Estatística e Probabilidade
Aula 3 - Medidas de dispersão

Professor: Josceli Tenório

Data: 24/03/2022

Introdução

Informar apenas a medida de tendência central fornece apenas informações parciais sobre um conjunto de dados ou uma distribuição. Diferentes amostras ou populações podem ter medidas de tendência central idênticas e apresentar diferenças entre si em outros aspectos importantes. A Figura 1 mostra as medidas de variação utilizadas.

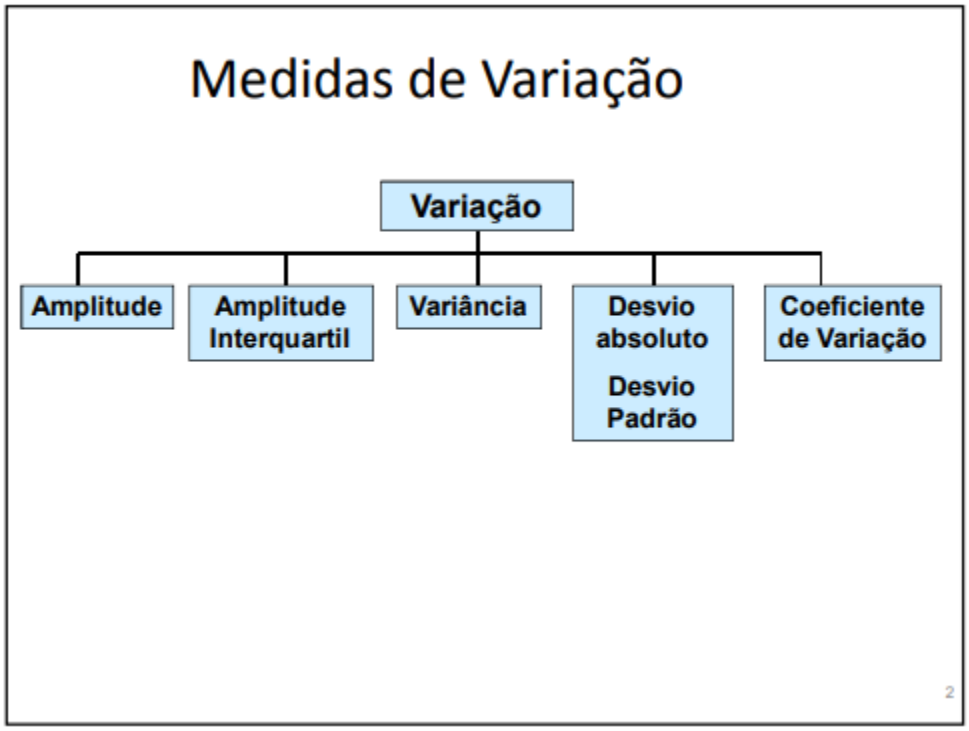


Figura 1. Medidas de variação

A medida de dispersão mais simples de uma amostra é a amplitude, a diferença entre o maior e o menor valores da amostra.

O que podemos afirmar acerca das medidas 1, 2 e 3 da Figura 2?

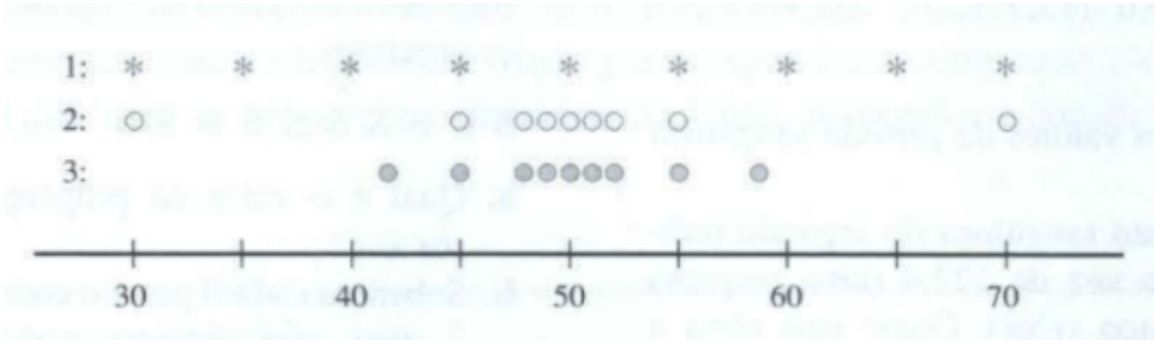


Figura 2. Medidas referentes a três amostras.

Nossa principal medida de dispersão envolve os desvios em relação à média, $x_1 - x_{médio}$, $x_2 - x_{médio}$, $x_n - x_{médio}$ .Ou seja, os desvios da média são obtidos pela subtração de $x_{médio}$

de cada uma das n observações da amostra. Uma forma simples de combinar os desvios em uma única quantidade é calcular a sua média (somá-los e dividi-los por n). Porém, poderemos ter um problema:

$$\text{somatória dos desvios} = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Uma solução para evitar esse efeito é considerar o valor absoluto e calcular o desvio médio absoluto. Como a operação em valor absoluto conduz a diversas dificuldades teóricas, vamos considerar para o cálculo os quadrados dos desvios. Porém, devem ser considerados dois cenários:

Quando a população é finita e consiste de n valores, a variância é calculada por:

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

Entretanto, se usássemos o divisor n na fórmula da variância da amostra, a quantidade resultante tenderia a subestimar a variância (gerar valores na média muito pequenos para a estimativa), enquanto a divisão pelo valor, ligeiramente menor, n - 1 corrige a subestimativa. A divisão por n -1 fornece uma estimativa não enviesada da variância populacional maior, o que é conhecida como correção de Bessel. Desta forma:

A **variância amostral**, representada por s^2, é dada por

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

O **desvio padrão amostral,** representado por s, é a raiz quadrada (positiva) da variância:

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

Para dados agrupados basta multiplicar: $(x_i - x_{médio})^2 \cdot f_i$ onde f é a frequência.

Coeficiente de Variação (CV)

- Mede a variação relativa à média
- Costuma-se ser dado em porcentagem (%)
- Pode ser utilizada para comparar duas ou mais séries de dados em unidades diferentes

$$CV = \left(\frac{S}{\bar{X}} \right) \cdot 100\%$$

Considere o preço e a variação de duas ações no mercado:

- Ação A:
  - Preço médio no último ano = $50
  - Desvio-padrão = $5

$$CV_A = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Ação B:
  - Preço médio no último ano = $100
  - Desvio-padrão = $5

$$CV_B = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

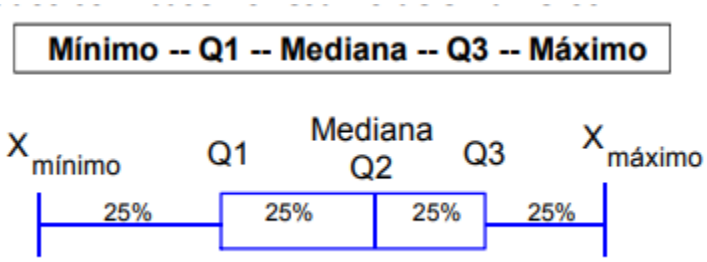
As duas ações possuem o mesmo desvio-padrão, mas o preço da ação B teve uma variação relativa menor.

Box Plot - variável numérica

Um resumo esquemático denominado boxplot é usado para descrever as características mais proeminentes de conjuntos de dados. Essas características incluem (1) centro, (2) dispersão, (3) a extensão e a natureza de qualquer desvio em relação à simetria e (4) a identificação de outliers, observações que normalmente estão distantes da maior parte dos dados.

Como apenas um outlier pode afetar drasticamente os valores de $x_{médio}$ e s, um boxplot é baseado em medidas "resistentes" à presença de alguns outliers: a mediana e uma medida de dispersão denominadas dispersão entre os quartos.

O boxplot mais simples tem base no seguinte resumo de cinco números:



Para criar o boxplot:

Ordene as n observações da menor para a maior e então separe a metade menor da maior. A mediana $\tilde{x}$ estará incluída em ambas as partes se n for ímpar. Então o **quarto inferior** será a mediana da metade menor e o **quarto superior** será a mediana da metade maior. Uma medida de dispersão resistente a *outliers* é a **dispersão entre os quartos** f_s, dada por

$$f_s = \text{quarto superior} - \text{quarto inferior}$$

Vantagens do boxplot

Ajudam na escolha de medidas numéricas descritivas apropriadas:

- O propósito para o qual o resumo descritivo dos dados é realizado.
- Facilidade de interpretação.
- O grau de sensibilidade a valores extremos.
- Potencial para uso em inferência estatística.

O ultra-som foi usado para obter informações sobre dados de corrosão na espessura da chapa do assoalho de um reservatório elevado usado para armazenar óleo bruto.

Cada observação é a maior profundidade do orifício na placa, expressa em milipolegadas.

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

O resumo de cinco números segue:

menor $x_i = 40$ quarto inferior = 72,5 $\tilde{x} = 90$ quarto superior = 96,5
maior $x_i = 125$

A Figura 3 mostra o boxplot obtido;

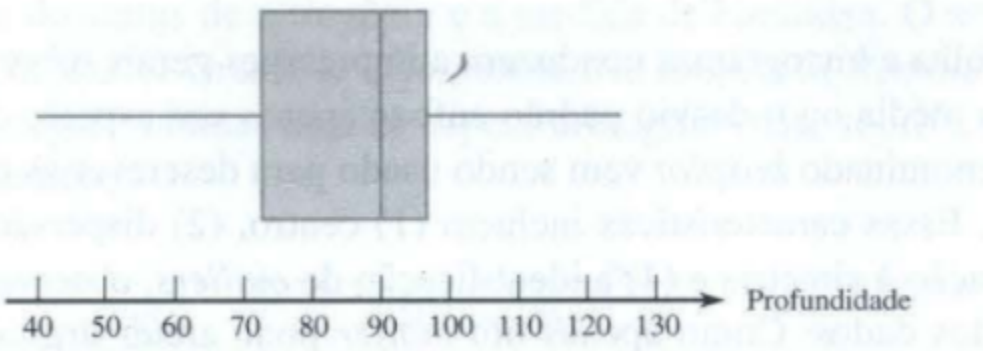


Figura 3. Box plot para os dados de corrosão.

Os dados também podem ser representados por meio de um diagrama de dispersão, como mostra a Figura 4.

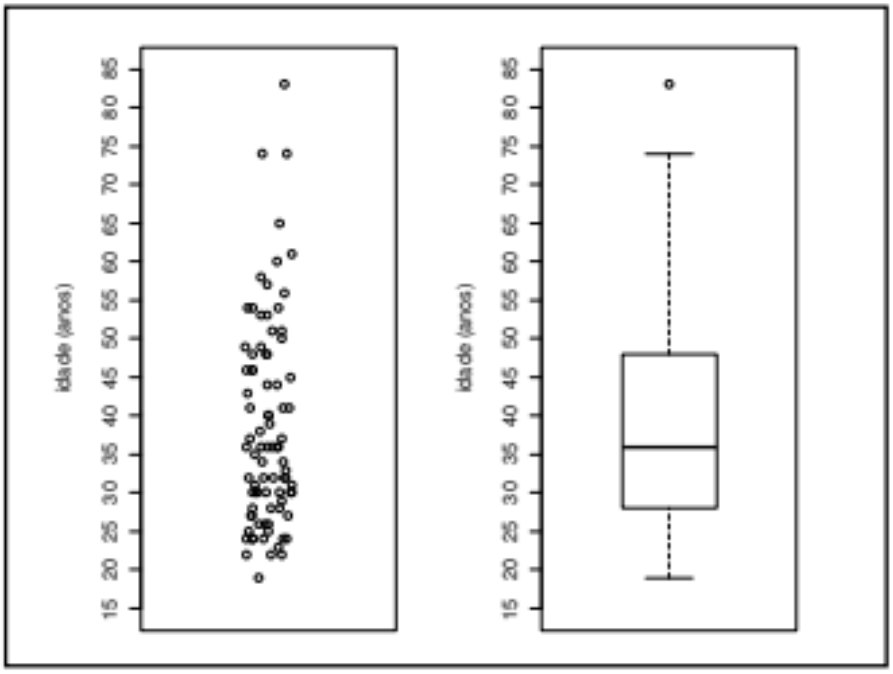


Figura 4. Diagrama de dispersão unidimensional da idade em anos.

Como obter os outliers?

Os efeitos de descargas parciais na degradação de materiais de cavidades isolantes têm importantes implicações na vida útil de componentes de alta voltagem. Consideremos a seguinte amostra de n = 25 larguras de pulso de descargas lentas em uma cavidade cilíndrica de polietileno.

5,3 8,2 13,8 74,1 85,3 88,0 90,2 91 ,5 92,4 92,9 93,6 94,3 94,8

94,9 95,5 95,8 95,9 96,6 96,7 98,1 99,0 101,4 103,7 106,0 113,5

Os indicadores relevantes são:

$x_{mediana} = 94,8$

quarto inferior = 90,2

quarto superior= 96,7

$f_s = 6,5$

$1,5f_s = 9,75$

$3f_s = 19,50$

Qualquer observação menor que 90,2 - 9,75 = 80,45 ou maior que 96,7 + 9,75 = 106,45 é

um outlier.

Como 90,2- 19,5 = 70,7, as três observações: 5,3, 8,2 e 13,8 são outliers extremos. Os outros dois outliers são moderados.

Os "bigodes" se estendem até 85,3 e 106,0, as observações mais extremas que não são outliers.

A Figura 5 mostra o boxplot

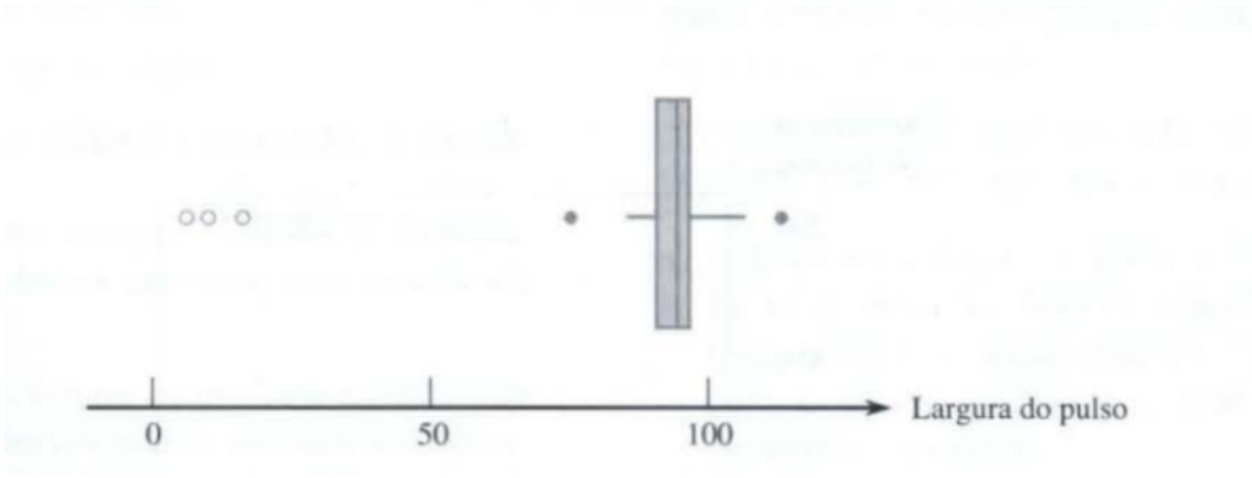


Figura 5. Boxplot dos dados de largura de pulso.

Aplicações

1. O valor do módulo de Young (GPa) foi determinado para chapas fundidas feitas de algumas substâncias metálicas, resultando nas observações a seguir

   116,4 115,9 114,6 115,2 115,8

   a. Calcule $x_{médio}$ e os desvios em relação à média.

   b. Use os desvios calculados na parte (a) para obter a variância amostral e o desvio padrão amostral.

c. Subtraia 100 de cada observação para obter uma amostra de valores transformados. Agora calcule a variância amostral desses valores transformados e a compare ao s^2 dos dados originais.

2. Um estudo da relação entre idade e diversas funções visuais (como precisão e percepção de profundidade) informou as seguintes observações da área de lâmina escleral (mm^2) nas extremidades do nervo óptico humano.

2,75	2,62	2,74	3,85	2,34	2,74	3,93	4,21	3,88
4,33	3,46	4,52	2,43	3,65	2,78	3,56	3,01	

a) Calcule os valores necessários à análise desse conjunto de dados.
b) Determine os quartos inferior e superior.
c) Calcule o valor da dispersão entre os quartos.
d) Se os dois maiores valores da amostra, 4,33 e 4,52, fossem 5,33 e 5,52, como f_s seria afetado? Explique.
e) Em quanto a observação 2,34 pode ser aumentada sem afetar f_s? Explique.
f) Se uma 18' observação, x_{18} = 4,60, fosse adicionada à amostra, qual seria o valor de f_s?

3. Foram selecionadas amostras de três tipos de corda e o limite de fadiga (MPa) foi determinado para cada amostra, resultando os dados a seguir.

Tipo 1: 350 350 350 358 370 370 370 371 371 372 372 384 391 391 392

Tipo 2: 350 354 359 363 365 368 369 371 373 374 376 380 383 388 392

Tipo 3: 350 361 362 364 364 365 366 371 377 377 377 379 380 380 392

a) Construa um boxplot comparativo e comente as semelhanças e diferenças.
b) Construa um gráfico de pontos (dotplot) comparativo (um dotplot para cada amostra com uma escala comum). Comente as semelhanças e diferenças.
c) O boxplot comparativo da parte (a) fornece uma avaliação informativa das semelhanças e diferenças? Explique seu raciocínio.

Referência

Devore J L. Probability and Statistics for Engineering and the Sciences. Thomson ed. 5a. ed. 284 p.