Estatística

IN

Aula 2

- Medidas de dispersão
- Medidas de posição.

IN

Conceitos Básicos

- **Medidas de posição**: Indicam um valor que melhor representa todo o conjunto de dados, ou seja, dão a tendência da concentração dos valores observados. **(média, mediana, moda)**

- **Medidas de dispersão:** Parâmetros estatísticos usados para determinar o grau de variabilidade dos dados de um conjunto de valores **(amplitude, variância, desvio padrão, coeficiente de variação)**

IN

Medidas de Posição e de Dispersão

Estes indicadores estatísticos são de extrema importância para a análise de dados, tanto os qualitativos como quantitativos.

É de extrema importância para o analista saber quando usar estes indicadores estatísticos a partir dos dados coletados.

Um erro de interpretação no uso de um indicador deste colocará toda a análise estatística em descrédito.

IN

A importância do uso correto de um indicador estatístico

Uma pesquisa foi realizada em 20 apartamentos de um prédio. Em um item da pesquisa foi perguntado ao morador quantos aparelhos smartphones haviam na sua residência.

A tabela abaixo mostra os dados da pesquisa (o número faltante na tabela é o 302):

101	102	103	104	201	202	203	204	301		303	304	401	402	403	404	501	502	503	504
2	3	3	2	4	4	3	3	2	3	3	3	2	105	3	2	4	3	2	4

A partir dos dados coletados, o analista toma a decisão de calcular a média aritmética dos dados. Este então apresenta o seguinte resultado:

Na média, cada apartamento do prédio tem **8** aparelhos smartphones.

Aí está o problema na metodologia adotada, pois não foi observado que o morador do ap. 402 apresenta um dado muito discrepante dos outros moradores.

Provavelmente, o morador do ap. 402 é um técnico que conserta smartphones no seu apartamento ou ele pode ser um colecionador de smartphones.

Então, apresentar o valor médio como indicador estatístico pode não ser o mais indicado como neste caso.

IN

Medidas de Posição: Valor Médio

O valor médio de uma amostra de n elementos corresponde à **média aritmética** ($\overline{x}$) desses n elementos.

$$\text{Média aritmética} = \frac{N1 + N2 + \ldots + Nn}{N}$$

Ex: Valor médio de {22, 43, 35, 30, 20} = (22+43+35+30+20)/5 = 150/5 = 30

Obs: o valor médio da população é representada pela letra grega μ

IN

Medidas de Posição: Mediana

A mediana é o valor central de um conjunto de números colocados por ordem de grandeza. **Trata-se do número que se encontra exatamente no centro, de modo que 50% dos números são superiores e 50% são inferiores a essa mediana**.

Vale ressaltar, no entanto, que, para identificar a mediana, devemos primeiramente **ordenar os elementos do conjunto**, o que chamamos de **rol** (geralmente a ordenação é do menor para o maior).

Por exemplo, o rol de {20 90 143 15 3 64 23} seria {3 15 20 23 64 90 143}

Para a identificação da mediana, como o número de amostras (n) pode ser positivo ou negativo, devemos pensar da seguinte forma:

P/n ímpar, a mediana é o elemento na posição n/2 (arredondada p/ cima).

{3 15 20 23 64 90 143}

7/2 = 3.5, portanto a mediana é o elemento na 4ª posição; perceba que o 23 dividiu o rol pela metade (3 elementos de um lado e 3 elementos do outro)

P/n par, a mediana é a média aritmética entre os elementos na posição n/2 e seu sucessor.

{3 7 15 20 23 64 90 143}

Mediana = (20 + 23) / 2 = 21.5; perceba que o 20 e o 23 dividiram o rol pela metade (3 elementos de um lado e 3 elementos do outro)

A mediana é fundamental quando temos um conjunto de dados com muitos "outliers", ou seja, dados extremamente discrepantes no conjunto.

Medidas de Posição: Moda

A moda de uma série de valores são os **valores de maior frequência absoluta**, ou seja, os valores que aparecem o maior número de vezes no conjunto.

Exemplo: A moda da série de dados: {1 22 33 41 33 23} é 33.

Um conjunto pode ser classificado de quatro diferentes formas dependendo da moda:

1. **Amodal:** Todos os elementos do conjunto têm a mesma frequência absoluta. Por exemplo: moda de {2 3 3 2 4 4} não existe

2. **Unimodal:** Há um único elemento do conjunto com a maior frequência absoluta. Por exemplo: moda de {1 2 3 3 4 5 6} é 3

3. **Bimodal:** Há exatamente dois elementos no conjunto que compartilham a maior frequência absoluta. Por exemplo: moda de (2 5 5 2 3) é (2 5)

4. **Multimodal:** Há mais de dois elementos no conjunto que compartilham a maior frequência absoluta. Por exemplo: moda de (2 5 5 2 3 9 3) é (2 5 3)

A moda é utilizada muitas vezes para **preencher dados indefinidos em variáveis categóricas**. Por exemplo, se numa tabela estiver faltando dados na coluna sexo onde lê-se M=0 e F=1, então normalmente opta-se pela moda para preencher os dados faltantes.

Medidas de Dispersão: Amplitude

É a **diferença entre o maior e o menor valor** em um conjunto de dados.

A amplitude fornece uma **medida bruta da variabilidade dos dados**.

Por exemplo: No dia 09/08/2023 foi previsto que a temperatura máxima na cidade de São Paulo seria de 24ºC, enquanto que a temperatura mínima seria de 16ºC, portanto, a **amplitude térmica** prevista para o dia foi de 8ºC.

IN

Medidas de Dispersão: Variância, Desvio Padrão e Coeficiente de Variação.

Todas as três medidas de dispersão acima nos dizem **o quão distante da média estão os nossos dados**.

Para calcular a **variância** de um conjunto de n elementos, temos que utilizar a fórmula:

$$\sum^2 = \frac{(X1 - \overline{X})^2 + (X2 - \overline{X})^2 + (X3 - \overline{X})^2 + \ldots + (Xn - \overline{X})^2}{n}$$

A notação da variância é uma letra ao quadrado (Σ²) para destacar que a unidade da variância é o quadrado da unidade dos dados. Por exemplo, dadas certas alturas (18m, 21m, etc.), a unidade de medida da variância entre elas seria m².

Xn representa um elemento do conjunto; $\overline{x}$ representa a média aritmética dos dados

Por exemplo:

Para {4,6,8,10,12}, primeiro calculamos $\overline{x}$

$\overline{x}$ = (4 + 6 + 8 + 10 + 12) / 5 = 8

Agora aplicamos a fórmula:

$V = ((4 – 8)^2 + (6 – 8)^2 + (8 – 8)^2 + (10 – 8)^2 + (12 – 8)^2) / 5$

V = (16 + 4 + 0 + 4 + 16) / 5 = 40 / 5 = 8

O desvio padrão é a raiz quadrada da variância

$$DP = \sqrt{\Sigma^2}$$

No exemplo anterior, o desvio padrão seria igual à $\sqrt{8} \approx 2.83$

O coeficiente de variação (%) é 100 vezes o desvio padrão dividido pela média aritmética

$$CV = 100 * DP / \overline{X}$$

Ou

CV = 100 * desvio padrão / média aritmética

No exemplo anterior, o coeficiente de variação seria 100*2.83/8 = 283/8 = 35.275**%**

Classificação de dados pelas diferenças entre si

- **Dados homogêneos: CV < 30%** (os dados não são tão diferentes entre si)
- **Dados heterogêneos: CV >= 30%** (os dados são muito diferentes entre si)

Observação: há acadêmicos que utilizam valores de até 50% para diferenciar dados homogêneos de heterogêneos, no entanto, para a disciplina, o ponto de corte será 30%