Unidade II – Estatística descritiva

2.1. Apresentação de dados
- 2.1.1 Séries estatísticas
- 2.1.2 Tabelas
- 2.1.3 Gráficos

2.2. Distribuições de freqüências e gráficos
- 2.2.1 Tabelas de classificação simples
- 2.2.2 Tabelas de classificação cruzada

2.3. Medidas descritivas
- 2.3.1 Medidas de localização ou tendência central
- 2.3.2 Medidas separatrizes
- 2.3.3 Medidas de variação ou dispersão
- 2.3.4 Medidas de formato

2.4. Análise exploratória de dados

Profª. Clause Piana

2. Medidas separatrizes

Profª. Clause Piana

Medidas separatrizes

Medidas descritivas que buscam dividir um conjunto de dados ordenado em proporções essencialmente iguais

Quantis → delimitam proporções de valores no conjunto ordenado

Mediana → divide o conjunto ordenado em **duas** partes

Quartis → dividem o conjunto ordenado em **quatro** partes

Decis → dividem o conjunto ordenado em **dez** partes

Percentis → dividem o conjunto ordenado em **cem** partes

Profª. Clause Piana

Quantis → delimitam proporções de valores no conjunto ordenado

x_p , sendo p = proporção

Mediana → divide o conjunto ordenado em **duas** partes

$Md = x_{.5} \rightarrow p = 0,5$

Quartis → dividem o conjunto ordenado em **quatro** partes

$Q_1 = x_{.25} \rightarrow p = 0,25$

$Q_2 = x_{.5} \rightarrow p = 0,5$

$Q_3 = x_{.75} \rightarrow p = 0,75$

Decis → dividem o conjunto ordenado em **dez** partes

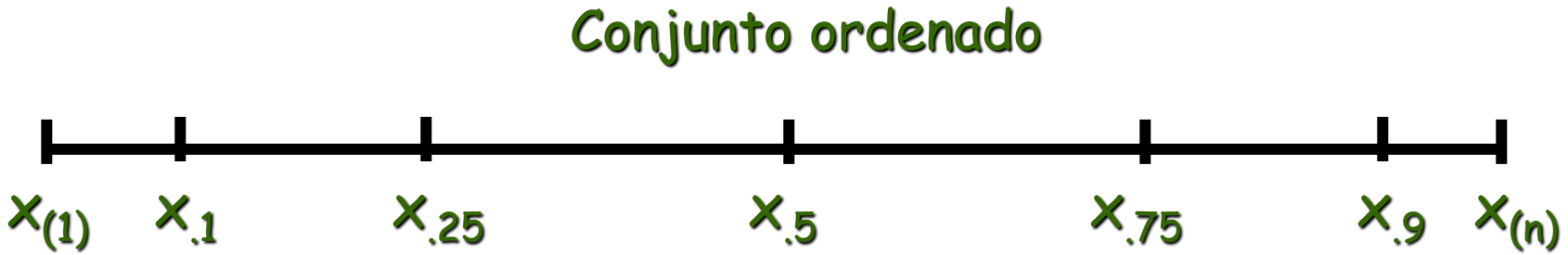
$D_1 = x_{.1} \rightarrow p = 0,1$ ... $D_5 = x_{.5} \rightarrow p = 0,5$ ... $D_9 = x_{.9} \rightarrow p = 0,9$

Percentis → dividem o conjunto ordenado em **cem** partes

$P_1 \rightarrow p = 0,01$ ... $P_{50} \rightarrow p = 0,5$ ... $P_{99} \rightarrow p = 0,99$

Profª. Clause Piana

Quantis (x_p)



$x_{.1}$: pelo menos 0,10 dos valores são menores ou iguais a $x_{.1}$

$x_{.25}$: pelo menos 0,25 dos valores são menores ou iguais a $x_{.25}$

$x_{.5}$: pelo menos 0,50 dos valores são menores ou iguais a $x_{.5}$

$x_{.75}$: pelo menos 0,75 dos valores são menores ou iguais a $x_{.75}$

$x_{.9}$: pelo menos 0,90 dos valores são menores ou iguais a $x_{.9}$

Determinação do quantil (x_p) pelo método de inversão da função de distribuição empírica (FDE):

1. Ordenar os dados

2. Multiplicar a proporção **(p)** pelo número de observações **(n)**

3. Fazer **np = j + f** , onde **j** = parte inteira e **f** = parte decimal

$$\begin{cases} \text{Se } f = 0 \Rightarrow x_p = \dfrac{x_{(j)} + x_{(j+1)}}{2} \\ \text{Se } f > 0 \Rightarrow x_p = x_{(j+1)} \end{cases}$$

Nem sempre há um valor do conjunto para a proporção desejada.

Profª. Clause Piana

Exercício proposto:

Foram registrados os tempos de frenagem (em décimos de segundos) para 21 motoristas que dirigiam a 30 milhas por hora. Os valores obtidos foram:

69 57 70 80 46 61 65 74 75 55 67

56 71 72 61 66 58 68 70 68 59

Para o conjunto de valores, calcule os quartis e interprete esses valores.

1 milha = 1,61 km

Exercício proposto:

Foram registrados os tempos de frenagem (em décimos de segundos) para 21 motoristas que dirigiam a 30 milhas por hora. Os valores obtidos foram:

46 55 56 57 58 59 61 61 65 66 67
68 68 69 70 70 71 72 74 75 80

Para o conjunto de valores, calcule os quartis e interprete esses valores.

1 milha = 1,61 km

Medidas descritivas para dados agrupados em classes

Profª. Clause Piana

Distribuições de variáveis discretas – medidas exatas

A partir da distribuição de frequências, calcule as medidas de localização (média, mediana e moda) e os quartis.

Número de erros em um conjuntos de caracteres (*string*) de 1.000 *bits*.

j	Classe	F_j	F'_j	f_j	f'_j
1	0	55	55	0,1571	0,1571
2	1	60	115	0,1714	0,3286
3	2	112	227	0,32	0,6486
4	3	82	309	0,2343	0,8829
5	4	31	340	0,0886	0,9714
6	5	8	348	0,0229	0,9943
7	6	2	350	0,0057	1,0000
	Σ	350	-	1,0000	-

$$\bar{x}_p = \frac{\sum c_j F_j}{\sum F_j} = \frac{0 \times 55 + 1 \times 60 + 2 \times 112 + \ldots + 6 \times 2}{350} = 2{,}017$$

$$\bar{x}_p = \sum c_j f_j = 0 \times 0{,}1571 + 1 \times 0{,}1714 + 2 \times 0{,}32 + \ldots + 6 \times 0{,}0057 = 2{,}017$$

Distribuições de variáveis discretas – medidas exatas

A partir da distribuição de frequências, calcule as medidas de localização (média, mediana e moda) e os quartis.

Número de erros em um conjuntos de caracteres (*string*) de 1.000 *bits*.

j	Classe	F_j	F'_j	f_j	f'_j
1	0	55	55	0,1571	0,1571
2	1	60	115	0,1714	0,3286
3	2	112	227	0,32	0,6486
4	3	82	309	0,2343	0,8829
5	4	31	340	0,0886	0,9714
6	5	8	348	0,0229	0,9943
7	6	2	350	0,0057	1,0000
	Σ	350	-	1,0000	-

Moda

Classe modal

Classe mediana

Classe que contém a mediana

Classe que contém o terceiro quartil

Profª. Clause Piana

Distribuições de variáveis discretas – medidas exatas

A partir da distribuição de frequências, calcule as medidas de localização (média, mediana e moda) e os quartis.

Número de erros em um conjuntos de caracteres (*string*) de 1.000 *bits*.

j	Classe	F_j	F'_j	f_j	f'_j
1	0	55	55	0,1571	0,1571
2	1	60	115	0,1714	0,3286
3	2	112	227	0,32	0,6486
4	3	82	309	0,2343	0,8829
5	4	31	340	0,0886	0,9714
6	5	8	348	0,0229	0,9943
7	6	2	350	0,0057	1,0000
	Σ	350	-	1,0000	-

Classe modal

Classe mediana

Média ponderada

$$\bar{x}_p = \frac{\sum c_j F_j}{\sum F_j} = 2{,}017$$

Moda = 2

Mediana = 2

Q_1 = 1

Q_2 = 2

Q_3 = 3

Profª. Clause Piana

Distribuições de variáveis contínuas – medidas aproximadas

A partir da distribuição de frequências, calcule as medidas de localização (média, mediana e moda) e os quartis.

Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

j	Classe	c_j	F_j	f_j	f_j'
1	3,11|— 16,00	9,56	8	0,16	0,16
2	16,00|— 28,89	22,45	20	0,4	0,56
3	28,89|— 41,78	35,34	6	0,12	0,68
4	41,78|— 54,67	48,23	8	0,16	0,84
5	54,67|— 67,56	61,12	3	0,06	0,9
6	67,56|— 80,45	74,01	1	0,02	0,92
7	80,45|—|93,34	86,90	4	0,08	1
	Σ	-	50	1	-

Média ponderada

$$\bar{x}_p = \frac{\sum c_j F_j}{\sum F_j} = \frac{9,56 \times 8 + 22,45 \times 20 + 35,34 \times 6 + \ldots + 86,90 \times 4}{50} = 37,57$$

$$\bar{x}_p = \sum c_j f_j = 9,56 \times 0,16 + 22,45 \times 0,4 + 35,34 \times 0,12 + \ldots + 86,90 \times 0,08 = 37,57$$

Distribuições de variáveis contínuas – medidas aproximadas

A partir da distribuição de frequências, calcule as medidas de localização (média, mediana e moda) e os quartis.

Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

j	Classe	c_j	F_j	f_j	f'_j
1	3,11|— 16,00	9,56	8	0,16	0,16
2	16,00|— 28,89	22,45	20	0,4	0,56
3	28,89|— 41,78	35,34	6	0,12	0,68
4	41,78|— 54,67	48,23	8	0,16	0,84
5	54,67|— 67,56	61,12	3	0,06	0,9
6	67,56|— 80,45	74,01	1	0,02	0,92
7	80,45|—|93,34	86,90	4	0,08	1
	Σ	-	50	1	-

Classe modal

Classe mediana

Classe que contém a mediana e o primeiro quartil

Classe que contém o terceiro quartil

Distribuições de variáveis contínuas – medidas aproximadas

A partir da distribuição de frequências, calcule as medidas de localização (média, mediana e moda) e os quartis.

Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

j	Classe	c_j	F_j	f_j	f'_j
1	3,11|— 16,00	9,56	8	0,16	0,16
2	16,00|— 28,89	22,45	20	0,4	0,56
3	28,89|— 41,78	35,34	6	0,12	0,68
4	41,78|— 54,67	48,23	8	0,16	0,84
5	54,67|— 67,56	61,12	3	0,06	0,9
6	67,56|— 80,45	74,01	1	0,02	0,92
7	80,45|—|93,34	86,90	4	0,08	1
	Σ	-	50	1	-

Média ponderada

$$\bar{x}_p = \frac{\sum c_j F_j}{\sum F_j} = 37{,}57$$

Classe modal = 2

Classe mediana = 2

Classe do Q_1 = 2

Classe do Q_2 = 2

Classe do Q_3 = 4

Profª. Clause Piana

3. Medidas de variação ou dispersão

Profª. Clause Piana

Observando os três conjuntos de dados abaixo, verificamos que uma medida de tendência central não é suficiente para diferenciá-los. Que característica dos dados poderia evidenciar que os conjuntos são diferentes?

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
$\sum$	20	20	20
Média	4	4	4

A variabilidade !!!

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

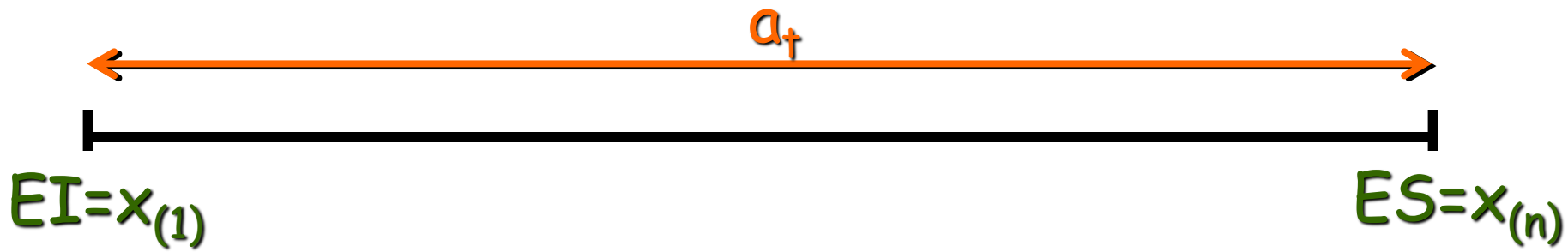
Medidas de variação mais utilizadas:

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

Profª. Clause Piana

Amplitude total (a_t)

- Fornece uma idéia rudimentar 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 = peso (kg)

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

$a_t = ES - EI = 10 - 4 = 6\,kg$

Significado: todos os valores do conjunto de dados diferem, no máximo, em 6kg

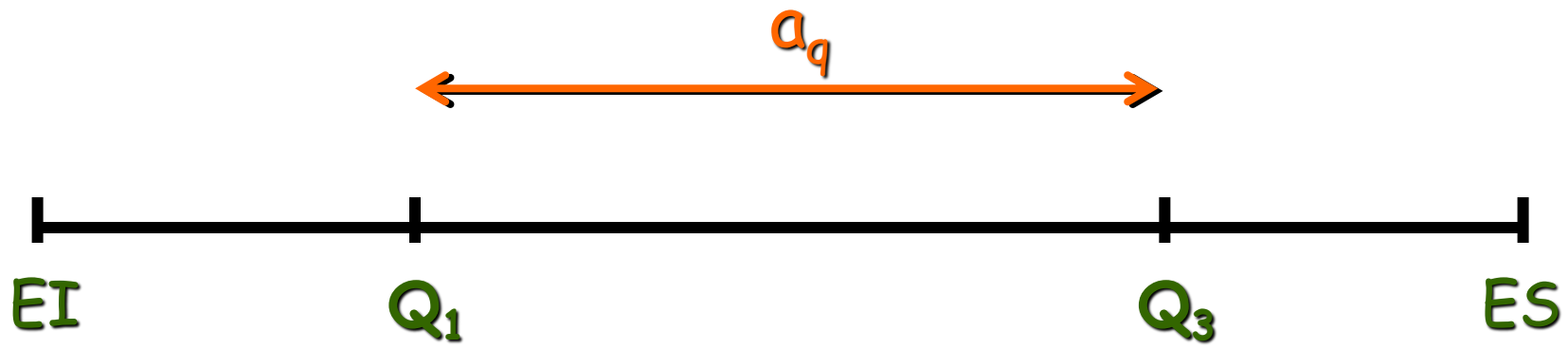
Desvantagens

- 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
$\sum$	20	20	20
Média	4	4	4
a_t	0	4	7

Amplitude interquartílica (a_q)

⇨ É obtida pela diferença entre o **terceiro** e o **primeiro quartis**



$$a_q = Q_3 - Q_1$$

Q_1 : primeiro quartil

Q_3 : terceiro quartil

Exemplo:

X = peso (kg)

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

Q_1 = 4 kg e Q_3 = 9 kg

$$a_q = Q_3 - Q_1 = 9 - 4 = 5 kg$$

Significado: 50% dos valores mais centrais do conjunto, diferem, no máximo, em 5 kg

Vantagem

- medida resistente (não é afetada por valores discrepantes)

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
- Amplitude interquartílica
- **Variância**
- Desvio padrão
- Coeficiente de variação

Variância (s^2)

⇨ Medida de variação mais utilizada:

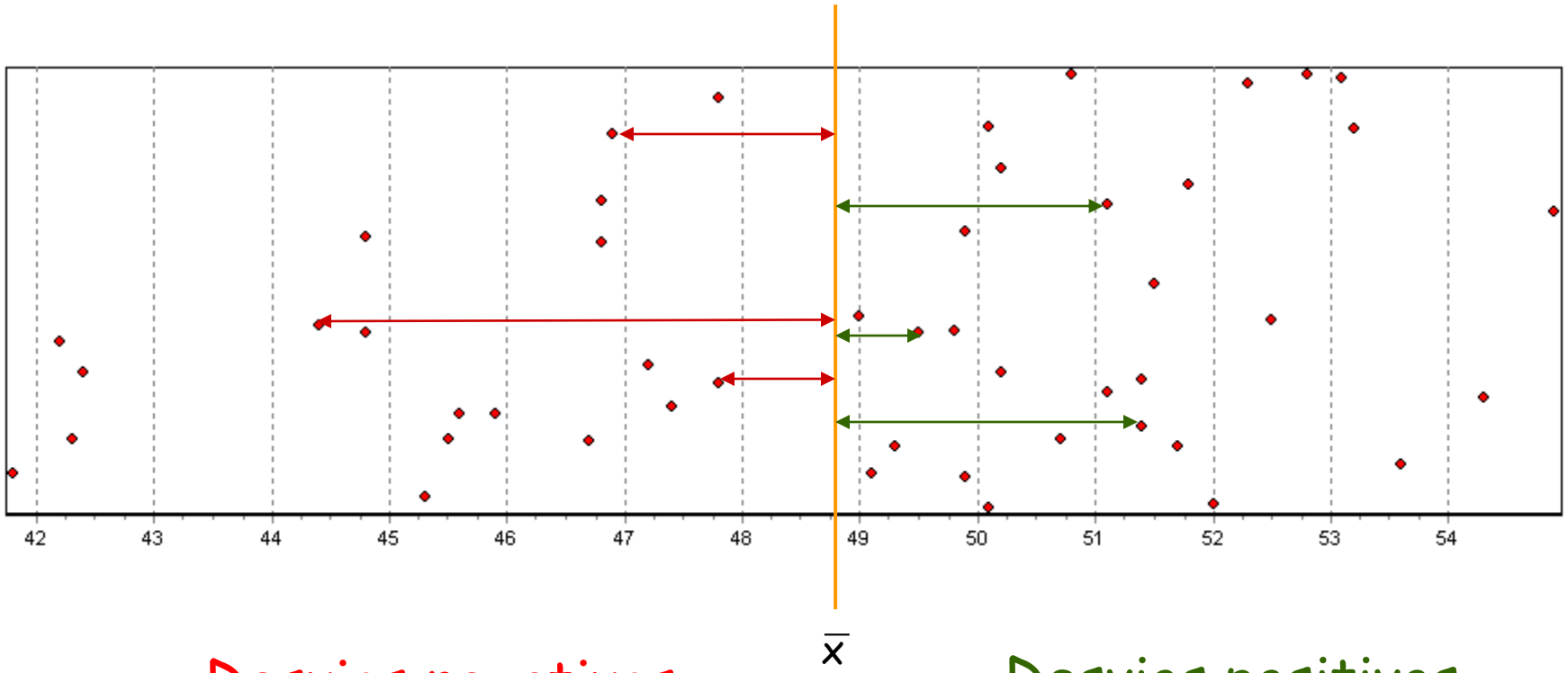
- **facilidade de compreensão**
- **propriedades matemáticas e estatísticas**

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

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

n=48

$$(x_i - \bar{x})$$



Desvios negativos

Desvios positivos

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 de cada x_i em relação à média

Média dos desvios → variação média do conjunto de valores

soma de todos os desvios →

total de desvios somados →

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

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

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 de cada x_i em relação à média

Média dos desvios → variação média do conjunto de valores

soma de todos os desvios →

total 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 → desvios negativos ficam positivos e podem ser somados

soma dos quadrados dos desvios

total de desvios somados

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

Média dos quadrados dos desvios

Variância: definida como a **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**.

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

- 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**.

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

Exemplo:

X = peso (kg)

$$x_i = 9, 7, 5, 10, 4 \rightarrow \bar{x} = 7 \text{ kg}$$

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

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

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

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

unidade de medida fica elevada ao quadrado

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

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

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

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

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

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

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

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

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

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

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

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

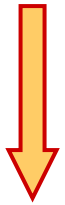
$$= \sum x_i^2 - 2\frac{\left(\sum x_i\right)^2}{n} + n\frac{\left(\sum x_i\right)^2}{n^2}$$

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

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

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

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



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

Propriedades algébricas da variância

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

$$x_i = 4, 4, 4, 4, 4 \qquad \overline{x} = 4$$

$$s^2 = \frac{(4-4)^2 + (4-4)^2 + (4-4)^2 + (4-4)^2 + (4-4)^2}{5-1}$$

$$s^2 = 0$$

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

Verificação numérica:

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

Somar c=2

$$x_i + 2 = 11, 9, 7, 12, 6 \quad \begin{cases} \nearrow \ \bar{x}_{x+2} = 9 \rightarrow \boxed{\bar{x}_{x+c} = \bar{x} + c} \\ \searrow \ s^2_{x+2} = ? \rightarrow \boxed{s^2_{x+c} = s^2} \end{cases}$$

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

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

Demonstração:

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

$$s^2_{x+c} = \frac{\sum \left[(x_i + c) - (\bar{x} + c)\right]^2}{n-1}$$

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

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

$$\boxed{s^2_{x+c} = s^2}$$

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

Verificação numérica:

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

Multiplicar por c=2

$$2x_i = 18, 14, 10, 20, 8 \rightarrow \bar{x}_{2x} = 14 \rightarrow \boxed{\bar{x}_{cx} = c\bar{x}}$$

$$s^2_{2x} = ? \rightarrow \boxed{s^2_{cx} = c^2 s^2}$$

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

$$= \frac{16+0+16+36+36}{4} = \frac{104}{4} = 26\,kg^2 = 2^2 \times 6{,}5$$

Demonstração:

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

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

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

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

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

$$\boxed{s_{cx}^2 = c^2 s^2}$$

Propriedades algébricas da variância

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

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

$$s^2_{x+c} = s^2$$

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

$$s^2_{cx} = c^2 s^2$$

Profª. Clause Piana

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
$\sum$	20	20	20
Média	4	4	4
Amplitude total	0	4	7
Variância	0	2,5	7,5

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{(2-4)^2 + (5-4)^2 + (4-4)^2 + (6-4)^2 + (3-4)^2}{5-1} = 2,5$$

$$s_z^2 = \frac{\sum (z_i - \bar{z})^2}{n-1} = \frac{(1-4)^2 + (8-4)^2 + (5-4)^2 + (2-4)^2 + (3-4)^2}{5-1} = 7,5$$

Exemplo: Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

$\bar{x} = 34,78$ reais

Calcule a variância desses dados.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(3,11-34,78)^2 + (8,88-34,78)^2 + \ldots + (93,34-34,78)^2}{50-1} = 471,14$$

$s^2 = 471,14 \text{ reais}^2$

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 = peso (kg)

x_i = 9, 7, 5, 10, 4

$\bar{x} = 7\ kg$

$s^2 = 6{,}5\ kg^2$

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

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

$$s = 2{,}55\ kg$$

Apresentação do desvio padrão:

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

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

Peso médio de 7 kg com uma variação média de 2,55 kg acima e abaixo da média.

Significado: variação média em torno da média aritmética

Vantagens

⇨ Facilidade de interpretação

⇨ É possível associar proporções de valores a intervalos entre a **média** e o **desvio padrão**

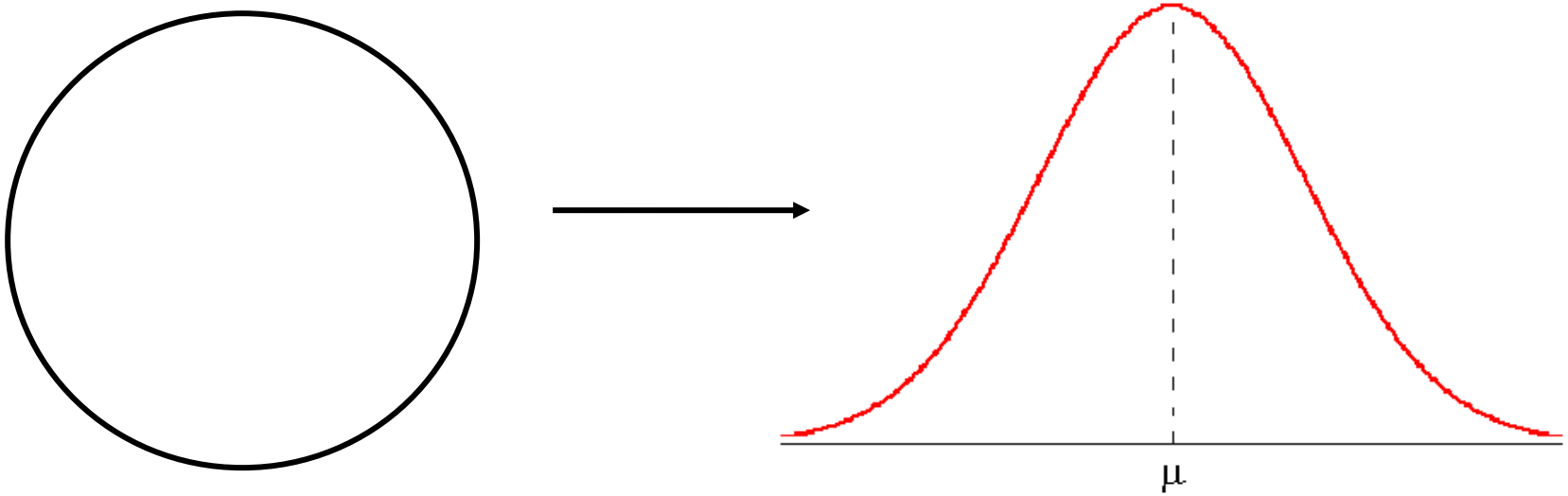
Numa distribuição **simétrica** e **unimodal**

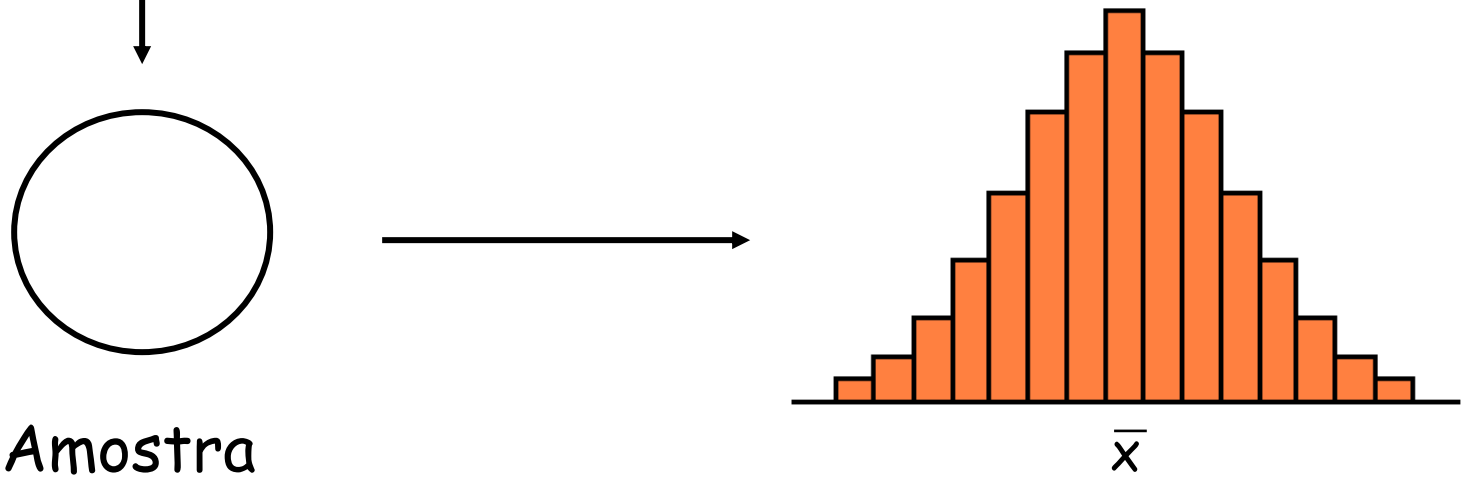
$$\bar{x} \pm s \rightarrow \cong 68\%$$

Profª. Clause Piana

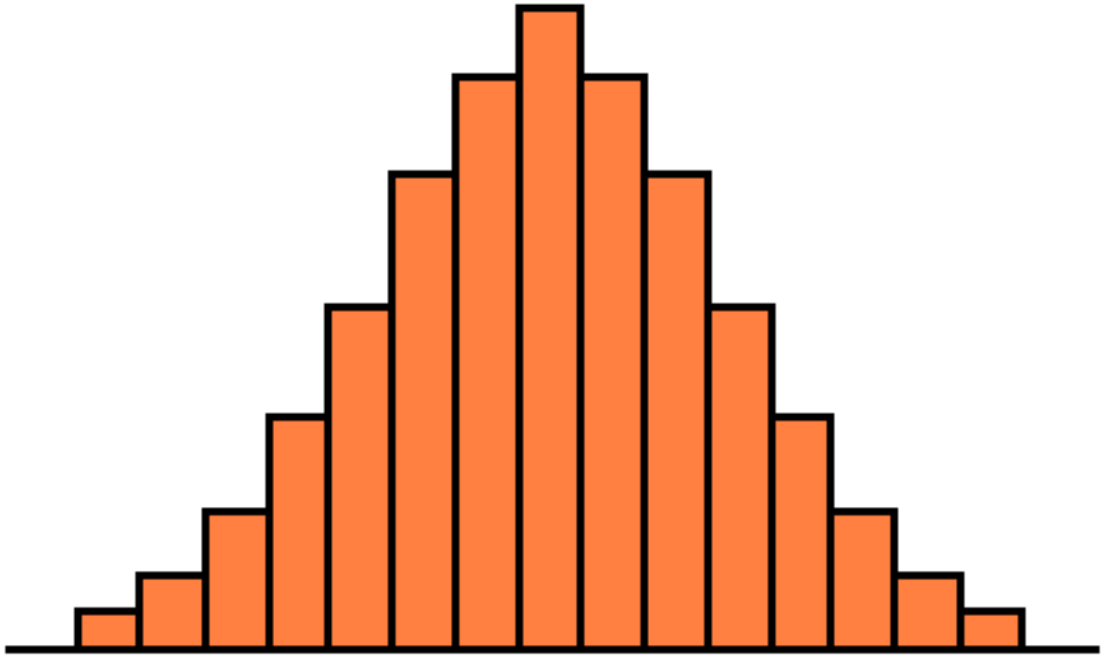
população

Comportamento normal





Amostra



$\bar{x} \pm s \rightarrow \cong 68\%$

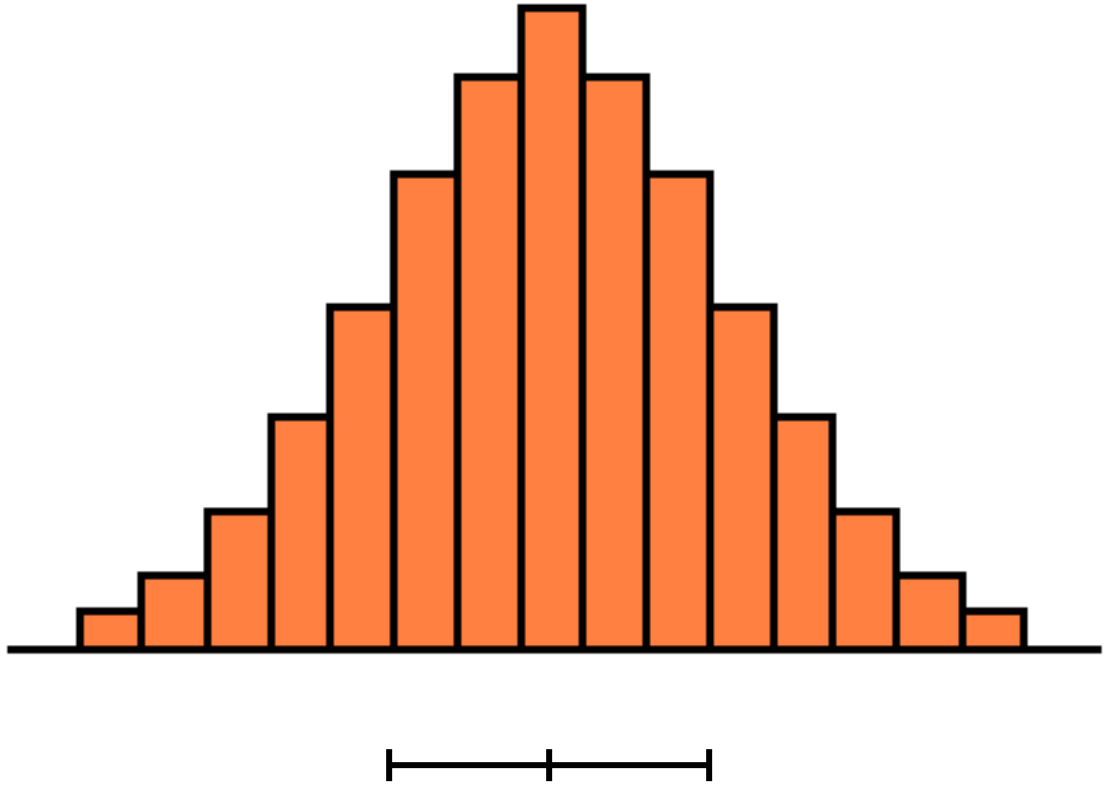
$\bar{x} \pm 2s \rightarrow \cong 95\%$

$\bar{x} \pm 3s \rightarrow \cong 99{,}7\%$

Profª. Clause Piana

No exemplo:

$7 \pm 2{,}55$



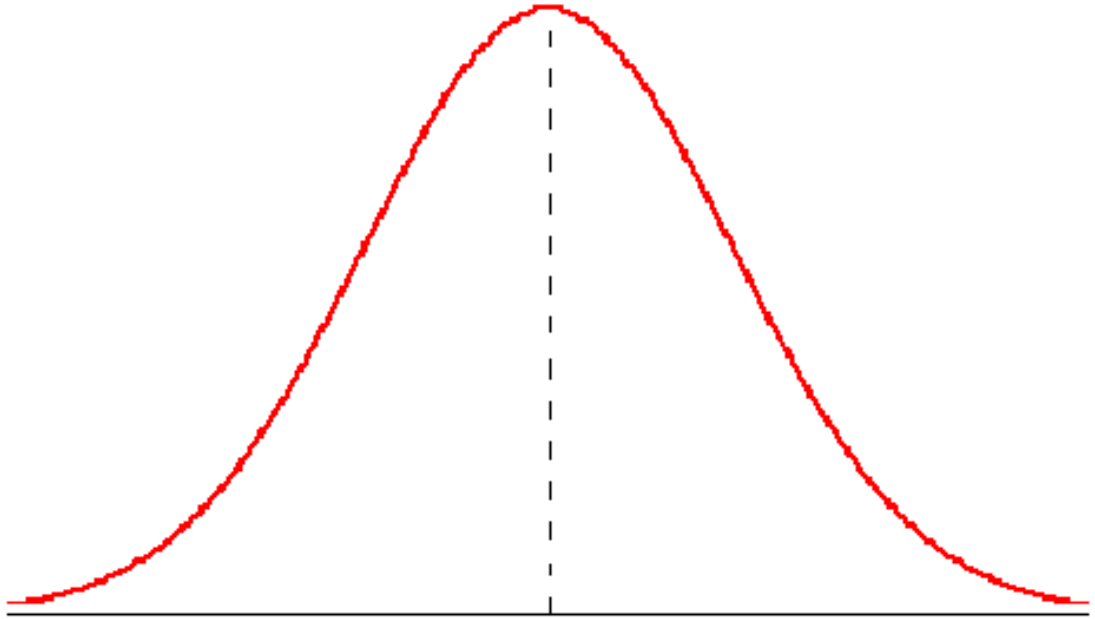
Intervalo: 4,45 a 9,55

$\bar{x} \pm s \rightarrow \cong 68\%$

Aproximadamente 68% dos valores do conjunto estão entre 4,45 kg e 9,55 kg.

μ = média

σ = desvio padrão



$\mu \pm \sigma \rightarrow 0{,}68$

$\mu \pm 2\sigma \rightarrow 0{,}95$

$\mu \pm 3\sigma \rightarrow 0{,}997$

Para a distribuição normal essas proporções são constantes

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
$\sum$	20	20	20
Média	4	4	4
Amplitude total	0	4	7
Variância	0	2,5	7,5
Desvio padrão	0	1,581	2,739

Exemplo: Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

$\overline{x} = 34{,}78$ reais

$s^2 = 471{,}14 \text{ reais}^2$

Calcule o desvio padrão desses dados.

$$s = \sqrt{s^2} = \sqrt{471{,}14 \text{ reais}^2} = 21{,}71 \text{ reais}$$

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 = peso (kg)

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

$\bar{x} = 7 \text{ kg}$

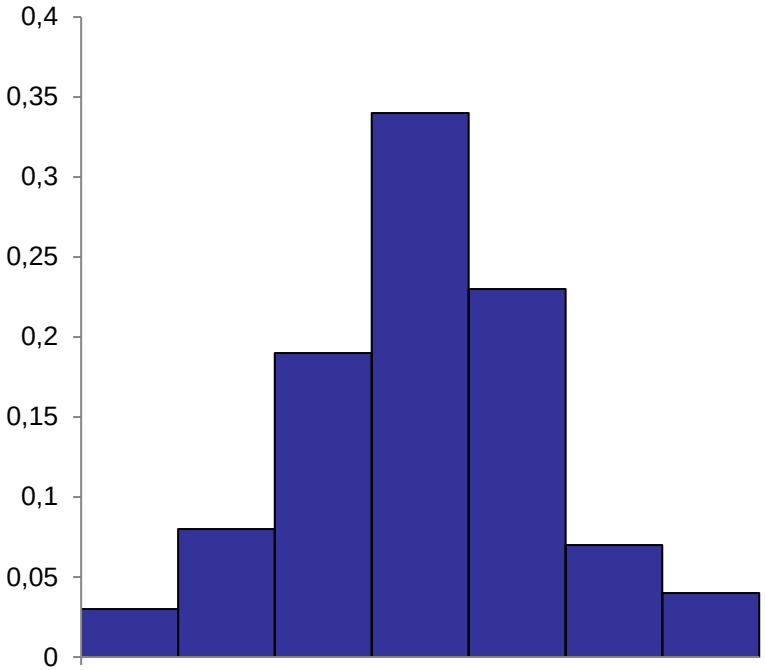
$s = 2{,}55 \text{ kg}$

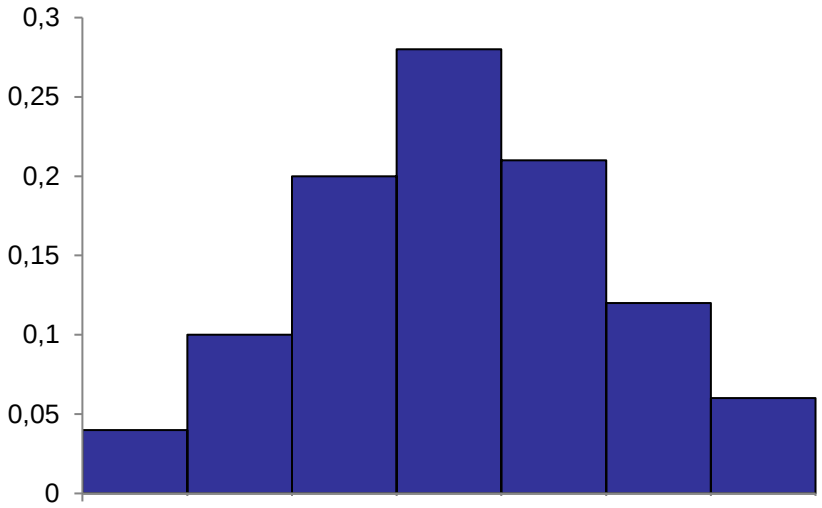
$$CV = \frac{2{,}55 \text{ kg}}{7 \text{ kg}} \; 100$$

$$CV = 36{,}4\%$$

⇨ É uma medida que somente tem sentido para variáveis mensuradas em **escala de razão** e cuja média não é zero nem está próxima de 0.

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





Exemplo 1:

Consideremos que x_{1i} e x_{2i} são conjuntos de valores referentes a produção diária de leite (em kg) de vacas das raças Jersey e Holandesa, para os quais foram obtidas as seguintes medidas:

Jersey (X_1): $\bar{x}_1 = 13$ kg

$s_1 = 3{,}4$ kg

Holandesa (X_2): $\bar{x}_2 = 25$ kg

$s_2 = 4{,}2$ kg

Qual grupo varia mais em relação à produção de leite?

$\bar{x}_1 = 13$ kg $\quad s_1 = 3{,}4$ kg $\quad\quad \bar{x}_2 = 25$ kg $\quad s_2 = 4{,}2$ kg

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

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

Quando as médias são diferentes, devemos usar o CV.

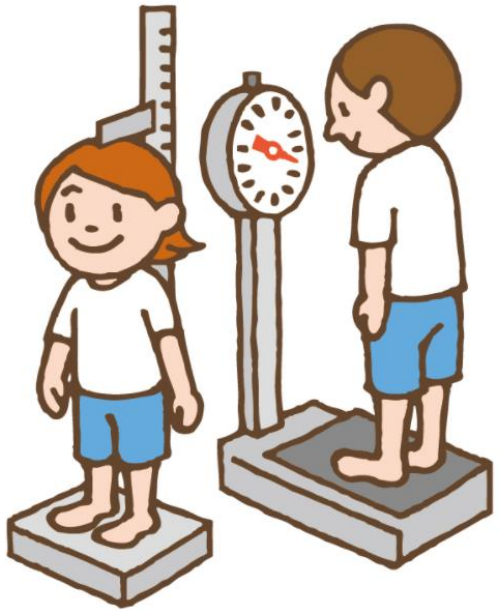
Profª. Clause Piana

Exemplo 2:

Consideremos, agora, que x_i e y_i são conjuntos de valores referentes a **alturas** (em cm) e **pesos** (em kg) de um grupo de estudantes, para os quais foram obtidas as seguintes medidas:

Altura (X): $\bar{x} = 158$ cm $\quad s_X = 12$ cm $\quad CV_X = 7{,}59\%$

Peso (Y): $\bar{y} = 52$ kg $\quad s_Y = 10$ kg $\quad CV_Y = 19{,}23\%$



Peso e altura não são grandezas comparáveis.

Quando as unidades de medida são diferentes, devemos usar o CV.

⇨ Para a comparação de variabilidades o uso do CV é particularmente recomendado em duas situações:

- **quando as médias dos conjuntos são diferentes**
- **quando as unidades de medida são diferentes**

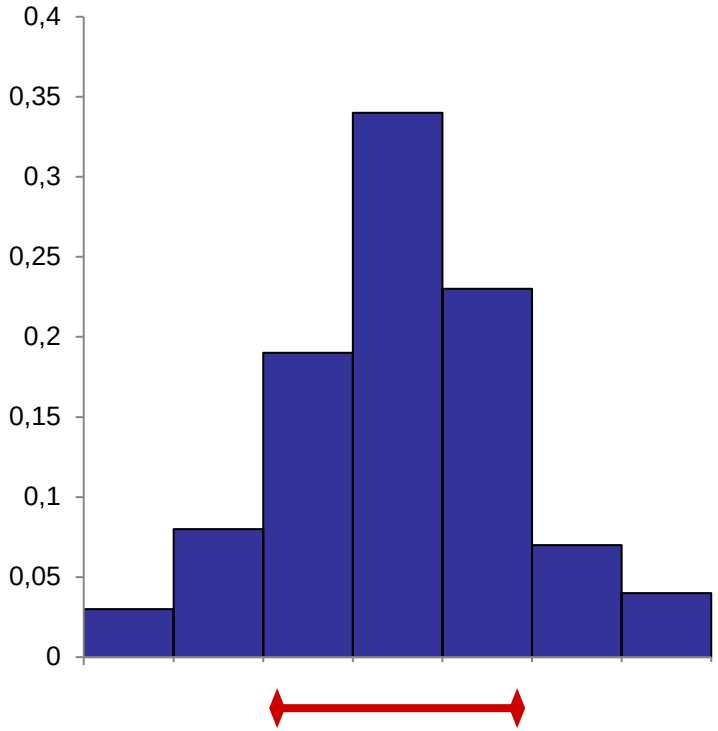
Se a média e a unidade de medida dos conjuntos são iguais, **qualquer medida de variação** pode ser utilizada para a comparação.

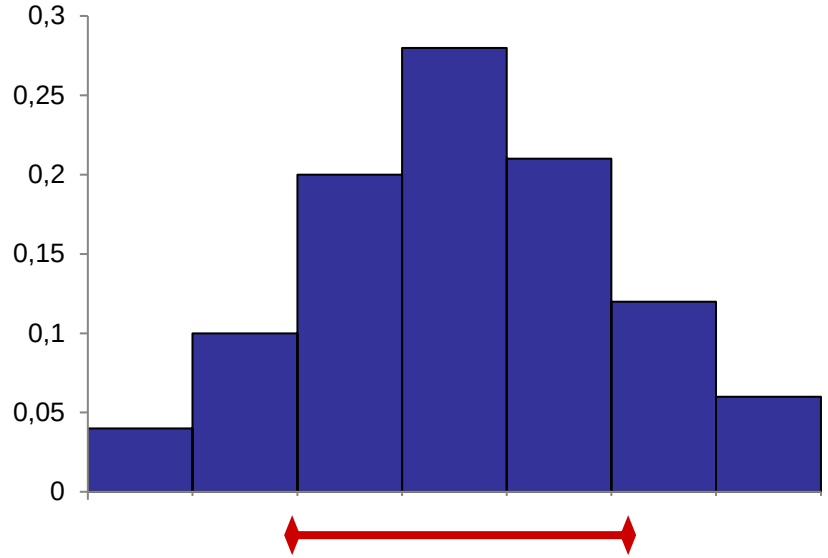
Vantagens:

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

Desvantagem

- O CV pode ter seu valor grandemente alterado enquanto a variabilidade permanece constante.





Em distribuições simétricas e unimodais observa-se que **quanto menor é o desvio padrão** em relação a média:

- ⇨ **menor é o intervalo para a mesma proporção de valores**
- ⇨ **menor é a amplitude do histograma**
- ⇨ **menor é o coeficiente de variação**

i	x_i (kg)	y_i (kg)	z_i (kg)
1	4	2	1
2	4	5	8
3	4	4	5
4	4	6	4
5	4	3	2
$\sum$	20	20	20
Média	4	4	4
Amplitude total	0	4	7
Variância	0	2,5	7,5
Desvio padrão	0	1,581	2,739
CV	0	39,53 %	68,46 %

Exemplo: Valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

$\overline{x} = 34,78$ reais

$s = 21,71$ reais

Calcule o coeficiente de variação desses dados.

$$CV = \frac{s}{\overline{x}}\ 100 = \frac{21,71}{34,78} \times 100 = 62,42\%$$

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 mediana, a moda e a média aritmética.

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

Bibliografia utilizada

SILVA, J. G. C. da. **Estatística Básica (versão preliminar)**. Universidade Federal de Pelotas.

Silveira Junior, P. ; Machado, A.A. ; Zonta, E.P.; Silva, J.B. da. **Curso de Estatística** v.1. Pelotas: Universidade Federal de Pelotas, 1992, 135p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br/topico.html