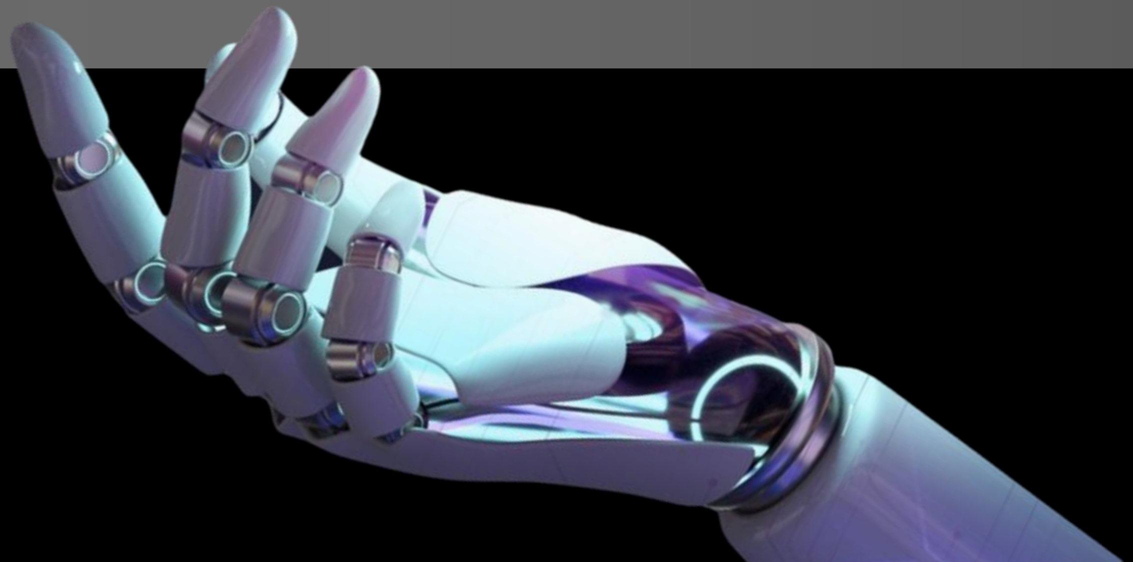
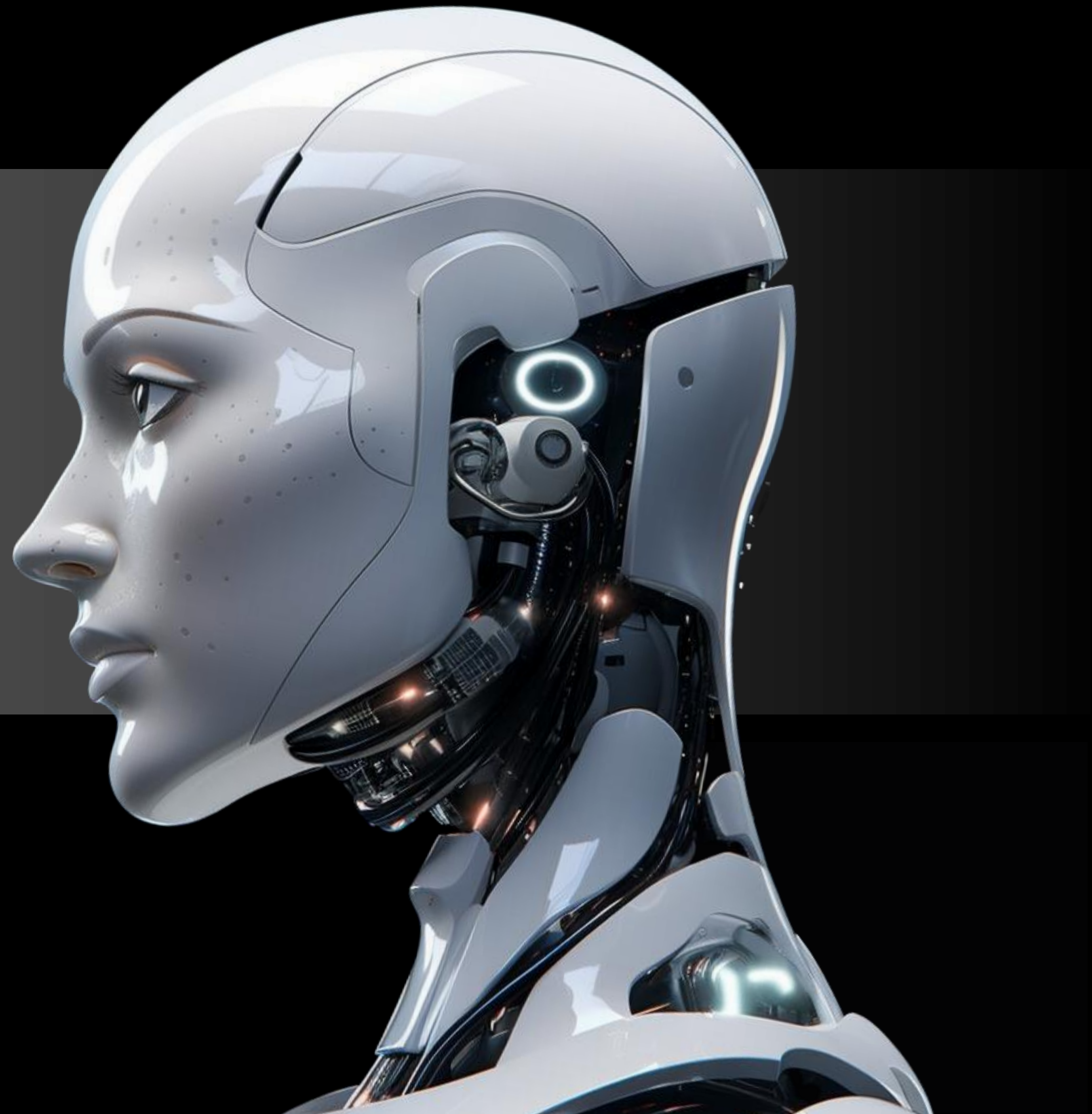
PROJETO </> EDUCAÇÃO

DO FUTURO

Data Science 2

AULA 02





AULA 02

CONTEÚDOS

- Introdução a estatística;
- População e amostra;
- Medidas de frequência;
- Medidas de tendência ;
- Medidas de dispersão;
- Intervalo Interquartílico e Outliers;
- Normalização e Padronização;
- Covariância e Correlação;

youth! </> space

Introdução à Estatística

Introdução a Estatística

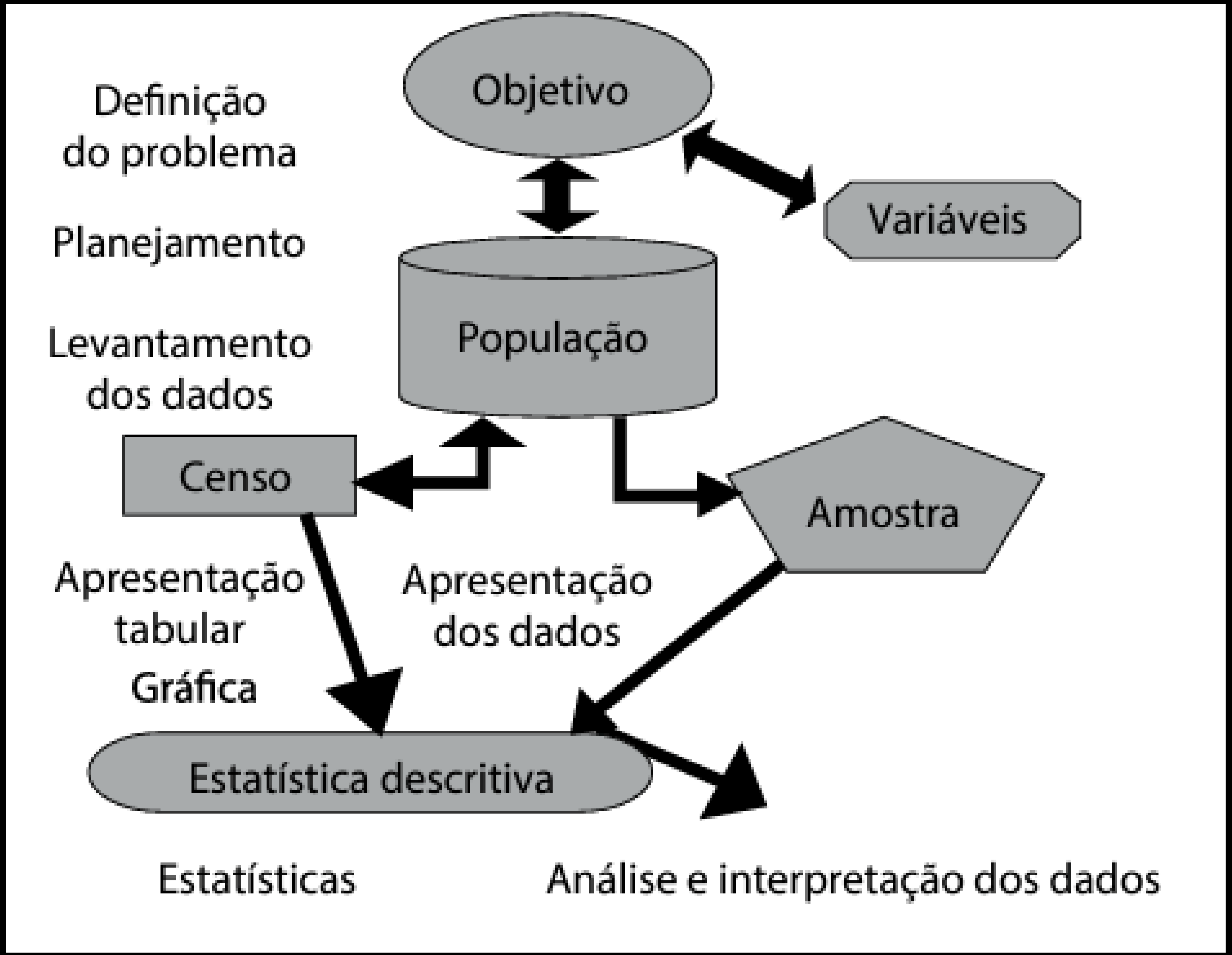
O que é Estatística?

Estatística: É a ciência que envolve a coleta, organização, análise, interpretação e apresentação de dados. É usada para compreender fenômenos e tomar decisões baseadas em dados.

A coleta, a organização e a descrição dos dados estão a cargo da Estatística **Descritiva**, enquanto a análise e a interpretação desses dados ficam a cargo da Estatística **Indutiva** ou **Inferencial**.

Introdução a Estatística

Visão geral do processo estatístico



População e amostra

População

É um conjunto de elementos com pelo menos uma característica em comum, que deve delimitar inequivocamente quais os elementos pertencem à população e quais não pertencem. Exemplos: os alunos de uma universidade, os clientes de um banco.

População e amostra

A População pode ser:

Finita: quando o número de unidades a observar pode ser contado e é limitado. Ex: alunos matriculados nas escolas públicas, pessoas que possuem aparelho telefone celular...

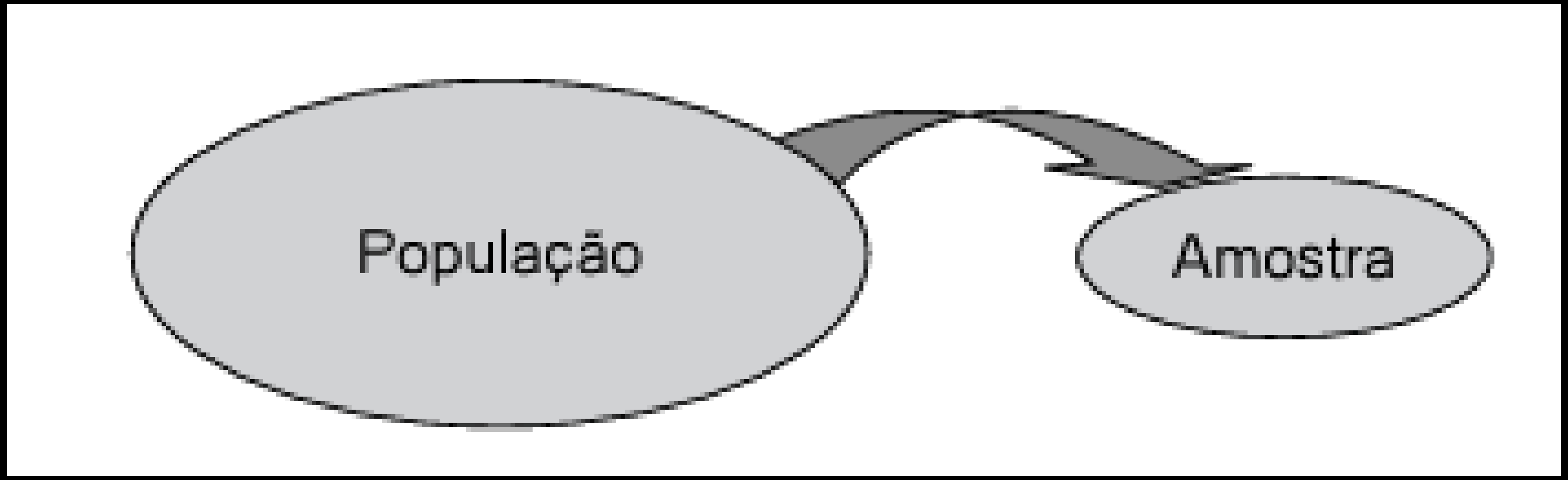
Infinita: quando a quantidade de observação é ilimitada ou quando as unidades da população não podem ser contadas. Ex: conjunto de medidas de determinado comprimento, gases, líquidos, em que as unidades não podem ser identificadas ou contadas.

Censo: É uma coleção de dados relativos a todos os elementos de uma população.

População e amostra

Amostra

É um subconjunto de uma população, necessariamente finito, pois todos os seus elementos serão examinados para efeito da realização do estudo estatístico desejado.



Amostragem é uma técnica especial usada para recolher amostras que garante o acaso na escolha de modo a garantir à amostra o caráter de representatividade.

População e amostra

Amostragem: Três principais tipos

Amostragem casual simples: composta de elementos retirados ao acaso da população, ou seja, consiste em selecionar a amostra através de um sorteio. Dessa maneira, todos os elementos da população terão igual probabilidade de serem escolhidos.

Amostragem sistemática: É utilizada quando a população está naturalmente ordenada, como listas telefônicas, fichas de cadastramento etc.

Amostragem estratificada: composta por elementos provenientes da divisão da população em subgrupos denominados estratos (por exemplo, por sexo, renda, bairro etc.)

Medidas de Frequência

Medidas de Frequência: Refletem a contagem de ocorrência de valores em um conjunto de dados.

Frequência Absoluta (f): Número de vezes que um valor ocorre.
Frequência Relativa (fr): Proporção ou porcentagem de vezes que um valor ocorre.

Frequência acumulada (Fi): é o total das frequências de todos os valores inferiores ao limite superior do intervalo de uma dada classe.

Frequência acumulada relativa (Fri): é a frequência acumulada da classe dividida pela frequência total da distribuição.

Medidas de Frequência

Vamos fazer a tabela de distribuição de frequência dos seguintes dados qulitativos:

Coca-cola	Sprite	Coca-cola	Pepsi-cola
Coca-cola zero	Coca-cola	Coca-cola	Coca-cola
Pepsi-cola	Sprite	Pepsi-cola	Pepsi-cola
Coca-cola	Pepsi-cola	Coca-cola	Sprite
Coca-cola zero	Coca-cola	Sprite	Coca-cola zero
Pepsi-cola	Pepsi-cola	Sprite	Coca-cola zero

Medidas de Frequência

Vamos fazer a tabela de distribuição de frequência dos seguintes dados:

DISTRIBUIÇÃO DE FREQUÊNCIA DAS COMPRAS DE REFRIGERANTES		
Refrigerantes	**Frequência Absoluta**	**Frequência Relativa**
Coca-cola	8	0,3333
Coca-cola zero	4	0,1667
Pepsi-cola	7	0,2917
Sprite	5	0,2083
Total	24	1

Medidas de Frequência

Se os dados forem quantitativos, podemos proceder de duas formas diferentes pois eles podem ser:

Quantitativo discreto: aquele que pode assumir apenas valores pertencentes a um conjunto enumerável.

Quantitativo contínuo: pode assumir qualquer valor em certo intervalo de variação.

Medidas de Frequência

Se os dados forem quantitativos discretos, podemos fazer a tabela de frequência da mesma forma que foi feito anteriormente, porém se ele forem contínuos , é melhor usar o **intervalo de classe**.

Vamos imaginar uma pesquisa referente às estaturas de quarenta alunos que compõem uma amostra dos alunos de uma universidade, o que resultou na tabela de valores a seguir:

ESTATURA DOS ALUNOS DE UMA UNIVERSIDADE									
166	160	161	150	162	160	165	167	164	160
162	168	161	163	156	173	160	155	164	168
155	152	163	160	155	155	169	151	170	164
154	161	156	172	153	157	156	158	158	161

Medidas de Frequência

Construindo a tabela de frequência absoluta teremos:

Distribuição de frequência da estatura de 40 alunos de uma universidade	
Estatura (cm)	Frequência
150	1
151	1
152	1
153	1
154	1
155	4
156	3
157	1
158	2
160	5
161	4
162	2
163	2
164	3
165	1
166	1
167	1
168	2
169	1
170	1
172	1
173	1
Total	**40**

Não é a visualização ideal.

Medidas de Frequência

A tabela a seguir foi construída considerando a frequência de uma classe o número de valores da variável pertencente à classe. Essa tabela é denominada "distribuição de frequência com intervalos de classe".

DISTRIBUIÇÃO DE FREQUÊNCIA DA ESTATURA DE 40 ALUNOS DE UMA UNIVERSIDADE	
Classes	**Frequência**
150 ├ 154	4
154 ├ 158[12]	9
158 ├ 162	11
162 ├ 166	8
166 ├ 170	5
170 ├ 174	3

Medidas de Frequência

Passos para construção das categorias em classes

1 Organize os dados brutos em um ROL.

ESTATURA DOS ALUNOS DE UMA UNIVERSIDADE									
166	160	161	150	162	160	165	167	164	160
162	168	161	163	156	173	160	155	164	168
155	152	163	160	155	155	169	151	170	164
154	161	156	172	153	157	156	158	158	161

ESTATURA DOS ALUNOS DE UMA UNIVERSIDADE									
150	154	155	157	160	161	162	164	166	169
151	155	156	158	160	161	162	164	167	170
152	155	156	158	160	161	163	164	168	172
153	155	156	160	160	161	163	165	168	173

Medidas de Frequência

Passos para construção das categorias em classes

2. Calcule a amplitude amostral AA
(no nosso exemplo, AA = 173 - 150 = 23)

3. Calcule o número de classes através da “Regra de Sturges”
o número de classes em função do tamanho da amostra $k \approx 1 + 3{,}3.\log(n)$

(no nosso exemplo: n = 40 dados, então, a princípio, a regra sugere a adoção de 6 classes) .

Medidas de Frequência

Passos para construção das categorias em classes

4. Decidido o número de classes, calcule a amplitude do intervalo de classe dividindo a amplitude total da amostra pelo número de classes $h \approx AA / k$

(no nosso exemplo $h \approx 23/6 = 3,8 \approx 4$)

Medidas de Frequência

Passos para construção das categorias em classes

5. Temos então o menor número da amostra, o número de classes e a amplitude do intervalo;

No nosso exemplo, o menor número da amostra = 150 + 4 = 154, logo a primeira classe será representada por 150—| 154

As classes seguintes respeitarão o mesmo procedimento.

Medidas de Frequência

Assim a tabela de frequência total ficaria dessa forma:

DISTRIBUIÇÃO DE FREQUÊNCIA DA ESTATURA DE 40 ALUNOS DE UMA UNIVERSIDADE						
i	Estaturas (cm)	f_i	x_i	fr_i	F_i	Fr_i
1	150 |– 154	4	152	0,100	4	0,100
2	154 |– 158	9	156	0,225	13	0,325
3	158 |– 162	11	160	0,275	24	0,600
4	162 |– 166	8	164	0,200	32	0,800
5	166 |– 170	5	168	0,125	37	0,925
6	170 |– 174	3	172	0,075	40	1,000
		$\sum = 40$		$\sum = 1{,}000$		

Medidas de Tendência Central

Medidas de Tendência Central

Média: Soma dos valores dividida pelo número de valores.

$$\bar{x} = \frac{\sum x}{n}$$

Mediana: Valor que divide o conjunto de dados ao meio.

Moda: Valor que ocorre com mais frequência.

Medidas de Dispersão

Medidas de Tendência Central

Medidas de Dispersão: As principais medidas de dispersão são:

- Amplitude total.
- Variância.
- Desvio padrão.
- Coeficiente de Variação

Medidas de Tendência Central

Amplitude total.

A amplitude total em dados não agrupados é a diferença entre o maior e o menor valor da série de dados, ou seja

$$AT = X_{máximo} - X_{minimo}$$

Medidas de Tendência Central

Variância.

A variância mede a dispersão dos dados em torno de sua média, levando em consideração a totalidade dos valores da variável em estudo, o que a torna um índice de variabilidade bastante estável.

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$$

Medidas de Tendência Central

Desvio Padrão.

É a medida de dispersão geralmente mais empregada, pois leva em consideração a totalidade dos valores da variável em estudo. O desvio padrão é uma medida de dispersão usada com a média. Mede a variabilidade dos valores à volta da média.

$$\sigma = \sqrt{\sigma^2}$$

Medidas de Tendência Central

Coeficiente de variação.

O desvio padrão é uma medida limitada se utilizada isoladamente. Por exemplo, um desvio padrão de 2 unidades pode ser considerado pequeno para uma série de valores cujo valor médio é 200; no entanto, se a média for igual a 20, o mesmo não pode ser dito.

Quando desejamos comparar duas ou mais unidades relativamente à sua dispersão ou variabilidade, o desvio padrão não é a medida mais indicada, visto que ele se encontra na mesma unidade dos dados.

Medidas de Tendência Central

Coeficiente de variação.

O coeficiente de variação de pearson é a razão entre o desvio padrão e a média referentes aos dados de uma mesma série.

$$CV = \frac{\sigma}{\bar{x}} \times 100\%$$

Intervalo Interquartílico e Outliers

Intervalo Interquartílico e Outliers

Intervalo Interquartílico (IQR)

Definição: O Intervalo Interquartílico (IQR) é uma medida de dispersão que descreve a amplitude do intervalo onde se concentra a parte central dos dados. Ele é calculado como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1).

$$IQR = Q3 - Q1$$

Q1 (Primeiro Quartil): O valor abaixo do qual 25% dos dados estão.
Q3 (Terceiro Quartil): O valor abaixo do qual 75% dos dados estão.

Intervalo Interquartílico e Outliers

Intervalo Interquartílico (IQR)

Exemplo:

Considere o conjunto de dados: [1, 3, 4, 6, 7, 8, 9, 11, 15, 18].

Intervalo Interquartílico e Outliers

Outliers

Definição: Outliers são valores atípicos que se encontram significativamente distantes do restante dos dados. Eles podem distorcer as estatísticas descritivas e, por isso, é importante identificá-los.

Identificação de Outliers: Outliers são identificados usando o IQR. Valores que estão fora do intervalo $[Q_1 – 1,5 \times IQR, Q_3 + 1,5 \times IQR]$

Normalização e Padronização

Normalização e Padronização

Normalização: Ajuste dos dados para que eles fiquem em uma escala comum, geralmente entre 0 e 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Normalização e Padronização

Padronização: Ajuste dos dados para que tenham média 0 e desvio padrão 1.

$$z = \frac{x - \mu}{\sigma}$$

z é o valor padronizado.
x é o valor original do dado.
μ é a média da amostra.
σ é o desvio padrão da amostra.

Covariância e Correlação

Covariância e Correlação

Covariância

Definição: A covariância é uma medida que indica a direção do relacionamento linear entre duas variáveis. Ela nos diz se as variáveis aumentam ou diminuem juntas.

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Cov(X,Y) é a covariância entre as variáveis X e Y.

x_i e y_i são os valores das variáveis x e y.

x barra é a média dos valores de x.

Y barra é a média dos valores de y.

Covariância e Correlação

Interpretação:

Cov(X, Y) > 0: As variáveis tendem a aumentar juntas.

Cov(X, Y) < 0: Uma variável tende a aumentar enquanto a outra tende a diminuir.

Cov(X, Y) = 0: Não há tendência linear aparente entre as variáveis.

Covariância e Correlação

Correlação

Definição: A correlação é uma medida que indica a força e a direção do relacionamento linear entre duas variáveis. A correlação é a covariância padronizada.

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

r é o coeficiente de correlação de Pearson.

Cov(X,Y) é a covariância entre as variáveis x e y.

σx e σy são os desvios padrão de x e y.

Covariância e Correlação

Interpretação:
$r = 1$: Correlação perfeita positiva.

$r = -1$: Correlação perfeita negativa.

$r = 0$: Nenhuma correlação linear.

(85) 98524-9935 youthspace

contato@youthidiomas.com.br

https://www.youthspace.com.br/

2023