

Estatística

1. Média

A média é o valor obtido pela soma de todos os números em um conjunto de dados, dividida pelo número total de dados. É frequentemente chamada de "média aritmética". A fórmula para calcular a média é:

$$\text{Média} = \frac{\text{Soma de todos os valores}}{\text{Número total de valores}}$$

Visualmente, você pode pensar na média como o ponto de equilíbrio de um conjunto de números, onde a soma de todos os valores é distribuída igualmente.

Suponha que tenhamos os seguintes números: 3, 5, 7, 9 e 11.

- Soma dos valores: $3 + 5 + 7 + 9 + 11 = 35$
- Número total de valores: 5
- Média: $\frac{35}{5} = 7$

2. Mediana

A mediana é o valor que divide o conjunto de dados em duas partes iguais. Para encontrar a mediana, primeiro você precisa ordenar os dados em ordem crescente ou decrescente e, em seguida, encontrar o valor do meio. Se houver um número par de dados, a mediana é a média dos dois valores do meio. Se houver um número ímpar de dados, a mediana é o valor do meio exato. Abaixo está a fórmula básica:

- Se o número de observações (n) for ímpar: A mediana é o valor na posição $(n + 1)/2$.
- Se o número de observações (n) for par: A mediana é a média dos valores nas posições $(n/2)$ e $(n/2 + 1)$.

Visualmente, a mediana é o ponto no meio dos dados quando estão organizados em ordem crescente.

Suponha que temos os seguintes números: 3, 7, 2, 8, 5, 4, 9.

- Primeiro, vamos ordená-los: 2, 3, 4, 5, 7, 8, 9.
- Como o número de dados é ímpar (7), a mediana é o valor na posição $(7 + 1)/2 = 4$.
- Portanto, a mediana é 5.

3. Moda

A moda é o valor que aparece com maior frequência em um conjunto de dados. Em outras palavras, é o valor que ocorre mais vezes. Por exemplo, considere o conjunto de dados: 2, 4, 4, 6, 6, 6, 7, 8, 8, 8. Aqui, o número 6 é o que mais se repete, então a moda deste conjunto de dados é 6.

A moda é útil quando queremos saber qual é o valor mais comum em um conjunto de dados, especialmente em distribuições de dados que podem não ser simétricas.

4. Desvio Padrão

O desvio padrão é uma medida de dispersão que indica o quanto os valores em um conjunto de dados estão distantes da média desse conjunto. Basicamente, ele nos diz o quão "espalhados" os valores estão em relação à média.

Por exemplo, considere dois conjuntos de dados:

Conjunto A: 2, 4, 4, 4, 5

Conjunto B: 2, 2, 2, 5, 5

Ambos os conjuntos têm a mesma média, que é 4. Porém, o conjunto B tem um desvio padrão menor do que o conjunto A. Isso ocorre porque os valores no conjunto B estão mais próximos da média do que os valores no conjunto A.

O desvio padrão é útil para entender a variabilidade ou a dispersão dos dados. Quanto maior o desvio padrão, mais os dados estão espalhados em relação à média, e quanto menor o desvio padrão, mais os dados estão agrupados em torno da média.

5. Utilização na Análise de Dados

É possível utilizar esses recursos para ter alguns insights na área de análise de dados. Abaixo listei alguns deles:

Resumo Estatístico:

- Ao realizar uma análise exploratória de dados, é comum calcular essas medidas para resumir e descrever o conjunto de dados. Elas fornecem uma visão geral das características centrais, da variabilidade e dos padrões nos dados.

Identificação de Tendências:

- A média é útil para identificar tendências centrais nos dados. Por exemplo, ao analisar dados de vendas ao longo do tempo, a média pode indicar se há uma tendência de crescimento, queda ou estabilidade nas vendas.

- A mediana é útil quando os dados são influenciados por valores extremos, pois ela não é tão sensível a esses outliers. Isso permite identificar a tendência central de forma mais robusta.
- A moda pode revelar padrões de repetição ou valores mais comuns nos dados, como o produto mais vendido em uma loja ou a cor favorita entre os clientes.

Comparação de Grupos:

- Ao comparar diferentes grupos ou categorias nos dados, é possível calcular essas medidas para cada grupo e comparar suas características centrais, variabilidade e distribuições.
- Por exemplo, ao analisar dados demográficos, pode-se calcular a média de idade em diferentes grupos de gênero ou calcular a moda da profissão mais comum em cada grupo.

Detecção de Outliers:

- O desvio padrão pode ser usado para identificar valores atípicos nos dados. Valores que estão a mais de um certo número de desvios padrão da média podem ser considerados outliers e merecem atenção especial durante a análise.
- A mediana é menos sensível a outliers do que a média, portanto, ao encontrar diferenças significativas entre a mediana e a média, isso pode indicar a presença de valores extremos.

Validação de Modelos:

- Ao construir modelos estatísticos ou de aprendizado de máquina, é comum usar essas medidas para avaliar a qualidade do modelo. Por exemplo, comparando a média dos valores observados com a média dos valores previstos pelo modelo.

Essas são apenas algumas maneiras de aplicar média, mediana, moda e desvio padrão na análise de dados. Elas são ferramentas fundamentais para entender a distribuição, a tendência central e a variabilidade dos dados, além de identificar padrões e outliers que podem influenciar as conclusões da análise.