# Matemática Computacional

## Teórica 4

Departamento de Matemática
Instituto Superior de Engenharia do Porto

2º Semestre 21-22

# Estatística descritiva-revisões

1 Conceitos gerais

2 Organização e Apresentação dos dados

3 Medidas Descritivas

4 Exercícios propostos

## Estatística

É uma área científica que se ocupa da obervação de um fenómeno, recolha de informação ou da produção da informação relevante, com o objetivo de a descrever e modelar para depois inferir e predizer, servindo-nos de guia em situações de incerteza.

## Estatística descritiva

É um conjunto de métodos com o objetivo de apresentar, analisar e interpretar um conjunto de dados com a utilização de ferramentas adequadas.

## Estatística inferencial

É um conjunto de métodos que permitem fazer estimativas e tirar conclusões sobre uma população a partir da informação contida num subconjunto dessa população (amostra).

## População

É o conjunto de todos os objetos (indivíduos) com uma característica comum

## Amostra

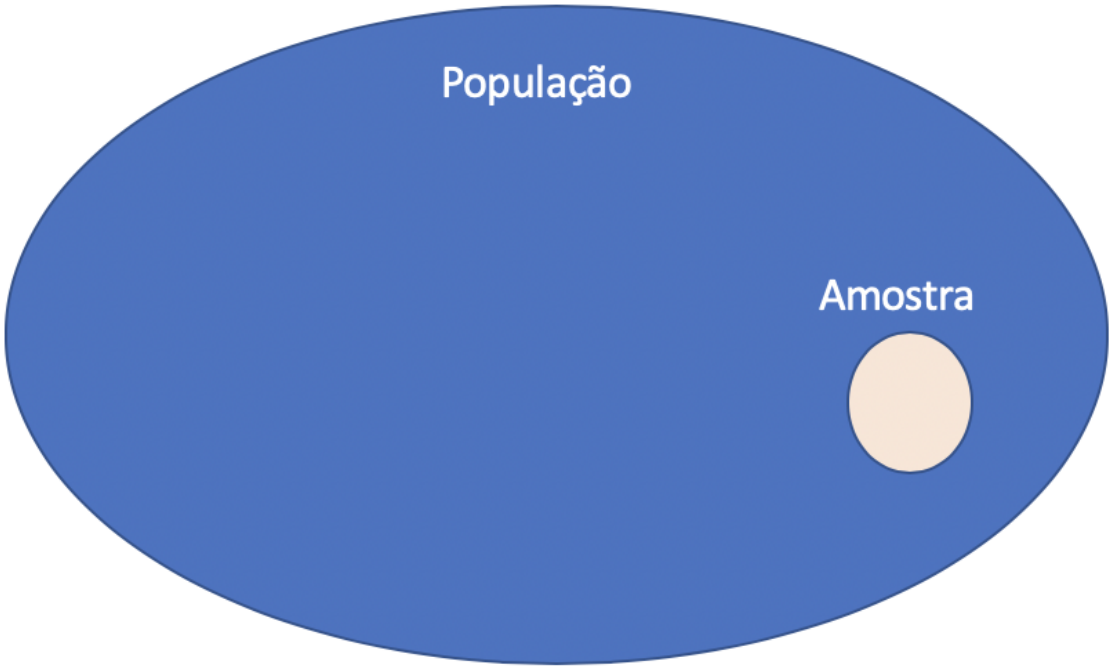
É um subconjunto finito da população que se supões representativo.

## Unidade estatística

Um elemento da população.

Exemplo:

- População: Conjunto dos alunos do ISEP
- Amostra: Conjunto dos alunos de uma turma de Matcp



Relativamente ao tamanho as populações podem ser:

- Finitas (número de filhos, Alunos do ISEP, ...)
- Infinitas (número de lançamentos de uma moeda até obter uma cara, tempo de espera por um autocarro)

**Censo**

É o estudo de todos os elementos de uma população.

**Sondagem**

É um estudo (da população) efetuado a partir da informação recolhida de uma amostra.

## Variável estatística

É uma variável que representa uma característica da população e passível de tomar diversos valores. Pode ser qualitativa ou quantitativa.

## Dado estatístico

É um valor particular de uma variável estatística

# Tipos de variáveis

## Variáveis qualitativas

São variáveis que traduzem uma característica não numérica

- variáveis nominais (dados nominais, categorias)
  exemplo:sexo, cor dos olhos, raça,...
- variáveis ordinais (dados ordinais).
  Exemplo: escolaridade (básica, secundária, superior), qualidade (má, média, boa)

# Tipos de variáveis

## Variáveis quantitativas (dados quantitativos)

São variáveis que resultam de processos de medição ou contagem

- Variáveis discretas (dados discretos)
  Podem tomar um conjunto numerável ou infinitamente numerável de valores.
- variáveis contínuas (dados contínuos)
  Podem tomar quaisquer valores num dado intervalo (podem assumir um conjunto não numerável de valores)

# Frequências

## Classe de uma variável estatística

É cada um dos diferentes valores que a variável pode tomar (dados quantitativos discretos), qualquer intervalo de valores (dados contínuos) ou categoria (dados qualitativos).

## Frequência absoluta da classe i $(n_i)$

É o número de observações associadas à classe $i$

$$\sum_{i=1}^{c} n_i = n$$

onde, $n$ o número total de observações e $c$ é o número de classes.

# Frequências

**Frequência relativa da classe i ($f_i$)**

$$f_i = \frac{n_i}{n}$$

Verifica-se que:

$$\sum_{i=1}^{c} f_i = 1$$

é usual exprimir-se em percentagem.

# Frequências

## Frequência acumulada até à classe $i$ ($N_i$)

É o número de observações de valor inferior ou igual ao valor que caracteriza a classe $i$.

$$N_i = \sum_{j=1}^{i} n_j.$$

# Frequências

## Frequência acumulada relativa até à classe $i$ ($F_i$)

É o número de observações de valor inferior ou igual ao valor que caracteriza a classe $i$.

$$F_i = \sum_{j=1}^{i} f_j.$$

# Dados quantitativos discretos

Dados quantitativos discretos são observações de variáveis quantitativas discretas, geralmente associadas a processos de **contagem**.

**Tabela de distribuição de frequências de dados quantitativos discretos**

É uma tabela com $k$ linhas (uma para cada valor distinto) e três colunas. A primeira coluna representa cada uma dos diferentes valores observados, a segunda a frequência absoluta e a terceira a frequência relativa.

# Dados quantitativos discretos

Exemplo: Observação de 50 peças de artesanato quanto ao número de defeitos

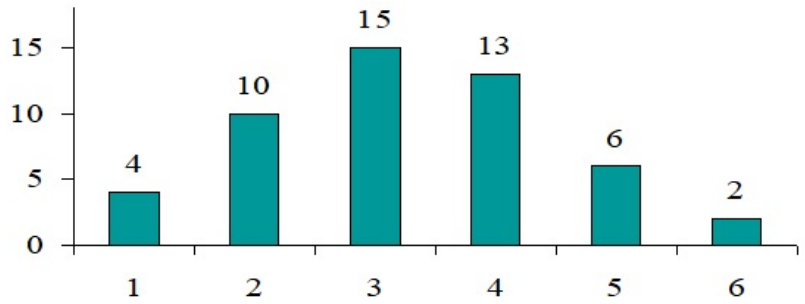
| Número de defeitos $(x_i)$ | frequência absoluta $(n_i)$ | frequência relativa $(f_i)$ |
|---|---|---|
| 0 | 23 | 0.46 |
| 1 | 16 | 0.32 |
| 2 | 7 | 0.14 |
| 3 | 4 | 0.08 |

Exemplo: Dados de um inquérito realizado em 50 habitações quanto ao número de elementos do agregado familiar.

Tabela de frequências

| Household members (xi) | Absolute frequency (ni) | Relative frequency (fi) | Cumulative relative frequency (Fi) |
|---|---|---|---|
| 1 | 4 | 0,08 | 0,08 |
| 2 | 10 | 0,2 | 0,28 |
| 3 | 15 | 0,3 | 0,58 |
| 4 | 13 | 0,26 | 0,84 |
| 5 | 6 | 0,12 | 0,96 |
| 6 | 2 | 0,04 | 1 |
| | | | |
| Total | 50 | 1 | |

Exemplo: Cont.
Gráfico de barras



# Dados quantitativos contínuos

- Os dados quantitativos contínuos são obtidos a partir da observação de uma variável numérica contínua de uma população.
- Os dados de uma variável contínua apresentam uma diversidade tal que é necessário agrupa-los em **classes** (intervalos).

# Dados quantitativos contínuos

## Classe de uma variável quantitativa contínua

É um intervalo na forma $[a, b[$ ou $]a, b]$ que representa um conjunto de valores que a variável pode tomar.

- **Amplitude da classe** $h$: $h = b - a$
- **Marca da classe** é o representante dessa classe para efeitos de cálculo.
  Por defeito considera-se $marca = x_i = \frac{a+b}{2}$

# Dados quantitativos contínuos

## Número de classes

Número de classes É habitual escolher entre 4 e 20 classes, dependendo do número de observações. Regra de Sturges: $c = int(1 + 3.3log(n))$.

- Amplitude aproximada da classe (aproximada)

$$\text{Amplitude classe} = \frac{\text{maior obs.} - \text{menor obs.}}{\text{nº de classes} = c}$$

- Limite inferior da 1ª classe - é qualquer número inferior ou igual à menor das observações

# Dados quantitativos contínuos

Exemplo: Dados observados de tempos de realização de 34 tarefas, em minutos.
Tabela de frequências

| amplitude classe ($h_i$) | Tempo classe $i$ | Marca da classe($x_i'$) | Freq. abs. ($n_i$) | Freq. rel. ($f_i$) | Freq. rel. acumulada ($F_i$) |
|---|---|---|---|---|---|
| 2 | [5, 7[ | 6 | 5 | 0.1471 | 0.1471 |
| 2 | [7, 9[ | 8 | 13 | 0.3824 | 0.5295 |
| 2 | [9, 11[ | 10 | 8 | 0.2353 | 0.7648 |
| 4 | [11, 15] | 13 | 8 | 0.2353 | 1.000 |
| | | Total | 34 | 1.000 | |

# Dados quantitativos contínuos

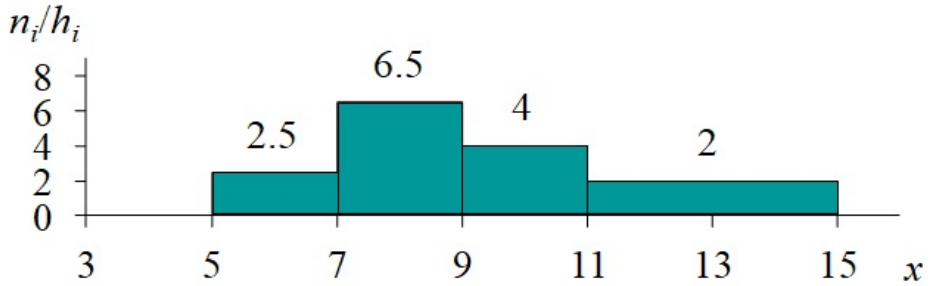
Representação gráfica

## Histograma

É uma representação gráfica dos dados em que se marcam as classes no eixo horizontal, as frequências no eixo vertical e em que se usam barras de área proporcional à frequência da classe correspondente. As barras contíguas têm uma fronteira comum.

# Dados quantitativos contínuos

Na figura seguinte representa-se um histograma referente ao tempo de realização de 34 tarefas, em minutos. Note-se que $h_i$ e $n_i$ representam, respetivamente, a amplitude e a frequência absoluta da classe $i : i = 1, 2, 3, 4$.



# Medidas de localização

## Média aritmética $\bar{x}$ para dados não classificados

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

## Média aritmética $\bar{x}$ para dados classificados

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{c} x_i n_i = \sum_{i=1}^{c} x_i f_i$$

onde $c$ é o número de classes, $n$ o número total de observações, $n_i$ é a frequência absoluta e $f_i$ a frequência relativa. No caso de dados contínuos classificados $x_i$ representa a marca da classe.

# Medidas de localização

## Definição: Moda, $Mo$

É a classe (ou classes) com maior frequência. Para dados discretos, ou contínuos não classificados, é o valor (ou valores) que apresenta(m) a maior frequência.

## Mediana, $Me$

É o valor que divide uma série de $n$ observações em duas partes iguais, tal que 50% das observações tem um valor inferior, ou igual, a $Me$

# Medidas de localização

- Mediana para dados não classificados em série ordenada

$$Me = \begin{cases} \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & , n \text{ par} \\ x_{(n+1)/2} & , n \text{ impar} \end{cases}$$

- Mediana para dados quantitativos contínuos classificados

$$Me = L_k + \left( \frac{0.5 - F_{k-1}}{f_k} \right) h_k$$

  - k - classe mediana (1ª classe onde se verifica $F_k \geq 50\%$).
  - $f_k$ - frequência relativa da classe mediana.
  - $F_{k-1}$ - frequência acumulada relativa da classe anterior.
  - $h_k$ - amplitude da classe mediana.
  - $L_k$ - limite inferior da classe mediana.

# Medidas de localização

- Moda para dados quantitativos contínuos classificados
  - Mo- identificada pela classe modal (classe de maior freq absoluta/relativa).
  - Mo - ponto médio da classe modal
  - Fórmula de Czuber

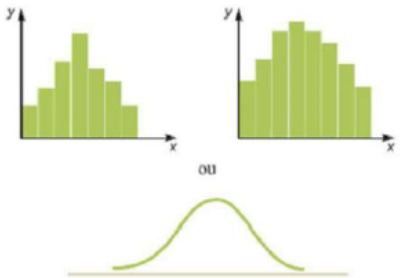
$$Mo = L_k + \left( \frac{n_k - n_{k-1}}{2n_k - n_{k+1} - n_{k-1}} \right) h_k$$

  - Fórmula de King

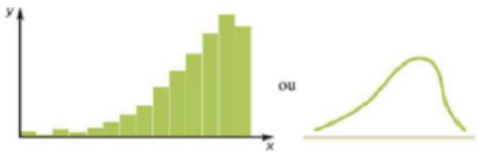
$$Mo = L_k + \left( \frac{n_{k+1}}{n_{k-1} + n_{k+1}} \right) h_k$$

    - k - classe modal (classe de maior freq absoluta/relativa).
    - $n_k$ - frequência absoluta da classe modal.
    - $n_{k-1}$ - frequência absoluta da classe anterior à classe modal.
    - $h_k$ - amplitude da classe modal.
    - $L_k$ - limite inferior da classe modal.

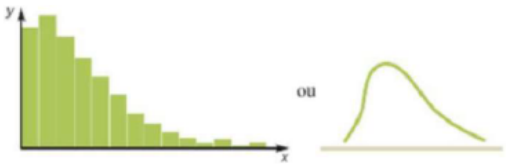
# Comparação entre a média moda e mediana (simetria)




Simétrica
Média=Mediana=Moda




Assimétrica à esquerda ou negativa
Média < Mediana < Moda




Assimétrica à direita ou positiva
Média > Mediana > Moda

Obs: Deve ser calculado também o coeficiente de assimetria (slide 40)

# Medidas de localização não central

## Quantil de ordem $\alpha$, $z_\alpha$

É o valor que divide uma série de $n$ observações em duas partes, tal que $\alpha$% das observações tem um valor menor, ou igual, a $z_\alpha$

- Quantil para dados não classificados em série ordenada

$$z_\alpha = x_k$$

onde $k$ é o maior inteiro menor que $n\alpha + 1$

- Quantil para dados quantitativos contínuos classificados

$$z_\alpha = L_k + \left( \frac{\alpha - F_{k-1}}{f_k} \right) h_k$$

onde, $k$ é a classe do quantil $\alpha$, tal que $F_{k-1} < \alpha$ e $F_k \geq \alpha$

# Medidas de localização

## Percentil de ordem $k$, $p_k$

$$p_k, (k = 1, 2, ..., 99) = z_{k/100}$$

## Decil de ordem $k$, $d_k$

$$d_k, (k = 1, 2, ..., 9) = z_{k/10}$$

# Medidas de localização

**Quartil de ordem $k$, $d_k$**

$$q_k, (k = 1, 2, 3) = z_{k/4}$$

De acordo com as definições anteriores, tem-se que:

$$Me = p_{50} = d_5 = q_2$$

# Medidas de dispersão ou variabilidade

## Amplitude total, "Range" $r$

É a diferença entre o maior e o menor valor de um conjunto de observações.

- Depende apenas das observações extremas.
- Depende o número de observações
- Para dados classificados, define-se como a diferença entre o valores máximo e mínimo das marcas da classe.

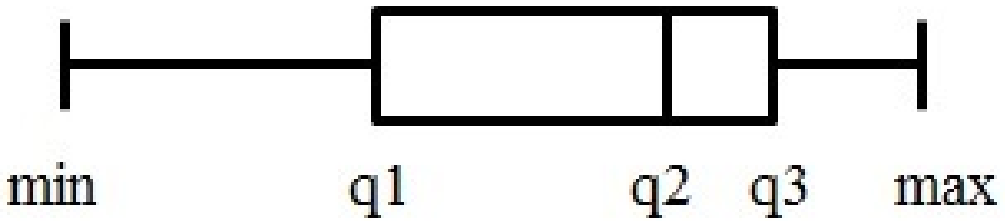
## Amplitude interquartil, "Interquartile Range" $r_q$

$$r_q = q_3 - q_1$$

# Medidas de dispersão

## Diagrama de extremos e quartis (Boxplot ou caixa de bigodes)

Consiste numa representação gráfica de dados baseado nos seus quartis, mínimo e máximo muito utilizada para analisar a dispersão dos dados.



# Outliers

## Outliers moderados

Um valor observado $x_i, i = 1, 2, ..., n$, é um candidato a outlier moderado se

$$x_i < q_1 - 1.5(q_3 - q_1)$$

ou

$$x_i > q_3 + 1.5(q_3 - q_1)$$

# Outliers

## Outliers severos

Um valor observado $x_i, i = 1, 2, ..., n$, é um candidato a outlier severo se

$$x_i < q_1 - 3(q_3 - q_1)$$

ou

$$x_i > q_3 + 3(q_3 - q_1)$$

- A remoção de outliers(valores anómalos) do conjunto de dados requer o conhecimento da área onde o estudo estatístico se insere.

# Medidas de dispersão

A variância e o desvio padrão são as medidas de variabilidade, ou dispersão mais utilizadas em estatística.

- Estas medidas têm em conta todos os valores observados.
- O desvio padrão indica a proximidade com que os valores observados se distribuem em torno da média.
- Um valor nulo do desvio padrão implica que todas as observações concentradas em torno do mesmo valor.
- Valores crescentes do desvio padrão indicam que os valores estão cada vez mais "espalhados"ou dispersos em relação à média.
- A variância é o quadrado do desvio padrão.

# Medidas de dispersão

## Variância de uma amostra $s^2$

Para dados não classificados: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Para dados classificados:

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

Nota: Para dados contínuos classificados $x_i$ representa a marca da classe.

# Medidas de dispersão

## Desvio padrão $s$

É a raiz quadrada positiva da variância

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

## Coeficiente de variação $s$

O coeficiente de variação de uma amostra é dado pela expressão

$$cv = \frac{s}{\bar{x}}$$

Trata-se de uma medida de dispersão relativa que tem em conta a magnitude dos valores observados.

# Medidas de forma

- Informação sobre a deformação dos dados
- Informação sobre o peso dos dados nas caudas

## Momento centrado de ordem $r$, $m_r$

Para dados não classificados:

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$$

Para dados classificados:

$$m_r = \frac{1}{n}\sum_{i=1}^{c}(x_i - \bar{x})^r n_i$$

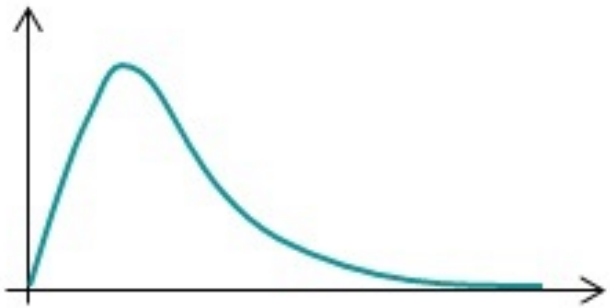
# Medidas de forma

## Coeficiente de assimetria amostral, $a_3$

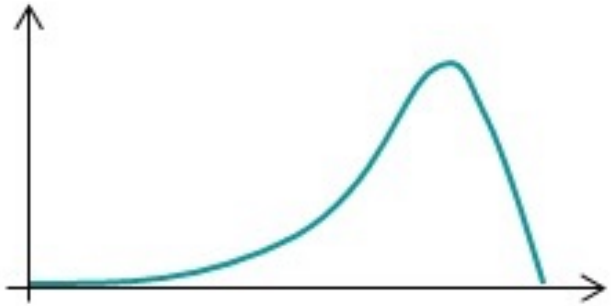
$$a_3 = \frac{m_3}{s^3}$$

- É uma medida adimensional
- Mede a assimetria da distribuição
- $a_3 > 0$ quando a cauda direita é mais comprida (enviesada à direita)
- $a_3 < 0$ quando a cauda esquerda é mais comprida (enviesada à esquerda)
- $a_3 = 0$ distribuição simétrica

# Medidas de forma



Positively skewed distribution



Negatively skewed distribution

# Medidas de forma

## Coeficiente de curtose amostral, $a_4$

$$a_4 = \frac{m_4}{s^4}$$

- É uma medida adimensional
- Mede o achatamento e o peso das caudas da distribuição.
- A distribuição normal tem $a_4 = 3$ - distribuição **mesocúrtica**
- $a_4 > 3$ quando a distribuição é mais esguia e as caudas mais pesadas do que a distribuição normal - distribuição **leptocúrtica**
- $a_4 < 3$ quando a distribuição é mais achatada e as caudas menos pesadas do que a distribuição normal - distribuição **platicúrtica**

## Exercício 1: Dados discretos

Considere a série estatística que representa as respostas quanto ao número de elementos do agregado familiar de uma amostra aleatória de 40 questionários:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 1 | 2 | 6 | 5 | 4 | 6 | 1 | 4 |
| 3 | 6 | 5 | 3 | 5 | 5 | 2 | 4 | 2 | 6 |
| 3 | 6 | 5 | 4 | 2 | 6 | 5 | 5 | 5 | 2 |
| 3 | 3 | 2 | 5 | 2 | 1 | 3 | 2 | 4 | 4 |

a) Construa a tabela de frequências.
b) Represente a distribuição dos dados num gráfico adequado
c) Calcule a média, a mediana e a moda do número observado de elementos do agregado familiar.
d) Calcule a variância e o desvio padrão da amostra.
e) Classifique os dados da amostra quanto à simetria.

## Exercício 2: Dados contínuos

Considere a série estatística que representa uma amostra aleatória dos tempos de jogo de 40 jogadores de basquetebol, em minutos.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10,52 | 9,87 | 3,28 | 10,82 | 32,42 | 30,55 | 4,17 | 9,62 | 15,8 | 14,27 |
| 11,4 | 5,63 | 12,88 | 10,8 | 22,68 | 23,32 | 30,1 | 21,78 | 7,98 | 17,7 |
| 10,35 | 14,77 | 10,42 | 9,57 | 31,95 | 32,92 | 33,28 | 27,97 | 11,03 | 18,02 |
| 17,78 | 0,75 | 14,75 | 21,02 | 28,43 | 30,33 | 29,67 | 28,05 | 26,1 | 23,28 |

a) Construa a tabela de frequências.
b) Represente a distribuição dos dados num histograma.
c) Calcule o tempo médio dos 40 jogadores com base na tabela de frequências.
d) Calcule a variância e o desvio padrão da amostra.
e) Classifique os dados da amostra quanto à simetria e curtose.