Medidas de variação ou dispersão

i	x_i	y_i	z_i
1	4	2	1
2	4	5	8
3	4	4	5
4	4	6	4
5	4	3	2
Σ	20	20	20
Média	4	4	4

Medidas de variação ou dispersão

Objetivo → indicar quanto os valores diferem entre si ou quanto eles se afastam da média

⇨ Complementam as medidas de tendência central

Medidas de variação mais utilizadas:

- Amplitude total
- Variância
- Desvio padrão
- Coeficiente de variação

Tendo a cabeça a arder e os pés enterrados no gelo, na média está tudo bem!

Amplitude total (a_t)

⇨ Fornece uma ideia inicial de variação

⇨ É obtida pela diferença entre o maior valor e o menor valor de um conjunto de dados

$$a_t = ES - EI$$

ES: extremo superior do conjunto de dados ordenado
EI: extremo inferior do conjunto de dados ordenado

$$a_t = x_{(n)} - x_{(1)}$$

Exemplo:

X = Tempo para realizar uma tarefa (min)

$$x_i = 3,\ 3,\ 4,\ 6,\ 7,\ 9,\ 9,\ 11,\ 12$$

$$a_t = ES - EI = 12 - 3 = 9\,\text{min}$$

EI 100% ES

Significado: todos os valores do conjunto de dados diferem, no máximo, em 9 minutos

Desvantagens

- pouco precisa
- extremamente influenciada por valores discrepantes

i	x_i	y_i	z_i
1	4	2	1
2	4	5	8
3	4	4	5
4	4	6	4
5	4	3	2
Σ	20	20	20
Média	4	4	4
a_t	0	4	7

Variância (s^2)

⇨ Medida de variação mais utilizada:

- facilidade de compreensão
- propriedades estatísticas importantes para a inferência

⇨ Leva em conta todos os valores do conjunto de dados

⇨ Considera o desvio da média como unidade básica da variação:

Desvio: $(x_i - \bar{x})$ ← mede quanto cada valor varia em relação à média

Exemplo:

$x_i = 4, 5, 7, 9, 10$

$\bar{x} = 7$

$$(x_i - \bar{x}) \begin{cases} 4 - 7 = -3 \\ 5 - 7 = -2 \\ 7 - 7 = 0 \\ 9 - 7 = 2 \\ 10 - 7 = 3 \end{cases}$$

variação do x_i em relação à média

Média dos desvios → variação média do conjunto de valores

soma de todos os desvios →

número de desvios somados →

$$\frac{\sum (x_i - \bar{x})}{n} = 0$$

4ª propriedade da média → $\sum (x_i - \bar{x}) = 0$

Solução: elevar os desvios ao quadrado $\rightarrow$ desvios negativos ficam positivos e podem ser somados

soma dos quadrados dos desvios

número de desvios somados

$$\frac{\sum (x_i - \bar{x})^2}{n}$$

Média dos quadrados dos desvios

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

número de graus de liberdade ou desvios independentes

Por que utilizar n-1 como denominador?

Porque este denominador confere à variância melhores propriedades estatísticas (importante na inferência estatística).

⇨ Quando o objetivo for apenas descrever a variação de um conjunto de valores, podemos usar o denominador n.

$$s_n^2 = \frac{\sum (x_i - \overline{x})^2}{n}$$

⇨ Quando o objetivo for estimar a variação de uma população por meio da variação de um conjunto de valores (amostra), devemos usar o denominador n-1.

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n - 1}$$

Exemplo:

X = tempo para realizar uma tarefa (min)

x_i = 4, 5, 7, 9, 10 $\longrightarrow$ $\bar{x} = 7$ min

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$= \frac{(4-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (10-7)^2}{5-1}$$

$$= \frac{9+4+0+4+9}{4} = \frac{26}{4} = 6{,}5$$

$s^2 = 6{,}5 \text{ min}^2$ $\longleftarrow$ unidade de medida fica elevada ao quadrado

Propriedades da variância

1ª propriedade: A variância de um conjunto de dados que não varia, ou seja, cujos valores são uma constante, é zero.

Verificação numérica:

$$x_i = 7, 7, 7, 7, 7 \longrightarrow \bar{x}=7$$

$$s^2 = \frac{(7-7)^2+(7-7)^2+(7-7)^2+(7-7)^2+(7-7)^2}{5-1} = 0$$

$$s_c^2 = \frac{\sum (c-c)^2}{n-1} = 0$$

2ª propriedade: Ao somar uma constante **c** a todos os valores de um conjunto de dados, a variância destes dados **não se altera**.

Verificação numérica:

$x_i = 4, 5, 7, 9, 10$ → $\bar{x} = 7$; $s^2 = 6{,}5$

Somar c=2

$x_i + 2 = 6, 7, 9, 11, 12$ → $\bar{x}_{x+2} = 9 \rightarrow \boxed{\bar{x}_{x+c} = \bar{x} + c}$; $s^2_{x+2} = 6{,}5 \rightarrow \boxed{s^2_{x+c} = s^2}$

$$s^2_{x+2} = \frac{(6-9)^2 + (7-9)^2 + (9-9)^2 + (11-9)^2 + (12-9)^2}{5-1}$$

$$= \frac{9+4+0+4+9}{4} = \frac{26}{4} = 6{,}5\,\text{min}^2$$

3ª propriedade: Ao multiplicar todos os valores de um conjunto de dados por uma constante **c**, a variância destes dados fica multiplicada pelo **quadrado** desta constante.

Verificação numérica:

$$x_i = 4,\ 5,\ 7,\ 9,\ 10 \quad \begin{cases} \bar{x} = 7 \\ s^2 = 6{,}5 \end{cases}$$

Multiplicar por c=2

$$2x_i = 8,\ 10,\ 14,\ 18,\ 20 \quad \begin{cases} \bar{x}_{2x} = 14 \rightarrow \boxed{\bar{x}_{cx} = c\bar{x}} \\ s^2_{2x} = 26 \rightarrow \boxed{s^2_{xc} = c^2 s^2} \end{cases}$$

$$s^2_{2x} = \frac{(8-14)^2 + (10-14)^2 + (14-14)^2 + (18-14)^2 + (20-14)^2}{5-1}$$

$$= \frac{36+16+0+16+36}{4} = \frac{104}{4} = 26\,\text{min}^2 = 2^2 \quad 6{,}5$$

Trabalhando com a expressão da variância, é possível encontrar uma fórmula mais prática de cálculo:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \Rightarrow \sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2$$

$$= \sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left(\frac{\sum x_i}{n}\right)^2$$

$$= \sum x_i^2 - 2\frac{\left(\sum x_i\right)^2}{n} + n\frac{\left(\sum x_i\right)^2}{n^2}$$

$$= \sum x_i^2 - 2\frac{\left(\sum x_i\right)^2}{n} + \frac{\left(\sum x_i\right)^2}{n}$$

$$\boxed{s^2 = \frac{\sum x_i^2 - n\,\bar{x}^2}{n-1}}$$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$$

Desvantagens da variância:

1. Como a variância é calculada a partir da média, é uma medida pouco resistente, ou seja, muito influenciada por valores atípicos.

2. Como a unidade de medida fica elevada ao quadrado, a interpretação da variância se torna mais difícil.

Para solucionar o problema de interpretação da variância surge outra medida: o desvio padrão.

Desvio padrão (s)

- É definido como a raiz quadrada positiva da variância

$$s = \sqrt{s^2}$$

Exemplo:

X = tempo para realizar uma tarefa (min)

$$\bar{x} = 7\,\text{min}$$

x_i = 4, 5, 7, 9, 10

$$s^2 = 6{,}5\,\text{min}^2$$

$$s = \sqrt{s^2}$$

$$s = \sqrt{6{,}5\,\text{min}^2}$$

$$s = 2{,}55\,\text{min}$$

Apresentação do desvio padrão:

$$\bar{x} \pm s$$

$$7 \pm 2{,}55$$

Tempo médio para realizar a tarefa de 7 minutos com uma variação média de 2,55 minutos acima e abaixo da média.

Significado: variação média em torno da média aritmética

ATENÇÃO: ver propriedades da variância e quais as adaptações necessárias para o desvio padrão!

Interpretação do Desvio Padrão

O desvio padrão é uma forma de medir distância entre as observações de um conjunto de dados em relação a sua média.

Exemplo:

Podemos comparar 85 em um exame de inglês com 80 em um exame de alemão?

Que nota é realmente mais alta?

Um raciocínio rápido mostra que isso depende do desempenho dos outros estudantes em cada turma, ou seja, da variabilidade das notas – do desvio padrão.

Esse mesmo raciocínio de comparação de escores em provas é utilizado nos concursos vestibulares.

Coeficiente de Variação (CV)

- O coeficiente de variação é definido como a proporção (ou percentual) da média representada pelo desvio padrão.

$$CV = \frac{s}{\bar{x}} \; 100\%$$

Exemplo:

X = tempo para realizar uma tarefa (min)

$$x_i = 4, 5, 7, 9, 10 \quad \begin{cases} \bar{x} = 7 \text{ min} \\ s = 2{,}55 \text{ min} \end{cases}$$

$$CV = \frac{s}{\bar{x}} \; 100\% = \frac{2{,}55 \text{ min}}{7 \text{ min}} \; 100\% = 36{,}4\%$$

- O CV é a medida mais utilizada para comparar variabilidades de diferentes conjuntos de dados

- Esta comparação não deve ser feita através de qualquer medida de variação em duas situações:
  - quando as médias dos conjuntos comparados são muito desiguais
  - quando as unidades de medida são diferentes

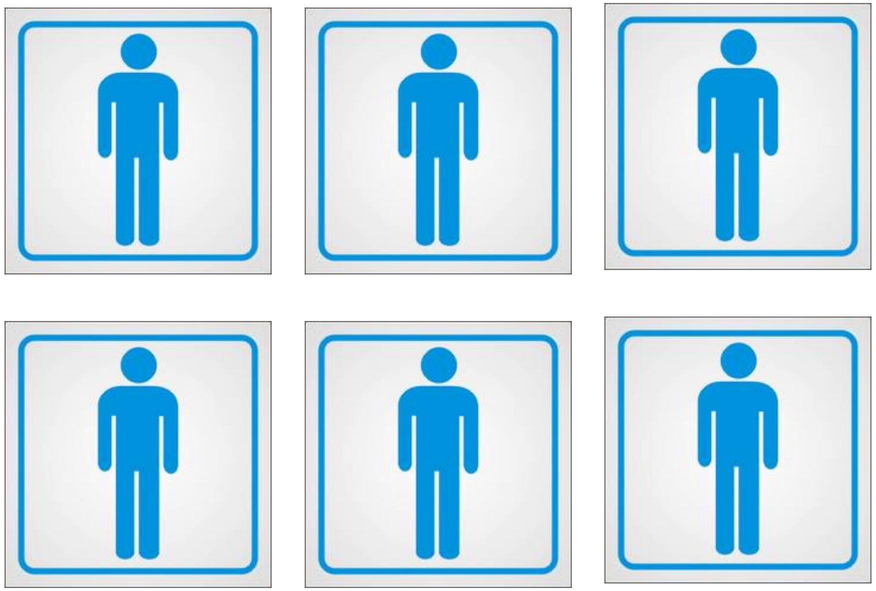
  Nessas situações devemos usar o CV.

Vantagens:

- O CV é desprovido de unidade de medida (expresso em percentagem)
- O CV é uma medida relativa, pois relaciona o desvio padrão com a sua respectiva média aritmética

Exemplo:

Consideremos que x_{1i} e x_{2i} são conjuntos de valores referentes aos salários (R$) de mulheres e homens , para os quais foram obtidas as seguintes medidas:



Mulheres (X_1): $\bar{x}_1 = 1.300$

$s_1 = 340$

Homens (X_2): $\bar{x}_2 = 2.500$

$s_2 = 420$

Qual grupo varia mais em relação aos salários?

$$\bar{x}_1 = 1300 \qquad \bar{x}_2 = 2500$$

$$s_1 = 340 \qquad s_2 = 420$$

$\boxed{CV_1 = 26{,}2\%}$ $\boxed{CV_2 = 16{,}8\%}$

O maior desvio padrão, quando comparado à sua média, representou menor variação.

Quando as médias são diferentes, devemos usar o CV.

Exercício proposto:

Contou-se o número de vendas de determinado produto durante os sete dias de uma semana, com os seguintes resultados:

14	20	20	20	15	16	18

a) Determine a média, a mediana e a moda.

b) Calcule a variância, o desvio padrão e o coeficiente de variação.

média = 17,6
mediana = 18
moda = 20
variância = 6,62
desvio padrão = 2,57
CV = 14,64%