temos um cálculo muito simples, basta realizar a contagem.
- **Mas para o caso de variáveis contínuas, como definir as classes? Como definir o limite superior e o limite inferior de cada uma das classes? Como classificar por classes valores que estão em regiões limites? O que mais podemos extrair? Quais os problemas podemos enfrentar?**
- Mais adiante vamos entender melhor tais questões, antes, vamos estudar em detalhes as medidas fundamentais da estatística: medidas de posição e medidas de dispersão dos dados.

Medidas de posição

Média Aritmética Simples	Média Ponderada	Mediana
"É a soma das observações observações dividida dividida pelo número de observações. Seus valores valores tendem a se localizar em um ponto central dentro de um conjunto conjunto de dados. **É a medida de posição mais utilizada"** $\bar{X} - \frac{\sum x_i}{n}$	"Nos cálculos envolvendo média aritmética simples, simples, todas as ocorrências têm exatamente a mesma importância ou o mesmo peso. No entanto, existem casos onde as ocorrências têm importância relativa diferente. Nestes casos, o cálculo da média deve levar em conta esta importância relativa ou peso relativo." $\bar{X}_p = \frac{X_1 \cdot p_1 + X_2 \cdot p_2 + \cdots + X_n \cdot p_n}{p_1 + p_2 + \cdots + p_n} = \frac{\sum_{i=1}^{n} X_i \cdot p_i}{\sum_{i=1}^{n} p_i}$	"Ocupa a posição posição central de uma série de observações ordenadas, ou seja, é o valor que divide os dados em duas partes iguais (isto é, em duas partes de 50% cada). 11 – 12 – 13 – (16) – 17 – 20 – 25 Me=16 11 – 12 – 13 – (16) - (17) – 20 – 25 – 26 Me=(16+17)/2 = 16,5

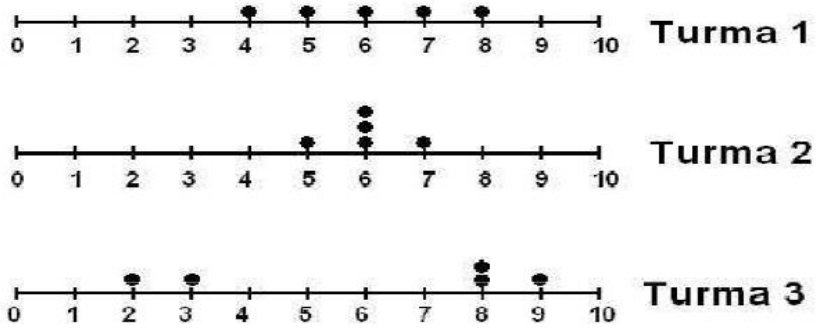
Material por: João Pedro Passos Pereira

Medidas de posição

Moda	**Quartis / Decis / Percentis**
É o valor mais frequente em um conjunto de dados. 1. Amodal: nenhum valor se repete mais que outros; 2. Unimodal: Um valor se repete mais do que outros; 3. Bimodal: dois valores se repetem mais; 4. Multimodal: dois valores que se repetem a mesma quantidade de vezes Usada em variáveis qualitativas	1. Quartis: dividem o conjunto de dados em 4 (quatro) partes iguais; 2. Decis: dividem o conjunto de dados em 10 (dez) partes iguais; 3. Percentis: dividem o conjunto de dados em 100 (cem) partes iguais; q2 = d5 = p50 = mediana

Material por: João Pedro Passos Pereira

Medidas de dispersão -> avaliação de variabilidade



Observações importantes

i) As três turmas possuem a mesma média.
ii) As notas estão distribuídas sob diferentes formas.
iii) A média resume o conjunto de dados apenas posição central.
iv) A média não fornece informações sobre a variabilidade dos dados.

Material por: João Pedro Passos Pereira

Medidas de dispersão -> avaliação de variabilidade

Amplitude Total (AT)	**Desvio médio**	**Variância**
"Verifica que a amplitude como medida de dispersão é limitada. Essa medida só depende dos valores valores extremos, ou seja, não é afetada pela dispersão dos valores internos" $AT = Maior\ valor - Menor\ valor$	"O desvio médio é uma medida de "VARIABILIDADE ABSOLUTA". Ela mede a variabilidade do conjunto em termos de desvios em relação à média aritmética. É uma quantidade sempre não negativa e expressa na mesma unidade de medida da variável." $DM = \frac{\sum \lvert X_i - \bar{X} \rvert}{n}$ Pouco usada.	"Mede a variabilidade do conjunto conjunto em termos de desvios quadrados quadrados em relação relação à média aritmética." $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$

Medidas de dispersão -> avaliação de variabilidade

Desvio padrão
É a raiz quadrada da variância. Mede o desvio em termos reais (na mesma medida que os dados coletados). Desvio Padrão = $\sqrt{Variância}$ (Raiz quadrada da Variância).

Material por: João Pedro Passos Pereira

Medidas de dispersão -> avaliação de variabilidade

Coeficiente de variação

É uma medida de "VARIABILIDADE RELATIVA", útil para comparar a variabilidade de observações com diferentes unidades de medida. É definida por:

$$CV = \frac{S}{\overline{X}} \times 100$$

Material por: João Pedro Passos Pereira

Medidas de dispersão -> avaliação de variabilidade

Exemplo

VALORES	MÉDIA	D.P.	C.V.
1 - 2 - 3	2	1	50 %
100 - 200 - 300	200	100	50 %
101 - 102 - 103	102	1	1 %

Material por: João Pedro Passos Pereira

Observações sobre “n”

Amostra	**População**
Assume-se variância maior e desvio padrão maior, portanto usa-se n-1 (n -1 graus de liberdade)	Assume-se variância e desvio padrão normalizados, portanto usa-se n.

Contatos

Meus links úteis

Instagram: @j0pewd2

Site: https://joaopedropereira.com.br

E-mail: contato@joaopedropereira.com.br

Linkedin: https://linkedin.com/in/joaopedrowd/

Material por: João Pedro Passos Pereira