ESTATÍSTICA

Michelle Hanne Soares de Andrade
michellehanne.andrade@gmail.com

Estatística Descritiva

Desvio Padrão

- O conceito de variância é bastante rico, contudo, deve ser utilizado com cautela já que trata do problema original em escala quadrática.
- O desvio padrão surge como uma alternativa para corrigir este detalhe e assim facilitar a análise dos resultados.
- Tal medida e dada pela raiz quadrada da variância amostral:

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

Desvio Padrão

Em geral, a variância amostral possui propriedades matemáticas melhores, enquanto o desvio padrão oferece interpretações mais razoáveis.

Desvio Padrão

Desvio Padrão (S)

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

x_i são os valores da variável

média

Exemplo:

Calcular o desvio padrão

da população

representada por:

{ - 4 , -3 , -2 , 3 , 5}

Xi	$\overline{X}$	(Xi - $\overline{X}$)	(Xi - $\overline{X}$)^2
- 4	- 0,2	- 3,8	14,44
- 3	- 0,2	- 2,8	7,84
- 2	- 0,2	- 1,8	3,24
3	- 0,2	3,2	10,24
5	- 0,2	5,2	27,04
		$\overline{E}$ =	62,8

Sabemos que n = 5 e 62,8 / 5 = 12,56.

A raiz quadrada de 12,56 é o desvio padrão = **3,54**

Intervalo Interquartil

- Variância e desvio padrão também são sensíveis a valores discrepantes por considerar os valores observados diretamente.
- Uma maneira alternativa de contornar tal problema e considerar a amplitude interquartil:

$$A_{IQ} = Q_3 - Q_2.$$

- Tal quantidade indica a faixa de variação dos 50% centrais das observações.
- A escala original dos dados é preservada neste caso.

Intervalo Interquartil

- Calcular o Intervalo Interquartil da sequência: 12,6; 12,9; 13,4; 12,3; 13,6; 13,5; 12,6; 13,1.
- Ordenando os dados, obtemos: 12,3; 12,6; 12,6; 12,9; 13,1; 13,4; 13,5; 13,6; Logo:

$$Q_3 = \frac{13,4 + 13,5}{2} = 13,45 \qquad \text{e} \qquad Q_1 = \frac{12,6 + 12,6}{2} = 12,6$$

e,

$$A_{IQ} = 13,45 - 12,6 = 0,85.$$

Coeficiente de Variação

- Trata-se de uma medida relativa de dispersão útil para a comparação em termos relativos do grau de concentração. O coeficiente de variação e a relação entre o desvio padrão (S) e a média x.

$$CV = \frac{S}{\overline{x}}$$

Baixa dispersão: CV $\leq$ 15%

Média dispersão: 15%< CV<30%

Alta dispersão: CV $\geq$ 30%

Coeficiente de Variação

- Exemplo: como comparar as dispersões de alturas de pessoas com pesos destas mesmas pessoas?

	$\bar{x}$	s
ESTATURAS	175 cm	5,0 cm
PESOS	68 kg	2,0 kg

$$CV_E = \frac{5}{175} \times 100 = 0{,}0285 \times 100 = 2{,}85\%$$

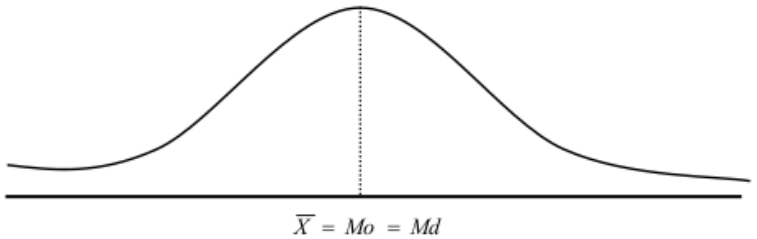
$$CV_P = \frac{2}{68} \times 100 = 0{,}0294 \times 100 = 2{,}94\%$$

Logo, nesse grupo de indivíduos, os pesos apresentam maior grau de dispersão que as estaturas.

Assimetria

- Estas medidas referem-se a forma da curva de uma distribuição de frequência, mais especificamente do polígono de frequência ou do histograma. Denomina-se assimetria o grau de afastamento de uma distribuição da unidade de simetria.
- Em uma distribuição simétrica, tem-se igualdade dos valores da média, mediana e moda.

Simetria

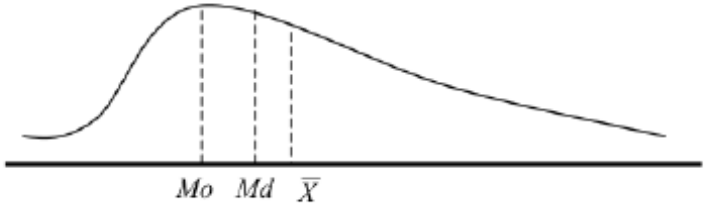


Assimetria

Assimetria

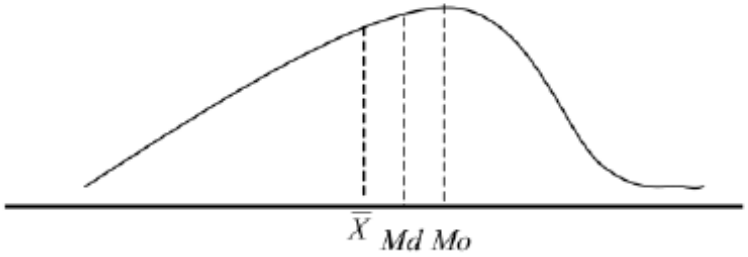
Assimetria à direita (ou positiva)

$Mo < Md < \overline{X}$



Assimetria à esquerda (ou negativa)

$\overline{X} < Md < Mo$



Assimetria

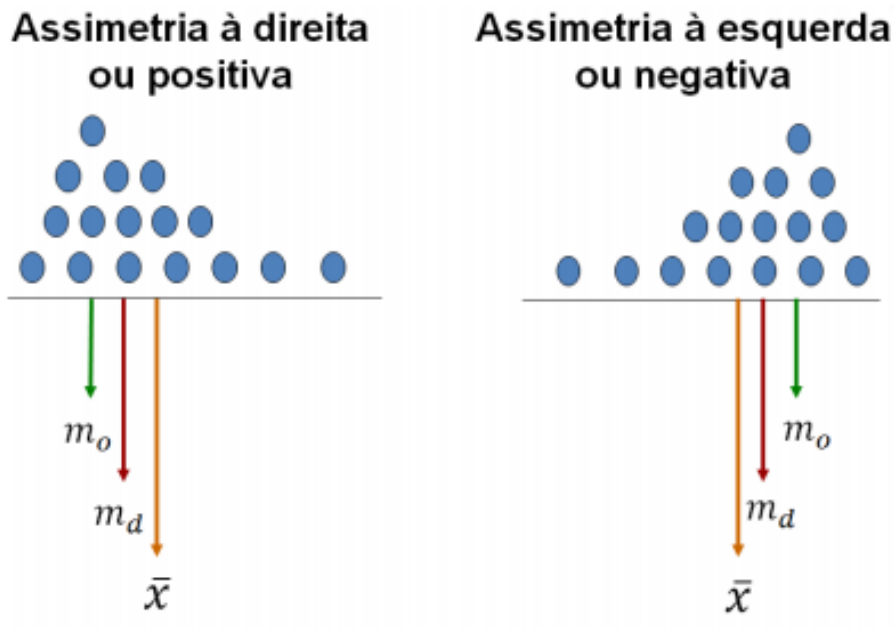


Figura 4: Posição relativa de medidas de tendência central sob assimetria dos dados

Assimetria

- Existem varias fórmulas para o cálculo do coeficiente de assimetria. As mais utilizadas são:
  - 1º Coeficiente de Pearson

$$AS = \frac{\bar{x} - Mo}{S}$$

M_o: valor modal (moda)
S : Desvio padrão
$\bar{X}$: Média

  - 2º coeficiente de Pearson

$$AS = \frac{Q_1 + Q_3 - 2Md}{Q_3 - Q_1}$$

Q_1 : valor do 1º Quartil
Q_3 : valor do 3º Quartil
M_d : valor da Mediana

Quando:
AS = 0, diz-se que a distribuição é simétrica.
AS > 0, diz-se que a distribuição é assimétrica positiva (à direita)
AS < 0, diz-se que a distribuição é assimétrica negativa (à esquerda)

Coeficiente de Assimetria de Pearson

$$A_P = \frac{\bar{x} - m_o}{S}.$$

Temos

a) Distribuições simétricas unimodais: $\bar{x} = m_d = m_o$; nesse caso, $A_P = 0$
b) Distribuições assimétricas positivas: $\bar{x} > m_d > m_o$; então $A_P > 0$
c) Distribuições assimétricas negativas: $\bar{x} < m_d < m_o$, fazendo com que $A_P < 0$.

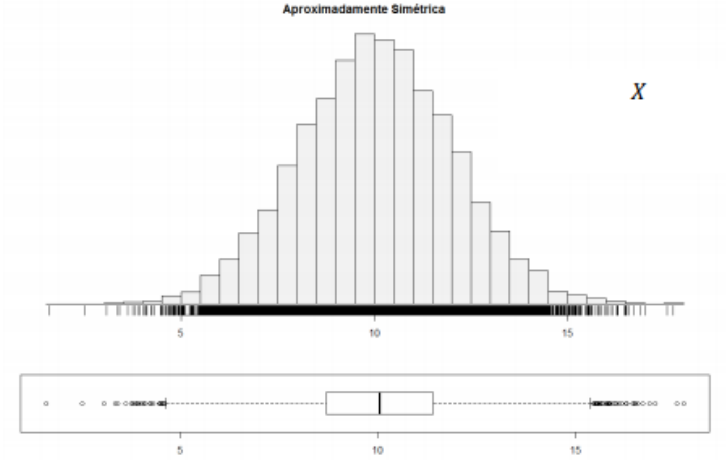
- O fato do denominador de ser o desvio-padrão faz com que essa medida seja adimensional, o que permite sua comparação mesmo quando se trabalha com dados em diferentes escalas (por exemplo, preços em reais ou em dólares).

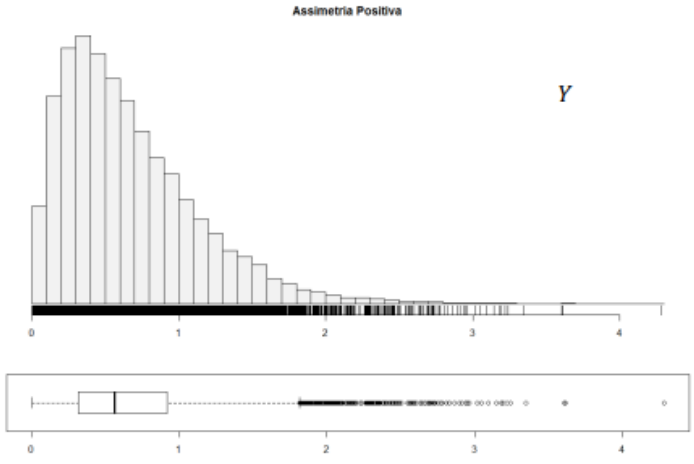
Coeficiente de Assimetria de Pearson

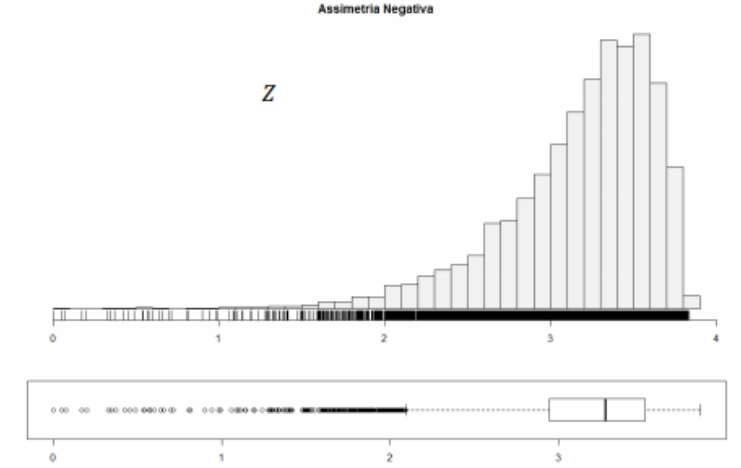
- A determinação da moda para dados contínuos não é trivial. Pode-se ter uma amostra de 1000 valores diferentes, por exemplo. **Isso requer o uso de algoritmos que levam a diferentes estimativas dessa medida**. Uma alternativa é utilizar o coeficiente

$$A_{P2} = \frac{\bar{x} - m_d}{S}.$$

Coeficiente de Assimetria de Pearson - Exemplo







Coeficiente de Assimetria de Pearson - Exemplo

Tabela 1: Estatísticas descritivas para os dados representados na Figura 3.

Estatística	X	Y	Z
Mínimo	1,592	0,004	0,000
Q_1: primeiro quartil	8,691	0,319	2,946
m_d: mediana	10,050	0,567	3,278
Q_3: terceiro quartil	11,400	0,918	3,514
Máximo	17,740	4,281	3,838
Média	10,050	0,675	3,172
S: desvio-padrão	1,993	0,477	0,469
m_o: moda[2]	10,040	0,294	3,386

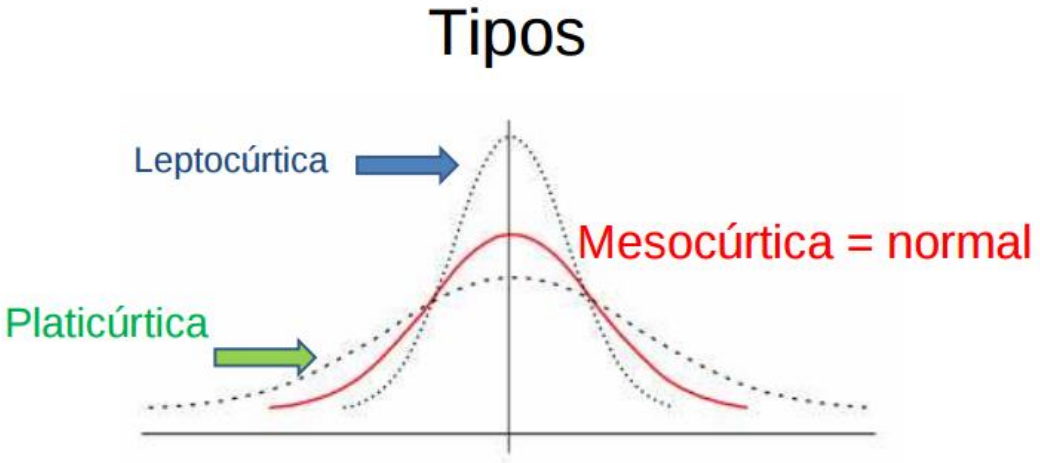
Tamanho das amostras = 10.000

Variável	A_P	A_{P2}
X	0,005	0,000
Y	0,799	0,226
Z	-0,456	-0,226

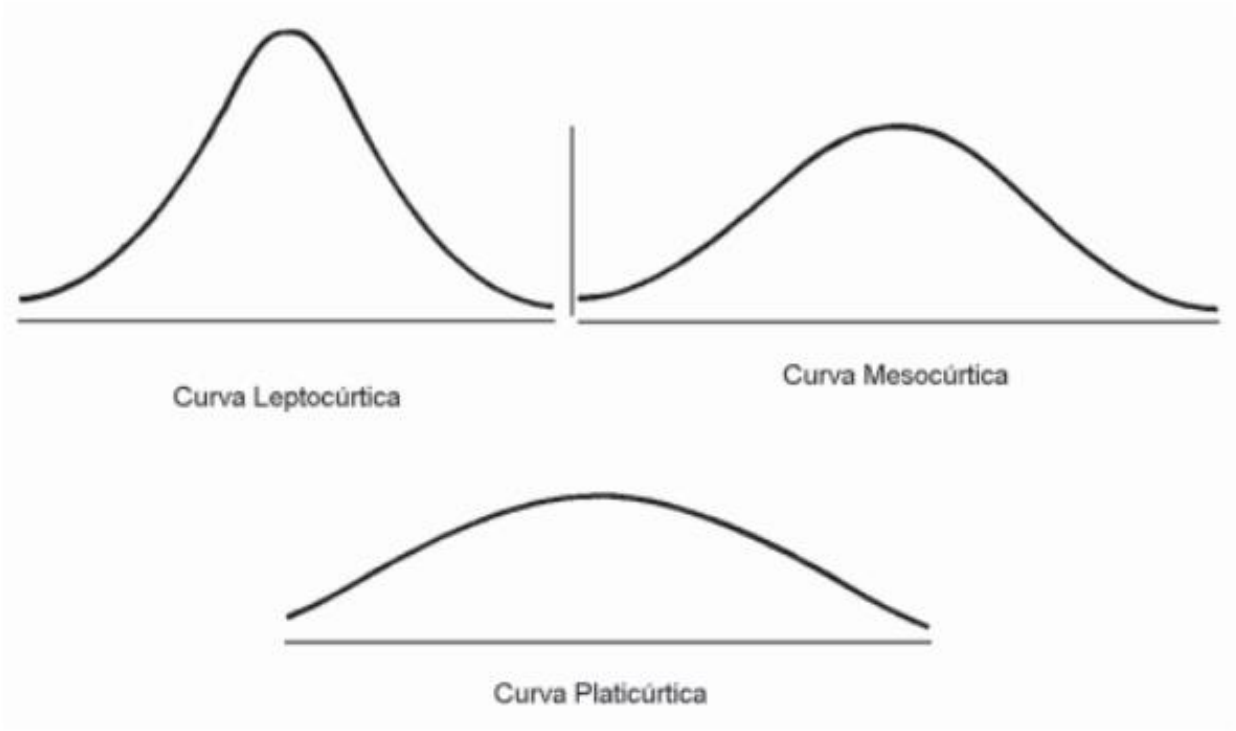
- Na Tabela 2, estão apresentados os coeficientes de Assimetria. Há indícios de assimetria fraca (quase simetria) para a variável X, assimetria positiva para Y e negativa para Z.

Curtose

- Curtose e o grau de achatamento (ou afilamento) de uma distribuição em comparação com uma distribuição padrão (chamada curva normal).
- De acordo com o grau de curtose, classificamos três tipos de curvas de frequência:



Curtose


Curva Leptocúrtica
Curva Mesocúrtica
Curva Platicúrtica

Curtose

- Para medir o grau de curtose utiliza-se o coeficiente:

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

Q_1 : valor do 1º Quartil
Q_3 : valor do 3º Quartil
P_{10} : valor do percentil 10
P_{90} : valor do percentil 90

- Se K= 0,263, diz-se que a curva correspondente à distribuição de frequência é mesocúrtica.
- Se K > 0,263, diz-se que a curva correspondente à distribuição de frequência é platicúrtica.
- Se K < 0,263, diz-se que a curva correspondente à distribuição de frequência é leptocúrtica.

Curtose - Exemplo

- Conclua a respeito do tipo de curva da distribuição da Tabela abaixo quanto à curtose

Tabela 6.2 – Distribuição de dados de uma variável “B” de determinada época

Intervalo de classe	Frequência	F_i
3 |— 8	5	5
8 |— 13	15	20
13 |— 18	20	40
18 |— 23	10	50

Fonte: Dados fictícios, apenas para fins ilustrativos.

Curtose - Exemplo

Para o cálculo do Coeficiente K, necessita-se calcular Q_1, Q_3, C_{10} e C_{90}:

$$Q_1 = li_k + A_k \frac{\frac{n}{4} - F_{k-1}}{f_k}$$

$$Q_1 = 8 + 5 \frac{12,5 - 5}{15} = 8 + 2,5 = 10,5$$

$$Q_3 = li_k + A_k \frac{\frac{3n}{4} - F_{k-1}}{f_k}$$

$$Q_3 = 13 + 5 \frac{37,5 - 20}{20} = 13 + 4,38 = 17,38$$

Curtose - Exemplo

$$C_{10} = li_k + A_k \frac{\frac{10n}{100} - F_{k-1}}{f_k}$$

$$C_{10} = 3 + 5\,\frac{5-0}{5} = 3 + 5 = 8$$

$$C_{90} = li_k + A_k \frac{\frac{90n}{100} - F_{k-1}}{f_k}$$

$$C_{90} = 18 + 5\,\frac{45-40}{10} = 18 + 2{,}5 = 20{,}5$$

$$K = \frac{Q_3 - Q_i}{2\left(C_{90} - C_{10}\right)} = \frac{17{,}38 - 10{,}5}{2\left(20{,}5 - 8\right)} = \frac{6{,}88}{25} = 0{,}27$$

Resposta:

Como **K > 0,263**, logo, a curva correspondente é suavemente **platicúrtica.**

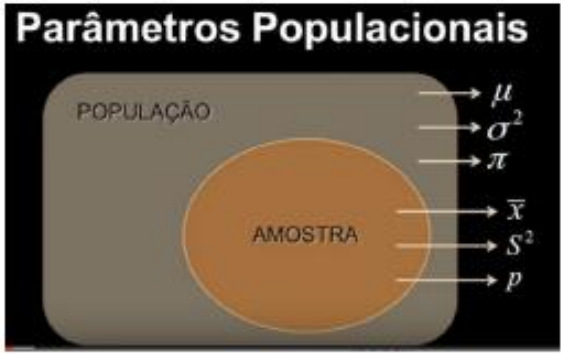
Resumo – População e Amostra

- **População e Amostra:** Ao examinar um grupo qualquer, considerando todos os seus elementos, estamos tratando da **população** ou **universo**. Nem sempre isso e possível. Nesse caso, examinamos uma pequena parte chamada **amostra**.
- Uma **população** pode ser finita ou infinita. Por exemplo:
  - a população dos alunos de sua escola é finita e a população constituída de todos os resultados (cara ou coroa) em sucessivos lances de uma moeda e infinita.
- Se uma **amostra** é representativa de uma população, podemos obter conclusões importantes sobre a população.

Resumo

Parâmetro (População)	
Valor médio	μ
Desvio padrão	σ
Proporção	p
Correlação	ρ

Estatística (Amostra)	
Média	$\bar{x}$
Desvio padrão	s
Proporção	$\hat{p}$
Correlação	r



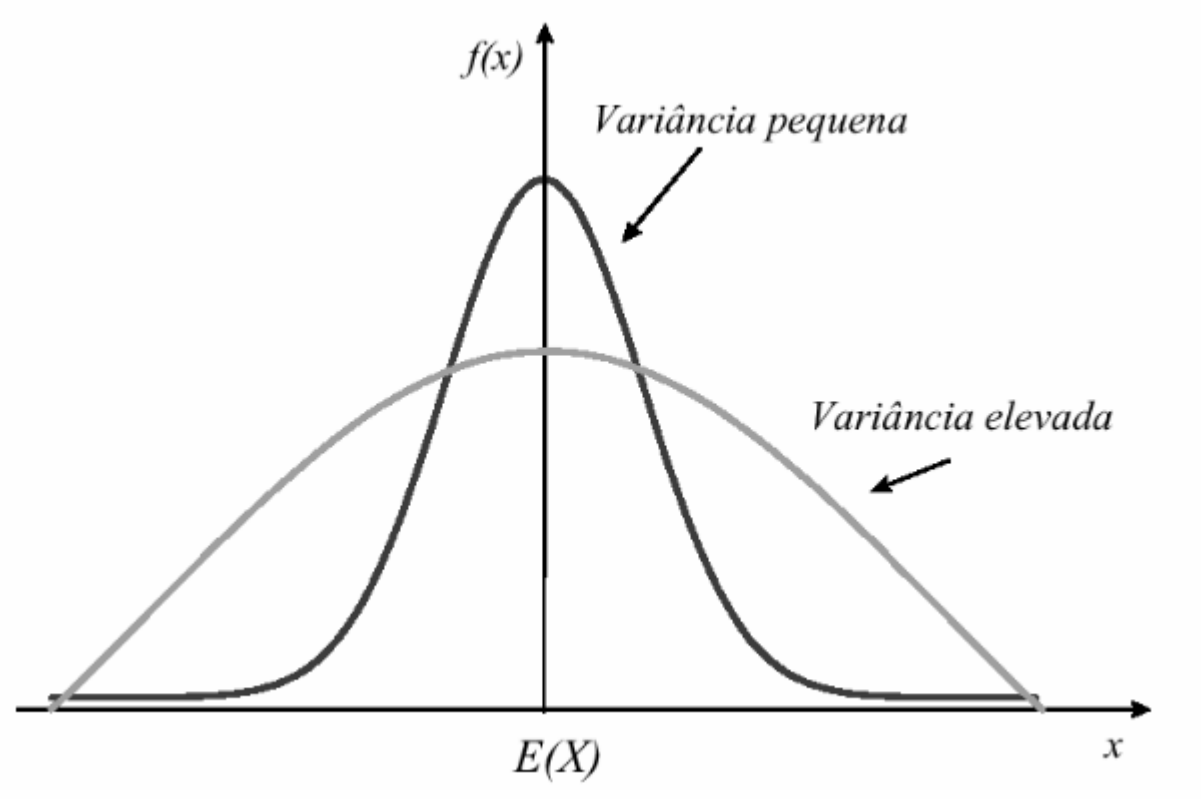
Resumo - Variância (S^2)

- **Variância (S^2)**
  - Sendo a variância calculada a partir dos quadrados dos desvios, ela é um número em unidade quadrada em relação a variável em questão, o que, sob o ponto de vista prático é um inconveniente; por isso, tem pouca utilidade na estatística descritiva, **mas é extremamente importante na inferência estatística e em combinações de amostras.**

Resumo - Desvio Padrão (S) x Variância (S^2)

- **Desvio Padrão (S) x Variância (S^2)**
  - O desvio padrão é a medida mais usada na comparação de diferenças entre conjuntos de dados, por ter grande precisão. O desvio padrão determina a dispersão dos valores em relação à média e é calculado por meio da raiz quadrada da variância.

Resumo – Variância (S^2)



Resumo – Desvio Padrão

- **O desvio padrão é uma medida que só pode assumir valores não negativos e quanto maior for, maior será a dispersão dos dados.**
- Algumas propriedades do desvio padrão, que resultam imediatamente da definição, são:
  - O desvio padrão é sempre não negativo e será tanto maior, quanta maior a variabilidade entre os dados.
- **Se S = 0, então não existe variabilidade, isto é, os dados são todos iguais.**

Resumo – Exercício

- Tendo por base uma amostra da altura de uma parcela da população apresentada na Tabela 5.2, determinar:

  a) A variância das alturas;

  b) O desvio-padrão das alturas.

Tabela 5.2 – Estatura de uma amostra de uma população A

Altura (cm)	Nº de pessoas
150 |— 158	5
158 |— 166	18
166 |— 174	42
174 |— 182	27
182 |— 190	8
∑	100

Fonte: Dados fictícios, apenas para fins ilustrativos.

Resumo – Exercício - Solução

Solução

a) A variância das alturas

Usando a fórmula 5.13 obtém-se o seguinte resultado:

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

Para calcular a variância, necessita-se conhecer as informações a seguir, cujos valores estão calculados na Tabela 5.3:

$\sum X_i f_i$

$\sum X_i^2 f_i$

Resumo – Exercício - Solução

Solução

a) A variância das alturas

Usando a fórmula 5.13 obtém-se o seguinte resultado:

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

Para calcular a variância, necessita-se conhecer as informações a seguir, cujos valores estão calculados na Tabela 5.3:

$\sum X_i f_i$

$\sum X_i^2 f_i$

Tabela 5.3 – Tabela auxiliar da Tabela 5.2

Altura (cm)	Nº de pessoas	X_i	$X_i f_i$	$X_{i2} f_i$
150 |— 158	5	154	770	118.580
158 |— 166	18	162	2916	472.392
166 |— 174	42	170	7140	1.213.800
174 |— 182	27	178	4806	855.468
182 |— 190	8	186	1.488	276.768
∑	100		17.120	2.937.008

Fonte: Dados fictícios, apenas para fins ilustrativos.

Resumo – Exercício - Solução

Substituindo os valores na fórmula 5.13 obtém-se os seguinte resultados:

$$s^2 = \frac{2937008 - \frac{17120^2}{100}}{100 - 1}$$

$$s^2 = \frac{2937008 - 2930944}{100 - 1} = \frac{6064}{99}$$

$s^2 = 61{,}25$

Resposta:

A variância das alturas é de 61,25 cm²

b) O desvio-padrão das alturas

Solução

Extrai-se a raiz quadrada da variância.

Assim,

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

$$s = \sqrt{61{,}25}$$

$s = 7{,}8263$

Resposta: As estaturas das pessoas estão dispersas em média 7,83 cm em relação à média da distribuição.