Estatística e Probabilidade

Carolina Silva Pena

Introdução

O resumo da informação contida nos dados utilizando uma única medida que representa a posição central dos dados não é capaz de capturar um aspecto muito importante: a variabilidade.

Por exemplo, suponha que quatro grupos de estudantes realizaram uma prova de estatística e obtiveram as seguintes notas:

- Grupo A: $\{2, 3, 4, 5, 6, 7, 8\}$.
- Grupo B: $\{5, 5, 5, 5, 5\}$.
- Grupo C: $\{0, 0, 0, 10, 10, 10\}$
- Grupo D: $\{2, 5, 5, 6, 7\}$

Note que $\overline{x}_A = \overline{x}_B = \overline{x}_C = \overline{x}_D = 5$.

Principais medidas de Dispersão

A identificação de tais séries de dados utilizando apenas a média não seria capaz de captar a diferença que existe entre elas. Isso pode ser feito utilizando as medidas de variabilidade.

As medidas de variabilidade que trataremos no curso são:

- Amplitude;
- Desvio Médio;
- Variância;
- Desvio Padrão;
- Distância interquartílica.

Amplitude

Amplitude $\triangle$: Diferença entre o maior e o menor valor do conjunto de dados.

- Exemplo: $\{2{,}79,\ 4{,}3,\ 4{,}46,\ 7{,}64,\ 7{,}7,\ 2{,}09,\ 4{,}94,\ 5{,}78,\ 8{,}33,\ 7{,}45,\ 5{,}28,\ 10,\ 7{,}8,\ 5{,}56,\ 4{,}15\}$
- Note que o Mínimo é igual a: **2,09**
- Note que o Máximo é igual a: **10**
- Amplitude = 10 - 2,09 = **7,91**

Distância em torno da média

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

Considere o seguinte conjunto de dados: $\{0, 2, 5, 4, 3\}$

- $\overline{x} = \{2{,}8\}$
- Distância:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x}) =& (0 - 2{,}8) + (2 - 2{,}8) + (5 - 2{,}8) + (4 - 2{,}8) + (3 - 2{,}8)\\ =& (\text{-}2{,}8) + (\text{-}0{,}8) + (2{,}2) + (1{,}2) + (0{,}2)\\ =& 0\end{aligned}$$

- Para qualquer conjunto de dados, a soma dos desvios é sempre igual a zero.

Desvio Médio (dm)

- Alternativa 1: considerar o valor absoluto da distância em torno da média.

$$dm = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

- Note que é conveniente trabalharmos em termos de distâncias médias, para que seja possível comparar conjuntos de dados com número de observações diferentes.

$$\begin{aligned} dm =& \frac{1}{5} \times \{|0 - 2{,}8| + |2 - 2{,}8| + |5 - 2{,}8| + |4 - 2{,}8| + |3 - 2{,}8|\} \\ =& \frac{1}{5} \times \{(2{,}8) + (0{,}8) + (2{,}2) + (1{,}2) + (0{,}2)\} \\ =& \frac{1}{5} \times 7{,}2 = 1{,}44 \end{aligned}$$

Variância (σ^2)

Alternativa 2: elevar a distância em torno da média ao quadrado.

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

- Nas situações em que a variância for utilizada apenas para descrever a variação de um conjunto de dados, ela será caculada conforme a equação (1).

Variância ($\hat{\sigma}^2$)

Nas situações em que a variação dos dados de uma amostra será utilizada para inferir sobre uma população, o denominador deve ser dividido por $n-1$, conforme mostrado na equação (2).

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

- o divisor $n-1$ faz com que a variância possua melhores propriedades estatísticas.
- Durante esse curso, a menos que seja dito o contrário, utilize a equação (2) para calcular a variância.

Variância ($\hat{\sigma}^2$) - Exemplo

Considere o seguinte conjunto de dados: $\{0, 2, 5, 4, 3\}$

- $\overline{x} = \{2{,}8\}$

$$
\begin{aligned}
\hat{\sigma^2} &= \frac{1}{5-1} \times \{(0-2{,}8)^2 + (2-2{,}8)^2 + (5-2{,}8)^2 + ... + (3-2{,}8)^2\} \\
&= \frac{1}{4} \times \{(7{,}84) + (0{,}64) + (4{,}84) + (1{,}44) + (0{,}04)\} \\
&= \frac{1}{4} \times 14{,}8 = 3{,}7
\end{aligned}
$$

Desvio padrão $\hat{\sigma}$

A variância é uma medida cuja dimensão é igual ao quadrado da dimensão dos dados. Por exemplo, se os dados forem expressos em *cm*, a variância será em cm^2. Isso pode gerar problemas de interpretação.

- O desvio padrão é então definido como a raiz quadrada da variância, sendo assim medido na escala original dos dados.

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

- Para o exemplo anterior, temos que o desvio padrão é:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{3{,}7} = 1{,}92$$

Coeficiente de Variação

O coeficiente de variação (CV) é muito utilizado para comparar grupos de dados que:

1. são medidos em escalas diferentes ou
2. quando as médias dos grupos são muito diferentes.

O CV é definido como a razão entre o desvio padrão ($\hat{\sigma}$) e a média amostral ($\overline{x}$):

$$CV = \frac{\hat{\sigma}}{\overline{x}} \times 100\%.$$

No caso do exemplo anterior, temos que: $CV = \frac{1{,}92}{2{,}8} \times 100 = 69\%$.

Percentil

A mediana também é conhecida como Percentil 50%, ou $q(50)$. De maneira mais ampla, podemos definir o conceito de percentil amostral.

- Percentil amostral: $q(p)$ é o valor tal que p% dos dados ordenados encontram-se abaixo dele e (100-p)% acima, em que $0 < p < 100$.

$$q(p) = \begin{cases} \frac{x_{(L)} + x_{(L+1)}}{2}, & \text{se } L \text{ é inteiro;} \\ x_{(\lceil L \rceil)}, & \text{caso contrário.} \end{cases}$$

em que:

- $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$.
- $L = \frac{p}{100} \times n$;
- $\lceil a \rceil$ é o menor inteiro maior que a.

Quartis

Os seguintes percentis são também conhecidos como Quartis:

1. $q(25) =$ Quartil 1 (Q_1);
2. $q(50) =$ Quartil 2 (Q_2);
3. $q(75) =$ Quartil 3 (Q_3).

Exemplo. Calcule o (Q_1) para o seguinte conjunto de dados:

$\{2{,}79,\ 4{,}3,\ 4{,}46,\ 7{,}64,\ 7{,}7,\ 2{,}09,\ 4{,}94,\ 5{,}78,\ 8{,}33,\ 7{,}45,\ 5{,}28,\ 10,\ 7{,}8,\ 5{,}56,\ 4{,}15\}$

- Passo 1: Ordenar

$\{2{,}09,\ 2{,}79,\ 4{,}15,\ 4{,}3,\ 4{,}46,\ 4{,}94,\ 5{,}28,\ 5{,}56,\ 5{,}78,\ 7{,}45,\ 7{,}64,\ 7{,}7,\ 7{,}8,\ 8{,}33,\ 10\}$

- Calcular $L = \frac{p}{100} \times n = \frac{25}{100} \times 15 = 3{,}75$.
- Logo $\lceil L \rceil = 4$ e $Q_1 = x_{(4)} = 4{,}3$

Distância interquatílica

Uma medida de dispersão alternativa, é a distância interquartílica.

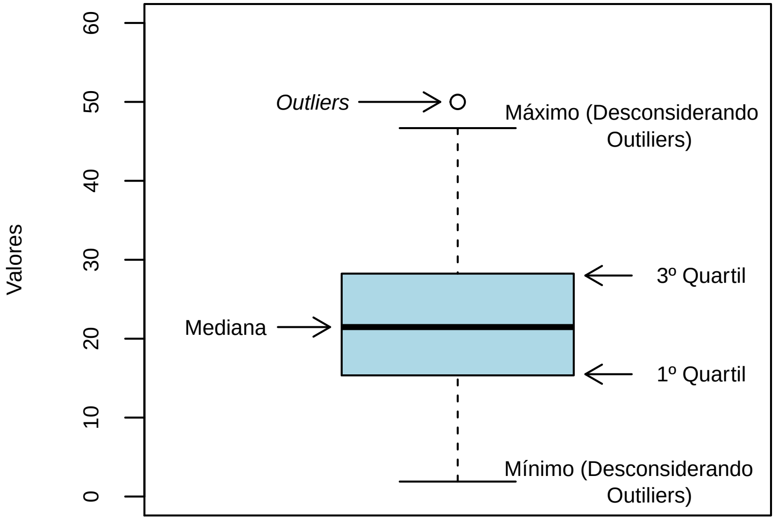
- Distância interquartílica: diferença entre o primeiro e terceiro quartil.

$$d_q = Q_3 - Q_1$$

Exemplo: Calcule a d_q para o conjunto de dados do slide anterior:

- Dados (já ordenados): {2,09, 2,79, 4,15, 4,3, 4,46, 4,94, 5,28, 5,56, 5,78, 7,45, 7,64, 7,7, 7,8, 8,33, 10}
- Anteriormente, mostramos que $Q_1 = x_{(4)} = 4{,}3$
- Para calcular Q_3 fazemos: $L = \frac{p}{100} \times n = \frac{75}{100} \times 15 = 11{,}25$.
  - Logo $\lceil L \rceil = 12$ e $Q_3 = x_{(12)} = 7{,}7$
- Por fim, temos que d_q = 7,7 - 4,3 = 3,4

Boxplot


Valores
0
10
20
30
40
50
60
Outliers
Máximo (Desconsiderando Outiliers)
3º Quartil
Mediana
1º Quartil
Mínimo (Desconsiderando Outiliers)

Boxplot

O Boxplot é um gráfico que traz informação sobre a dispersão e o nível de assimetria da amostra.

- 1º Intervalo: $Q_1 - x_{(min)}$;
- 2º Intervalo: $Q_2 - Q_1$;
- 3º Intervalo: $Q_3 - Q_2$;
- 4º Intervalo: $x_{(max)} - Q_3$;
- Valores atípicos (*Outliers*):
    - valores abaixo de $Q_1 - 1,5 \times (Q_3 - Q_1)$ ou
    - valores acima de $Q_3 + 1,5 \times (Q_3 - Q_1)$.