

Introdução à Estatística Descritiva

A **Estatística Descritiva** é a primeira etapa para entender o que os números podem nos dizer sobre a realidade.

Ela nos permite **resumir, organizar e apresentar** dados complexos de forma clara e concisa.

Uma "Foto" dos Dados

A estatística descritiva é como tirar uma fotografia de um conjunto de dados para entender suas características principais, sem tirar conclusões sobre a população inteira.

Resumo em Poucas Medidas

Transforma grandes volumes de dados em poucas medidas que capturam a essência das informações, facilitando a interpretação e análise.

Base para Análises Avançadas

É o primeiro passo antes de aplicar técnicas mais avançadas de inferência estatística ou machine learning.

66

"O que os números podem nos dizer sobre a realidade?"



DATA SNAPSHOT

Medidas de Tendência Central

As **medidas de tendência central** são valores que representam o centro ou o valor típico de um conjunto de dados.



Média

É a soma de todos os valores dividida pelo número de observações.

$$\text{Média} = (x_1 + x_2 + \dots + x_n) / n = \Sigma x / n$$

Exemplo: Para os valores [5, 10, 15, 20, 25]

$$\text{Média} = (5 + 10 + 15 + 20 + 25) / 5 = 75 / 5 = 15$$

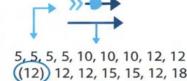


MEAN



$$\text{MEAN} = \frac{\Sigma x}{n}$$

MEDIAN



MODE



Mediana

É o valor que está exatamente no meio de um conjunto de dados ordenado. Menos afetada por valores extremos (outliers).

Exemplo: Para os valores [5, 10, 15, 20, 25]

Mediana = 15 (valor central)

Exemplo 2: Para os valores [5, 10, 15, 20]

Mediana = $(10 + 15) / 2 = 12.5$ (média dos dois valores centrais)



Moda

É o valor que aparece com maior frequência em um conjunto de dados.

Exemplo: Para os valores [5, 10, 10, 15, 20, 10, 25]

Moda = 10 (aparece 3 vezes)

Atividade A: Prática Inicial

1 Cálculo Manual e em Planilha (10 min)

Trabalho em grupos

1 Dados para Análise

Cada grupo receberá uma lista de 10 salários fictícios:

R\$ 1.500, R\$ 2.200, R\$ 1.800, R\$ 3.500, R\$ 2.200,
R\$ 1.800, R\$ 4.500, R\$ 2.800, R\$ 2.200, R\$ 3.000

2 Cálculo Manual

Calcule manualmente:

Média dos salários

Mediana (ordene os valores primeiro)

Moda (valor mais frequente)

3 Cálculo em Planilha

Utilize o Google Sheets ou Excel para verificar seus resultados:

Média: função **=MÉDIA(A1:A10)**

Mediana: função **=MED(A1:A10)**

Moda: função **=MODO(A1:A10)**



Medidas de Dispersão

As **medidas de dispersão** mostram o quanto "espalhados" os dados estão em relação à média ou entre si.

↔ Desvio Padrão e Variância

Indicam o quanto os valores se afastam da média. Um desvio padrão alto significa dados muito distantes da média.

$$\text{Variância} = \sum(x - \mu)^2 / n$$

$$\text{Desvio Padrão} = \sqrt{\text{Variância}}$$



STANDARD
DEVIATION

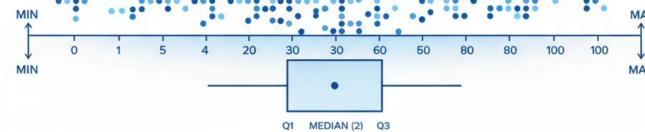


VARIANCE

■ Mínimo e Máximo

Os valores extremos do conjunto de dados, que definem a amplitude total.

$$\text{Amplitude} = \text{Máximo} - \text{Mínimo}$$



☰ Quartis

Dividem o conjunto de dados em quatro partes iguais:

Q1 (25%): primeiro quartil

Q2 (50%): mediana

Q3 (75%): terceiro quartil

O **Intervalo Interquartil (IIQ)** = Q3 - Q1, ajuda a identificar outliers.



QUARTILES
RANGE

Medidas de Frequência

As **medidas de frequência** nos mostram quantas vezes cada valor ou categoria aparece em um conjunto de dados.

Frequência Absoluta

É o número de vezes que um valor aparece no conjunto de dados.

Exemplo: Em uma turma de 30 alunos:

Nota 7: 12 alunos (frequência absoluta = 12)

Nota 8: 10 alunos (frequência absoluta = 10)



% Frequência Relativa

É o percentual de vezes que um valor aparece no conjunto de dados.

Frequência Relativa = (Frequência Absoluta / Total) × 100%

Exemplo: Para a nota 7: $12/30 = 40\%$

Para a nota 8: $10/30 = 33,3\%$

Em Python, podemos calcular facilmente as frequências usando o método `value_counts()` do Pandas.

ABSOLUTE FREQUENCY



A: 15	⊕
B: 10	⊕
C: 5	⊕
Total: 30	

RELATIVE FREQUENCY



A: 50%	÷
B: 33.3%	÷
C: 16.7%	÷
Total: 100%	

Atividade B: Automação em Python e Planilhas



Automação com Python (50 min)

Trabalho em grupos com dataset real

1 Carregar os Dados

Abra o Google Colab e carregue o dataset de salários do CAGED:

```
import pandas as pd

# Carregar o dataset
df = pd.read_csv('salarios_caged.csv')
```



2 Calcular as Medidas

Use os comandos do Pandas para calcular as estatísticas:

```
# Medidas de tendência central
media = df['salario'].mean()
mediana = df['salario'].median()
moda = df['salario'].mode()[0]
```

```
# Medidas de dispersão
desvio = df['salario'].std()
minimo = df['salario'].min()
maximo = df['salario'].max()
```

```
# Resumo completo
df['salario'].describe()
```

3 Calcular Frequências

Para variáveis categóricas (como gênero ou região):

```
# Frequência absoluta
freq_abs = df['genero'].value_counts()
```

```
# Frequência relativa
freq_rel = df['genero'].value_counts(normalize= True) * 100
```

Cálculos Estatísticos com Pandas

O **Pandas** oferece funções simples para calcular todas as medidas estatísticas:

↳ Medidas de Tendência Central

```
media = df['salario'].mean()  
mediana = df['salario'].median()  
moda = df['salario'].mode()[0]
```

Média: 2550.00 | Mediana: 2200.00 | Moda: 2200.00

↔ Medidas de Dispersão

```
desvio = df['salario'].std()  
variancia = df['salario'].var()  
minimo = df['salario'].min()  
maximo = df['salario'].max()
```

▣ Resumo Completo com describe()

```
df['salario'].describe()
```

count: 10.0 | mean: 2550.0 | std: 950.7
min: 1500.0 | 25%: 1800.0 | 50%: 2200.0
75%: 3250.0 | max: 4500.0

Statistical Measures with Pandas

Importing Library & Data:

```
df = pd.read_csv('data.csv')
```

Central Tendency

```
meanval = df['Value'].mean()  
print(f'Mean: {meanval}')
```



Motion & Spread

```
medianval = df['Value'].median()  
print(f'Median: {medianval}')
```



Dispersion & Quartiles

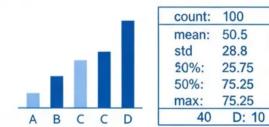
```
std-dev = df['Value'].std()  
std-dev = std()  
variance = df['Value'].var()
```



Frequency Distribution

```
df['Value'].describe()
```

```
df['Category'].value_counts()
```



Visualização de Dados

A **visualização de dados** é essencial para compreender padrões, tendências e relações que podem não ser óbvias nos dados brutos.

☰ Histograma

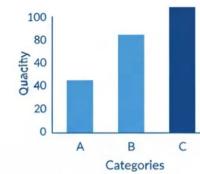
Mostra a **distribuição de frequência** de uma variável numérica, dividindo os dados em intervalos (bins) e contando quantos valores caem em cada intervalo.

Ideal para visualizar a distribuição de idades, salários, notas de prova, etc.

Histogram: Frequency Distribution



Bar Chart: Category Comparison

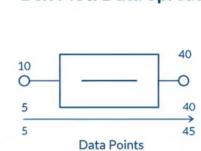


☰ Boxplot

Representa graficamente os **quartis** e identifica **outliers**. Mostra o mínimo, primeiro quartil (Q1), mediana, terceiro quartil (Q3) e máximo.

Perfeito para comparar distribuições entre diferentes grupos e identificar valores atípicos.

Box Plot: Data Spread



☰ Gráfico de Barras

Utilizado para representar **frequências de variáveis categóricas**, com barras proporcionais à quantidade que representam.

Útil para visualizar contagens por categoria, como gênero, região, tipo de produto, etc.

Código para Visualizações

Utilizando **Matplotlib** e **Seaborn** para criar visualizações eficientes:

Histograma

```
import matplotlib.pyplot as plt
import seaborn as sns

# Configuração visual
plt.figure(figsize=(10, 6))
sns.set_style('whitegrid')

# Criando o histograma
plt.hist(df['salario'], bins=10, color="#0066cc", alpha=0.7)
plt.title('Distribuição de Salários')
plt.xlabel('Salário (R$)')
plt.ylabel('Frequência')
plt.grid(True, alpha=0.3)
plt.show()
```

Boxplot

```
# Criando o boxplot
plt.figure(figsize=(10, 6))
sns.boxplot(x=df['salario'], color="#0066cc")
plt.title('Boxplot de Salários')
plt.xlabel('Salário (R$)')
plt.grid(True, alpha=0.3)
plt.show()
```

Gráfico de Barras

```
# Contagem de categorias
contagem = df['categoria'].value_counts()

# Criando o gráfico de barras
plt.figure(figsize=(10, 6))
contagem.plot(kind='bar', color="#0066cc")
plt.title('Frequência por Categoria')
plt.xlabel('Categoria')
plt.ylabel('Frequência')
plt.grid(True, alpha=0.3)
plt.show()
```

Estudo de Caso: Diferenças Salariais

Análise de Disparidades Salariais no Brasil

Aplicando estatística descritiva em dados reais

Contexto do Estudo

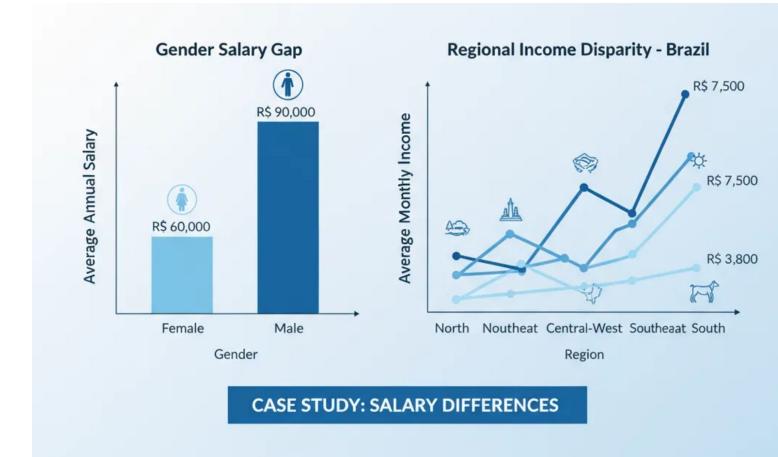
Utilizando dados do CAGED (Cadastro Geral de Empregados e Desempregados), vamos analisar as diferenças salariais entre gêneros e regiões do Brasil.

Este estudo de caso demonstra como a estatística descritiva pode revelar padrões importantes em dados socioeconômicos.

Tarefas de Análise

Para cada grupo (gênero e região), calcule e compare:

- Média e mediana salarial
- Desvio padrão e variância
- Quartis (Q1, Q2, Q3)
- Identificação de outliers



Perguntas para Discussão:

- Existe uma diferença significativa entre a média salarial de homens e mulheres?
- Qual região apresenta maior variabilidade salarial?
- A mediana é uma medida mais adequada que a média para comparar os salários? Por quê?
- Como você explicaria as diferenças encontradas para um público não técnico?

Redação de Análise Estatística



Comunicando Resultados Estatísticos

Transformando números em insights comprehensíveis

1

Estrutura da Análise

Introdução: Contexto e objetivos da análise

Metodologia: Dados utilizados e métodos aplicados

Resultados: Apresentação das estatísticas e visualizações

Interpretação: O que os números significam

Conclusão: Principais insights e recomendações

2

Linguagem Acessível

Traduza conceitos estatísticos para uma linguagem que qualquer pessoa possa entender:

Evite jargões técnicos desnecessários

Explique o significado prático das medidas

Use analogias e exemplos do cotidiano

Destaque as implicações dos resultados



Dica Importante

Sempre inclua visualizações junto com o texto. Um gráfico bem elaborado pode comunicar informações complexas de forma muito mais eficiente que parágrafos de texto.

Quiz de Fixação

Teste seus Conhecimentos

Responda às questões para fixar os conceitos aprendidos

Questão 1

Qual medida de tendência central é menos afetada por valores extremos (outliers)?

- A Média
- B Mediana
- C Moda

Resposta correta: B - Mediana

A mediana é menos sensível a valores extremos porque considera apenas a posição central dos dados ordenados.

Questão 2

Qual visualização é mais adequada para mostrar a distribuição de uma variável numérica?

- A Gráfico de barras
- C Histograma

Resposta correta: C - Histograma

O histograma é ideal para mostrar a distribuição de frequência de variáveis numéricas contínuas.

