PUC-SP

Mineração de Dados

Prof. Dr. Daniel Rodrigues da SIlva

PUC-SP

Medidas de Variação

Texto baseado no livro:

Estatística Aplicada - Larson / Farber – Editora Pearson – 2010

Variação

A diferença entre as entradas máxima e mínima em um conjunto de dados

Os dados precisam ser quantitativos

Variação = (Entrada máx.) – (Entrada mín.)

Exemplo: encontrando a variação

PUC-SP

Uma corporação contratou 10 graduados. Os salários iniciais de cada um são demonstrados abaixo. Encontre a variação dos salários iniciais.

Salários iniciais (milhares de dólares)

41 38 39 45 47 41 44 41 37 42

PUC-SP

- Ordenar os dados ajuda a encontrar o menor e o maior salário

37 38 39 41 41 41 42 44 45 47

Mínimo   Máximo

- Variação = (Entrada máx.) – (Entrada mín.)

= 47 – 37 = 10

A variação dos salários iniciais é igual a 10.

Desvio, variância e desvio padrão

PUC-SP

Desvio

- A diferença entre a entrada de dados, x, e a média do conjunto de dados

- Conjunto de dados da população:

  Desvio populacional: $x - \mu$

- Conjunto de dados da amostra:

  Desvio amostral: $x - \bar{x}$

Exemplo: encontrando o desvio

PUC-SP

Uma corporação contratou 10 graduados. Os salários iniciais de cada um são demonstrados abaixo. Encontre a variação dos salários iniciais.

Salários iniciais (milhares de dólares)									
41	38	39	45	47	41	44	41	37	42

Solução:

- Primeiro, determine a média dos salários iniciais.

$$\mu = \frac{\Sigma x}{N} = \frac{415}{10} = 41.5$$

PUC-SP

Solução: encontre o desvio amostral

Determine o desvio para cada entrada.

Salário ($ 1.000s), x	Desvio: $x – \mu$
41	41 – 41,5 = –0,5
38	38 – 41,5 = –3,5
39	39 – 41,5 = –2,5
45	45 – 41,5 = 3,5
47	47 – 41,5 = 5,5
41	41 – 41,5 = –0,5
44	44 – 41,5 = 2,5
41	41 – 41,5 = –0,5
37	37 – 41,5 = –4,5
42	42 – 41,5 = 0,5
$\Sigma x = 415$	$\Sigma(x – \mu) = 0$

Variância e Desvio Padrão

PUC-SP

Variância da população

- $$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$ ⟵ Soma dos quadrados, SQ_x

Desvio padrão da população

- $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

PUC-SP

Em palavras	*Em símbolos*
1. Encontre a média do conjunto de dados da população.	$\mu = \frac{\Sigma x}{N}$
2. Encontre o desvio de cada entrada.	$x - \mu$
3. Eleve os desvios ao quadrado.	$(x - \mu)^2$
4. Some para obter a soma dos quadrados.	$SQ_x = \Sigma(x - \mu)^2$

PUC-SP

Em palavras	*Em símbolos*
5. Divida por *N* para obter **a variância populacional**.	$\sigma^2 = \dfrac{\Sigma(x-\mu)^2}{N}$
6. Encontre a raiz quadrada para obter **o desvio padrão populacional**.	$\sigma = \sqrt{\dfrac{\Sigma(x-\mu)^2}{N}}$

Exemplo: encontrando o desvio padrão da população

PUC-SP

Uma corporação contratou 10 graduados. Os salários iniciais de cada um são demonstrados abaixo. Encontre a variação dos salários iniciais.

Salários iniciais (milhares de dólares)

41 38 39 45 47 41 44 41 37 42

Lembrar $\mu = 41{,}5$.

Solução: encontrando o desvio padrão da população

PUC-SP

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

$\Sigma(x - \mu) = 0$ $\qquad$ $SQ_x = 88{,}5$

PUC-SP

Variância da população

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = \frac{88.5}{10} \approx 8.9$$

Desvio padrão da população

$$\sigma = \sqrt{\sigma^2} = \sqrt{8.85} \approx 3.0$$

O desvio padrão da população é cerca de 3,0 ou $ 3.000.

Desvio, variância e desvio padrão

PUC-SP

Variância da amostra

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

Desvio padrão da amostra

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

Encontrando a variância e o desvio padrão da amostra

PUC-SP

Em palavras	*Em símbolos*
1. Encontre a média do conjunto de dados da amostra.	$\overline{x} = \frac{\Sigma x}{n}$
2. Encontre o desvio de cada entrada.	$x - \overline{x}$
3. Eleve cada desvio ao quadrado.	$(x - \overline{x})^2$
4. Some-os para obter a soma dos quadrados.	$\Sigma(x - \overline{x})^2$

PUC-SP

Em palavras	*Em símbolos*
5. Divida por n – 1 para obter **a variância da amostra**.	$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$
6. Encontre a raiz quadrada para obter **o desvio padrão da amostra**.	$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

Exemplo: encontre o desvio padrão da amostra

PUC-SP

Os salários inicias são para uma filial da empresa em Chicago. A empresa tem várias outras filiais e você planeja usar os salários iniciais de Chicago para estimar os salários iniciais da população maior. Encontre o desvio padrão dos salários iniciais da amostra.

Salários iniciais (milhares de dólares)

41 38 39 45 47 41 44 41 37 42

Solução: encontrando o desvio padrão da população

PUC-SP

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

$\Sigma(x - \mu) = 0$ $\qquad$ $SQ_x = 88{,}5$

PUC-SP

Variância da amostra

- $$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{88.5}{10-1} \approx 9.8$$

Desvio padrão da amostra

- $$s = \sqrt{s^2} = \sqrt{\frac{88.5}{9}} \approx 3.1$$

O desvio padrão da amostra é de aproximadamente 3,1 ou $ 3.100.

Exemplo: usando tecnologia para encontrar o desvio padrão

PUC-SP

A amostra dos aluguéis de escritórios (em dólares por metro quadrado ao ano) no distrito comercial central de uma cidade é exibida na tabela. Use uma calculadora ou um computador para encontrar a média dos aluguéis e o desvio padrão da amostra.

Preço dos aluguéis		
35,00	33,50	37,00
23,75	26,50	31,25
36,50	40,00	32,00
39,25	37,50	34,75
37,75	37,25	36,75
27,00	35,75	26,00
37,00	29,00	40,50
24,50	33,00	38,00

Solução: usando tecnologia para encontrar o desvio padrão

PUC-SP

EXCEL

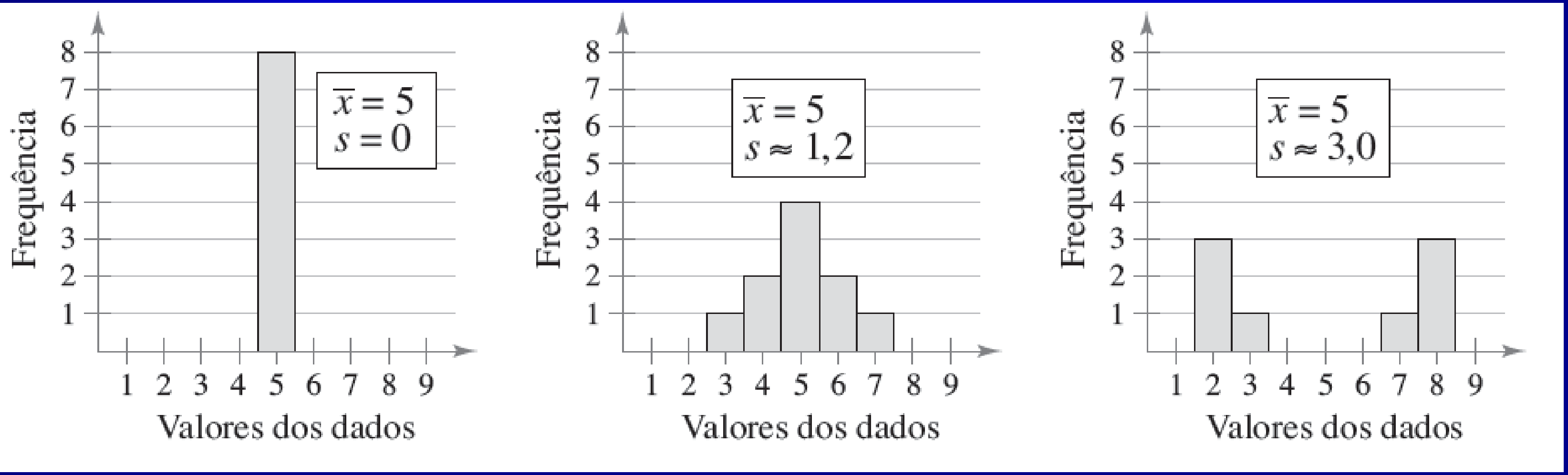
	A	B
1	Média	33,72917
2	Erro padrão	1,038864
3	Média	35,375
4	Moda	37
5	Desvio padrão	5,089373
6	Variância da amostra	25,90172
7	Kurtosis	-0,74282
8	Skewness	-0,70345
9	Extensão	16,75
10	Mínimo	23,75
11	Máximo	40,5
12	Soma	809,5
13	Conta	24

Interpretando o desvio padrão

PUC-SP

Desvio padrão é a medida do valor típico que uma entrada desvia da média

Quanto mais as entradas estão espalhadas, maior o desvio padrão



Desvio padrão para dados agrupados

PUC-SP

Desvio padrão de uma amostra para uma distribuição de frequência

- $$s = \sqrt{\frac{\Sigma(x - \overline{x})^2 f}{n - 1}}$$ em que $n = \Sigma f$ (o número de entradas no conjunto de dados)

- Quando uma distribuição de frequência tem classes, estime a média da amostra e o desvio padrão usando o ponto médio de cada classe.

Exemplo: encontrando o desvio padrão para dados agrupados

PUC-SP

Você coleta uma amostragem aleatória do número de crianças por casa em uma região. Encontre a média da amostra e o desvio padrão da amostra do conjunto de dados.

Número de crianças em 50 casas				
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

Solução: encontrando o desvio padrão para dados agrupados

PUC-SP

Primeiro, construa a distribuição da frequência

Encontre a média da distribuição da frequência

$$\bar{x} = \frac{\Sigma xf}{n} = \frac{91}{50} \approx 1.8$$

A média da amostra é de cerca de 1,8 criança.

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

Determine a soma dos quadrados.

x	f	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 f$
0	10	0 – 1,8 = –1,8	$(-1,8)^2 = 3,24$	3,24(10) = 32,40
1	19	1 – 1,8 = –0,8	$(-0,8)^2 = 0,64$	0,64(19) = 12,16
2	7	2 – 1,8 = 0,2	$(0,2)^2 = 0,04$	0,04(7) = 0,28
3	7	3 – 1,8 = 1,2	$(1,2)^2 = 1,44$	1,44(7) = 10,08
4	2	4 – 1,8 = 2,2	$(2,2)^2 = 4,84$	4,84(2) = 9,68
5	1	5 – 1,8 = 3,2	$(3,2)^2 = 10,24$	10,24(1) = 10,24
6	4	6 – 1,8 = 4,2	$(4,2)^2 = 17,64$	17,64(4) = 70,56

$$\Sigma(x-\bar{x})^2 f = 145.40$$

PUC-SP

PUC-SP

Encontre o desvio padrão da amostra.

$$s = \sqrt{\frac{\Sigma(x-\overline{x})^2 f}{n-1}} = \sqrt{\frac{145.40}{50-1}} \approx 1.7$$

O desvio padrão é de cerca de 1,7 criança.

Considerando o dataset "banco.csv", determine:

- **A variância**
- **O desvio Padrão**

Para todas as colunas numéricas do arquivo

PUC-SP

PUC-SP

```
import numpy as np

df = pd.read_csv('D:/A - PUC/Mineração/Dados/banco.csv' )

var = np.var(df['balance'])

desvio = np.std(df['balance'])

print('Variância = ' , var , '\nDesvio Padrão = ' , desvio )
```

Medidas de Posição

PUC-SP

Quartis

PUC-SP

- **Fractis** são números que particionam (dividem) um conjunto de dados ordenados em partes iguais
- **Quartis** dividem dados ordenados em quatro partes aproximadamente iguais

  Primeiro quartil, Q_1: Cerca de um quarto dos dados cai em ou abaixo de Q_1

  Segundo quartil, Q_2: Cerca de metade dos dados caem em ou abaixo de Q_2 (mediana)

  Terceiro quartil, Q_3: Cerca de três quartos dos dados caem em ou abaixo de Q_3

Exemplo: encontrando quartis

PUC-SP

As pontuações dos testes de 15 empregados matriculados em um curso de primeiros socorros são listadas. Encontre o primeiro, o segundo e o terceiro quartil das pontuações dos testes.

13 9 18 15 14 21 7 10 11 20 5 18 37 16 17

Solução:

- Q_2 divide o conjunto de dados em duas metades

$$\underbrace{5\ \ 7\ \ 9\ \ 10\ \ 11\ \ 13\ \ 14}_{\text{Metade inferior}}\ \ \underset{\uparrow \atop Q_2}{15}\ \ \underbrace{16\ \ 17\ \ 18\ \ 18\ \ 20\ \ 21\ \ 37}_{\text{Metade superior}}$$

PUC-SP

Solução: encontrando quartis

O primeiro e o terceiro quartis são as medianas das metades inferior e superior do conjunto de dados

$$\overbrace{5\quad 7\quad 9\quad \underset{\substack{\uparrow \\ \boldsymbol{Q_1}}}{10}\quad 11\quad 13\quad 14}^{\text{Metade inferior}}\quad \underset{\substack{\uparrow \\ \boldsymbol{Q_2}}}{15}\quad \overbrace{16\quad 17\quad 18\quad \underset{\substack{\uparrow \\ \boldsymbol{Q_3}}}{18}\quad 20\quad 21\quad 37}^{\text{Metade superior}}$$

Cerca de um quarto dos funcionários obteve nota 10 ou menor; cerca de metade deles obteve 15 ou menor; e cerca de três quartos obteve 18 ou menor.

Amplitude interquartil

PUC-SP

Encontrando a amplitude interquartil (VIQ)

- A diferença entre o terceiro e o primeiro quartis
- VIQ = $Q_3 - Q_1$

Exemplo: encontrando a amplitude interquartil

PUC-SP

Encontre a amplitude interquartil das notas dos testes.

Lembre-se: $Q_1 = 10$, $Q_2 = 15$ e $Q_3 = 18$

Solução:

- VIQ = $Q_3 - Q_1 = 18 - 10 = 8$

As notas dos testes na porção do meio do conjunto de dados variam no máximo em 8 pontos.

PUC-SP

Considerando o dataset “banco.csv”, determine *Os quartis p*ara todas as colunas numéricas do arquivo

```
q1 = np.quantile(df['age'], 0.25)

q2 = np.quantile(df['age'], 0.5)

q3 = np.quantile(df['age'], 0.75)

print('Quartil 1 = ',q1,'\nQuartil 2 = ',q2,
'\nQuartil 3 = ',q3)
```

PUC-SP

Gráfico de caixa-e-bigodes (Boxplot)

PUC-SP

Ferramenta exploratória de análise de dados

Destaca qualidades importantes do conjunto de dados

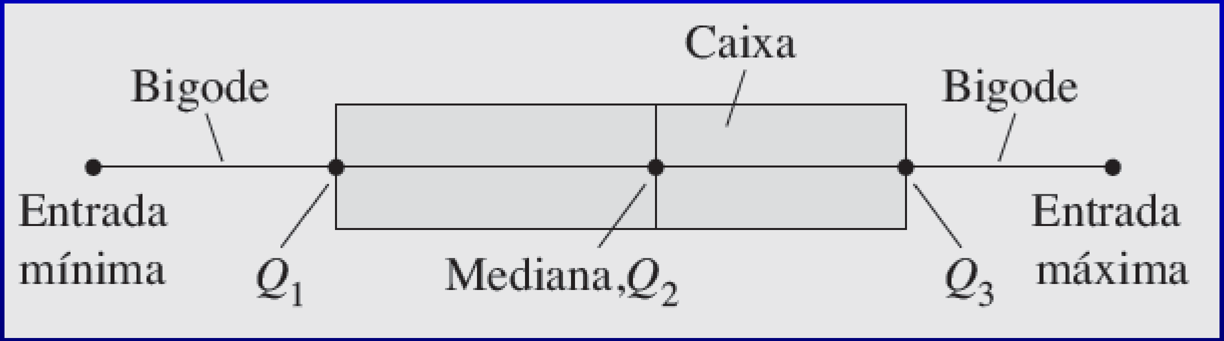
Requer (**sumário de cinco números**):

- Entrada mínima
- Primeiro quartil Q_1
- Mediana Q_2
- Terceiro quartil Q_3
- Entrada máxima

Desenhando um gráfico de caixa-e-bigodes

PUC-SP

1. Encontre o sumário dos cinco números do conjunto de dados.
2. Construa uma escala horizontal que cubra a variância dos dados.
3. Ponha os cinco números acima da escala horizontal.
4. Desenhe uma caixa acima da escala horizontal de Q_1 até Q_3 e desenhe uma linha vertical na caixa em Q_2.
5. Desenhe bigodes saindo da caixa para as entradas mínima e máxima.



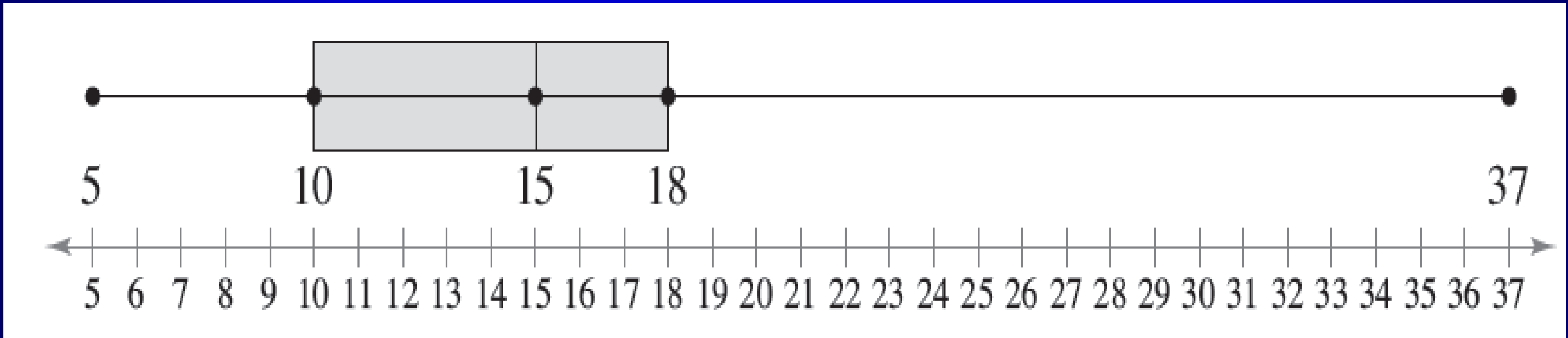
Exemplo: desenhando um gráfico de caixa-e-bigodes

PUC-SP

Desenhe um gráfico de caixa-e-bigodes que represente as 15 pontuações dos testes.

Lembre-se: Mín. = 5 $Q_1 = 10$ $Q_2 = 15$ $Q_3 = 18$ e Máx. = 37

Solução:



Cerca de metade das notas estão entre 10 e 18. Olhando para o comprimento do bigode direito, pode-se concluir que 37 é um possível valor discrepante.

PUC-SP

Considerando o dataset “banco.csv”, determine *BoxPlot p*ara todas as colunas numéricas do arquivo

```
import matplotlib.pyplot as plt

plt.boxplot(df['a'])

plt.title("Basic Box Plot")

plt.show()
```

PUC-SP

Percentis e outros fractis

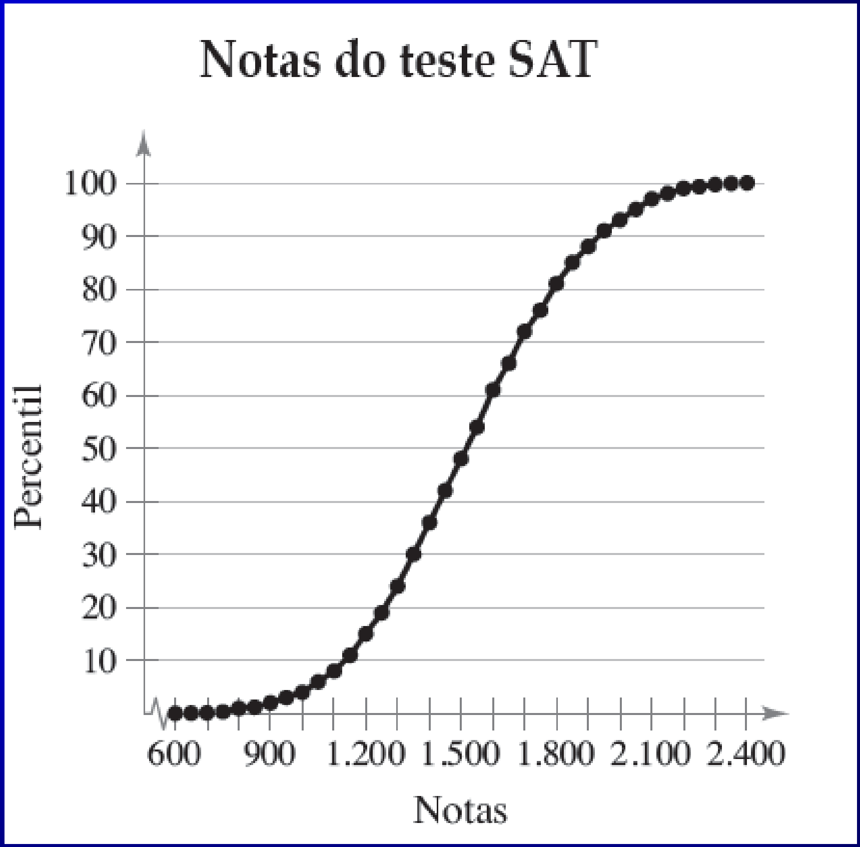
PUC-SP

Fractis	Sumário	Símbolos
Quartis	Divide os dados em 4 partes iguais	Q_1, Q_2, Q_3
Decis	Divide os dados em 10 partes iguais	$D_1, D_2, D_3, \ldots, D_9$
Percentis	Divide os dados em 100 partes iguais	$P_1, P_2, P_3, \ldots, P_{99}$

Exemplo: interpretando percentis

PUC-SP

A ogiva representa a distribuição de frequência cumulativa para as provas do SAT (vestibular dos EUA) de estudantes em uma ano recente. Qual nota representa o 72º percentil? Como você deve interpretar isso? (*Fonte: College Board Online*.)

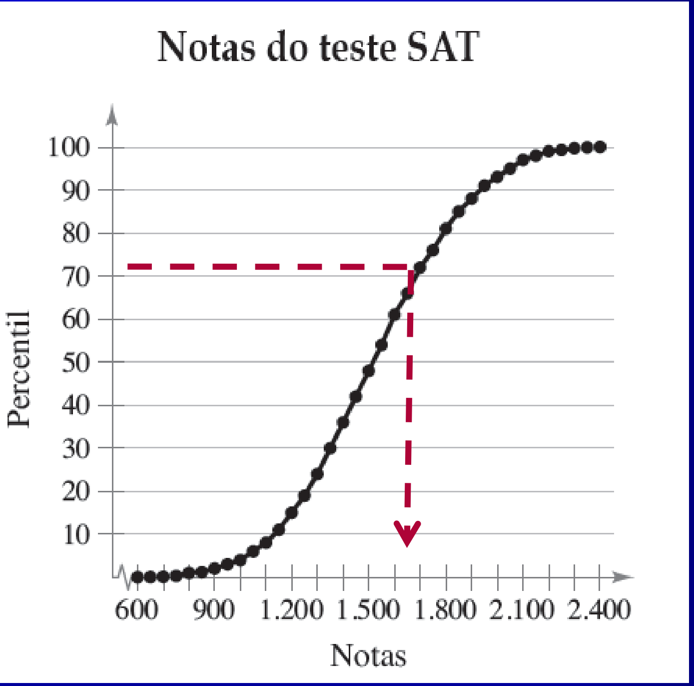


Solução: interpretando percentis

PUC-SP

O 72º percentil corresponde à nota 1.700.

Isso significa que 72% dos alunos obtiveram resultados de 1.700 ou menos.



O escore padrão

PUC-SP

Escore padrão (escore *z*)

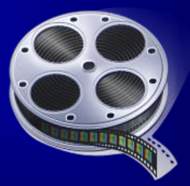
- Representa o número de desvios padrão que um dado valor *x* está da média μ.

$$\boxed{z = \frac{valor\ x\ -\ média}{desvio\ padrão} = \frac{x - \mu}{\sigma}}$$

PUC-SP

Exemplo: comparando escores *z* de diferentes conjuntos de dados

Em 2007, o ator Forest Whitaker ganhou o Oscar de melhor ator, aos 45 anos de idade, por sua atuação no filme *O Último Rei da Escócia*. A atriz Helen Mirren ganhou o prêmio de melhor atriz aos 61 anos por seu papel em *A Rainha*. A idade média para todos os vencedores do prêmio de melhor ator é 43,7, com desvio padrão de 8,8. A idade média para as vencedoras do prêmio de melhor atriz é 36, com desvio padrão de 11,5. Encontre o escore *z* que corresponda à idade de cada ator ou atriz. Depois, compare os resultados.



Solução: comparando escores z de diferentes conjuntos de dados

PUC-SP

Forest Whitaker:

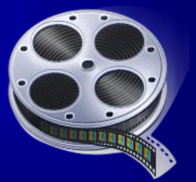
$$z = \frac{x - \mu}{\sigma} = \frac{45 - 43{,}7}{8{,}8} \approx 0{,}15$$

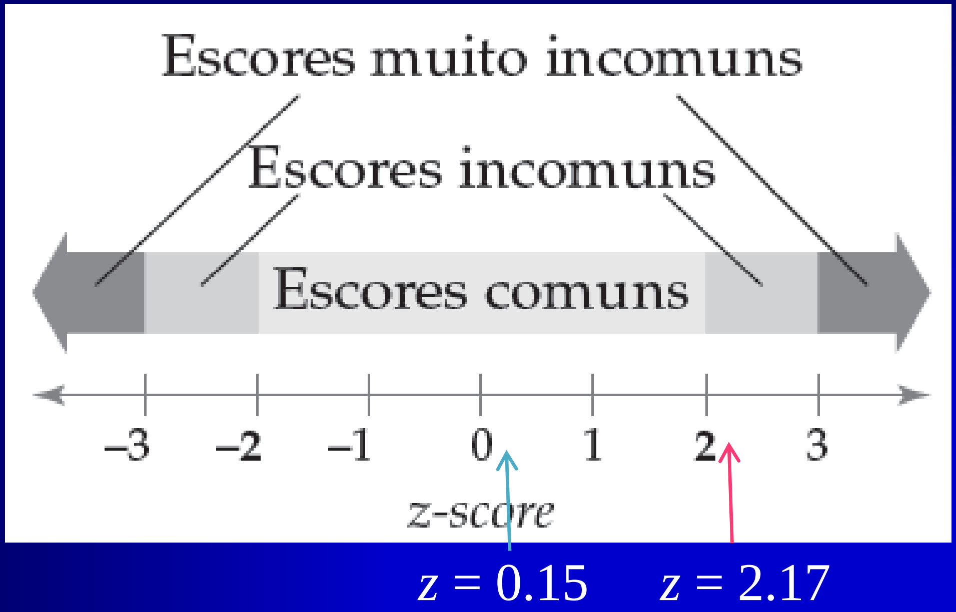
Desvio padrão 0,15 acima da média

- Helen Mirren:

$$z = \frac{x - \mu}{\sigma} = \frac{61 - 36}{11{,}5} \approx 2{,}17$$

Desvio padrão 2,17 acima da média





Escores muito incomuns

Escores incomuns

Escores comuns

Escore *z*

PUC-SP

O escore *z* correspondente à idade de Helen Mirren é mais de dois desvios padrão da média, então é considerado incomum. Comparado a outras vencedoras do prêmio de melhor atriz, ela é relativamente mais velha, enquanto a idade de Forest Whitaker é pouco acima da média dos ganhadores do prêmio de melhor ator.