Medidas de variação

Definição

A **amplitude** de um conjunto de dados é a diferença entre as entradas máximas e mínimas no conjunto. Para encontrar a amplitude, os dados devem ser quantitativos.

Amplitude = (Entrada máxima de dados) − (Entrada mínima de dados)

Exemplo 1

Encontrando a amplitude de um conjunto de dados

Duas corporações contrataram 10 formandos cada. O salário inicial para cada formando é mostrado a seguir. Encontre a amplitude dos salários iniciais para a Empresa A.

Salários iniciais para a Empresa A (milhares de dólares)

Salários	41	38	39	45	47	41	44	41	37	42

Salários iniciais para a Empresa B (milhares de dólares)

Salários	40	23	41	50	49	32	41	29	52	58

Solução

Ordenar os dados ajuda a encontrar os salários mínimos e máximos.

$$\underset{\text{Mínimo}}{37} \quad 38 \quad 39 \quad 41 \quad 41 \quad 41 \quad 42 \quad 44 \quad 45 \quad \underset{\text{Máximo}}{47}$$

$$\begin{aligned} \text{Amplitude} &= (\text{Salário máximo}) - (\text{Salário mínimo}) \\ &= 47 - 37 \\ &= 10. \end{aligned}$$

Então, a amplitude dos salários iniciais para a Empresa A é 10 ou \$ 10.000.

Desvio, variância e desvio padrão

Como uma medida de variação, a amplitude tem como vantagem ser fácil de calcular. A desvantagem, entretanto, é que a amplitude usa somente duas entradas do conjunto de dados. Duas medidas de variação que usam todas as entradas do conjunto de dados são a variância e o desvio padrão. Porém, antes de aprendermos essas medidas, precisamos entender o que chamamos desvio padrão e entrada no conjunto de dados.

Definição

O **desvio** de uma entrada x em uma população é a diferença entre a entrada e a média μ do conjunto de dados.

$$\text{Desvio de } x = x - \mu.$$

Exemplo 2

Encontrando os desvios de um conjunto de dados

Encontre o desvio de cada salário inicial para a Empresa A dado no Exemplo 1.

Solução

A média dos salários iniciais é $\mu = 415/10 = 41{,}5$. Para encontrar o quanto um salário desvia da média, subtraia 41,5 do salário. Por exemplo, o desvio de 41 (ou \$ 41.000) é:

$$\underset{\uparrow\, x}{41} - \underset{\uparrow\, \mu}{41{,}5} = 0{,}5 \text{ (ou } -\$\ 500). \qquad \textit{Desvio de } x = x - \mu$$

Salário, (milhares de dólares) x	Desvio (milhares de dólares) $x - \mu$
41	–0,5
38	–3,5
39	–2,5
45	3,5
47	5,5
41	–0,5
44	2,5
41	–0,5
37	–4,5
42	0,5
$\Sigma x = 415$	$\Sigma(x - \mu) = 0$

No Exemplo 2, note que a soma dos desvios é zero. Em razão de isso ser verdadeiro para qualquer conjunto de dados, não faz sentido encontrar a média dos desvios. Para superar esse problema, você pode fazer o quadrado de cada desvio. Quando adicionamos os quadrados dos desvios, calculamos a quantidade chamada **soma dos quadrados**, denotada para SS_x. Em um conjunto de dados populacional, a média dos quadrados dos desvios é chamada de **variância populacional**.

Definição

A **variância populacional** do conjunto de dados populacional de N entradas é:

$$\text{Variância populacional} = \sigma^2 = \frac{\sum(x-\mu)^2}{N}.$$

Definição

O **desvio padrão populacional** de um conjunto de dados populacional de N entradas é a raiz quadrada da variância populacional.

$$\text{Desvio padrão populacional} = \sigma^2 = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Encontrando a variância populacional e o desvio padrão

Em palavras

1. Encontre a média do conjunto de dados populacional.
2. Encontre o desvio de cada entrada.
3. Faça o quadrado de cada desvio.
4. Adicione para obter a **soma dos quadrados.**
5. Divida por N para obter a **variação populacional.**
6. Encontre a raiz quadrada da variância para obter **o desvio padrão populacional.**

Em símbolos

1. $\mu = \frac{\sum x}{N}$
2. $x - \mu$
3. $(x - \mu)^2$
4. $SS_x = \Sigma (x - \mu)^2$
5. $\sigma^2 = \frac{\sum (x-\mu)^2}{N}$
6. $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$

Encontrando o desvio padrão populacional

Encontre a variância populacional e o desvio padrão dos salários iniciais para a Empresa A dados no Exemplo 1.

Soma dos quadrados dos salários iniciais da Empresa A

Salário x	Desvio $x-\mu$	Quadrados $(x-\mu)^2$
41	–0,5	0,25
38	–3,5	12,25
39	–2,5	6,25
45	3,5	12,25
47	5,5	30,25
41	–0,5	0,25
44	2,5	6,25
41	–0,5	0,25
37	–4,5	20,25
42	0,5	0,25
	$\Sigma=0$	$SS_x=88,5$

$$SS_x = 88,5, \qquad N = 10,$$

$$\sigma^2 = \frac{88,5}{10} \approx 8,9, \qquad \sigma = \sqrt{8,85} \approx 3,0$$

3000 dólares

Definição

A **variância amostral** e o **desvio padrão amostral** de cada conjunto de dados amostral de n entradas estão listados a seguir.

$$\text{Variância da amostra} = s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

$$\text{Desvio padrão amostral} = s = \sqrt{s^2} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Dica de estudo

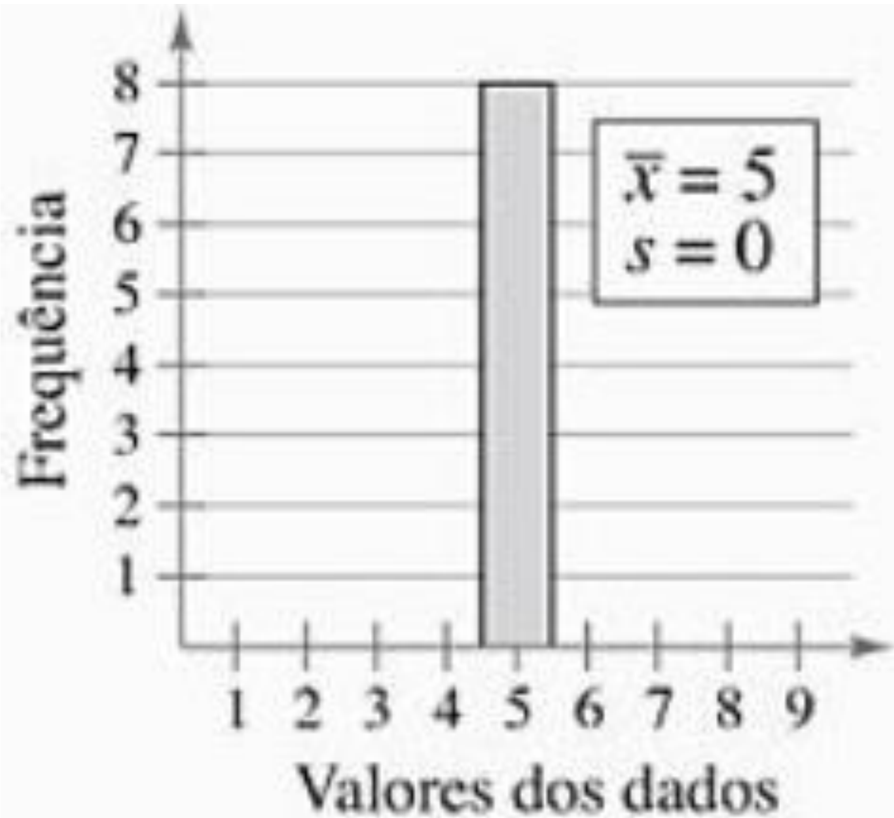
Note que quando encontramos a variância *populacional*, dividimos por N o número de entradas, mas por razões técnicas, quando encontramos a variância *amostral* dividimos por n − 1, um número a menos que o número de entradas.

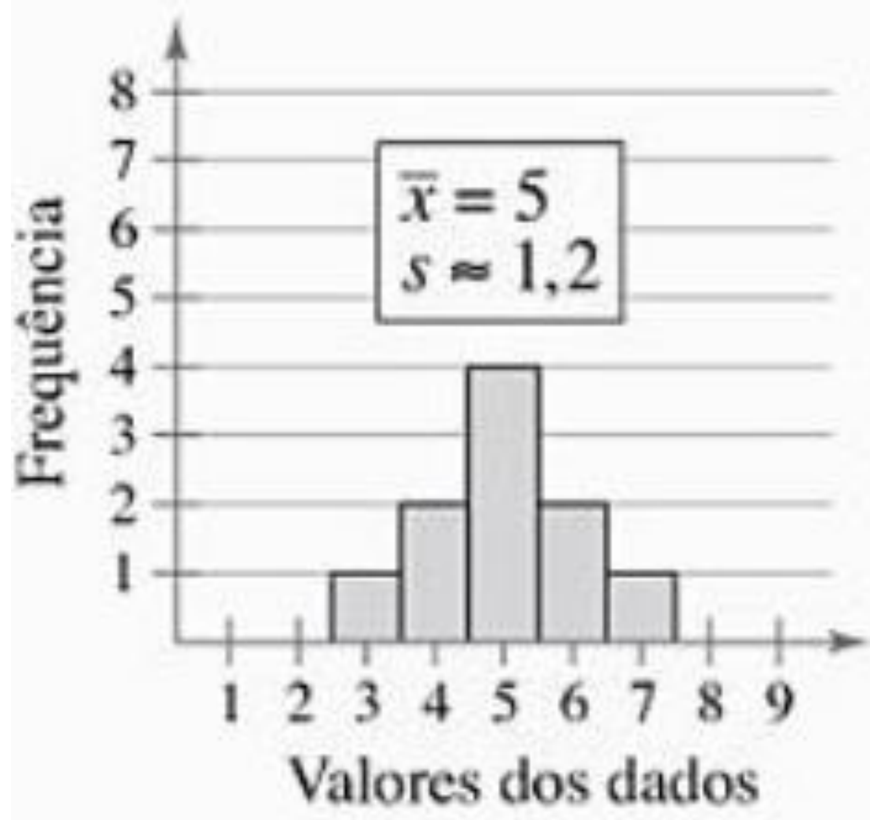
Símbolos nas fórmulas da variância e do desvio padrão

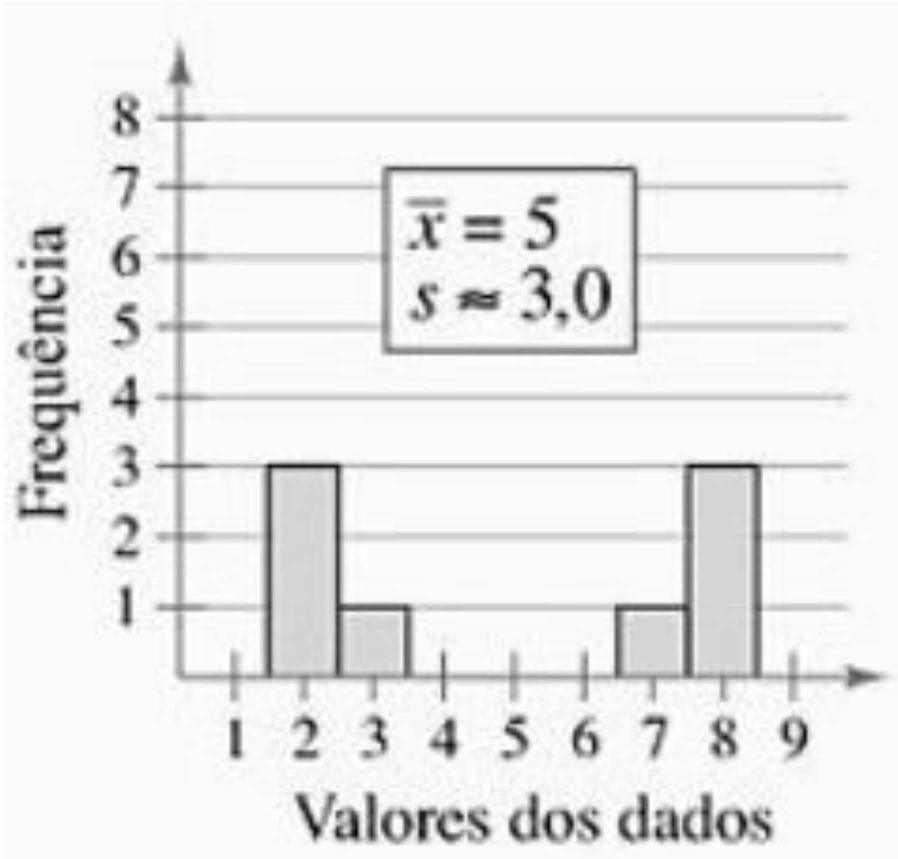
	População	Amostra
Variância	σ^2	s^2
Desvio padrão	σ	s
Média	μ	$\bar{x}$
Número de entradas	N	n
Desvio	$x - \mu$	$x - \bar{x}$
Soma dos quadrados	$\Sigma(x - \mu)^2$	$\Sigma(x - \bar{x})^2$

Interpretando o desvio padrão

Na interpretação do desvio padrão, lembre-se de que ele é a medida de quanto uma entrada típica se desvia da média. Quanto mais espalhadas estiverem as entradas, maior será o desvio padrão.



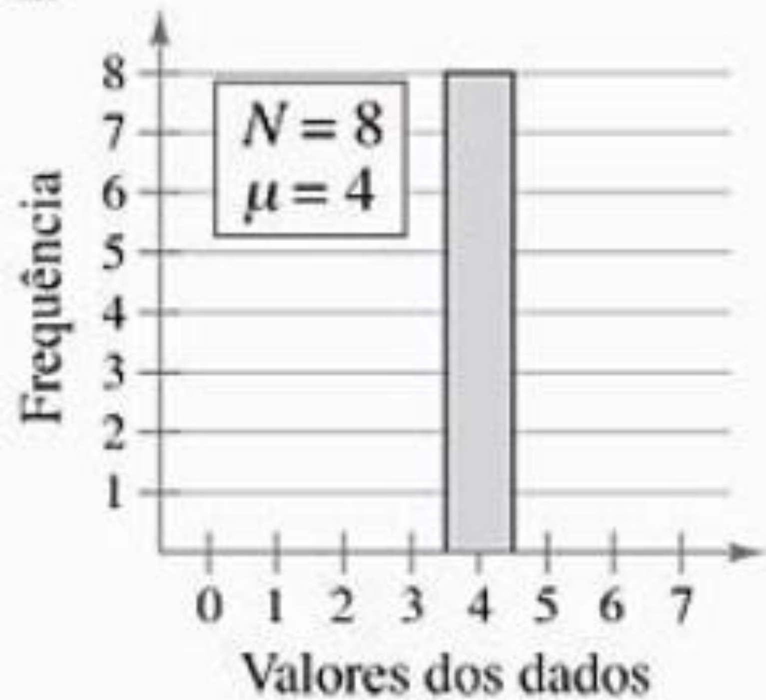



$\bar{x} = 5$
$s \approx 3,0$
Frequência
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9
Valores dos dados

Estimando o desvio padrão

Sem calcular, estime o desvio padrão populacional de cada conjunto de dados.

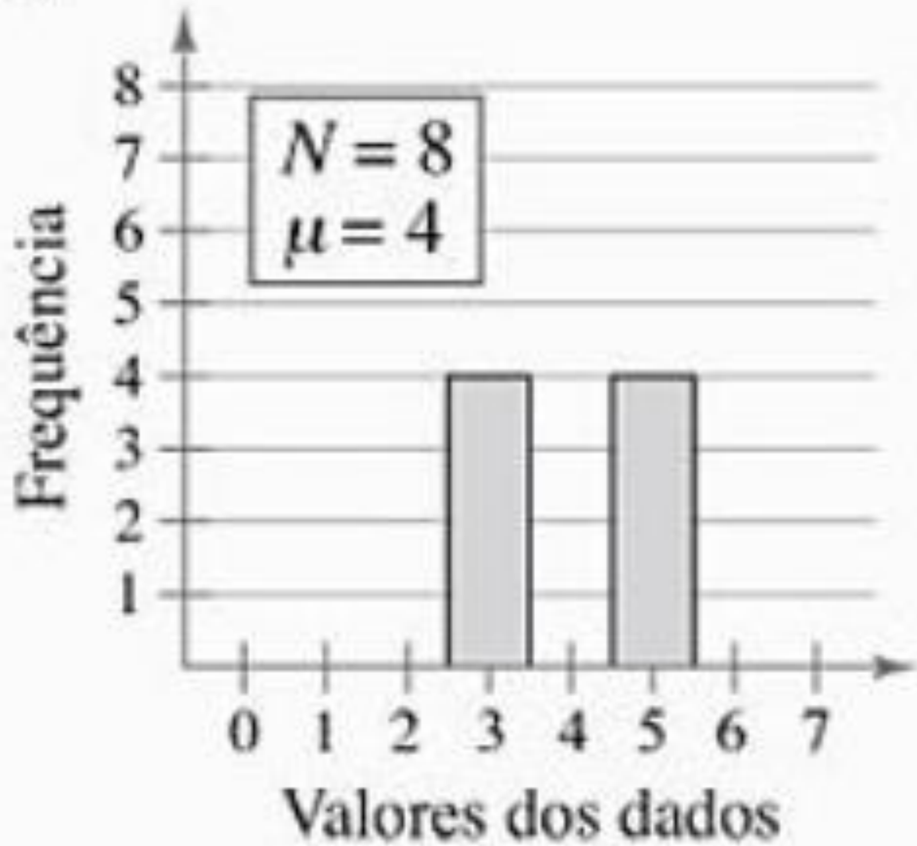
1.



Solução

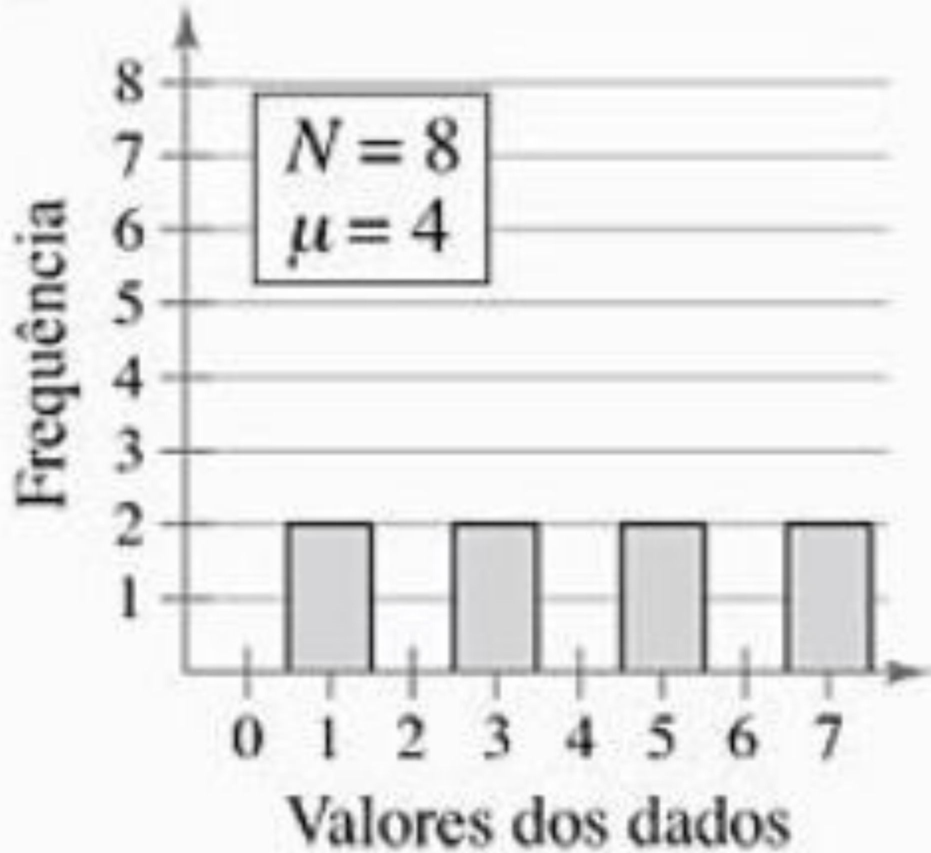
1. Cada uma das oito entradas é 4. Então, cada desvio é 0, o que implica $\sigma = 0$.

2.



2. Cada uma das oito entradas tem desvio de ± 1. Então, o desvio padrão populacional deveria ser 1. Calculando, você pode ver que $\sigma = 1$

3.



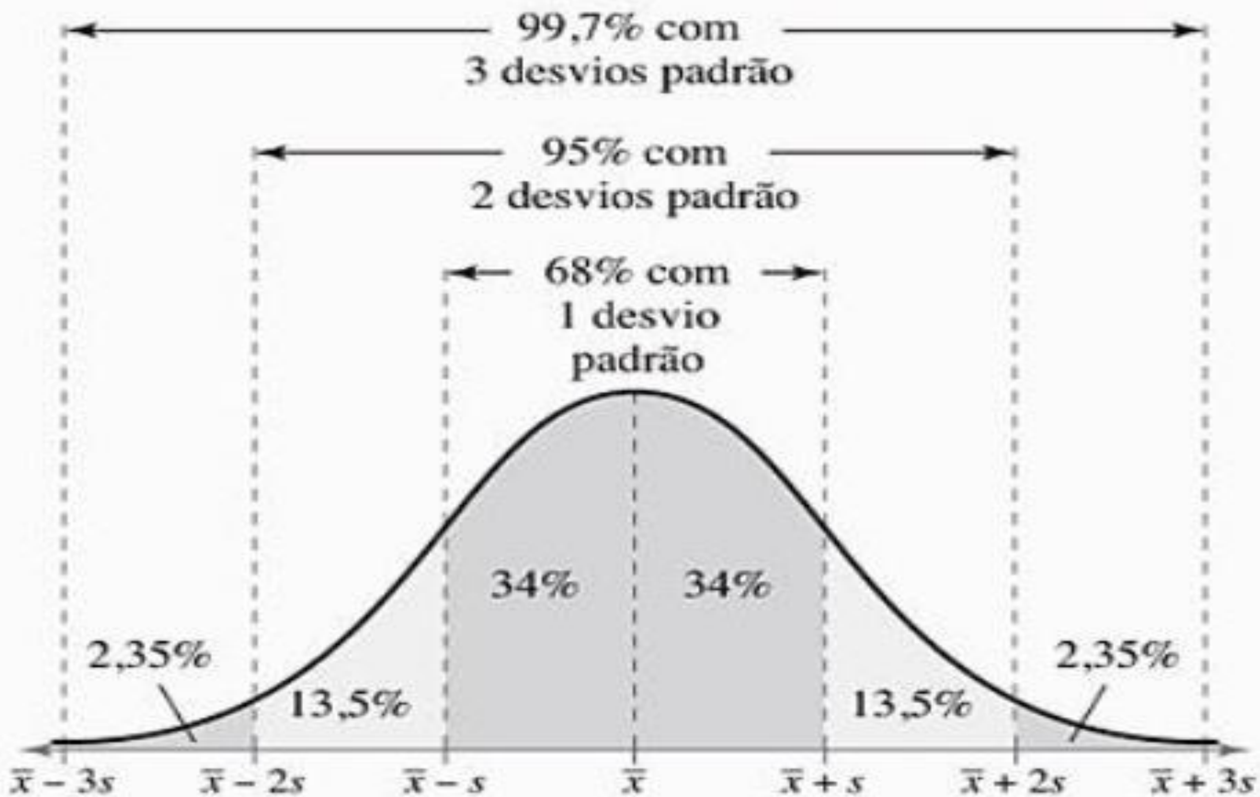
3. Cada uma das oito entradas tem desvio de ± 1 ou ± 3. Então, o desvio padrão populacional deveria ser 2. Calculando, você pode ver que $\sigma \approx 2{,}24$.

Regra empírica (ou regra 68-95-99,7)

Para os dados com distribuição (simétrica) com formato de curva, o desvio padrão tem as características a seguir.

1. Em torno de 68% dos dados está dentro de um desvio padrão em relação à média.
2. Em torno de 95% dos dados está dentro de dois desvios padrão em relação à média.
3. Em torno de 99,7% dos dados está dentro de três desvios padrão em relação à média.

Distribuição em forma de sino



Exemplo 7

Usando a regra empírica

Em uma pesquisa conduzida pelo National Center for Health Statistics (Centro Nacional de Estatísticas Médicas), uma amostra das alturas médias das mulheres nos EUA (idades entre 20 e 29) era de 64 polegadas, com desvio padrão amostral de 2,71 polegadas. Estime a porcentagem de mulheres cujas alturas estão entre 64 e 69,42 polegadas.

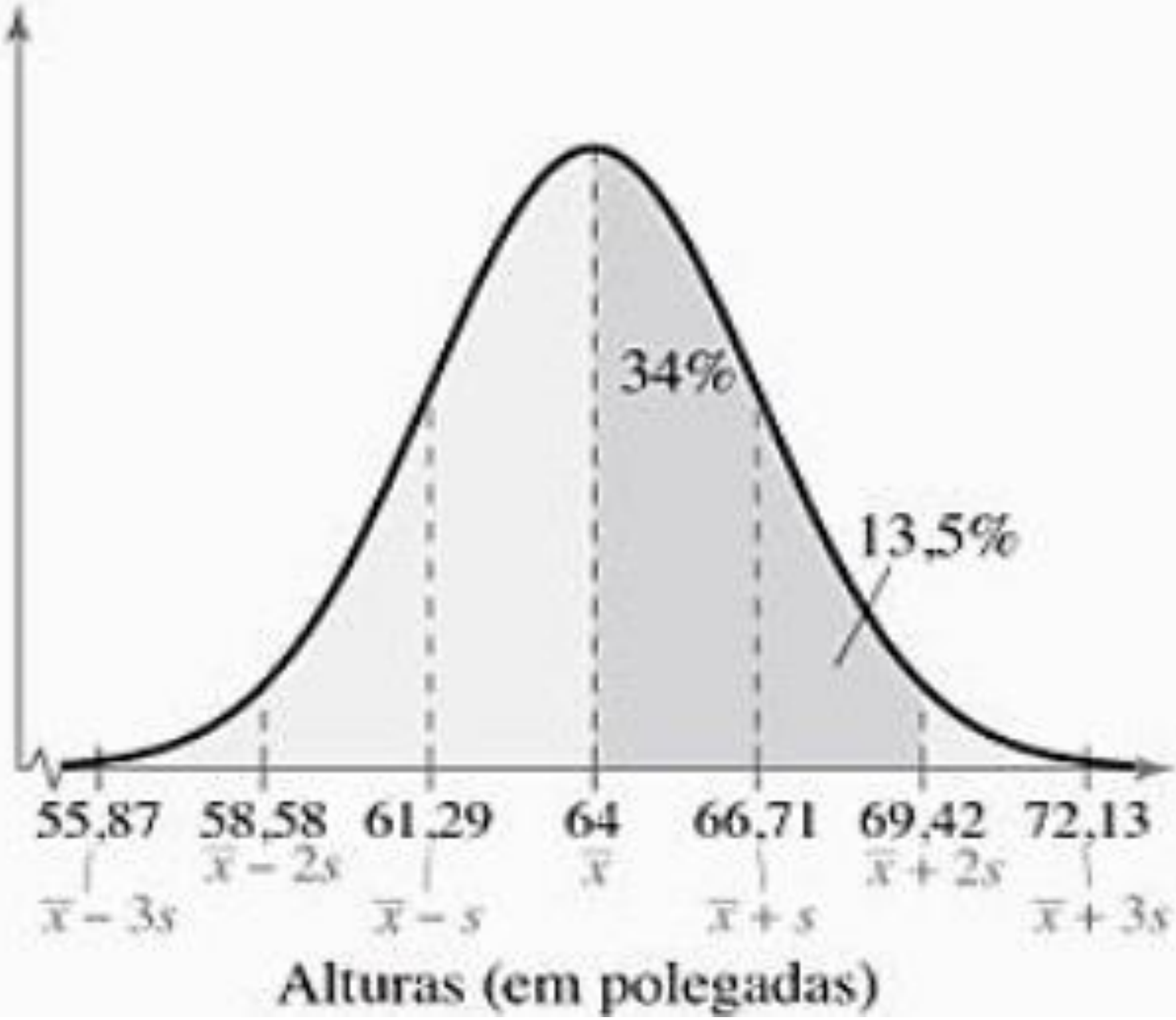
Solução

A distribuição das alturas das mulheres é mostrada no gráfico. Em razão de a distribuição ter formato de sino, você pode usar a regra empírica. A altura média é 64, então, quando você adicionar dois desvios padrão à altura média, você obtém:

$$\bar{x} + 2s = 64 + 2\,(2{,}71) = 69{,}42.$$

Já que 62,42 é dois desvios padrão acima da altura média, a porcentagem das alturas entre 64 polegadas e 69,42 polegadas são 34% + 13,5% = 47,5%.

Alturas das mulheres nos EUA com idades entre 20 e 29



A regra empírica se aplica somente às distribuições em forma de sino (simétricas). Mas e se a distribuição não for em forma de sino ou se a forma da distribuição for desconhecida? O teorema a seguir fornece uma afirmação de desigualdade que se aplica a todas as distribuições. Seu nome é em homenagem ao estatístico russo Pafnuti Chebychev (1821–1894).

Teorema de Chebychev

A porção de qualquer conjunto de dados que estejam dentro de k desvios padrão ($k > 1$) da média é, pelo menos:

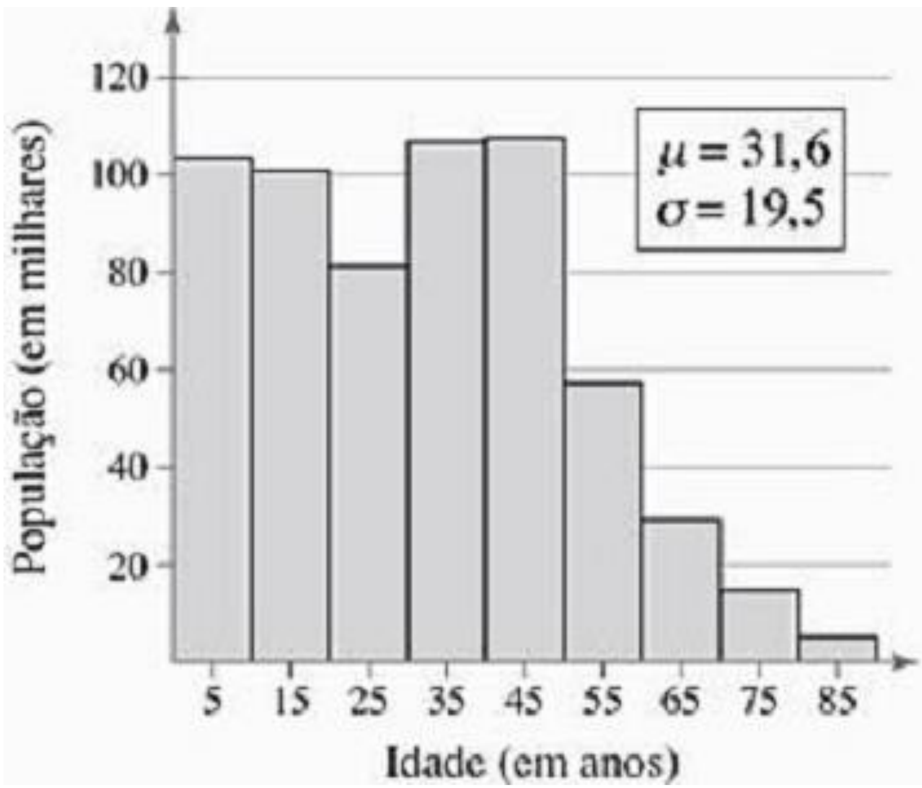
$$1 - \frac{1}{k^2}.$$

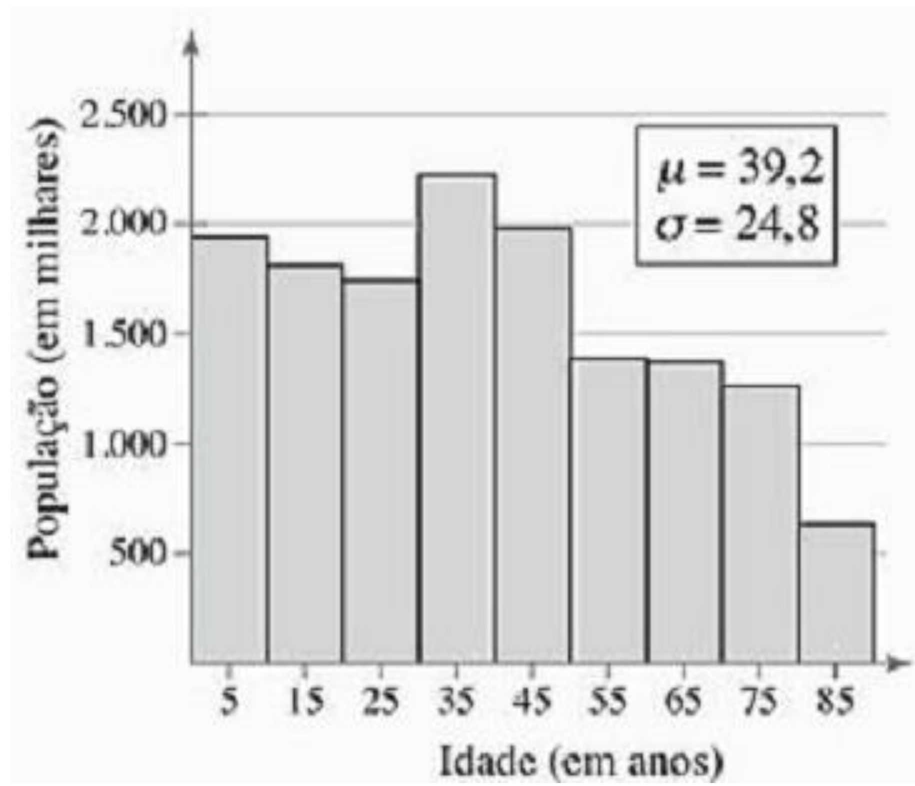
- $k = 2$: em qualquer conjunto de dados, pelo menos $1 - \frac{1}{2^2} = \frac{3}{4}$ ou 75% dos dados estão dentro de 2 desvios padrão em relação à média.
- $k = 3$: em qualquer conjunto de dados, pelo menos $1 - \frac{1}{3^2} = \frac{8}{9}$ ou 88,9% dos dados estão dentro de 3 desvios padrão em relação à média.

Exemplo 8

Usando o teorema de Chebychev

As distribuições das idades para o Alasca e a Flórida são mostradas no histograma. Decida qual é qual. Aplique o teorema de Chebychev para os dados da Flórida usando $k = 2$. O que podemos concluir?





Solução

O histograma da direita mostra a distribuição de idades na Flórida. Podemos afirmar isso, pois a população é maior e mais velha. Movendo dois desvios padrão para a esquerda da média chega-se em 0, pois $\mu - 2\sigma = 39{,}2 - 2(24{,}8) = -10{,}4$. Mover dois desvios padrão para a direita da média nos coloca em $\mu + 2\sigma = 39{,}2 + 2(24{,}8) = 88{,}8$. Pelo teorema de Chebychev podemos dizer que pelo menos 75% da população da Flórida está entre 0 e 88,8 anos de idade.

Com relação a população do Alasca podemos dizer que pelo menos 75% da população está entre 0 e 70,6 anos (ver cálculos abaixo para k=2).

$\mu - 2\sigma = 31{,}6 - 2(19{,}5) = -7{,}4$ e $\mu + 2\sigma = 31{,}6 + 2(19{,}5) = 70{,}6$

Desvio padrão para dados agrupados

Na Seção 2.1 aprendemos que conjuntos de dados grandes são melhores representados por uma distribuição de frequência. A fórmula para o desvio padrão amostral para uma distribuição de frequência é:

$$\text{Desvio padrão amostral} = s = \sqrt{\frac{\sum (x-\bar{x})^2 f}{n-1}},$$

onde $n = \Sigma f$ é o número de entradas no conjunto de dados.

Encontrando o desvio padrão para dados agrupados

Você coletou uma amostra aleatória do número de crianças por residência em certa região. Os resultados são mostrados na tabela à direita. Encontre a média amostral e o desvio padrão amostral para o conjunto de dados.

Número de crianças em 50 residências				
1	3	1	1	1
1	2	2	1	0
1	1	0	0	0
1	5	0	3	6
3	0	3	1	1
1	1	6	0	1
3	6	6	1	2
2	3	0	1	1
4	1	1	2	2
0	3	0	2	4

x	f	xf
0	10	0
1	19	19
2	7	14
3	7	21
4	2	8
5	1	5
6	4	24
	$\Sigma = 50$	$\Sigma = 91$

$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
−1,8	3,24	32,40
−0,8	0,64	12,16
0,2	0,04	0,28
1,2	1,44	10,08
2,2	4,84	9,68
3,2	10,24	10,24
4,2	17,64	70,56
		$S = 145,40$

$$\bar{x} = \frac{\sum xf}{n} = \frac{91}{50} \approx 1,8$$

Média amostral

Use a soma dos quadrados para encontrar o desvio padrão amostral.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n-1}} = \sqrt{\frac{145,4}{49}} \approx 1,7$$

Desvio padrão amostral

Usando pontos médios das classes

O gráfico circular à esquerda mostra os resultados de uma pesquisa na qual 1.000 adultos foram questionados sobre o quanto gastam na preparação para viagens pessoais no ano. Faça uma distribuição de frequência para os dados. Então use a tabela para estimar o desvio padrão amostral do conjunto de dados. (Ad

Classe	x	f	xf
0–99	49,5	380	18.810
100–199	149,5	230	34.385
200–299	249,5	210	52.395
300–399	349,5	50	17.475
400–499	449,5	60	26.970
500+	599,5	70	41.965
		$\Sigma = 1.000$	$\Sigma = 192.000$

$$\bar{x} = \frac{\sum xf}{n} = \frac{192.000}{1.000} = 192$$

$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
– 142,5	20.306,25	7.716.375,0
– 42,5	1.806,25	415.437,5
57,5	3.306,25	694.312,5
157,5	24.806,25	1.240.312,5
257,5	66.306,25	3.978.375,5
407,5	166.056,25	11.623.937,5
		$\Sigma = 25.668.750,0$

Média amostral

Então, a média amostral é \$ 192 por ano e o desvio padrão amostral é aproximadamente \$ 160,3 por ano.