Disciplina: 5170 - Probabilidade e Estatística

Aula 03

Prof. George Lucas Moraes Pezzott
glmpezzott2@uem.br

Departamento de Estatística - UEM
Sala: 222 - Bloco: E-90

Introdução

- **Variável:** Qualquer característica de interesse associada aos elementos de uma população.

Classificação de variáveis

- **Qualitativa**
  - **Nominal** (Ex: cor, sexo, estado civil)
  - **Ordinal** (Ex: grau de escolaridade, classe social, porte de empresa)
- **Quantitativa**
  - **Discreta** (Ex: nº de pessoas, nº de defeitos)
  - **Contínua** (Ex: peso, temperatura, preço, tempo,)

Introdução

Suponha que X representa uma variável (altura de indivíduo, tempo de resposta de um sistema, número de chegadas em uma fila, ...).

Desejamos conhecer o "comportamento" da variável X.

Para isso, considere as seguintes definições:

- **Tamanho da Amostra:** n
- **Amostra observada** ou **conjunto de observações** ou **dados:** n valores (respostas) observados da variável X: $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.
- **Amostra Aleatória:** É uma sequência de realizações da variável de interesse X sem observá-la, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$;
  $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ é uma observação de $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$;

Apresentação dos Dados

O que fazer com os dados coletados?

- **Etapa inicial:** Estatística Descritiva (Análise Exploratória): é o ramo que trata da organização, resumo e apresentação dos dados. Ela consiste em calcular medidas descritivas, gráficos e tabelas.

Para variáveis quantitativas, podemos calcular:

- **Medidas de Posição:** Encontrar um valor que descreva a posição dos dados sobre a reta.
  - moda, média, mediana (medidas de tendência central), quantis (quartis, decis, percentis).
- **Medidas de Dispersão:** Encontrar um valor que resuma a variabilidade de um conjunto de dados.
  - amplitude, intervalo interquartil, variância, desvio-padrão, coeficiente de variação.

Medidas de posição

- **Dados Brutos:** É o conjunto de observações de uma variável, apresentados da forma em que foram coletados.
  Ex: 3, 7, 2, 1, 5, 7

- **Rol:** É o conjunto de observações de uma variável, apresentados da forma ordenada (crescente ou decrescente).
  Ex: 1, 2, 3, 5, 7, 7

Em uma notação geral, podemos denotar os **dados brutos** e o **rol** das seguintes formas:

Dados Brutos: $x_1, x_2, x_3, \ldots, x_n$

Rol: $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$, tal que:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(n)} \quad \text{(crescente)}$$

ou

$$x_{(1)} \geq x_{(2)} \geq x_{(3)} \geq \ldots \geq x_{(n)} \quad \text{(decrescente)},$$

em que n representa o tamanho da amostra ou o número total de elementos na amostra.

Estatísticas Descritivas Medidas de Posição

- **Moda:** É a realização mais frequente do conjunto de valores observados.
  Notação: Mo.

  Ex 1: Dados: 2; 3; 5; 3; 1

  $Mo = 3$ (unimodal)

  Ex 2: Dados: 4; 1; 7; 4; 7

  $Mo = 4$ e 7 (bimodal)

  Ex 3: Dados: 8; 1; 7; 4; 0

  Não existe moda (amodal)

Estatísticas Descritivas Medidas de Posição

- **Média Aritmética (média amostral):** É a soma das observações dividido pelo número total delas.
  Notação: $\overline{x}$.

  Descrição matemática:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ex 1: Dados: 2; 3; 5; 3; 1

$\overline{x} = \frac{2+3+5+3+1}{5} = 2,8$

Estatísticas Descritivas Medidas de Posição

Propriedades da Média

1. A média de uma série de valores constantes é essa constante.

   Dados: 10, 10, 10, 10, 10, 10, 10 $\Rightarrow \bar{x} = 10$

Estatísticas Descritivas Medidas de Posição

Propriedades da Média

1. Se aumentamos ou diminuímos todos os valores segundo uma quantidade fixa c, então a média também fica aumentada ou diminuída de c.

   Dados modificados:
   $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$
   $\Rightarrow \bar{y} = \bar{x} + c$

   $y_1 = x_1 - c, y_2 = x_2 - c, \ldots, y_n = x_n - c$
   $\Rightarrow \bar{y} = \bar{x} - c$

Estatísticas Descritivas Medidas de Posição

Propriedades da Média

1. Se multiplicamos ou dividimos todos os valores segundo uma quantidade fixa c, então a média também fica multiplicada ou dividida por c.

   Dados modificados:
   $y_1 = x_1/c, y_2 = x_2/c, \ldots, y_n = x_n/c$
   $\Rightarrow \bar{y} = \bar{x}/c$

   $y_1 = x_1 c, y_2 = x_2 c, \ldots, y_n = x_n c$
   $\Rightarrow \bar{y} = \bar{x} c$

Estatísticas Descritivas Medidas de Posição

Desvantagem da Média Aritmética: Sofre influência de valores extremos (*outliers*).

Ex 2: Dados: 0; 10; 12; 12; 16

$\bar{x} = \frac{0+10+12+12+16}{5} = 10$

Ex 3: Dados: 1; 3; 5; 6; 35

$\bar{x} = \frac{1+3+5+6+35}{5} = 10$

Estatísticas Descritivas Medidas de Posição

- **Mediana:** É a observação que ocupa a posição central do conjunto de observações (amostra), quando estão ordenados em de forma crescente (ou decrescente).
  Notação: Md.

Como encontrar a mediana?

Passo 1: Ordenar os dados, ou seja, construir o rol.

Passo 2: Encontrar a posição da mediana,

$$P = \frac{n}{2}.$$

Estatísticas Descritivas Medidas de Posição

A mediana será o elemento, de tal forma que

$$Md = \begin{cases} \frac{x_{(P)} + x_{(P+1)}}{2}, & \text{se } P \text{ é inteiro} \\ x_{(P^*+1)}, & \text{se } P \text{ não é inteiro} \end{cases},$$

em que P^* é a parte inteira do número P.

Ex: $P = 3,5 \Rightarrow P^* = 3$

Estatísticas Descritivas Medidas de Posição

Ex 1: Dados: 1; 3; 5; 4; 0

Rol: 0; 1; 3; 4; 5

$P = \frac{5}{2} = 2,5 \Rightarrow P^* = 2 \Rightarrow Md = x_{(2+1)} = x_{(3)} = 3$

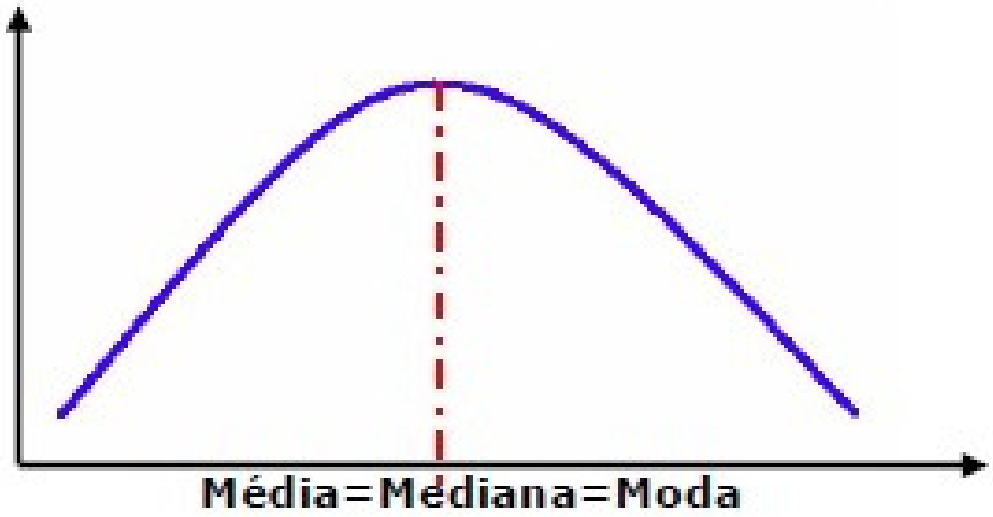
Ex 2: Rol: 1; 3; 5; 6; 6; 35

$P = \frac{6}{2} = 3 \Rightarrow Md = \frac{x_{(3)} + x_{(4)}}{2} = \frac{5+6}{2} = 5,5$

Estatísticas Descritivas Medidas de Posição

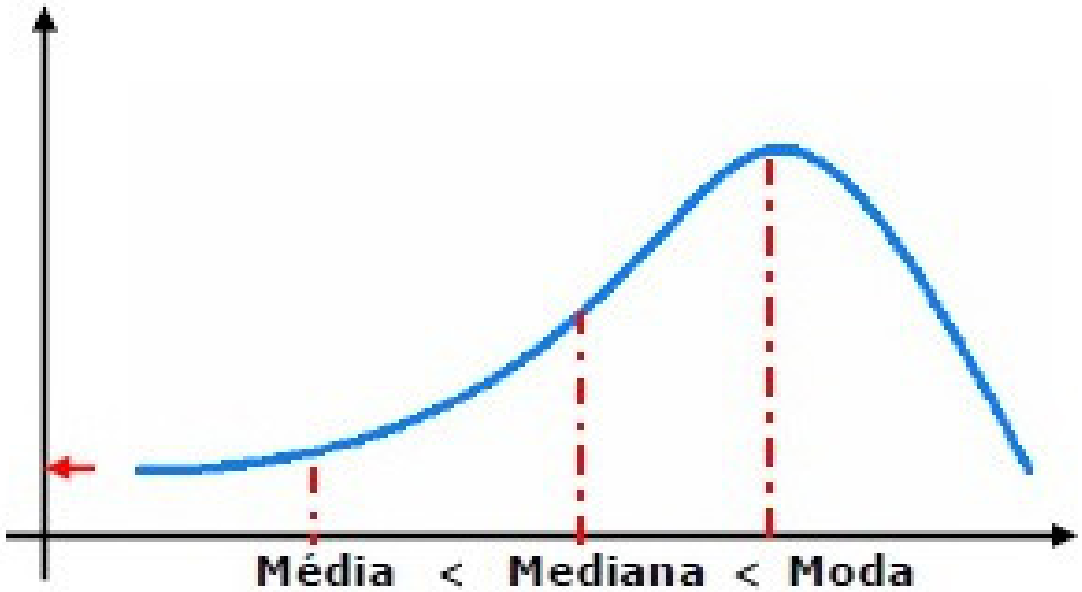
Relação entre a Média, Mediana e Moda:

a) Simetria: $\overline{x} = Md = Mo$



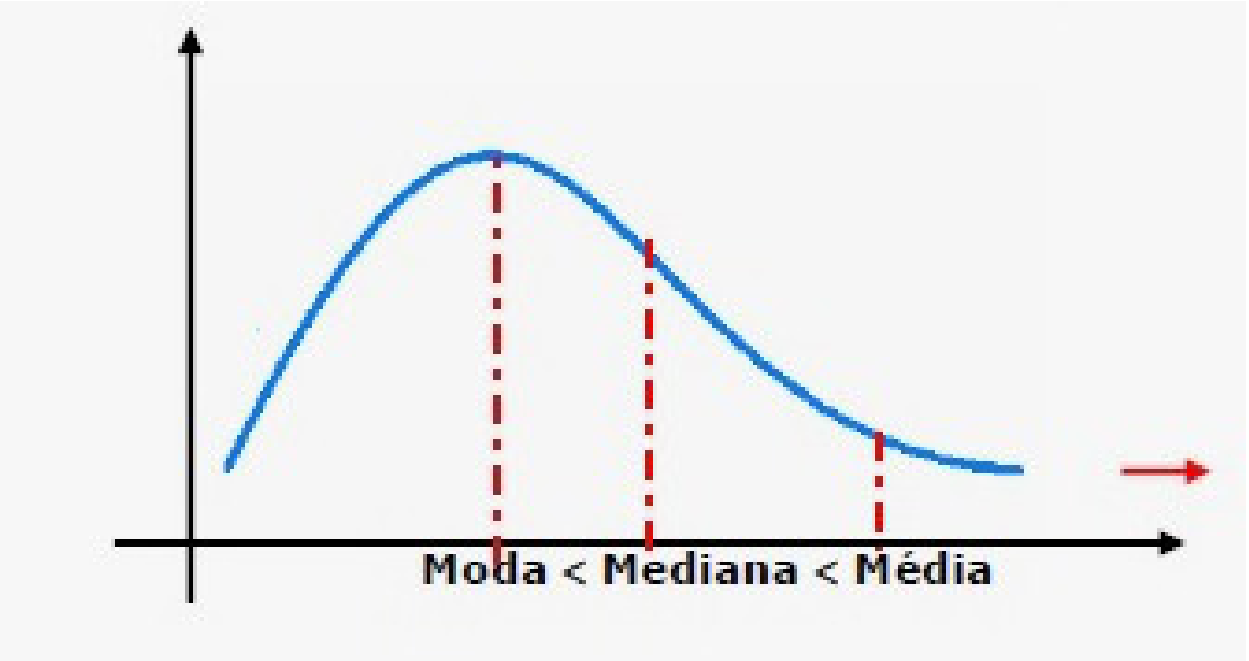
Estatísticas Descritivas Medidas de Posição

b) Assimetria à esquerda: $\overline{x} < Md < Mo$



Estatísticas Descritivas Medidas de Posição

c) Assimetria à direita: $\overline{x} > Md > Mo$



Estatísticas Descritivas Medidas de Posição

- **Quantis (quartil, decil, percentil):** O quantil de ordem p ou o $p-$quantil, indicado por $q(p)$, sendo p é uma proporção $(0 < p < 1)$, é uma medida em que $100p\%$ das observações sejam menores do que $q(p)$.
  Notação: $q(p)$.

a) **Quartis:** divide o rol em 4 partes iguais.
Notação: $Q_1 = q(25\%)$; $Q_2 = q(50\%)$; $Q_3 = q(75\%)$.

b) **Decis:** divide o rol em 10 partes iguais.
Notação: $D_1 = q(10\%)$; $D_2 = q(20\%)$; $\cdots$; $D_9 = q(90\%)$.

c) **Percentis ou Centis:** divide o rol em 100 partes iguais.
Notação: $P_1 = q(1\%)$; $P_2 = q(2\%)$; $\cdots$; $P_{99} = q(99\%)$.

Alguns casos particulares:

$$Q_1 : 1^o\text{quartil} = P_{25} : 25^o\text{percentil}$$

$$Q_2 : 2^o\text{quartil} = Md : \text{mediana} = D_5 : 5^o\text{decil} = P_{50} : 50^o\text{percentil}$$

$$Q_3 : 3^o\text{quartil} = P_{75} : 75^o\text{percentil}$$

$$D_1 : 1^o\text{decil} = P_{10} : 10^o\text{percentil}$$

Estatísticas Descritivas Medidas de Posição

Como encontrar o quantil de ordem p?

Passo 1: Ordenar os dados, ou seja, construir o rol.

Passo 2: Encontrar a posição do $p-$quantil,

$$P = p \cdot n.$$

O quantil de ordem p ou o $p-$quantil será o elemento, de tal forma que

$$q(p) = \begin{cases} \frac{x_{(P)} + x_{(P+1)}}{2}, & \text{se } P \text{ é inteiro} \\ x_{(P^*+1)}, & \text{se } P \text{ não é inteiro} \end{cases}.$$

onde P^* é a parte inteira de P.

Estatísticas Descritivas Medidas de Posição

Ex 1 (Primeiro Quartil, Q_1):
Dados: 3; 2; 8; 5; 3; 7; 3; 9; 2
Rol: 2; 2; 3; 3; 3; 5; 7; 8; 9
$n = 9$ e $p = 0,25$ ou 25%
$P = n \cdot p = 9 \times 0,25 = 2,25 \Rightarrow P^* = 2$
$Q_1 = q(25\%) = x_{(P^*+1)} = x_{(2+1)} = x_{(3)} = 3$

Ex 2: (Terceiro Quartil, Q_3): Dados: 18; 13; 12; 10; 17; 10
Rol: 10; 10; 12; 13; 17; 18
$n = 6$ e $p = 0,75$ ou 75%
$P = n \cdot p = 6 \times 0,75 = 4,5 \Rightarrow P^* = 4$
$Q_3 = q(75\%) = x_{(P^*+1)} = x_{(4+1)} = x_{(5)} = 17$

Exercício

Por exemplo, seja a estatura (em cm) observada em duas amostras de adolescentes saudáveis. Denotamos essas amostras por A e B.

As estaturas dos adolescente da amostra A são: 149, 156, 157, 158, 159, 160, 161 e 164.

As estaturas dos adolescentes da amostra B são: 132, 138, 152, 157, 160, 171, 176 e 178.
Encontre a média e a mediana das duas amostras.

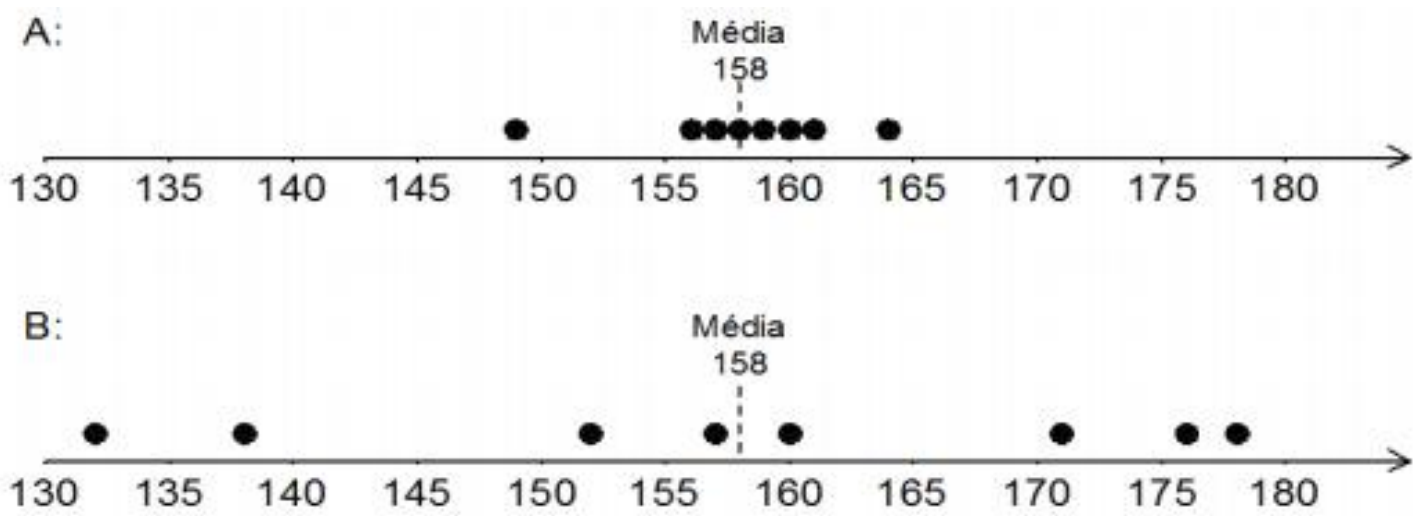
Exercício

- Amostra A: Média 158 e mediana 158,5
- Amostra B: Média 158 e mediana 158,5

Os adolescente da amostra A e B são semelhantes em relação à estatura?

Exercício

Os adolescente da amostra A e B são semelhantes em relação à estatura?



- A amostra A possue uma dispersão menor em relação à média.
- As medidas de posição fornecem um resumo incompleto do comportamento dos dados.
- Torna-se importante a apresentação de medidas de variabilidade dos dados.

Medidas de dispersão

- **Amplitude:** É a diferença entre o maior e o menor valor presentes em um conjunto de dados.
  Notação: A.

$$A = max - min$$

- Quanto maior a amplitude amostral, maior tende a ser a dispersão de nossos dados.

Amplitude

Exemplo: Sejam os tempos de processamento de $n = 13$ aplicativos, em segundos:

$$28, 20, 29, 27, 23, 27, 24, 23, 24, 25, 28, 29, 21$$

Como min = 20 e max = 29, a amplitude amostral é:

$$A = 29 - 20 = 9 \text{ segundos}$$

- Nenhum aplicativo demorou mais tempo que outro com diferença superior a nove segundos.

Amplitude

Observação: A amplitude amostral deve ser utilizada com cautela, pois é fortemente influenciada por valores atípicos.

Exemplo: Suponha que no exemplo anterior tenhamos o tempo de 130 segundos para um 14^o aplicativo (com problemas):

$$28, 20, 29, 27, 23, 27, 24, 23, 24, 25, 28, 29, 30, 130$$

Amplitude amostral é, portanto

$$A = 130 - 20 = 110 \text{ segundos}$$

- **Intervalo ou Amplitude Interquartil:** É a diferença entre o terceiro quartil e o primeiro quartil.
  Notação: d_Q.

$$d_Q = Q_3 - Q_1$$

- **Exemplo:** Sejam o rol dos tempos de processamento de $n = 14$ aplicativos, em segundos:

$$20, 21, 23, 23, 24, 24, 25, 27, 27, 28, 28, 29, 29, 130$$

$n = 14$, $p = 0,25$ e $p = 0,75$
$Q_1 = q(25\%) = x_{(3+1)} = x_{(4)} = 23$
$Q_3 = q(75\%) = x_{(10+1)} = x_{(11)} = 28$

$$d_Q = Q_3 - Q_1 = 28 - 23 = 5$$

Os 50% centrais dos dados estão concentrados dentro de uma amplitude de 5 segundos.

- **Exemplo:** Sejam o rol dos tempos de processamento de $n = 14$ aplicativos, em segundos:

$$20, 21, 23, 23, 24, 24, 25, 27, 27, 28, 28, 29, 29, 130000$$

$n = 14$, $p = 0,25$ e $p = 0,75$
$Q_1 = q(25\%) = x_{(3+1)} = x_{(4)} = 23$
$Q_3 = q(75\%) = x_{(10+1)} = x_{(11)} = 28$

$$d_Q = Q_3 - Q_1 = 28 - 23 = 5$$

Desvio médio

Desvio: A diferença ("distância") entre cada uma das observações amostrais e a média amostral $\overline{x}$.

Os desvios para cada uma das observações x_i são dados por:

$$x_i - \overline{x}$$

- Podemos considerar que a obtenção dessa média consiste em somar os desvios e dividir o resultado pelo tamanho da amostra?

Desvio médio

Desvio: A diferença ("distância") entre cada uma das observações amostrais e a média amostral $\overline{x}$.

Os desvios para cada uma das observações x_i são dados por:

$$x_i - \overline{x}$$

- Podemos considerar que a obtenção dessa média consiste em somar os desvios e dividir o resultado pelo tamanho da amostra?
- Não, pois $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.

Desvio médio: A média dos desvios em valores absolutos.

$$DM = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

Desvio médio

Exemplo: Considere os tempos em segundos de 8 aparelhos: 38, 40, 49, 67, 33, 57, 54 e 64.

x_i	$x_i - \overline{x}$	$\lvert x_i - \overline{x} \rvert$
38	38 - 50,25 = -12,25	12,24
40	40 - 50,25 = -10,25	10,25
49	49 - 50,25 = -1,25	1,25
67	67 - 50,25 = 16,75	16,75
33	33 - 50,25 = -17, 25	17,25
57	57 - 50,25 = 6,75	6,75
54	54 - 50,25 = 3,75	3,75
64	64 - 50,25 = 13,75	13,75

- Desvio médio = 10,25
- O tempo das 8 aparelhos que compõe a amostra distanciam em média 10,25 segundos da média amostral de 50,25.

Variância

Variância: é a divisão entre a soma dos quadrados dos desvios e o tamanho da amostra menos 1.

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

Variância

Variância: é a divisão entre a soma dos quadrados dos desvios e o tamanho da amostra menos 1.

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

- Por que $n-1$ e não n?
- Graus de liberdade: termos independentes contidos na soma
- Podemos obter o resultado dessa soma se estiver "faltando"um termo.

Exemplo: Considere os tempos em segundos de 8 aparelhos: 38, 40, 49, 67, 33, 57, 54 e 64.

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
38	38 - 50,25 = -12,25	150,06
40	40 - 50,25 = -10,25	105,06
49	49 - 50,25 = -1,25	1,56
67	67 - 50,25 = 16,75	280,56
33	33 - 50,25 = -17, 25	297,56
57	57 - 50,25 = 6,75	45,56
54	54 - 50,25 = 3,75	14,06
64	64 - 50,25 = 13,75	189,06

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{1083,5}{7} = 154,8 \text{segundos}^2$$

- O que é um segundo ao quadrado?
- Não tem interpretação. Por isso é conveniente eliminarmos esse termo "ao quadrado".

Desvio padrão

Desvio padrão: é a raiz quadrada da variância.

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

- Quanto maior o desvio padrão, maior é a dispersão dos nossos dados.
- É na mesma unidade de medida da variável orignal.
- No exemplo anterior $s = \sqrt{154,8} = 12,44$ segundos

Coeficiente de variação

Coeficiente de variação: Mede a variabilidade dos dados em termos relativos à média:

$$CV = \frac{s}{\overline{x}} \times 100\%$$

Obs: O coeficiente de variação deve ser computado apenas para dados que assumem apenas valores não negativos.

Coeficiente de variação

O coeficiente de variação é bastante útil quando queremos comparar a variabilidade de duas ou mais variáveis que possuem unidades diferentes, pois é uma medida adimensional.
Exemplo: Os dados da Tabela abaixo referem-se ao peso (em kg) e ao comprimento (em cm) de 10 cabras.

Tabela: Peso e comprimento de 10 cabras.

Cabra	1	2	3	4	5	6	7	8	9	10
Peso	37,5	46,6	42,3	39,4	40,4	47,4	38,3	39,9	32,5	35,9
Comp.	66,7	74,3	71,7	62,9	61,1	69,5	70,9	72,1	61,4	73,4

Qual dessas variáveis tem uma maior dispersão?

Coeficiente de variação

- A média do peso das cabras é 40 kg e o desvio padrão 4,55 kg.
- A média do comprimento é 68,4 cm e o desvio padrão 5,03 cm.

$$CV_{\text{peso}} = \frac{4,55kg}{40,02kg} \times 100\% = 11,4\%$$

$$CV_{\text{comp.}} = \frac{5,03cm}{68,4cm} \times 100\% = 7,3\%$$

O peso das cabras possui uma dispersão maior que o comprimento.

Medida de assimetria de Pearson

Medidas de assimetria de Pearson: Utiliza a relação entre moda e média em distribuições **unimodais**.

$$A_p = \frac{\overline{x} - Mo}{s}$$

- Distribuições simétricas unimodais: $\overline{x}$=Md=Mo; então $A_p = 0$;
- Distribuições assimétricas positivas: $\overline{x}$>Md>Mo; então $A_p > 0$;
- Distribuições assimétricas negativas: $\overline{x}$< Md<Mo, então $A_p < 0$

Coeficiente de assimetria baseado em quartis

Para distribuições simétricas, temos que $(Q_3 - Md) = (Md - Q_1)$

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

- Se a distribuição for simétrica, então $A_Q = 0$,
- Se a distribuição for assimétrica positiva, então $A_Q > 0$,
- Se a distribuição for assimétrica negativa, então $A_Q < 0$

Curtose

Curtose é uma medida de dispersão que caracteriza o "achatamento" da curva da função de distribuição.

$$b_2 = \frac{1}{n}\sum_{i-1}^{n}\left[\frac{x_i - \overline{x}}{s}\right]^4 - 3$$

- Se $b_2 = 0$ chamamos de Mesocúrtica.
- Se $b_2 > 0$ chamamos de leptocúrtica.
- se $b_2 < 0$ chamamos de platicúrtica.

Curtose