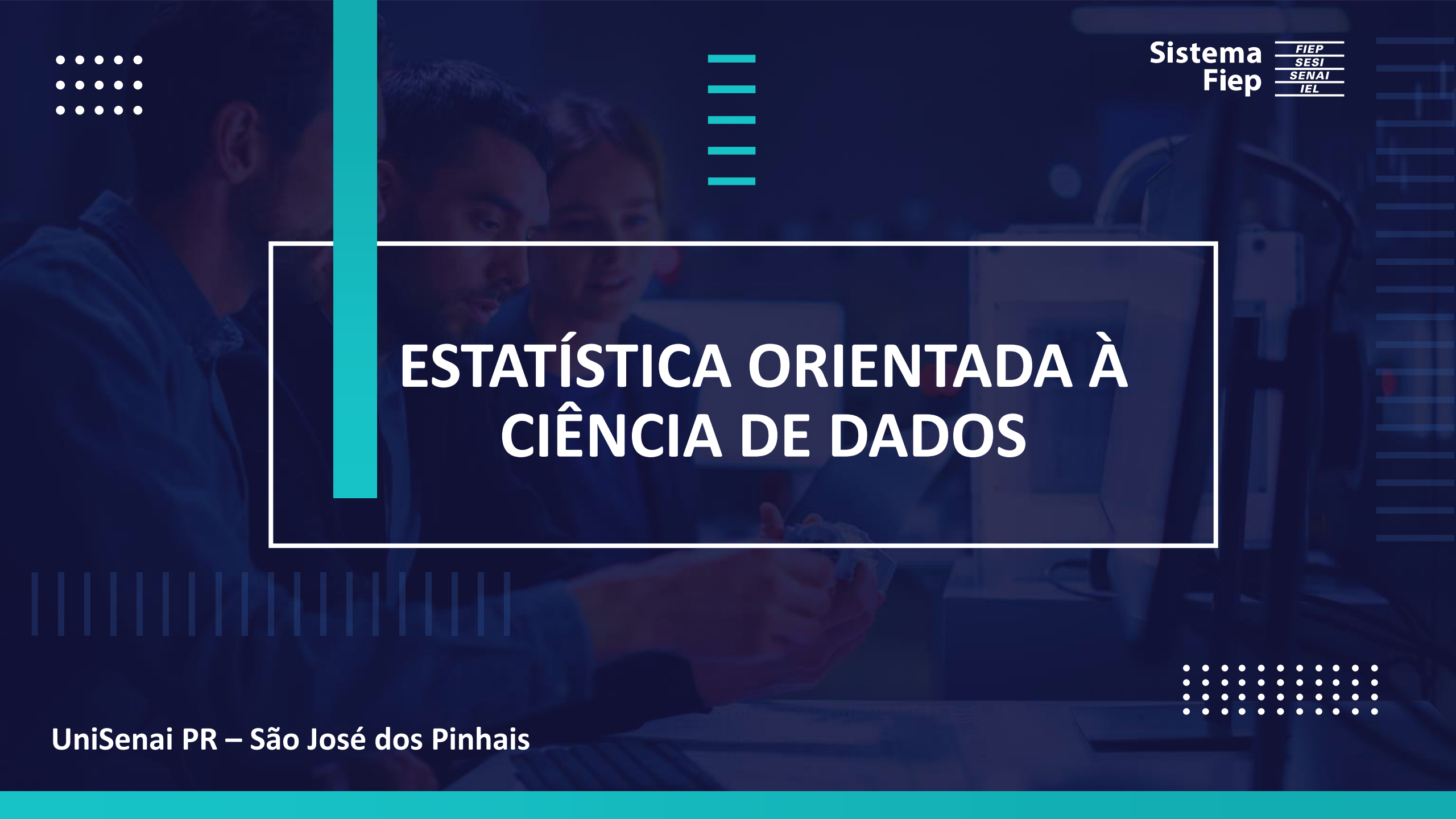


Sistema Fiep

FIEP SESI SENAI IEL

ESTATÍSTICA ORIENTADA À CIÊNCIA DE DADOS

UniSenai PR – São José dos Pinhais

Sistema Fiep
FIEP SESI SENAI IEL

Introdução à Estatística: Conceitos Básicos

Estatística

Dedica-se a **coletar, organizar, analisar e interpretar dados**;

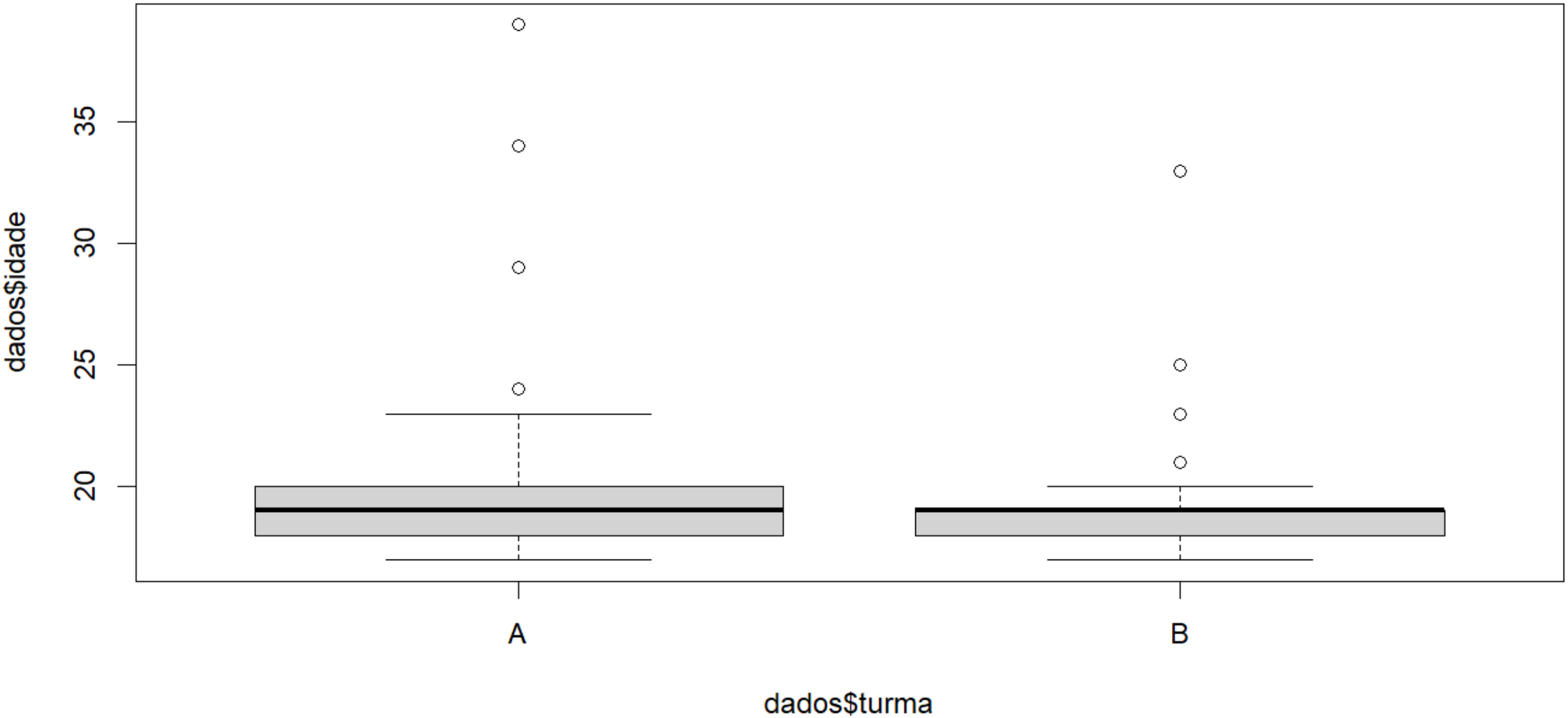
O **principal objetivo é extrair informações relevantes dos dados, permitindo tomar decisões**;

Estatística descritiva

Ao aplicar técnicas estatísticas descritivas, **pode-se obter uma visão geral dos padrões e características dos dados**, **facilitando a identificação de tendências, variações e relações entre variáveis**.

IDADE (TURMA A)	
Mínimo	17
Média	20,33
Desvio padrão	4,40
Mediana	19
Máximo	39
Moda	-
n	42

IDADE (TURMA B)	
Mínimo	17
Média	19,33
Desvio padrão	2,98
Mediana	19
Máximo	33
Moda	-
n	33


dados$idade
20
25
30
35
A
B
dados$turma

Estatística inferencial

Área da estatística que se **preocupa com a análise e interpretação de dados amostrais a fim de fazer inferências ou generalizações sobre a população a partir da qual a amostra foi obtida**.

População

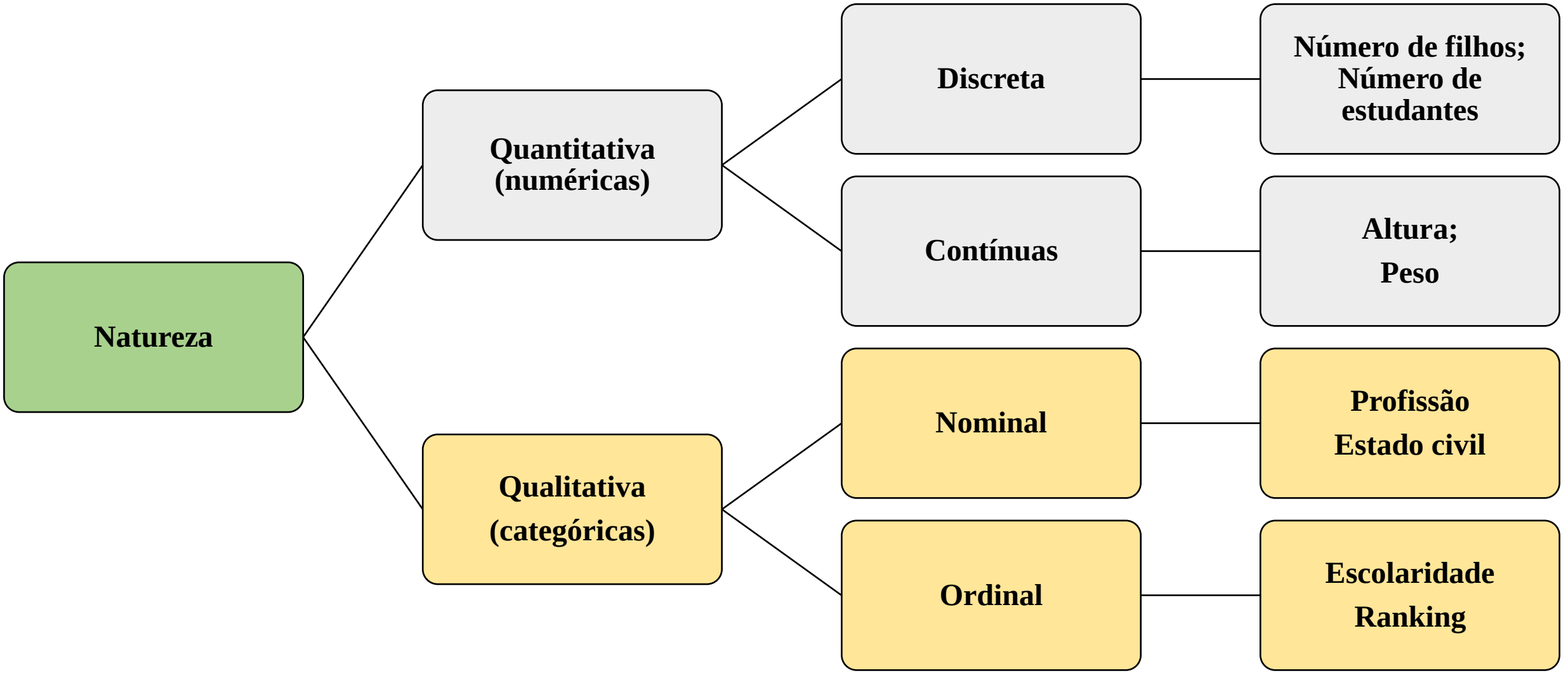
A população **refere-se ao conjunto completo de elementos ou indivíduos que compartilham características específicas de interesse em um estudo**.

Amostra

A amostra **é um subconjunto representativo selecionado da população, usado para fazer inferências e generalizações sobre a população como um todo**;

A escolha de uma amostra adequada **é crucial para garantir que os resultados obtidos sejam confiáveis e aplicáveis à população** de interesse.

Tipos de variáveis/dados


Natureza
Quantitativa (numéricas)
Qualitativa (categóricas)
Discreta
Contínuas
Nominal
Ordinal
Número de filhos; Número de estudantes
Altura; Peso
Profissão Estado civil
Escolaridade Ranking

Variáveis Qualitativas

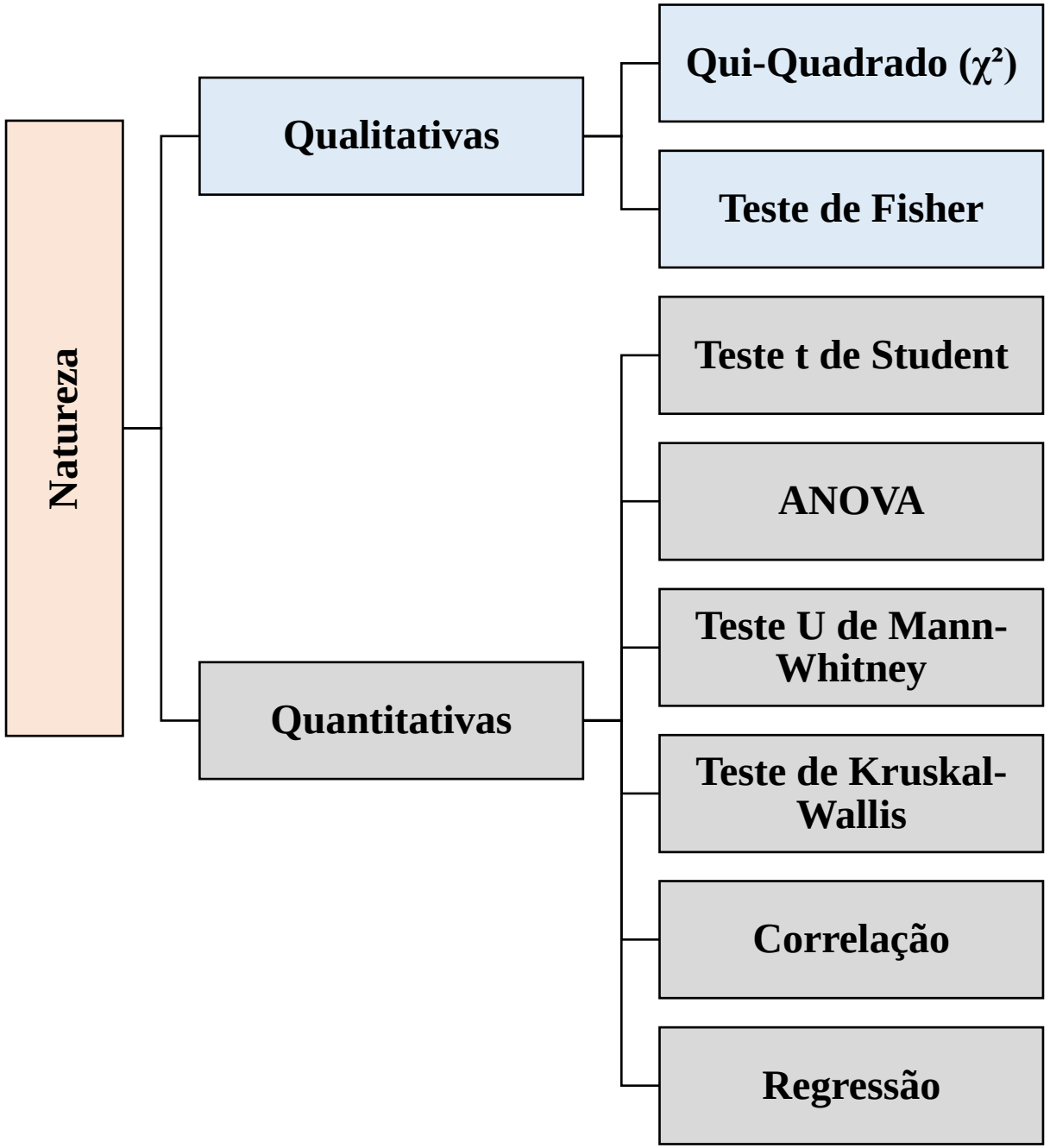
Representam **características, atributos ou categorias, e não podem ser medidas em uma escala numérica contínua**;

1. **Variáveis Nominais:** categorias sem ordem específica.
2. **Variáveis Ordinais:** categorias com uma ordem específica, mas não possuem igual intervalo entre os valores.

Variáveis Quantitativas

Representam **quantidades ou valores numéricos que podem ser submetidos a cálculos matemáticos**;

1. **Variáveis Discretas:** valores individuais são contados ou enumerados.
2. **Variáveis Ordinais:** medem quantidades com intervalos iguais entre os valores, mas não possuem um zero absoluto.


Natureza
Qualitativas
Qui-Quadrado (χ^2)
Teste de Fisher
Quantitativas
Teste t de Student
ANOVA
Teste U de Mann-Whitney
Teste de Kruskal-Wallis
Correlação
Regressão

Medidas de Tendência Central

As medidas de tendência central são **utilizadas para representar o valor central de um conjunto de dados**;

Fornecem uma ideia sobre onde a maioria dos valores está localizada e são essenciais para entender a distribuição dos dados.

Medida	Definição	Cálculo
Média	Soma de todos os valores em um conjunto de dados dividida pelo número total de observações	Amostra: $\bar{x} = \frac{(\sum_{i=1}^{n} x_i)}{n}$ População: $\mu = \frac{(\sum_{i=1}^{n} x_i)}{\mathrm{N}}$

$\bar{x}$: Média para uma amostra;

μ: Média para a população;

$\sum_{i=1}^{n} x$: Soma de todos os valores no conjunto de dados;

n: Número de observações na amostra;

N: Número total de elementos na população;

Considere o conjunto de dados: 10, 15, 18, 12, 10, 22, 18, 14, 20, 25

$$\bar{x} = \frac{(10 + 15 + 18 + 12 + 10 + 22 + 18 + 14 + 20 + 25)}{10} = 16{,}4$$

Medida	Definição	Cálculo
Mediana	Valor central de um conjunto de dados ordenado	Ordenar os dados em ordem crescente ou decrescente e, caso haja um **número ímpar de observações, a mediana é o valor do meio**; caso haja um **número par de observações, a mediana é a média dos dois valores centrais**.

Ordenar os dados: 10, 10, 12, 14, **15, 18**, 18, 20, 22, 25

$$mediana = \frac{(15 + 18)}{2} = 16{,}5$$

Medida	Definição	Cálculo
Moda	Valor que ocorre com maior frequência em um conjunto de dados. Pode haver uma única moda (moda unimodal), mais de uma moda (moda bimodal) ou nenhum valor repetido (sem moda).	É obtida identificando o valor com a maior frequência no conjunto de dados.

Considere o seguinte conjunto: 10, 10, 12, 14, 15, 18, 18, 20, 22, 25

$$moda = 10\ e\ 18$$

Medida	Fórmula para Amostras	Fórmula para População
Média	$\bar{x} = \frac{(\sum_{i=1}^{n} x_i)}{n}$	$\mu = \frac{(\sum_{i=1}^{n} x_i)}{\mathrm{N}}$
Mediana	Ordenar os dados e encontrar o valor do meio	Ordenar os dados e encontrar o valor do meio
Moda	Identificar o valor mais frequente	Identificar o valor mais frequente

Medida	Resumo
Média	É sensível a valores extremos, o que pode afetar significativamente o resultado final.
Mediana	Menos afetada por valores extremos.
Moda	Útil para dados qualitativos e pode não ser adequada para representar a tendência central em distribuições contínuas

Medidas de Dispersão

Estatísticas ajudam a **entender o quão “espalhados” ou concentrados os valores estão em torno da medida central (média, mediana, moda);**

São essenciais para **compreender a variabilidade dos dados e auxiliam na interpretação e análise mais completa das informações**.

Medida	Definição	Cálculo
Amplitude	Diferença entre o maior e o menor valor em um conjunto de dados.	$valor_{máximo} - valor_{mínimo}$

Considere o conjunto de dados: 10, 15, 18, 12, 10, 22, 18, 14, 20, 25

$$amplitude = 25 - 10 = 15$$

Medida	Definição	Cálculo
Variância	Mede a dispersão dos dados em relação à média. É a média dos quadrados dos desvios dos valores em relação à média.	População: $\sigma^2 = \frac{(\sum_{i=1}^{n} x_i - \mu)^2}{\mathrm{N}}$ Amostra: $s^2 = \frac{(\sum_{i=1}^{n} x_i - \bar{x})^2}{\mathrm{n-1}}$

Explicando a fórmula da variância (População)

$x_i =$ cada valor do conjunto de dados;

$\mu =$ média da população;

N = total de elementos da população;

$\sum_{i=1}^{n} x$ = somatório dos valores individuais no conjunto de dados;

$(x_i - \mu)^2$= subtração do valor individual em relação à média da população e depois elevado ao quadrado.

Explicando a fórmula da variância (Amostra)

x_i = cada valor do conjunto de dados;

$\bar{x}$ = média da amostra;

n = total de elementos da amostra;

$\sum_{i=1}^{n} x$ = somatório dos valores individuais no conjunto de dados;

$(x_i - \bar{x})^2$= subtração do valor individual em relação à média da população e depois elevado ao quadrado;

(n-1): graus de liberdade corrigido.

Conjunto de dados: 10, 15, 18, 12, 10, 22, 18, 14, 20, 25

Primeiro passo: calcular média amostral $\bar{x}$

$$\bar{x} = \frac{(10 + 15 + 18 + 12 + 10 + 22 + 18 + 14 + 20 + 25)}{10} = 16{,}4$$

Segundo passo: Calcular a soma dos quadrados dos desvios em relação à média $(x_i - \bar{x})^2$

(10 – 16,4) = -6,4
(15 – 16,4) = -1,4
(18 – 16,4) = 1,6
(12 – 16,4) = -4,4
(10 – 16,4) = -6,4
(22 – 16,4) = 5,6
(18 – 16,4) = 1,6
(14 – 16,4) = -2,4
(20 – 16,4) = 3,6
(25 – 16,4) = 8,6

$(-6,4)^2 + (-1,4)^2 + (1,6)^2 + (-4,4)^2 + (-6,4)^2 + (5,6)^2 + (1,6)^2 + (-2,4)^2 + (3,6)^2 + (8,6)^2 = 172,4$

Terceiro passo: Calcular a variância da amostra (s^2)

$$s^2 = \frac{(\sum_{i=1}^{n} x_i - \bar{x})^2}{\mathrm{n} - 1}$$

$$s^2 = \frac{172{,}4}{10 - 1}$$

$$s^2 = \frac{172{,}4}{9} = 19{,}16$$

Medida	Definição	Cálculo
Desvio padrão	Indica a dispersão dos dados em relação à média. Quanto maior o desvio padrão, maior a dispersão dos dados	População: $\sigma = \sqrt{\frac{(\sum_{i=1}^{n} x_i - \mu)^2}{\mathrm{N}}}$ Amostra: $s = \sqrt{\frac{(\sum_{i=1}^{n} x_i - \bar{x})^2}{\mathrm{n-1}}}$

Tirar a raiz quadrada da variância

$$s = \sqrt{19{,}16} = 4{,}38$$

O desvio padrão é frequentemente preferido em vez da variância para descrever a dispersão dos dados **porque ele tem a mesma unidade de medida dos dados originais, enquanto a variância tem unidades ao quadrado**.

Na variância, os **desvios dos valores em relação à média são elevados ao quadrado,** como forma de evitar que desvios positivos e negativos se anulem quando somados.

Medida	Definição	Cálculo
Coeficiente de variação	Medida relativa de dispersão utilizado para comparar a variabilidade entre diferentes conjuntos de dados, independentemente de suas escalas.	$cv = \left(\frac{Desvio\ padrão}{média}\right) * 100$

Conjunto de dados: 10, 15, 18, 12, 10, 22, 18, 14, 20, 25

$$cv = \left(\frac{4{,}38}{16{,}4}\right) * 100 = 26{,}71\%$$

Isso nos indica que a variabilidade relativa dos valores em relação à média é de cerca de 26,71%

Medida	Resumo	Utilidade
Amplitude	Medida simples da extensão total dos valores no conjunto de dados.	Não fornece informações sobre a dispersão dos valores em relação à média.
Variância	Medida da dispersão dos valores em relação à média.	Indica a variabilidade dos dados; Maior valor indica maior dispersão; Menor valor indica menor dispersão.
Desvio padrão	Raiz quadrada da variância.	Mede a dispersão dos dados em unidades originais; Facilita a interpretação comparada com os dados originais.
Coeficiente de variação	Medida relativa de dispersão.	Permite comparar a variabilidade entre diferentes conjuntos de dados, independentemente de suas escalas ou médias.

Exercícios