Funções e cálculos estatísticos

Estatística Descritiva

Sub-grupo do conteúdo estatístico

Ana Clara Lacerda da Silva, Emanuel Gonçalves Menezes, João Pedro Inacio Porto Vidigal, Luiz Gustavo Alves Alencar e Uigor Teodoro Martins.

Distribuição de frequências

- **Pontual, sem perda de informação**

A construção de uma distribuição de freqüência pontual é equivalente à construção de uma tabela simples, onde se listam os diferentes valores observados da variável, com suas freqüências absolutas, denotadas por Fi, onde o índice i corresponde ao número de linhas da tabela,

- **frequência relativa**

$$f_i = \frac{F_i}{n}$$

onde n é o tamanho da amostra, devendo ser substituída por N se os dados forem populacionais. A soma das freqüências relativas de todas as categorias é igual a 1;

- **relativa em percentual**

$$f_i\% = \frac{F_i}{n} \cdot 100 ,$$

representando o percentual de observações que pertencem àquela categoria. A soma das freqüências deve, agora, ser igual a 100%;

- **absoluta acumulada**

  A frequência absoluta acumulada é a soma das frequências absolutas ao decorrer das linhas da tabela. Essa frequência é bastante útil para obter alguns dados de determinada tabela.

- **acumulada relativa**

$$f_{a_i}\% = \frac{F_{a_i}}{n} 100$$

A frequência relativa acumulada é o acúmulo da frequência relativa. Para encontrar a frequência relativa acumulada, acrescentamos uma nova coluna à tabela. Copiamos a primeira frequência relativa na primeira linha, a segunda linha será a soma da frequência relativa da linha com a frequência acumulada da linha anterior, e assim sucessivamente.

- **Em classes, com perda de informações.**

O menor valor da classe é denominado limite inferior (li) e o maior valor da classe é denominado limite superior (Li). O intervalo ou classe pode ser representado das seguintes maneiras:

a) li | ____ Li, onde o limite inferior da classe é incluído na contagem da freqüência absoluta mas o superior não;

b) li ____| Li, onde o limite superior da classe é incluído na contagem mas o inferior não;

c) li | ____| Li, onde tanto o limite inferior quanto o superior são incluídos na contagem;

d) li ____ Li, onde os limites não fazem parte da contagem.

Pode-se escolher qualquer uma destas opções sendo o importante tornar claro no texto ou na tabela qual está sendo usada.

Milone (2004, p.36) apresenta os seguintes critérios para a determinação do número de intervalos, denotado por k:

1. Raiz quadrada: k n =

2. Log (Sturges): k =1+ 3,3log n

3. ln (Milone): k 1 2 ln n =− + ·

4. k 1 10 AT d = + ,

onde n é o número de elementos da amostra, AT é a amplitude total dos dados e d é o número de decimais de seus elementos.

Medidas Descritivas

Medidas de tendência central.

- **Média aritmética (amostral e populacional).**

  A Média Aritmética de um conjunto de dados é obtida somando todos os valores e dividindo o valor encontrado pelo número de dados desse conjunto.

  A média também pode ser simbolizada pelo somatório:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i-1}^{n} x_i$$

- **Moda.**

  Para calcular a moda de um conjunto de dados só é preciso observar os dados que aparecem com maior frequência no conjunto.

Exemplos:

Considere o conjunto de dados abaixo:

A = {2, 23, 4, 2, 5}

A moda para esse conjunto é: **M_o = 2**. É o número que aparece o maior número de vezes.

B = {17, 21, 2, 21, 8, 2}

Neste exemplo, a moda é: **M_o = 2 ou 21**. Então, podemos dizer que o conjunto **B** é bimodal (possui duas modas).

- **Mediana.**

A Mediana (**M_d**) é o valor de centro de um conjunto de dados. Para calcular, primeiro devemos ordenar o conjunto de dados.

Para calcular a mediana:

- Devemos ordenar o conjunto de dados em ordem crescente;
- Se o número de elementos for par, então a mediana é a média dos dois valores centrais. Soma os dois valores centrais e divide o resultado por 2: **(a + b)/2**.
- Se o número de elementos for ímpar, então a mediana é o valor central.

Medidas Separatrizes

- **Quartil.**

  Os quartis dividem o conjunto de dados em quatro partes iguais.

  $$p = \frac{n}{4}k \text{ , com } k = 1, 2, 3\text{, para determinação dos quartis;}$$

- **Decil.**

  Os decis dividem o conjunto de dados em dez partes iguais.

  $$p = \frac{n}{10}k \text{ , } k = 1, 2, \cdots, 9 \text{ para o cálculo dos decis;}$$

- **Percentil. (cálculo complexo)**

  Os percentis dividem o conjunto de dados em cem partes iguais.

  $$p = \frac{n}{100}k\text{, } k = 1, 2, \cdots, 99 \text{ para os percentis;}$$

Medidas de Dispersão

- **Amplitude total.**

  A amplitude total de um conjunto de dados é a diferença entre o maior e o menor valor observado. A medida de dispersão não levar em consideração os valores intermediários perdendo a informação de como os dados estão distribuídos e/ou concentrados.

  $$At = x_{max} - x_{min}$$

- **Amplitude interquartílica.**

  A amplitude interquartílica é a diferença entre o terceiro e o primeiro quartil. Esta medida é mais estável que a amplitude total por não considerar os valores mais extremos. Esta medida abrange 50% dos dados e é útil para detectar valores discrepantes. Por outras palavras, é a distância entre o terceiro quartil e o primeiro quartil.

$$A_q = Q_3 - Q_1$$

- **Desvio médio.**

  Ao somar todos os desvios, ou seja, ao somar todas as diferenças de cada valor observado em relação a média, o resultado é igual a zero (propriedade 5 da média). Isto significa que esta medida não mede a variabilidade dos dados. Para resolver este problema, pode-se desconsiderar o sinal da diferença, considerando-as em módulo e a média destas diferenças em módulo é denominada desvio médio.

$$d_m = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \quad \text{ou} \quad d_m = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- **Variância populacional.**

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Onde,

σ^2: variância
x_i: valor analisado
$\bar{X}$: média aritmética do conjunto
n: número de dados do conjunto

- **Variância amostral.**

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

- **Desvio Padrão.**

$$DP = \sqrt{\sigma^2} \text{ ou } DP = \sqrt{s^2}$$

- Coeficiente de variação.

O coeficiente de variação é uma medida de dispersão relativa definida como a razão entre o desvio padrão e a média:

$$CV = \frac{\sigma}{\mu}100 \quad \text{ou} \quad CV = \frac{S}{\overline{X}}100,$$

- **Medidas de assimetria.**

$$A_s = \frac{\mu - M_o}{\sigma} \quad \text{ou} \quad A_s = \frac{\overline{X} - M_o}{S}$$

para dados populacionais e amostrais, respectivamente.

Uma distribuição é classificada como:

simétrica se média = mediana = moda ou As = 0;

assimétrica negativa se média ≤ mediana ≤ moda ou As < 0.

- **Medidas de curtose.**

A medida de curtose é o grau de achatamento da distribuição, é um indicador da forma desta distribuição. É definido como:

$$K = \frac{(Q_3 - Q_1)}{2(P_{90} - P_{10})}$$