Medidas Numéricas Descritivas:

Medidas de centro

A notação para variáveis e para a soma

O valores de uma variável x em um conjunto de dados pode ser representado simbolicamente por:

$$x_1, x_2, x_3, \ldots, x_n$$

onde n é o número de medidas sobre a variável x no conjunto dos dados.

Os valores x_1 , x_2 , ……, x_n representam a 1ª observação,

a 2ª observação, …, n-ésima observação.

notação de Somatória $\sum$:

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

soma de todos os x com i variando de 1 até n

Exemplo:

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

$$\sum_{i=1}^{4} (2 - x_i) = (2 - x_1) + (2 - x_2) + (2 - x_3) + (2 - x_4)$$

se $x_1 = 2{,}5 \quad x_2 = 2{,}0 \quad x_3 = 3{,}2 \quad x_4 = 2{,}5 \quad x_5 = 3$

temos: $\sum_{i=1}^{3} x_i = 5{,}7$ e $\sum_{i=1}^{4} (2 - x_i) = -2{,}2$

propriedade básicas da somatória :

se a e b são constantes

$$\sum_{i=1}^{n} a = na$$

$$\sum_{i=1}^{n} bx_i = b \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (bx_i + a) = \sum_{i=1}^{n} bx_i + na$$

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} x_i^{\ 2} - 2a \sum_{i=1}^{n} x_i + na^2$$

Exercício: Tente mostrar a propriedades acima usando definição de somatória

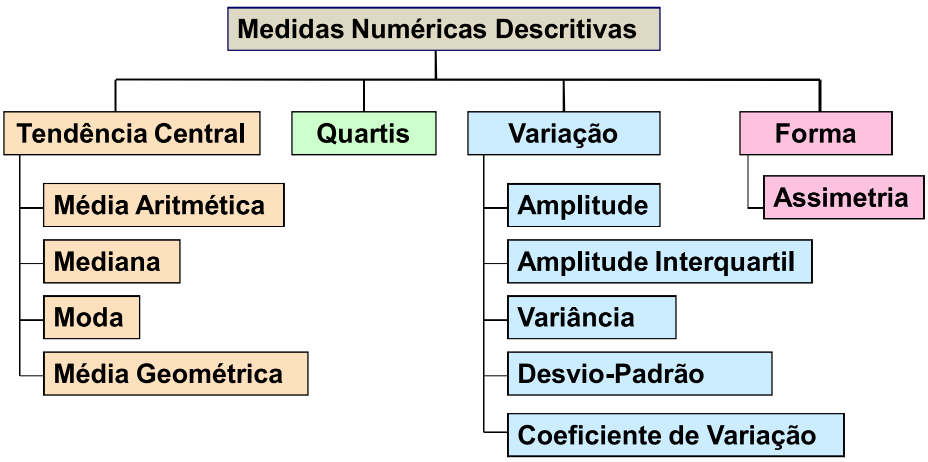
Observação:

Se a somatória se estender a **todos os valores dos dados** é comum não colocar o índice na variável e os limites da soma, ou seja,

$$\sum_{i=1}^{n} x_i = \sum x$$

Se n for o número de elementos do conjunto de dados

Medidas Numéricas Descritivas

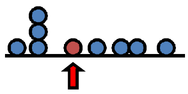


Medidas de Tendência Central

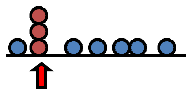
Tendência Central

- **Média Aritmética**
- **Mediana**
- **Moda**
- **Média Geométrica**

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$



Ponto médio dos valores classificados



Valor observado com maior frequência

$$\bar{X}_G = (X_1 \times X_2 \times \Lambda \times X_n)^{1/n}$$

Média aritmética ($\bar{x}$)

A média aritmética é a medida de tendência central mais usual.

Para uma amostra de tamanho n:

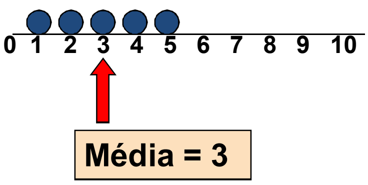
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

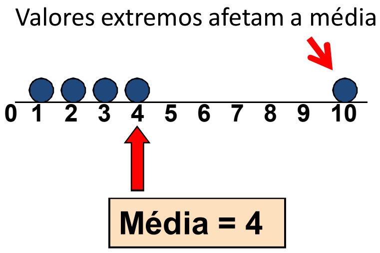
Tamanho da amostra

Valores observados

Média aritmética ($\overline{x}$)

- Média = soma dos valores dividida pelo total de termos
- Afetada por valores extremos (outliers)





$$\overline{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\overline{x} = \frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Média para dados tabulados

lembrando que $\sum_{i=1}^{n} f_i = n$, onde f_i a frequência do valor x_i

podemos calcular a média pela expressão :

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

ou em termos da frequencia relativa f_{ri}, ou da porcentagem $(f_r\,\%)_i$:

$$\overline{x} = \sum_{i=1}^{n} f_{ri} x_i$$

onde usamos o fato que $\sum_{i=1}^{n} f_{ri} = 1$

Exemplo: 24, 24, 24, 27, 27, 27, 27, 30, 30, 38

Tabela de frequências

x_i	f_i
24	3
27	4
30	2
38	1
total	10

$$\text{temos}: \bar{x} = \frac{\sum_{i=1}^{4} f_i x_i}{\sum_{i=1}^{4} f_i} = \frac{3.24 + 4.27 + 2.30 + 1.38}{10}$$

$$\text{e portanto} \quad \bar{x} = 27{,}8$$

Média para dados agrupados em classes

Neste caso utiliza-se a frequências das classes e o valor x_i da variável é substituido pelo ponto médio PM_i de cada classe, ou seja,

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i \, . \, PM_i}{\sum_{i=1}^{n} f_i}$$

n: número de classes

f_i: frequência na classe i

PM_i: ponto médio da classe i

Ou em termos da frequencia relativa fr (ou da fr%):

$$\bar{x} = \sum_{i=1}^{K} f_{ri} \, . \, PM_i$$

Exemplo: vamos usar os dados sobre a temperatura (utilizados das aulas anteriores)

classes	PM	f
10 |- 20	15	3
20 |- 30	25	6
30 |- 40	35	5
40 |- 50	45	4
50 |- 60	55	2
total		20

Temos 5 classes, então:

$$\bar{x} = \frac{\sum_{i=1}^{5} f_i \, . \, PM_i}{20} = \frac{3.15 + 6.25 + 5.35 + 4.45 + 2.55}{20}$$

$$\boxed{\bar{x} = 33,5}$$

Note que se calcularmos a média usando a soma sobre todos os valores vamos obter: $\bar{x}$ = 32,4

A diferença ocorre pois ao agruparmos os dados em classes perdemos informação sobre os valores nas classes, que são substituidas pelo ponto médio

Média ponderada: a idéia de classes pode ser generalizada para incluir grupos ou categorias onde a média das categorias correspondem a uma média ponderada da média de cada grupo ou categoria

Considere k- grupos ou categorias

n_i : o número de elementos (frequência) do grupo ou categoria i.

$\bar{x}_i$: média na categoria i

média ponderada (média das médias) : $$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$

No lugar de n_i pode ser usado a frequência relativa (fr) ou a porcentagem (fr%) em cada grupo ou categoria

Exemplo: a tabela mostra o número de indivíduos em cada grupo de aplicações financeiras e valor médio aplicado (em milhares de reais) por grupo.

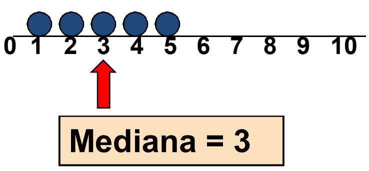
Aplicação	No de pessoas (ni)	Média de R$ no grupo (xi)
poupança	83	320
ações	26	800
Renda fixa	47	550
total	**156**	

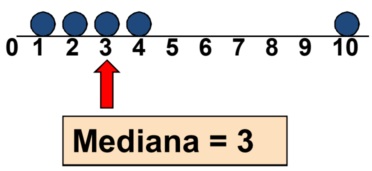
média geral de dinheiro aplicado corresponde a méda ponderada:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{3} n_i \bar{x}_i}{\sum_{i=1}^{3} n_i} = \frac{83.320 + 25.800 + 47.550}{156} = 464{,}16$$

Mediana (Md)

Em uma disposição ordenada, a mediana é o valor do meio (50% dos valores é maior, 50% é menor)





- Não é afetada por valores extremos

Encontrando a Mediana

Dado um conjunto de valores x1, x2, x3….xn:

(1) Ordene os valores
(2) Se o número de valores é ímpar, a mediana é o elemento central
(3) Se o número de valores é par, a mediana é a média dos dois valores centrais

Exemplo (**número par de dados**): 27, 24, 24, 38, 30, 30, 32, 24, 31,27

Encontrando a mediana:

Ordenar os dados: 24, 24, 24, 27, 27, 30, 30, 31, 32, 38

A posição da mediana:

$$n = 10 \text{ é par : a mediana é média dos valores centrais.}$$

Temos: 24 24 24 27 27 30 30 31 32 38

50% dos dados 50% dos dados

$$\boxed{\mathrm{Md} = \frac{27+30}{2} = 28{,}5}$$

Exemplo (**número ímpar de dados**): 27, 24, 24, 38, 30, 30, 32, 24, 31

Encontrando a mediana:

Ordenar os dados: 24, 24, 24, 27, 30, 30, 31, 32, 38

A posição da mediana:

$$n = 5 \quad \text{é impar : a mediana é o elemento central}$$

Temos: 24 24 24 27 30 30 31 32 38

50% dos dados 50% dos dados

$$\boxed{\mathrm{Md} = 30}$$

Mediana para dados agrupados: a frequência acumulada dos dados pode ajudar.
Basta localizar o elemento cuja a frequência acumulada superar pela primeira vez 50% do número de elementos

xi	fi	fA
24	3	3
27	4	7
30	2	9
38	1	10
total	10	

Vamos fazer um exemplo: no caso temos n=10 e a mediana está na posição 5,5. Logo a mediana é um elemento que está na segunda frequência acumulada.
O 5 elemento é 27, o sexto elemento é 27 também. Logo a mediana é o ponto médio destes dois valores, ou seja,

Md=27

Mediana para dados agrupados em classes:

$$Md = l + \frac{h\left(\frac{f_T}{2} - f_{Aant}\right)}{f_{md}}$$

$l =$ limite inferior da classe da mediana

$h =$ intervalo da classe

$f_T =$ frequencia total

$f_{Aant} =$ freq. acumulada da classe anterior a da classe da mediana

$f_{md} =$ frequência na classe da mediana

As frequencias podem ser substituidas pela frequência relativa (fr) ou pela porcentagem (fr%) .

Exemplo:

classes	PM	f	fA
10 |- 20	15	3	3
20 |- 30	25	6	9
30 |- 40	35	5	14
40 |- 50	45	4	18
50 |- 60	55	2	20
total		20	

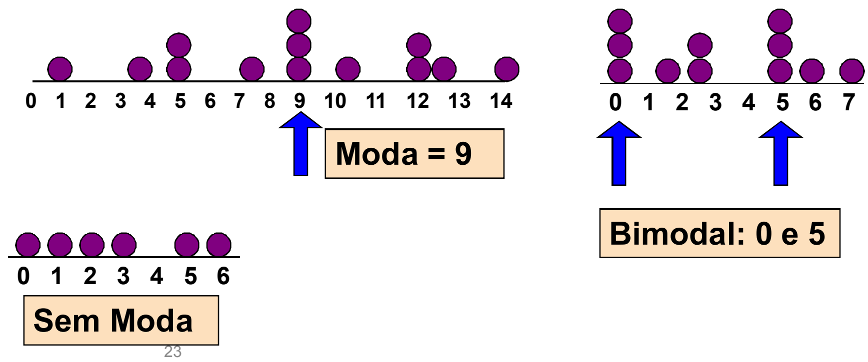
Como n=20 a mediana estara na posição 10,5 (onde se acumulam metade dos dados) e assim no intervalo de classe [30,40)

$$Md = l + \frac{h\left(\frac{f_T}{2} - f_{Aant}\right)}{f_{md}} = 30 + \frac{10\left(\frac{20}{2} - 9\right)}{5}$$

$$Md = 33$$

Moda: Valor que ocorre com mais frequência

- Não é afetada por valores extremos
- Utilizada para dados numéricos ou categóricos
- Pode ser que não exista moda
- Pode ser que existam várias modas (multimodal)



Outras medidas de centro:

Média Geométrica

$$\overline{X}_G = (X_1 \times X_2 \times \Lambda \times X_n)^{1/n}$$

Média Harmônica:

$$\overline{X}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

Ponto Médio (PM):

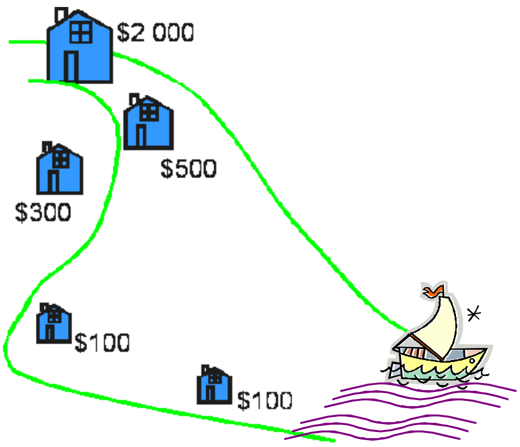
$$AM = \frac{\text{maior valor} + \text{menor valor}}{2}$$

Exemplo

- Cinco casas em um morro perto da praia

Preços das casas:

$2.000.000
$500.000
$300.000
$100.000
$100.000



Exemplo

Preços das casas:

$2.000.000
$500.000
$300.000
$100.000
$100.000

Soma $3.000.000

- **Média:** ($3.000/5) = **$600.000**
- **Mediana:** valor do meio = **$300.000**
- **Moda:** valor mais frequente = **$100.000**

Não existe uma medida de centro que seja melhor que a outra. O problema em estudo é que poderá definir qual a mais adequada. Em muitas situações, elas podem ser complementares.

- A **Média** é a mais utilizada, mas pode não ser adequada em casos nos quais existem valores extremos
- A **Mediana** pode ser mais adequada quando existem valores extremos

Note que: localizar valores extremos (discrepantes) é importante para verificar se estes são erros de medida ou de digitação nos dados. Neste caso podemos eliminá-los.

Caso contrário, devemos mantê-los como parte dos dados serem estudados. Neste caso, pode ser útil medidas menos sensíveis a eles, tal como a mediana.

Tempo de Espera de Clientes em Diferentes Bancos

em minutos

Banco A	6,5	6,6	6,7	6,8	7,1	7,3	7,4	7,7	7,7	7,7
Banco B	4,2	5,4	5,8	6,2	6,7	7,7	7,7	8,5	9,3	10,0

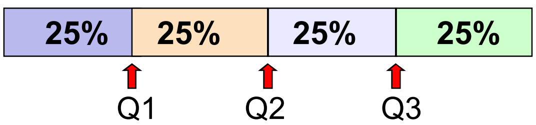
	Banco A	Banco B
Média	7.15	7,15
Mediana	7.20	7,20
Moda	7.7	7,7
Ponto médio	7.10	7,10

Já que as medidas centrais são a mesmas como diferenciar os bancos?

Medidas de ordenamento

Quartis

São os valores que dividem os **dados ordenados** em 4 partes iguais.



- 25% das observações são menores e 75% são maiores para o 1º quartil, Q_1
- Q_2 é o mesmo que a mediana (50% são menores, 50% são maiores)
- Apenas 25% das observações são maiores que 3º quartil

Como obter os quartis

(1) Ordenar os dados

(2) Posição dos quartis:

Posição 1º Quartil: $\mathbf{L_1 = n/4}$

Posição 2º Quartil: $\mathbf{L_2 = n/2}$

Posição 3º Quartil: $\mathbf{L_3 = 3n/4}$

onde **n** é o número de valores observados.

Note que L é a posição do quartil. O quartil corresponde ao elemento na posição L.

(3) Adotaremos a seguinte aproximação:

L inteiro: *O quartil sera a média entre o elemento na posição L e o elemento na posição posterior.*

L não inteiro: *arredonde L para o próximo inteiro. O quartil será o elemento nessa posição.*

Podemos definir de modo geral o número de partes que divimos os dados ordenados como quantis:

Temos 3 quartis dividindo os dados em 4 partes:
Q1, Q2, Q3

Temos 9 decis dividindo os dados em 10 partes:
Q1, Q2, Q3, Q4, ….., Q9

Temos 99 percentis dividindo os dados e 100 partes
Q1, Q2,......, Q97, Q98, Q99

E de forma geral:

Temos (n-1) quantis dividindo os dados em n partes.

Cálculo dos quantis (fórmula geral)

$$\boxed{\begin{aligned} Q_{n_q} &= x_{L_q} \\ L_q &= \frac{n_q . n}{N_q} \end{aligned}}$$

Q_{n_q} : quantil que se deseja obter

n_q : número do quantil (1, 2,3,..., (n-1))

n : tamanho da amostra

x : elemento dos dados ordenados

N_q : número de *divisões* dos dados

L_q : Localização do quantil

A fórmula acima é lida assim:

O quantil Q_{n_q} de número n_q é o valor x dos dados ordenados ocupando a posição $L_q = \frac{n_q . n}{N_q}$.

E aproximaremos:

L inteiro: O quantil sera a média entre o elemento na posição L e o elemento na posição posterior.

L não inteiro: arredonde L para o próximo inteiro. O quantil será o elemento nessa posição.

Quartis

- Exemplo: **12 11 17 14 16 16 18 21 22**

Disposição ordenada : 11 12 14 16 16 17 18 21 22

$$1^o \text{ quartil}: \quad Q_1 = x_{\left(\frac{1.9}{4}\right)} = x_{2,25}$$

temos que encontrar o elemento de valor x com localização L = 2,25.
L = 2,25 não é inteiro, pela nossa convesão Q1 é o elemento na próxima localização inteira L = 3. Nos dados ordenados temos:

$$Q1 = 14$$

2^o quartil (mediana): $Q_2 = x_{\left(\frac{2.(9)}{4}\right)} = x_{4,5}$

temos que encontrar o x na posição $L = 4,5$ que não é inteiro.
Tomando a próxima posição inteira $L = 5$ cujo elemento é:
$Q_2 = 16$

3^o quartil (mediana): $Q_3 = x_{\left(\frac{3.9}{4}\right)} = x_{6,75}$

temos que encontrar o x na posição localizada em $L = 6,75$.
que não é inteiro. Tomando a próxima posição inteira $L = 7$
cujo elemento é: $Q_3 = 18$