#867 #DesenvolveDados Seg • Qua • Sex

Estatística I

Conteúdo Exercícios

1 Introdução a Estatística

1. Definições

Estatística Descritiva: Primeira etapa inicial da análise, quando ainda não conhecemos a forma do dado, com o objetivo de tirar informações prévias de modo informal e direto, quando obtemos grande volume de dados, precisamos de informações que *resumam* nosso conjuto de dados a fim de que possamos tirar conclusões sobre nossos dados

Probabilidade: pode ser pensada como a teoria matemática utilizada para se estudar a incerteza oriundas de *fenômenos* de caráter aleatório.

Inferência estatística: É o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados, denominado *população*, obtidos a partir de um conjunto extraido sobre esta denominada *amostra*.

1.2. Tipos de Variáveis

- **Variavél qualitativa**
    - *Nominal*: Valores que expressam atributos sem nenhum tipo de ordem. Ex: sexo, estado civil, país de origem
    - *Ordinal*: Valores que expressam atributos, porém com algum tipo de ordem ou grau. Ex: escolaridade, resposta de um paciente (piora, igual, melhora), classe social (alta, média, baixa)
- **Variavél quantitativa**
    - *Discreta*: Valores que expressam atributos nos valores inteiros. Ex: idade, numero de banheiros, numero de filhos.
    - *Contínua* : Valores que expressam atributos nos valores reais. Ex: Salário, temperatura

2. Medidas de Posição

2.1. Média

Define-se média como sendo:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

2.2. Mediana

A **mediana** de uma variável é um número tal que há o mesmo número de observações maiores e menores do que ele, ocupando assim a posição central da série de observações.

Exemplo

- [3, 4, 7, 8, 8] $\rightarrow$ **mediana = 7**
- [3, 4, 7, 8, 8, 9] $\rightarrow$ **mediana =** $\frac{7+8}{2} = 7,5$

Logo, podemos definir :

- *Se tamanho da amostra ímpar*, mediana será $X = X_{(\frac{n+1}{2})}$
- *Se tamanho da amostra par*, mediana será $X = \frac{X_{(\frac{n}{2})}+X_{(\frac{n+1}{2})}}{2}$

Como os valores de **índice no python** começam em *0*, devemos nos atentar que a equação acima deve ficar como:

- *Se tamanho da amostra ímpar*, mediana será $X = X_{(\frac{n}{2})}$
- *Se tamanho da amostra par*, mediana será $X = \frac{X_{(\frac{n-1}{2})}+X_{(\frac{n}{2})}}{2}$

2.3. Moda

A **moda** é o valor que ocorre com **mais frequência** em um conjunto de dados.

Dependendo do conjunto de dados, ele pode ser:

- Sem moda: quando nenhum valor se repete;
- Unimodal: quando existe apenas um valor repetido com maior frequência;
- Bimodal: quando existem dois valores com a mesma frequência.

3. Medidas de Dispersão

Medidas de variabilidade indicam o quanto as observações variam ao redor da medida de centralidade. Em outras palavras, indicam o quão longe podemos esperar que uma observação esteja do valor típico para aquela variável. Existem diversas medidas de variabilidade, algumas das quais apresentamos a seguir.

3.1. Amplitude

A amplitude é a diferença entre o maior e o menor valor observado. Esta medida de variabilidade é fortemente influenciada por valores extremos nas observações, como outliers.

3.2. Variância

A variância está relacionada ao quanto os valores se encontram distantes em relação à média, por isso seus valores são calculas o valor quadrático da diferença $X_i - \overline{x}$.

$$\sum_{i=1}^{n} \frac{(x - \overline{x})^2}{n}$$

3.3. Assimetria

É o **grau de desvio ou afastamento** da simetria de uma distribuição.Quando a curva é simétrica, a *média*, a *mediana* e a *moda* coincidem num mesmo ponto, de ordenada máxima, havendo um perfeito equilíbrio na distribuição. Quando o equilíbrio não acontece, isto é, a média, a mediana e a moda recaem em pontos diferentes da distribuição, esta será assimétrica, enviesada a direita ou esquerda.

4. Tabelas de Frequência

A distribuição de frequências é um agrupamento de dados em classes, de tal forma que contabilizamos o número de ocorrências em cada classe. O número de ocorrências de uma determinada classe recebe o nome de frequência absoluta. O objetivo é apresentar os dados de uma maneira mais concisa e que nos permita extrair informação sobre seu comportamento. A seguir, apresentamos algumas definições necessárias à construção da distribuição de frequências.

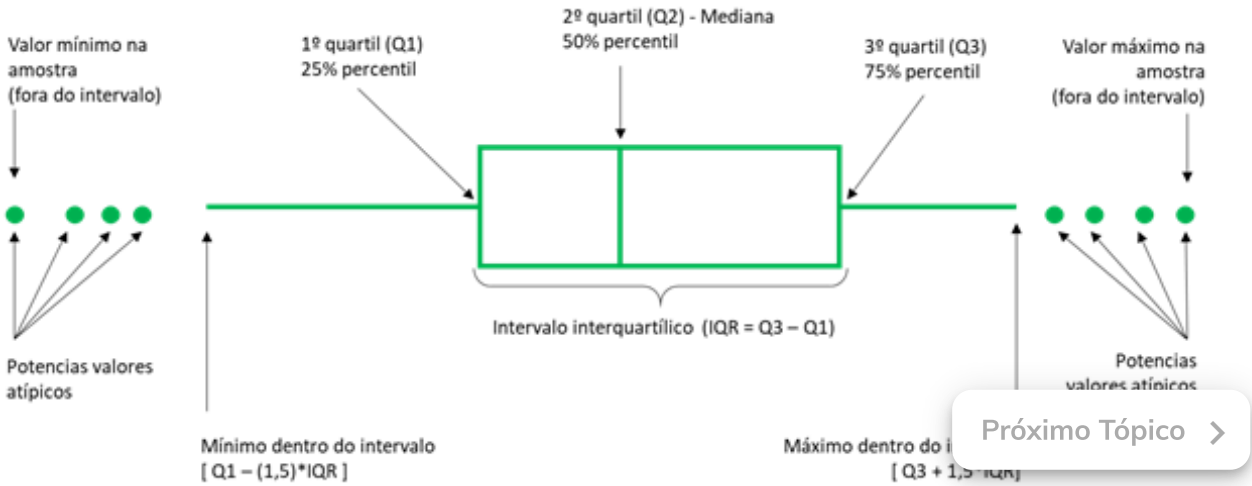
- **Frequência absoluta (fi):** É o número de observações correspondente a cada classe. A frequência absoluta é, geralmente, chamada apenas de frequência.
- **Frequência relativa ($f$ri):** É o quociente entre a frequência absoluta da classe correspondente e a soma das frequências (total observado), isto é, $f_{ri} = \frac{f_i}{\sum_j f_j}$, onde n representa o número total de observações.
- **Frequência percentual (pi):** É obtida multiplicando a frequência relativa por 100%.
- **Frequência acumulada:** É o total acumulado (soma) de todas as classes anteriores até a classe atual. Pode ser: frequência acumulada absoluta, frequência acumulada relativa, ou frequência acumulada percentual.

5. Boxplot

O boxplot é um gráfico utilizado para avaliar a distribuição empírica do dados. O boxplot é formado pelo primeiro e terceiro quartil, além da mediana. As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior ao limite inferior e do quartil superior até o maior valor não superior ao limite superior. Os limites são calculados da forma abaixo

- Limite inferior: $\max\{\min(\text{dados}); Q_1 - 1,5(Q_3 - Q_1)\}$.
- Limite superior: $\min\{\max(\text{dados}); Q_3 + 1,5(Q_3 - Q_1)\}$.

Para este caso, os pontos fora destes limites são considerados valores discrepantes (outliers) e são denotados por asterisco (*). A Figura a seguir apresenta um exemplo do formato de um boxplot.


Valor mínimo na amostra (fora do intervalo)
1º quartil (Q1) 25% percentil
2º quartil (Q2) - Mediana 50% percentil
3º quartil (Q3) 75% percentil
Valor máximo na amostra (fora do intervalo)
Intervalo interquartílico (IQR = Q3 – Q1)
Potencias valores atípicos
Potencias
Mínimo dentro do intervalo [Q1 – (1,5)*IQR]
Próximo Tópico