Estatística e Probabilidade

Bacharelado em Sistemas de Informação

Aula 2: Resumindo dados: Medidas de Tendência

Prof. Dr. Samuel Sanches

MEDIDAS DE TENDÊNCIA

- ★ Cada caso é um caso!
- ★ Podemos estar satisfeitos com a apresentação utilizando alguma das maneiras apresentadas na Aula 1.
- ★ Porém, muitas vezes, precisaremos resumir ainda mais o nosso conjunto de valores.
- ★ As **medidas de tendência central**, possuem informações a respeito dos dados como um todo, dando um panorama geral que muitas vezes pode revelar informações surpreendentes.

POPULAÇÕES E AMOSTRAS

★ **População**: todas observações possíveis do fenômeno, como todos os estudantes da universidade com idades entre 25 e 35 anos.

★ **Amostra**: conjunto que possui apenas uma parte da população, como todos os estudantes da universidade com idades entre 25 e 35 anos de uma pesquisa que envolva todos os estudantes.

★ Vamos comprar 400 peças de cerâmica e escolhemos 20 para verificar sua resistência, quem é amostra e quem é população? Fornecimento de dezenas de milhares dessas peças, se escolhermos 400 dessas peças, quem é amostra e quem é população?

MÉDIA ARITMÉTICA

- Comumente só média (existem média geométrica e média harmônica), é definida como: **A média de n números é sua soma dividida por n**.

- **Exemplo**: O total de heroína apreendida por várias agências de polícia dos EUA foi de 1794, 3030, 2551, 3514 e 2824 quilogramas, de 1990 a 1994. Encontre a média combinada de heroína apreendida por essas agências no dado período de cinco anos.

  O total é: 1794 + 3030 + 2551 + 3514 + 2824 = 13713 quilogramas, sua média então 13713/5 = 2742,6 quilogramas.

MÉDIA ARITMÉTICA

★ **Exemplo**: No início da primeira sessão anual do congresso, durante oito anos consecutivos havia 67, 71, 78, 82, 96, 110, 104 e 92 congressistas com pelo menos 60 anos de idade. Encontre a média.

O total é: 67 + 71 + 78 + 82 + 96 + 110 + 104 + 92 = 700, sua média então 700/8 = 87,5.

Planilha Aula 2:
https://docs.google.com/spreadsheets/d/1x87nOgoUdY8Bgu1kczaSLr6BfHVj9_tJo4CjIWkXszw/edit?usp=share_link

MÉDIA ARITMÉTICA

- ★ Número de dados de uma amostra (**tamanho da amostra**) denotado por **n**, **valores** da nossa amostra denotados por **x1, x2, x3, …., xn**:

$$\text{média de amostra} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

- ★ Ou (mais formal), temos a média amostral:

$$\bar{x} = \frac{\sum x}{n}$$

- ★ Média populacional, se troca n -> N, tamanho de população:

$$\mu = \frac{\sum x}{N}$$

SOMATÓRIA

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

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$$

$$\sum_{j=1}^{m} x_j y_j = x_1 y_1 + x_2 y_2 + \cdots + x_m y_m$$

$$\sum_{i=1}^{n} x_i^2 f_i = x_1^2 f_1 + x_2^2 f_2 + \cdots + x_n^2 f_n$$

$$\sum_{j=1}^{3}\sum_{i=1}^{4} x_{ij} = \sum_{j=1}^{3}(x_{1j} + x_{2j} + x_{3j} + x_{4j})$$
$$= x_{11} + x_{21} + x_{31} + x_{41} + x_{12} + x_{22} + x_{32} + x_{42}$$
$$+ x_{13} + x_{23} + x_{33} + x_{43}$$

Regra A : $\sum_{i=1}^{n}(x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$

Regra B : $\sum_{i=1}^{n} k \cdot x_i = k \cdot \sum_{i=1}^{n} x_i$

Regra C : $\sum_{i=1}^{n} k = k \cdot n$

MÉDIA ARITMÉTICA

★ Descrições de **amostras**, são referidas como **estatística**, utilizando o alfabeto latino.

★ Descrições de **populações**, são referidas como **parâmetro**, utilizando o alfabeto grego.

MÉDIA ARITMÉTICA

- **Exemplo**: Se as lâmpadas da amostra duram 967, 949, 952, 940 e 922 horas de uso continuado, o que podemos concluir sobre a duração média das 40000 lâmpadas do lote?

  Aqui n = 5, então a média da amostra é:

$$\bar{x} = \frac{967 + 949 + 952 + 940 + 922}{5} = 946 \text{ horas}$$

  Supondo que os dados são uma amostra (ou seja, podemos generalizar), é estimado que as 40000 lâmpadas tem duração média de $\mu = 946$ horas.

MÉDIA ARITMÉTICA

★ **Exemplo**: Se o salário anual médio pago a três jogadores de basquete na temporada 2001-2002 foi de 3.650.000 dólares, pode: **a)** algum deles ter recebido 6.000.000 dólares? **b)** dois deles terem recebido, cada um, 6.000.000 dólares?

O salário combinado dos três jogadores é (**n*média**) = 3*3.650.000 = 10.950.000 dólares, ou seja, os salários somados não podem ultrapassar este valor!

a) Se 1 recebeu 6.000.000, então 10.950.000 - 6.000.000 = 4.950.000 para os outros dois, então é possível.

b) Se 2 receberem 6.000.000, somados temos 12.000.000, que ultrapassa o que foi pago aos três, então não é possível.

MÉDIA ARITMÉTICA

★ **Exemplo**: Se seis alunos fizeram uma média de 57 pontos em um teste, quantos deles, no máximo, podem ter obtido 72 ou mais?

Aqui n = 6 e média = 57, então 6*57 = 342, que pode ser fatorado como 342 = 4*72 + 54, assim no máximo quatro podem ter obtido 72 ou mais.

MÉDIA ARITMÉTICA

- ★ **Centro** do conjunto de dados.
- ★ Pode ser calculada para qualquer conjunto de dados numéricos, s**empre existe**.
- ★ Conjunto de dados tem somente uma média, ela é **única**.
- ★ Média de vários conjuntos de dados, podem ser combinados em uma média global de todos os dados (**média das médias**).
- ★ **Confiável**, médias de amostras repetidas extraídas da mesma população geralmente não flutuam, como outras medidas estatísticas usadas para estimar a média de uma população.
- ★ **Utiliza todos os elementos do conjunto de dados**, salvo os outliers.

MÉDIA ARITMÉTICA

★ **Exemplo**: Um laboratório fez a calorimetria de fatias de pizzas de calabresa grande. Obtendo os valores 265, 332, 340, 225, 238 e 346. **a)** Calcule a média. **b)** Suponha que ao calcular a média, foi cometido um erro, digitando 832, em vez de, 238. Qual será o tamanho do erro cometido?

a) $\bar{x} = \dfrac{265 + 332 + 340 + 225 + 238 + 346}{6}$

$= 291$

b) $\bar{x} = \dfrac{265 + 332 + 340 + 225 + 832 + 346}{6}$

$= 390$

$390 - 291 = 99$

MÉDIA ARITMÉTICA

★ **Exemplo**: Nove alunos que foram a uma excursão, têm idades, 18, 16, 16, 17, 18, 15, 17, 17 e 17 anos e o professor 49. Qual a idade média das dez pessoas que foram na excursão?

$$\bar{x} = \frac{18 + 16 + \cdots + 17 + 49}{10}$$
$$= 20$$

Note que nenhum aluno tem 20 anos, esse valor (um pouco estranho), é devido ao número 49, que para esse conjunto de dados é um **dado estranho**, que pode prejudicar nossas interpretações, por isso temos outras medidas, que juntas nos ajudam a evitar interpretações erradas.

MÉDIA PONDERADA

★ Quando os dados não possuem a mesma importância (ou significância), a média aritmética, pode estar informando um resultado inútil.

★ Alguns casos, nossos dados, possuem um **peso de importância relativa** e então calcular a **média ponderada** dos valores, chamada de xw, dos dados, x1, x2, ... com importância w1, w2, ...

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum w \cdot x}{\sum w}$$

MÉDIA PONDERADA

★ **Exemplo**: Calcular a média combinada dos rebates de jogadores de beisebol, um deles possui média 0,357 (35,7%), outro 0,351, outro 0,350, outro 0,348 e por último 0,341. O número de vezes que cada um rebateu foi 350, 388, 400, 368 e 413.

$$\bar{x}_w = \frac{(350)(0{,}357) + (388)(0{,}351) + \cdots + (413)(0{,}341)}{350 + 388 + 400 + 368 + 413}$$

$$= \frac{670{,}035}{1.919} = 0{,}349$$

GRANDE MÉDIA DE DADOS COMBINADOS

★ Determinados casos se pode calcular a **média global** (ou grande média) de k conjuntos de dados, onde cada possui média x1, x2, ... com n1, n2, … medidas ou observações (n, que é o peso, é o tamanho da amostra), aqui o numerador é o total de todas as medidas e o denominador é o número total de itens nas amostras combinadas:

$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum n \cdot \bar{x}}{\sum n}$$

GRANDE MÉDIA DE DADOS COMBINADOS

★ **Exemplo**: Em uma turma da universidade, há 14 calouros, 25 alunos de segundo e 16 alunos de terceiro ano. Dado que num exame os calouros obtiveram a média 76, os alunos do segundo ano a média 83 e os alunos do terceiro ano a média 89, qual é a grande média para toda a classe?

Aqui, a quantidade de alunos é o tamanho da amostra:

$$\bar{\bar{x}} = \frac{14 \cdot 76 + 25 \cdot 83 + 16 \cdot 89}{14 + 25 + 16}$$

$$= 82,96$$

Arredondando (para manter **coerência** com nossos dados), a média é 83.

MEDIANA

- ★ Uma maneira de contornar problemas que dados grandes ou pequenos podem trazer, utilizamos o valor **central** ou **meio** do conjunto de dados. Para isso é obrigatório a **ordenação** desses valores!

- ★ Vale lembrar, ela **sempre** existe!

- ★ Não temos mediana global.

- ★ **A mediana é o valor do elemento do meio se n for ímpar, e a média dos dois valores do meio se n é par.**

MEDIANA

- ★ **Exemplo**: Uma determinada cidade registrou 14, 17, 20, 22 e 17 arrombamentos em residências. Encontre a mediana desses dados.

  Veja que o meio é 20 (n é ímpar), porém os dados não estão ordenados, então, 14, 17, 17, 20 e 22, assim, encontramos a mediana: **17**.

- ★ **Exemplo**: 12 cursos foram frequentados por 37, 32, 28, 40, 35, 38, 40, 24, 30, 37, 32 e 40 pessoas, encontre a mediana dos dados.

  Ordenando: 24, 28, 30, 32, 32, 35, 37, 37, 38, 40, 40, 40, temos n par, os valores do meio são 35 e 37, assim (35 + 37)/2 = **36**.

POSIÇÃO MEDIANA

- ★ A posição de n dados, é dada por (n + 1)/2

- ★ **Exemplo**: Encontre a posição da mediana para **a)** n = 17 e **b)** n = 41.
  a) (n + 1)/2 = (17 + 1)/2 = 9, mediana é o **9º item**.
  b) (n + 1)/2 = (41 + 1)/2 = 21, mediana é o **21º item**.

- ★ **Exemplo**: Encontre a posição da mediana para **a)** n = 16 e **b)** n = 50.
  a) (n + 1)/2 = (16 + 1)/2 = 8,5, mediana é a **média dos valores do 8º e 9º** itens.
  b) (n + 1)/2 = (50 + 1)/2 = 25,5, mediana é a **média dos valores do 25º e 26º** itens.

MEDIANA

★ **Exemplo**: Na aula 1 obtivemos o diagrama de folhas e ramos para ocupação de quartos em um hotel. A partir dele, encontre a mediana dos valores.

Temos n = 30, posição:
(n + 1)/2 = (30 + 1)/2 = **15,5 (entre 15º e 16º)**

Pelo diagrama 15º –> 53 e 16º –> 54, então:
(53 + 54)/2 = **53,5**

A média que calculamos foi 54,4

3	5	7					
4	0	0	2	3			
4	5	6	6	6	8	9	9
5	2	3	4				
5	5	6	7	8	9		
6	1	2	2	4			
6	9						
7	2	3					
7	8						
8	1						

OUTROS QUANTIS

- A **mediana** é apenas um quantil, ela divide os dados em duas partes iguais.
- **Quartis**: dividir o conjunto de dados em 4 partes.
- **Decis**: dividir o conjunto de dados em 10 partes.
- **Percentis**: dividir o conjunto de dados em 100 partes.
- **Q1**: é a mediana de todos os valores inferiores à mediana de todo o conjunto de dados.
- **Q2**: é a mediana.
- **Q3**: é a mediana de todos os valores superiores à mediana de todo o conjunto de dados.

OUTROS QUANTIS

★ **Exemplo**: Verifique que há tantos valores inferiores a Q1 e a mediana, entre a mediana e Q3 e superiores a Q3 para: **a)** n = 12; **b)** n = 13; **c)** n = 14; **d)** n = 15.

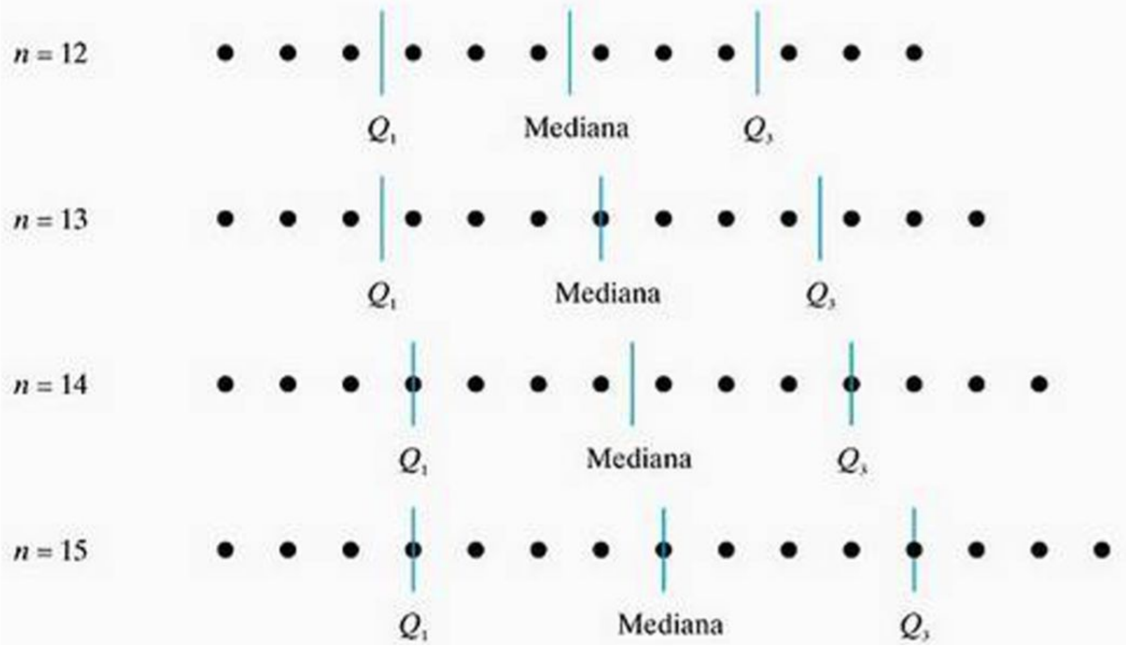


GRÁFICO DE CAIXA

★ As informações que temos com a mediana, Q1, Q3, maior e menor valor, pode ser utilizada para fazer o gráfico de caixa (box plot ou candlestick).

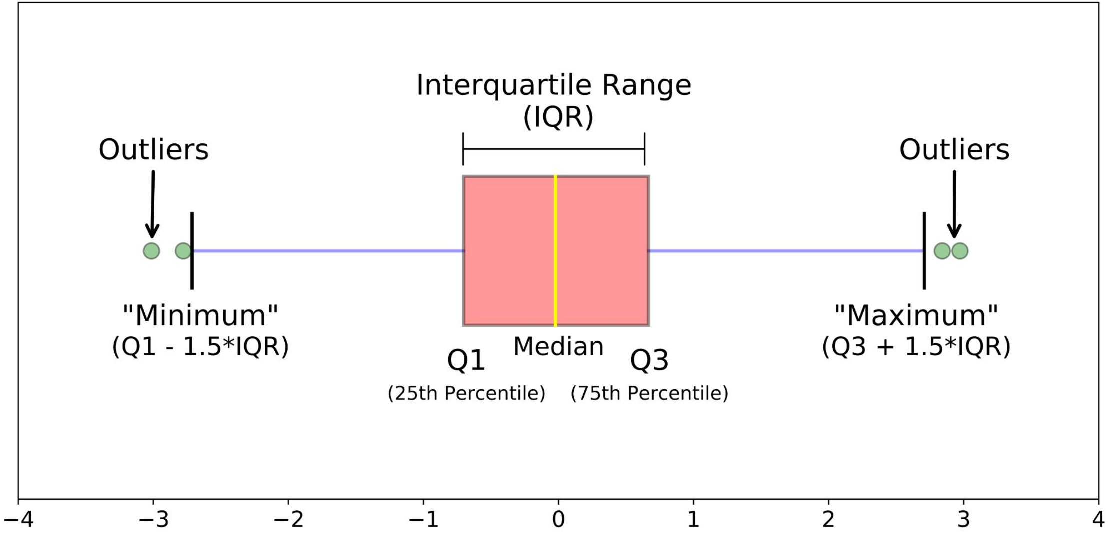


GRÁFICO DE CAIXA

★ **Exemplo**: Na aula 1 obtivemos o diagrama de folhas e ramos para ocupação de quartos em um hotel. Onde a mediana foi 53,5.

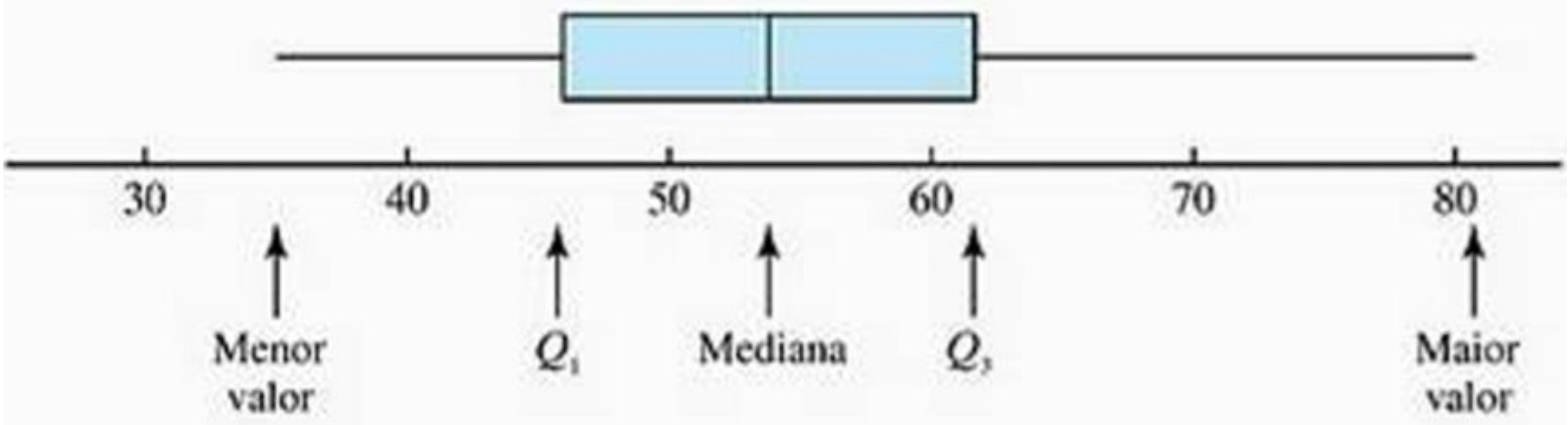
a) Encontre o menor e o maior;
b) Encontre Q1 e Q3;
c) Esboce o gráfico de caixa.

a) Pelo diagrama: Menor = 35; Maior = 81.
b) n = 30, posição mediana = 15,5, para os 15 valores inferiores a 53,5 temos (15 + 1)/2 = 8. Então Q1 é 8º valor = 46, Q3 é o 8º valor a partir da outra extremidade = 62.

3	5	7					
4	0	0	2	3			
4	5	6	6	6	8	9	9
5	2	3	4				
5	5	6	7	8	9		
6	1	2	2	4			
6	9						
7	2	3					
7	8						
8	1						

GRÁFICO DE CAIXA

★ **c)** Menor = 35; Maior = 81; Mediana 53,5; Q1 = 46; Q3 = 62.



Planilha Aula 2:
https://docs.google.com/spreadsheets/d/1x87nOgoUdY8Bgu1kczaSLr6BfHVj9_tJo4CjIWkXszw/edit?usp=share_link

MODA

- ★ Esta é outra medida para nos informar sobre o centro ou meio dos dados, definida como o valor que ocorre com maior frequência e mais de uma vez, caso dois ou mais dados tenham a mesma frequência, todos são moda.

- ★ **Exemplo**: As 21 reuniões de um clube foram frequentadas por 22, 24, 23, 24, 27, 25, 24, 20, 24, 26, 28, 26, 23, 21, 24, 25, 23, 28, 24, 26 e 25. Encontre a moda.

  Os dados 20, 21, 22 e 27 ocorrem só 1 vez, 28 ocorre 2 vezes, 23, 25 e 26 ocorrem, 3 vezes e 24 ocorre 5 vezes, então a **24 é a frequência modal (ou moda) dos dados**.

DADOS AGRUPADOS

★ Muitas vezes teremos somente as frequências dos dados, como vimos na aula anterior. Sendo f a frequência dos n dados, a **Média de Dados Agrupados** é:

$$\bar{x} = \frac{\sum x \cdot f}{n} \qquad \mu = \frac{\sum x \cdot f}{N}$$

DADOS AGRUPADOS

Tempo de espera entre erupções (minutos)	*Freqüência*
30–39	2
40–49	2
50–59	4
60–69	19
70–79	24
80–89	39
90–99	15
100–109	3
110–119	2
	110

★ **Exemplo**: Na aula passada foi montado o agrupamento ao lado, encontre a média dos valores.
Então:

Pontos médios de classe x	*Freqüência* f	$x \cdot f$
34,5	2	69,0
44,5	2	89,0
54,5	4	218,0
64,5	19	1.225,5
74,5	24	1.788,0
84,5	39	3.295,5
94,5	15	1.417,5
104,5	3	313,5
114,5	2	229,0
	110	8.645,0

$$\bar{x} = \frac{8.645,0}{110} = 78,59$$

O erro devido ao agrupamento em classes, pode ser verificado obtendo a média a partir dos dados originais, que é 78,27, assim o **erro de agrupamento** nesse caso é 78,59 - 78,27 = 0,27 (pequeno).

DADOS AGRUPADOS

★ A **mediana** pode ser obtida a partir de um histograma, sendo que a área total dos retângulos à sua esquerda é igual à área total dos retângulos à sua direita.

★ Podemos escrever, L é a fronteira inferior da classe onde a mediana cairá, f a frequência, c o seu intervalo de classe e j o número de itens que ainda faltam para atingir L, assim a **Mediana de Dados Agrupados** é:

$$\tilde{x} = L + \frac{j}{f} \cdot c$$

★ O mesmo princípio vale para os outros quartis, decis e percentis.

DADOS AGRUPADOS

Tempo de espera entre erupções (minutos)	*Freqüência*
30–39	2
40–49	2
50–59	4
60–69	19
70–79	24
80–89	39
90–99	15
100–109	3
110–119	2
	110

★ **Exemplo**: Na aula passada foi montado o agrupamento ao lado, encontre a mediana dos valores.

Total 110, então 110/2 = 55, contar 55 dos itens de alguma das extremidades (vamos fazer com os menores), então, 2 + 2 + 4 + 19 + 24 = 51 caem nas primeiras 5 classes, assim, 55 - 51 = 4 (i) valores a mais dentre os que estão na sexta classe. Então os 39 (f) valores restantes que estão distribuídos igualmente, somamos 4/39 do intervalo da classe que é 10 (c), a sua fronteira inferior (6º classe) que é 79,5 (L)

$$\tilde{x} = 79{,}5 + \frac{4}{39} \cdot 10 = 80{,}53$$

DADOS AGRUPADOS

Tempo de espera entre erupções (minutos)	*Freqüência*
30–39	2
40–49	2
50–59	4
60–69	19
70–79	24
80–89	39
90–99	15
100–109	3
110–119	2
	110

- **Exemplo**: Ainda com os dados, encontre Q1, Q3, D2 e P5

Total 110, então 110/4 = 27,5, então 2 + 2 + 4 + 19 = 27 nas 4 primeiras classes, sobra 27,5 - 27 = 0,5 dos 24 valores da 5º classe, então:

$$Q_1 = 69{,}5 + \frac{0{,}5}{24} \cdot 10 \approx 69{,}71$$

Total 110, então 110/4 = 27,5, então (maior para menor) 2 + 3 + 15 = 20 nas 3 últimas classes, sobra 27,5 - 20 = 7,5 dos 39 valores da 6º classe, então:

$$Q_3 = 89{,}5 - \frac{7{,}5}{39} \cdot 10 \approx 87{,}58$$

DADOS AGRUPADOS

Tempo de espera entre erupções (minutos)	*Freqüência*
30–39	2
40–49	2
50–59	4
60–69	19
70–79	24
80–89	39
90–99	15
100–109	3
110–119	2
	110

★ **Exemplo**: Ainda com os dados, encontre Q1, Q3, D2 e P5

Total 110, então 110*(2/10) = 22, então 2 + 2 + 4 = 8 nas 3 primeiras classes, sobra 22 - 8 = 14 dos 19 valores da 4º classe, então:

$$D_2 = 59,5 + \frac{14}{19} \cdot 10 \approx 66,87$$

Total 110, então 110*(5/100) = 22, então 2 + 3 + 15 = 20 nas 3 últimas classes, sobra 22 - 20 = 2 dos 39 valores da 6º classe, então:

$$P_8 = 89,5 - \frac{2}{39} \cdot 10 \approx 88,99$$

EXERCÍCIOS

★ **Lista 1 de Exercícios → Parte 1**

https://drive.google.com/file/d/1f755nMDqpEIkcHnFF49CjbAfw8CvWlCS/view?usp=share_link

- Muito obrigado pela atenção!