instituto politécnico de gestão e tecnologia

CURSO: GESTÃO DE RECURSOS HUMANOS

Unidade Curricular: Métodos Quantitativos

(1º ano)

CAPÍTULO 1 – ESTATÍSTICA DESCRITIVA

1. INTRODUÇÃO

1.1 Nota Histórica

Desde as antigas civilizações que existem registos históricos da presença de processos estatísticos. No Egipto há registos datados de 2900 a. C. que dizem respeito ao recenseamento de pessoas. Contudo, a estatística como ciência tem a sua origem no século XVIII. O alemão Gottfried Achemmel (1719-1772) é considerado por muitos autores o "pai" da palavra estatística. Na sua origem está a palavra «statu» (estado) pois eram os estados que mandavam realizar estudos para recolha de informações sobre as populações e os seus territórios.
De acordo com a Bíblia, Maria e José foram a Belém para efeitos de um estudo estatístico: recenseamento da população por ordem do imperador de Roma.
Nesta altura, os estudos estatísticos tinham como principal objetivo o lançamento de impostos e o alistamento de homens para as guerras.

A estatística em Portugal

No primeiro ano da nossa era, o imperador César Augusto mandou realizar um levantamento populacional relativo à região entre Douro e Guadiana.
Na Idade Média, os árabes realizaram vários recenseamentos relativos à Península Ibérica.
O Instituto Nacional de Estatística (INE) foi fundado em 1935 e é onde se centraliza toda a atividade estatística oficial.

1.2 População e Amostra

Em qualquer estudo estatístico parte-se da observação de um conjunto de elementos.

População é um conjunto de todos os elementos cujas caraterísticas se pretendem estudar.
Uma população pode ser finita ou infinita.
Exemplo de uma população finita: número de alunos do Isla no ano letivo 2013/2014.
Exemplo de uma população infinita: conjunto dos números reais.

Amostra é um subconjunto representativo e recolhido aleatoriamente da população, que se estuda com o objetivo de tirar conclusões para a população de onde foi recolhida.
A cada um dos elementos da população dá-se o nome de *indivíduo* ou *unidade estatística*.

Num estudo estatístico é sempre melhor usar a população em vez de uma amostra, mas nem sempre isto é possível. Normalmente, recorre-se ao uso de uma amostra por a população ser infinita, economia de dinheiro e tempo, comodidade, etc.

É necessário ter muito cuidado na escolha da amostra (as suas caraterísticas devem aproximar-se tanto quanto possível das da população) pois se esta não for bem escolhida, o estudo pode conduzir a conclusões erradas. Na escolha de uma amostra deve-se ter em conta a imparcialidade, a representatividade e o tamanho.
Ao processo de seleção de uma amostra de uma dada população chama-se amostragem.
Há vários métodos de amostragem cuja aplicação pode não ser fácil tendo em conta que se pretende obter sempre uma amostra representativa da população.
Um estudo estatístico baseado numa amostra que não seja representativa da população – *amostra enviesada* – conduz a conclusões erradas, que se afastam da realidade.

Para que uma amostra seja representativa da população de onde é retirada deve satisfazer determinadas caraterísticas de entre as quais se destacam:

- A amostra deve ser escolhida de forma aleatória;
- A dimensão da amostra deve ser tal que as caraterísticas da população sejam evidenciadas.

O processo de amostragem a adotar depende, entre outros fatores, das caraterísticas da população e dos objetivos do estudo. Consideram-se dois processos para se obter a amostra:

- Aleatório: qualquer um dos elementos da população pode fazer parte da amostra com uma probabilidade conhecida. Possibilita estimar as margens de erro dos resultados devidos à amostragem, evitando o enviesamento das amostras (facto que surge sempre que se usam métodos empíricos para selecionar os indivíduos que integram a amostra).
- Determinístico: a opinião de quem faz o estudo e a sua experiência é determinante na seleção dos indivíduos a incluir na amostra. A noção de representatividade da população muda mediante o analista, sendo portanto impossível estimar a probabilidade dum elemento ser selecionado para a amostra.

Tipos de amostragem aleatória:

- *Amostragem aleatória simples*

  O processo de amostragem aleatória simples garante que todos os elementos da população têm a mesma probabilidade de ser incluídos na amostra. Todas as possíveis amostras de n elementos, retiradas de uma população de dimensão N, têm a mesma probabilidade de serem selecionadas.

- *Amostragem aleatória estratificada*

  Para a recolha de uma amostra estratificada há que previamente proceder à estratificação da população, ou seja dividi-la em grupos (estratos) de tal modo que dentro de cada estrato haja a maior uniformidade possível e, em que as diferenças entre os vários estratos sejam tão grandes quanto possível. Estando a população classificada em estratos, a amostra é recolhida através de amostragens aleatórias simples dentro de cada estrato.

- *Amostragem aleatória por clusters*

  Este tipo de amostragem é útil quando a população se encontra dividida num reduzido número de grupos (clusters) onde, as diferenças entre os elementos dentro de cada cluster devem ser as maiores possíveis. A amostra é recolhida escolhendo por amostragem aleatória simples um ou mais clusters.

- *Amostragem sistemática*

  Este método é designado por quase aleatório pois não dá a todas as amostras, que se podem retirar de uma população, a mesma probabilidade de ocorrência. Para aplicar este método é necessário calcular o quociente $k = \frac{N}{n}$. Em seguida, escolhe-se aleatoriamente um número, no intervalo [1; K], que servirá como ponto de partida e primeiro elemento da amostra. Adicionando k ao primeiro valor obtido (arredondando o resultado por defeito), obtém-se o segundo elemento. A adição sucessiva do valor k permite encontrar os restantes elementos da amostra. Apenas o primeiro elemento é escolhido aleatoriamente enquanto os restantes são determinados de modo sistemático pelo valor de k.

1.3 Parâmetro e Estatística

Parâmetro é uma medida numérica que descreve uma caraterística de uma população.

Estatística é uma medida numérica que descreve uma caraterística de uma amostra.

1.4 Recenseamento e Sondagem

Recenseamento ou *censo* é quando toda a população é observada tendo em atenção o objetivo do estudo.

Sondagem é quando, num estudo estatístico, se observa apenas uma parte da população.

1.5 Estatística Descritiva e Estatística Indutiva

A *estatística descritiva* consiste na recolha, apresentação, análise e interpretação de dados numéricos através da criação de instrumentos adequados: quadros, gráficos e indicadores numéricos.
A estatística descritiva pretende medir ou contar os dados obtidos na tomada de observações de determinada fonte e consiste na análise preliminar das observações.

Na *estatística indutiva* o objetivo é generalizar os resultados obtidos em determinado conjunto de elementos a um outro conjunto mais numeroso. Nos cálculos da estatística indutiva entra-se com a teoria das probabilidades.

A utilidade da Estatística pode ser resumida do seguinte modo:

- Permite descrever e compreender relações entre as variáveis.
- Permite a tomada de melhores e mais rápidas decisões.
- Facilita a tomada de decisões para fazer face à mudança.

1.6 As etapas do método estatístico

Para construir uma tabela ou um gráfico que transmita informação de interesse para a população é necessário elaborar um estudo estatístico.

As principais etapas do método estatístico são:

- identificação do problema ou situação;
- recolha de dados;
- análise e crítica dos dados;
- apresentação dos dados;
- interpretação dos resultados.

- <u>Identificação do problema ou situação</u>

  Esta etapa requer algum conhecimento estatístico pois os métodos a aplicar não são, de modo nenhum, independentes da informação que se pretende recolher. Uma identificação incorreta do problema torna todas as etapas seguintes inúteis.

- <u>Recolha de dados</u>

  Uma vez identificado o problema, a etapa seguinte consiste na recolha de dados necessários, apropriados, tão completos quanto possível e, sobretudo, pertinentes para a situação que se pretende analisar.
  A recolha de toda a informação necessária pode ser feita diretamente quando os dados são obtidos de fonte originária ou de forma indireta quando os dados recolhidos provêm já de uma recolha direta. Aos primeiros chamam-se dados primários e aos segundos dados secundários.

  A recolha dos dados poder ser classificada como:
  - ✓ Contínua → realiza-se permanentemente;
  - ✓ Periódica → realiza-se em intervalos de tempo;
  - ✓ Ocasional → realiza-se de modo esporádico.

- <u>Análise e Crítica dos dados</u>

  Depois dos dados recolhidos é necessário proceder-se a uma revisão crítica de modo a suprimir valores estranhos ou eliminar erros capazes de provocar enganos de apresentação e análise ou mesmo de enviesar as conclusões obtidas.

- Apresentação dos dados

  Após a recolha e a crítica, convém organizar os dados de maneira prática e racional, para um melhor entendimento do fenómeno que se pretende estudar. Torna-se necessário criar instrumentos necessários para classificar e apresentar conjuntos de dados numéricos de tal modo que a informação neles contida seja apreendida mais fácil e rapidamente.
  O processo de classificação consiste na identificação de unidades de informação com caraterísticas comuns e no seu agrupamento em classes. A informação é sintetizada através de quadros, gráficos e valores numéricos descritivos.

- Interpretação dos resultados

  Por último é necessário interpretar os resultados obtidos. Esta interpretação será mais fácil se nas etapas anteriores tiverem sido escolhidos os instrumentos mais apropriados para a representação e análise do tipo de dados recolhidos.

1.7. Dados Qualitativos e Dados Quantitativos

Ao estudar uma população ou uma amostra dessa população, o que se pretende é conhecer as suas caraterísticas ou atributos para que, posteriormente, seja possível tomar decisões com base nesse conhecimento. Tendo em conta as caraterísticas ou atributos dos elementos de uma população, os dados podem ser:

- *Qualitativos*: representam a informação que identifica alguma qualidade, categoria ou caraterística, não suscetível de medida, mas de classificação.
  Exemplo: O estado civil de um indivíduo é um dado qualitativo, assumindo as categorias: solteiro, casado, viúvo e divorciado.

  Para sumariar dados qualitativos numericamente utilizam-se contagens, proporções, percentagens...
  Por exemplo, se 40 de 160 estudantes de Matemática são homens, pode-se traduzir este facto dizendo que 25% dos estudantes de Matemática são homens.

  Os dados qualitativos podem ainda ser classificados em:

  - ✓ *Nominais*
    Exemplo: sexo (masculino, feminino).
  - ✓ *Ordinais*
    Exemplo: desempenho (baixo, médio, alto).

- *Quantitativos*: representam informação resultante de caraterísticas suscetíveis de serem medidas, apresentando-se com diferentes intensidades, que podem ser de natureza discreta ou contínua.
  Exemplo: Idade dos alunos da turma de Estatística.

  Os dados quantitativos podem ser de dois tipos:

  - ✓ *Discretos*: são contáveis; tomam apenas valores inteiros. São aqueles em que o número de valores possíveis é finito ou "enumerável". Exemplo*:* n.º de admissões num hospital, nº de funcionários,…
  - ✓ *Contínuos*: medem-se; podem tomar qualquer valor num intervalo relevante especificado. Os dados contínuos resultam de infinitos valores possíveis numa escala contínua. Exemplo: peso; altura.

As distinções dos tipos de dados são menos rígidas do que a descrição apresenta. Por exemplo, em geral a idade das pessoas de uma amostra seria tratada como uma variável contínua, mas se a idade for apresentada pelo número de anos concluídos, pode ser tratada como discreta, e se separar a amostra, em função da classificação, em "crianças", "jovens", "pessoas de meia-idade" e "velhos", então a idade pode ser assumida como uma variável qualitativa.

2. ORGANIZAÇÃO E INTERPRETAÇÃO DOS DADOS

2.1 Introdução

Quando se realizam estudos estatísticos, a recolha de dados por si só não permite dar resposta a várias questões. Na verdade a resposta a essas questões é encontrada se os dados estiverem organizados recorrendo a tabelas ou gráficos. A forma de organizar e apresentar a informação depende do tipo de variável em estudo.

2.2 Tabela de Frequências

Variáveis discretas

Variável é um símbolo que representa determinada caraterística de uma população ou amostra.

Considere-se uma população (N) ou uma amostra (n) de indivíduos com a caraterística que apresenta p valores observados $X_1, X_2, X_3, ..., X_p$. Dá-se o nome de distribuição de frequências ao conjunto de todos os valores da variável e da frequência ou número de ocorrências correspondente.

A *frequência absoluta* (F_i) é o número de vezes que cada valor da variável se repete na amostra ou população.

Exemplo:

Tipos de fábricas de um concelho

Tipos de fábricas	Contagem	F_i
Conservas	~~IIII~~ I	6
Calçado	~~IIII~~ III	8
Têxtil	III	3
Cerâmica	I	1

A *frequência relativa* (f_i) de um valor da variável é dada por $f_i = \frac{F_i}{n}$, isto é, o número de vezes que esse valor ocorre (F_i) relativamente ao total da amostra (n) ou população (N).

A *frequência acumulada* de cada valor da variável obtém-se adicionando à sua frequência as frequências dos valores anteriores.

Exemplo:

Nº de faltas ao trabalho de 100 trabalhadores de uma empresa

X_i	F_i	f_i	cumF_i	cumf_i
0	10	0,10	10	0,10
1	15	0,15	25	0,25
2	25	0,25	50	0,50
3	40	0,40	90	0,90
4	10	0,10	100	1,00
Total	100	1,00		

Variáveis contínuas

As variáveis contínuas, por poderem assumir um número infinito não-numerável de valores, obrigam à definição de classes de valores. O número de classes não deverá ser muito grande para que não introduza irregularidades que poderão não existir na população, mas também não deverá ser muito pequeno para que não haja perda de informação.

A seleção do número de classes ou intervalos não constitui nenhum método rigoroso e científico, no entanto existem algumas regras básicas que devem ser seguidas na sua construção:

- Em geral, o número de classes (K) deverá estar compreendido entre quatro e catorze;
- Nenhuma classe deverá ter frequência nula;
- As classes deverão ter, sempre que possível, amplitudes iguais;
- Os pontos médios das classes devem ser números de cálculo fácil;
- Classes abertas devem ser evitadas embora nem sempre seja possível fazê-lo;
- Os limites das classes são definidos de modo a que cada valor da variável é incluído num e num só intervalo.

Tendo em conta estas regras básicas, é frequente calcular o número de classes através da Fórmula de Sturges:

$$K = 1 + 3{,}22 \log_{10}(n)$$

Geralmente as classes têm todas a mesma *amplitude*, mas não é obrigatório.

Se as classes têm todas a mesma amplitude, então:

$$a = \frac{Max(X_i) - Min(X_i)}{K}, \quad K = \text{n.º de classes}$$

O limite inferior da classe mais baixa tem que ser menor ou igual ao mínimo e o limite superior da classe mais alta tem que ser maior que o máximo.

Exemplo:

Experimentou-se uma marca de gasolina em 25 automóveis de marca diferente.
Em cada automóvel foram colocados 10 litros de gasolina e observou-se a distância percorrida até acabar o combustível. Registaram-se os seguintes dados:

33,2	34,7	33,7	31,8
36,2	34,2	30,6	32,4
33,3	33,1	33,2	31,8
33,1	31,4	30,9	34,0
31,2	34,9	31,6	
35,1	31,8	34,1	
35,9	32,8	32,7	

Tem-se que: $Min(X_i) = 30{,}6$ e $Max(X_i) = 36{,}2$

Logo, $K = 5{,}501... \approx 6$ e $a = \frac{36{,}2 - 30{,}6}{6} = 0{,}9 \approx 1$.

Pode-se então escolher a primeira classe a começar em 30 até 31 ou em 30,5 até 31,5, etc...

Classes	Ponto médio	Frequência absoluta (F_i)	Frequência relativa (f_i)
[30,5;31,5[	31	4	0,16
[31,5;32,5[	32	5	0,20
[32,5;33,5[	33	7	0,28
[33,5;34,5[	34	4	0,16
[34,5;35,5[	35	3	0,12
[35,5;36,5[	36	2	0,08
		25	1

Uma vez definidas as classes e as frequências absolutas de cada classe, o cálculo das frequências acumuladas é idêntico ao das variáveis discretas.

2.3 Representação gráfica das distribuições de frequências

A representação gráfica utilizada depende do tipo de variável (discreta ou contínua) e se a frequência é simples ou acumulada.

2.3.1 Variáveis discretas

Diagrama diferencial

Para representar a distribuição de frequências relativas de uma variável discreta utiliza-se o gráfico de barras ou diagrama diferencial.

Exemplo:

Nº de faltas ao trabalho de 100 trabalhadores de uma empresa

X_i	F_i	f_i	cumF_i	cumf_i
0	10	0,10	10	0,10
1	15	0,15	25	0,25
2	25	0,25	50	0,50
3	40	0,40	90	0,90
4	10	0,10	100	1,00
Total	100	1,00		

A representação gráfica das frequências simples é a seguinte:

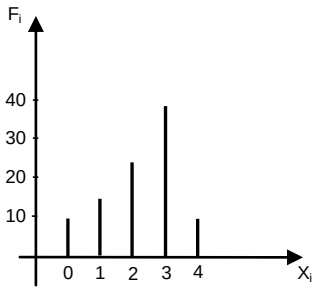
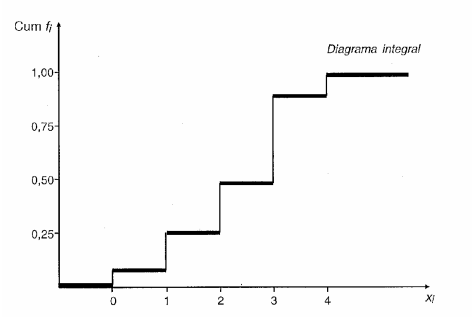


Diagrama integral

Como as variáveis discretas assumem valores separados uns dos outros, a representação gráfica das frequências acumuladas vai ter o aspecto de uma série de degraus: é o chamado diagrama integral.



Este gráfico dá uma frequência acumulada nula até ao ponto zero. A frequência acumulada é de 10 % até a variável atingir o valor 1, mas excluindo este valor. O salto dado pelas frequências acumuladas em cada ponto corresponde exactamente à frequência simples para esse mesmo ponto.

Separador de frequências

Há uma forma de organizar os dados estatísticos designada por separador de frequências (também conhecido por diagrama de caule-e-folha).
Esta forma de organização de dados tem a particularidade de permitir ao observador uma perceção do aspecto global da distribuição dos dados sem que, ao mesmo tempo, se perca informação contida na coleção inicial dos dados.

Exemplo:
Considerem-se as classificações dos alunos de uma turma do 8º ano num teste de Português numa escala de 0 a 100:

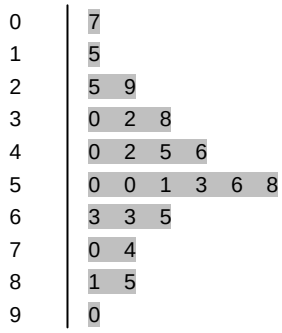
Classificações Teste de Português (0 – 100)				
29	85	56	74	70
65	42	32	46	63
50	38	25	30	45
81	51	58	40	15
53	50	63	90	7

Tomando cada um dos dados, isto é, cada uma das classificações, separa-se o algarismo das dezenas, a que se chama caule, do algarismo das unidades, a que se chama folha.
Pode-se, então, dispor os dados da seguinte forma:

Separador	
Caule	Folha
0	7
1	5
2	5 9
3	0 2 8
4	0 2 5 6
5	0 0 1 3 6 8
6	3 3 5
7	0 4
8	1 5
9	0

Nas folhas, os números podem aparecer por ordem crescente ou decrescente.
Relativamente a outras formas de organização de dados, o separador de frequências apresenta algumas vantagens, nomeadamente:

- Não implica a definição de classes arbitrárias;
- Todos os dados observados estão presentes;
- Observando o separador de frequências é possível imaginar o gráfico resultante do grupo de dados.



Construção de um separador de frequências:

- ✓ Escreve-se primeiro a lista de todos os caules;
- ✓ Seguidamente, ao lado de cada caule, escrevem-se as folhas que lhe correspondem.

2.3.2 Variáveis contínuas

Histograma

A representação gráfica adequada para as frequências simples deste tipo de variáveis é o histograma.

Para construir um histograma deve atender-se ao seguinte:

- o gráfico deve ter um título;
- os dados devem estar agrupados em classes;
- no eixo horizontal devem ser colocadas as classes e no eixo vertical as frequências absolutas ou relativas;
- as barras são desenhadas verticalmente e são adjacentes;
- cada barra corresponde a uma das classes em que os dados foram agrupados;
- a área das barras é proporcional à frequência;
- se as classes têm todas a mesma amplitude, basta desenhar os retângulos (barras) com alturas proporcionais às frequências;
- se as classes têm amplitudes diferentes entre si convém normalizar todas as frequências para que a proporcionalidade das áreas se verifique. Isto pode ser feito dividindo as frequências das classes pelas respetivas amplitudes e construindo as barras com alturas proporcionais a estes novos valores.

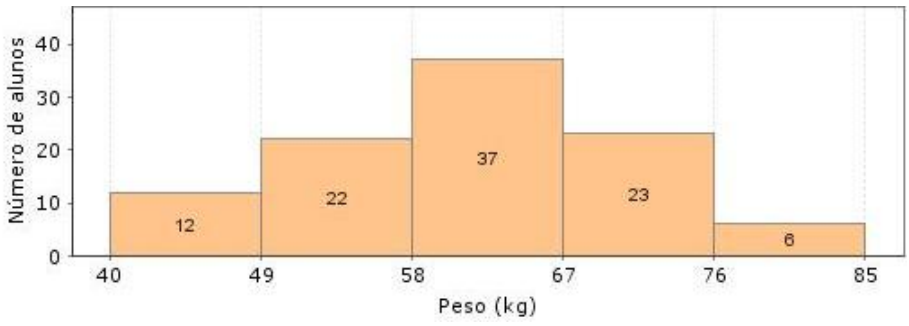
Exemplo:

Distribuição do Peso dos Alunos de uma Escola

Classes de Peso (kg)	Frequência Simples		Frequência Acumulada	
	Absoluta (F_i)	Relativa (f_i) (%)	Absoluta (F_i)	Relativa (f_i) (%)
[40;49[	12	12,00	12	12,00
[49;58[	22	22,00	34	34,00
[58;67[	37	37,00	71	71,00
[67;76[	23	23,00	94	94,00
[76;85[	6	6,00	100	100,00
	100	100,00		

Histograma

Distribuição do Peso dos Alunos de uma Escola



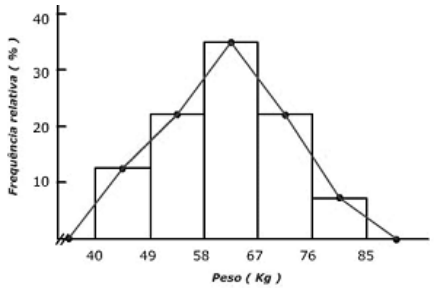
Polígono de frequências

O polígono de frequências é uma variação do histograma, construído através da ligação por meio de segmentos de reta dos pontos médios dos topos das caixas do histograma.
De acordo com o princípio da proporcionalidade, a área definida pela soma das áreas dos retângulos do histograma deverá ser igual à área sob o polígono de frequências, o que obriga a um cuidado especial quando as classes têm amplitudes diferentes.
Para fechar o polígono é necessário criar uma classe adicional em cada um dos extremos do histograma, com amplitude idêntica à das classes adjacentes e com frequência nula.

Exemplo:

Histograma e polígono de frequências

Distribuição do Peso dos Alunos de uma Escola



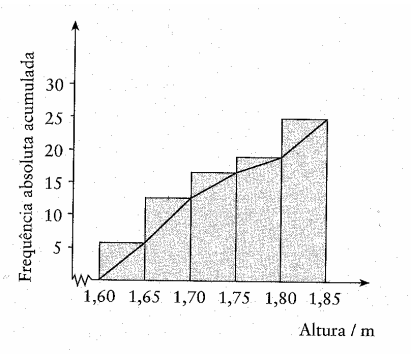
Os polígonos de frequências são mais úteis que os histogramas quando queremos comparar duas ou mais distribuições de frequências por sobreposição de umas sobre as outras.

Uma outra maneira útil de apresentação gráfica é a que utiliza as frequências acumuladas. Para estas, pode traçar-se um polígono de frequências acumuladas pressupondo que a distribuição dos elementos dentro da classe se faz de uma forma uniforme o que origina, no intervalo de valores de uma mesma classe, uma representação do tipo linear.

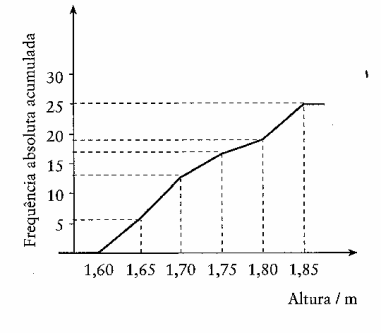
Exemplo:

Altura dos jogadores da equipa X

Altura dos jogadores da equipa X (X_i)	F_i	f_i	cum F_i	cum f_i
[1,60 ; 1,65[	6	0,24	6	0,24
[1,65 ; 1,70[	7	0,28	13	0,52
[1,70 ; 1,75[	4	0,16	17	0,68
[1,75 ; 1,80[	2	0,08	19	0,76
[1,80 ; 1,85[	6	0,24	25	1,00
	25	1		



À linha desenhada a unir os vértices dos retângulos chama-se polígono de frequências acumuladas (Ogivas de Galton).



3. MEDIDAS DE LOCALIZAÇÃO

3.1 Medidas de tendência central

Nesta secção dar-se-á particular atenção às medidas de estatística descritiva que dão uma indicação do valor central ou média dos dados observados – as **medidas de tendência central**.

As medidas de tendência central incluem, além da média, a moda e a mediana.

3.1.1 Média Aritmética

Definição:

Sejam $X_1, X_2, \ldots, X_n$ os valores de uma variável quantitativa.

Designa-se por *média aritmética* de uma amostra de uma população, e representa-se por $\overline{X}$, ao valor:

- Para dados desagregados:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \Leftrightarrow \overline{X} = \frac{1}{n}\sum X_i$$

- Para dados agregados:

$$\overline{X} = \frac{\sum X_i \cdot F_i}{n} = \sum X_i \cdot f_i$$

- Para dados agregados em classes:

$$\overline{X} = \frac{\sum C_i \cdot F_i}{n} = \sum C_i \cdot f_i$$

onde:

- F_i é a frequência absoluta da classe i.
- f_i é a frequência relativa da classe i.
- X_i é o valor correspondente à classe i.
- C_i é o ponto médio da classe i.
- n é o tamanho da amostra.
- N é o tamanho da população.

Exemplos:

1. *Cálculo da média em dados desagregados*

Um criador de cães registou o número de crias em várias ninhadas durante um ano, como mostra o seguinte quadro.

3	4	4	7	7	8	8	12	12
3	4	12	12	7	7	8	12	4
7	8	4	12	7	7	7	12	

$$\overline{X} = \frac{3+4+4+7+\ldots+7+12}{26} \Leftrightarrow \overline{X} = \frac{198}{26} \Leftrightarrow \overline{X} = 7{,}6$$

Logo, em média, nasceram 7,6 crias em cada ninhada.

2. *Cálculo da média em dados agregados*

Considerem-se os dados da tabela, os quais se referem ao exemplo anterior:

X_i	F_i
3	2
4	5
7	8
8	4
12	7

$$\overline{X} = \frac{3\times 2+4\times 5+7\times 8+8\times 4+12\times 72}{26} \Leftrightarrow \overline{X} = \frac{198}{26} \Leftrightarrow \overline{X} = 7{,}6$$

Logo, em média, nasceram 7,6 crias em cada ninhada.

3. *Cálculo da média em dados agregados em classes*

Considere-se a variável “descontos para a segurança social” dos trabalhadores de uma empresa, cuja distribuição é definida pela tabela:

Categorias dos descontos	Número de empregados
$[180{,}185[$	2
$[185{,}190[$	4
$[190{,}195[$	1
$[195{,}200[$	2
$[200{,}205[$	3
$[205{,}210[$	2

Neste caso, como em todos em que os dados estão agrupados em classes que são intervalos, recorre-se ao cálculo do valor central (ou médio) da classe.

O valor central de uma classe ou marca da classe calcula-se determinando a média dos extremos.

Classes	F_i	C_i (Marca)	F_iC_i
$[180,185[$	2	182,5	365
$[185,190[$	4	187,5	750
$[190,195[$	1	192,5	192,5
$[195,200[$	2	197,5	395
$[200,205[$	3	202,5	607,5
$[205,210[$	2	207,5	415
	$\sum F_i = 14$		$\sum F_iC_i = 2725$

$$\overline{X} = \frac{\sum F_i C_i}{\sum F_i} \Leftrightarrow \overline{X} = \frac{2725}{14} \Leftrightarrow \overline{X} \approx 194{,}643$$

O desconto médio é de 194,6 euros.

Para estes casos a média passa a ser um valor aproximado da verdadeira média, que se obteria a partir dos dados desagregados, uma vez que se substituíram os valores exatos individuais incluídos em cada classe pelo ponto médio ou marca dessa classe. Como se compreenderá, uma diferente divisão em classes implicaria diferentes pontos médios e, portanto, um valor para a média, provavelmente, um pouco diferente do anterior. Em qualquer caso, sempre que se trabalha com amostras, as estatísticas são estimativas dos verdadeiros parâmetros da população, sujeitas a erros de medida e de amostragem.

Propriedades da Média Aritmética

Propriedade 1: Adicionando um valor constante a cada um dos elementos de um conjunto de números, a média vem adicionada a essa constante.

Propriedade 2: Multiplicando cada um dos elementos de um conjunto de números por uma constante, a média vem multiplicada por essa constante.

Propriedade 3: Dados dois conjuntos de números, com um mesmo número de elementos, de média $\overline{X}$ e $\overline{X}'$ respetivamente, o conjunto obtido pela soma dos elementos dos conjuntos dados, um a um, é um conjunto de n números de média $\overline{X} + \overline{X}'$.

Caraterísticas importantes da Média Aritmética:

- É única. Para um determinado conjunto de dados existe uma e uma só média aritmética.
- É simples de calcular.
- Como todos os valores dos dados entram no cálculo da média, os valores extremos têm grande influência no cálculo desta e, em alguns casos, podem distorcê-la de tal maneira que não se torna uma boa medida de localização.
- A média é a medida de tendência central mais eficiente quando se trata de inferir sobre uma população, a partir de dados recolhidos apenas para uma amostra.

Observações:

1. A soma das diferenças entre os valores observados e a média é nula.

2. A soma dos quadrados das diferenças entre os valores observados e a média observada é um valor mínimo.

3.1.2 Mediana

É uma medida da localização do centro da distribuição dos dados que os divide ao meio.

Analogamente ao caso da média, também no cálculo da mediana se deve ter em conta a forma como os dados estão apresentados.

Variáveis discretas

- Se o número de observações n é ímpar, a mediana é o elemento médio $\left(\text{corresponde ao elemento de ordem } \frac{n+1}{2}\right)$ quando todas as observações estão ordenadas: $Me = X_{\frac{n+1}{2}}$

- Se o número de observações n é par, não há um valor do meio, mas dois $\left(\text{os elementos de ordem } \frac{n}{2} \text{ e } \frac{n+2}{2}\right)$. Neste caso, a mediana será o ponto médio destes dois valores, depois de todas as observações estarem ordenadas: $Me = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2}$

Exemplos:

1. *Cálculo da mediana em dados desagregados*

Os dados seguintes referem-se ao número de passageiros transportados pelos autocarros A e B em cada um dos percursos efetuados durante um dia.

Autocarro A						
30	78	38	44	65	52	28
34	46	50	61	84	44	

Autocarro B				
41	83	60	54	33
39	48	72	90	47

Para determinar as medianas dos números de passageiros transportados por cada um dos autocarros, deve começar-se por ordenar os dados e contar o número de percursos efetuados.

Para o autocarro A, com n = 13, vem:

28 30 34 38 44 44 **46** 50 52 61 65 78 84

Como n é ímpar, $\frac{n+1}{2} = \frac{13+1}{2} = 7$

Logo,

$$Me = X_7 = 46$$

Para o autocarro B, com n = 10, vem:

33 39 41 47 **48 54** 60 72 83 90

Como n é par, $\frac{n}{2} = \frac{10}{2} = 5$ e $\frac{n+2}{2} = \frac{12}{2} = 6$

Logo,

$$Me = \frac{X_5 + X_6}{2} = \frac{48+54}{2} = 51$$

2. *Cálculo da mediana em dados agregados*

Determinar a mediana das classificações em Francês dos 21 alunos da turma B.

- n ímpar

Classificação (níveis de 1 a 5)	F_i	cum F_i
1	1	1
2	5	6
3	10	16
4	3	19
5	2	21
	21	

$$\frac{n+1}{2}=\frac{21+1}{2}=\frac{22}{2}=11$$

A mediana é o valor de ordem 11. Procura-se o número 11 na coluna das frequências absolutas acumuladas. Como este valor não aparece na referida coluna e o imediatamente superior é 16, então:

$$Me = X_{11} = 3$$

Determinar a mediana das classificações em Francês dos 30 alunos da turma C.

- n par

Classificação (níveis de 1 a 5)	F_i	cum F_i
1	4	4
2	11	15
3	12	27
4	2	29
5	1	30
	30	

$$\frac{n}{2}=\frac{30}{2}=15 \text{ e } \frac{n+2}{2}=\frac{32}{2}=16$$

A mediana é a semi soma dos valores de ordem 15 e 16. Observando a tabela:

$$Me = \frac{X_{15}+X_{16}}{2}=\frac{2+3}{2}=2{,}5$$

<u>Variáveis contínuas</u>

O procedimento para calcular a mediana em dados agregados em classes é o seguinte:

1.º) Calcula-se a ordem $\frac{n}{2}$.

2.º) Pelas frequências acumuladas identifica-se a classe que contém a mediana e que será a classe mediana.

A classe mediana é a classe em que $\text{cum}F_{me} \geq \frac{\sum F_i}{2}$.

3.º) Calcula-se o valor exato da mediana através da seguinte fórmula:

$$Me = li_{Me} + \frac{n/2 - cumF_{Me-1}}{F_{Me}} \cdot a_{Me}$$

em que:

li_{Me} = limite inferior da classe mediana

$cumF_{Me-1}$ = frequência absoluta acumulada da classe anterior à classe mediana

F_{Me} = frequência absoluta da classe mediana

a_{Me} = amplitude da classe mediana

A aplicação desta fórmula pressupõe que as frequências se distribuem uniformemente dentro de cada classe, ou seja, que as frequências acumuladas variam linearmente dentro das classes.

Exemplo:

Considere-se a seguinte tabela de frequências:

Classe	Frequência absoluta (F_i)	Frequência absoluta acumulada $(cumF_i)$
35-45	5	5
45-55	12	17
55-65	18	35
65-75	14	49
75-85	6	55
85-95	3	58
	$\sum = 58$	

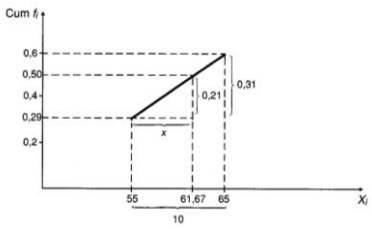
As classes já se encontram ordenadas.

Como $\frac{n}{2} = 29$, a classe mediana é [55, 65[.

Supondo que nesta classe os valores se encontram uniformemente distribuídos, tem-se:

$$Me = 55 + \frac{29 - 17}{18} \times 10 = 61{,}67$$

Graficamente:



Classe mediana: [55, 65[

10 ---------- 0,31

X ---------- 0,21

$$X = \frac{0{,}21 \times 10}{0{,}31} = 6{,}67$$

Logo, $Me = 55 + X = 55 + 6{,}67 = 61{,}67$

Caraterísticas importantes da Mediana:

- É única.
- É simples de calcular.
- É determinada pelo número de observações e não pelo seu valor.
- Não é drasticamente afetada por valores extremos como acontece com a média.

3.1.3 Moda

Sejam $X_1, X_2, \ldots, X_n$ os n valores de uma variável estatística.

Designa-se por moda, e representa-se por Mo, ao valor que ocorre com maior frequência.

Para um conjunto de dados, pode existir mais do que uma moda ou até nem existir.

- Se um conjunto de dados tiver uma única moda, esse conjunto diz-se *unimodal*.
- Se um conjunto de dados tiver duas modas, esse conjunto diz-se *bimodal*.
- Se um conjunto de dados tiver mais de duas modas, esse conjunto diz-se *multimodal*.
- Se um conjunto de dados não tiver moda, esse conjunto diz-se *amodal*.

Tal como no caso da média ou da mediana, na indicação da moda deve-se ter em conta o facto de os dados estarem ou não classificados.

Exemplos:

1.

Dados	Moda
8 8 9 3 3 5 8 9 9	8 e 9 (bimodal)
0 6 4 5 8 0 4 6 5 8	Não tem moda

2.

Classificação	F_i
1	4
2	11
3	12
4	2
5	1

Neste caso, a moda é 3 pois é o valor da variável que regista maior frequência absoluta.

No caso de variáveis contínuas faz-se referência frequentemente à classe modal. Pode-se, neste caso, determinar a moda graficamente e também através da expressão:

$$Mo = li_{Mo} + \frac{F_{Mo+1}}{F_{Mo-1} + F_{Mo+1}} a_{Mo}$$

em que:

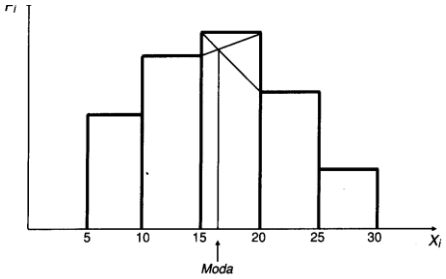
li_{Mo} = limite inferior da classe modal

F_{Mo+1} = frequência absoluta da classe seguinte à classe modal

F_{Mo-1} = frequência absoluta da classe anterior à classe modal

a_{Mo} = amplitude da classe modal

A moda pode ser determinada graficamente. Para tal é necessário construir o histograma da distribuição, identificar a classe modal e unir os vértices superiores do retângulo da classe modal com os vértices das classes contíguas, como se indica na figura:



Existem algumas fórmulas para localizar a moda dentro da classe modal, mas todas elas dão valores aproximados e não exatos.
A moda pode ser utilizada para descrever dados qualitativos.

Caraterísticas importantes da Moda:

- A Moda é fácil de calcular e interpretar.
- Não é afetada por valores extremos.
- Não pode ser definida com rigor e o seu valor exato, muitas vezes é incerto.
- Pode ser determinada em qualquer situação.
- Pode ser utilizada para descrever dados qualitativos. Por exemplo, numa fábrica permite determinar o tipo de produto mais procurado.

3.2 Medidas de tendência não central

Para além das medidas de tendência central existem outras medidas que dão a localização dos valores da variável. São designados *quantis* e podem ser:

- Quartis
- Decis
- Percentis

Os *quartis* são os valores da variável que dividem a distribuição de frequências em quatro partes iguais.
O primeiro quartil será um valor da variável tal que o número de observações para valores inferiores será 25%, e superiores, 75%.

O segundo quartil, terá metade das observações à sua direita e outra metade à sua esquerda, logo coincide com a mediana.
O terceiro quartil será um valor tal que à sua esquerda concentrar-se-ão 75% das observações e à sua direita os restantes 25%.

O método de cálculo dos quartis é exatamente o mesmo utilizado para o cálculo da mediana e dependerá de se tratar de uma variável discreta ou contínua.

Existem fórmulas para localizar os quartis num conjunto de dados:

- Se n par:

$$Q_1 = X_{\frac{n+2}{4}} \quad \text{e} \quad Q_3 = X_{\frac{3n+2}{4}}$$

- Se n ímpar:

$$Q_1 = X_{\frac{n+1}{4}} \quad \text{e} \quad Q_3 = X_{3\frac{n+1}{4}}$$

Exemplos:

Calcular Q_1 e Q_3 para os dados seguintes:

- 28 30 34 38 44 44 46 50 52 61 65

  n=11 (n ímpar)

$$Q_1 = X_{\frac{n+1}{4}} = X_{\frac{12}{4}} = X_3 = 34 \quad \text{e} \quad Q_3 = X_{3\frac{n+1}{4}} = X_{\frac{36}{4}} = X_9 = 52$$

- 28 30 34 38 44 44 46 50 52 61

  n=10 (n par)

$$Q_1 = X_{\frac{n+2}{4}} = X_{\frac{12}{4}} = X_3 = 34 \quad \text{e} \quad Q_3 = X_{\frac{3n+2}{4}} = X_{\frac{32}{4}} = X_8 = 50$$

Os *decis* são os valores da variável que dividem a distribuição em dez partes iguais enquanto que os *percentis* a dividem em 100 partes iguais.
O número de decis é nove e o de percentis noventa e nove.

Para o cálculo do valor exato dos quartis, decis e percentis quando as variáveis são contínuas e estão agregadas em classes, tem-se uma fórmula idêntica à apresentada para o cálculo da mediana:

$$Q_i/D_i/P_i = lc_i + \frac{n_i - cumF_{i-1}}{F_i} \cdot a_{C_i}$$

em que:

lc_i = limite inferior da classe escolhida

n_i = número de observações acumuladas antes do quartil considerado

$cumF_{i-1}$ = frequência acumulada da classe anterior à classe escolhida

F_i = frequência absoluta da classe escolhida

a_{C_i} = amplitude da classe escolhida.

Exemplo:

Calcular o 4º decil e o 72º percentil da seguinte distribuição:

Classes	F_i	cumF_i	f_i(%)	cumf_i
4 – 9	8	8	20	20
9 – 14	12	20	30	50
14 – 19	17	37	42,5	92,5
19 – 24	3	40	7,5	100
	$\sum = 40$		$\sum = 100$	

O 4º decil é o valor que divide a amostra em duas partes, uma com 40% dos elementos e outra com 60%. Assim o 4º decil deverá corresponder àquele que acumula 16 observações uma vez que $0{,}40 \times 40 = 16$.

A classe que contém o 4º decil é 9 – 14 logo, o 4º decil é igual a:

$$D_4 = 9 + \left(\frac{16-8}{12}\right) \times 5 = 12{,}33$$

O 72º percentil será o valor da variável que acumula $0{,}72 \times 40 = 28{,}8$ das observações.

A classe escolhida é 14 – 19 logo, o 72º percentil será:

$$P_{72} = 14 + \left[\frac{28{,}8-20}{17}\right] \times 5 = 16{,}89$$

Portanto, nesta distribuição, o valor 12,33 divide a amostra em duas partes: uma com 40% dos elementos e outra com 60%. O valor 16,89 indica que 72% das observações estão abaixo deste valor e 28% acima dele.

Diagrama de extremos e quartis

É um meio útil para resumir toda a informação contida nas observações.
A observação do diagrama de extremos e quartis para um conjunto de observações revela informação respeitante à quantidade de dispersão, localização de concentração e simetria dos dados.
É muito útil quando se pretende comparar duas ou mais amostras.

A construção do diagrama de extremos e quartis implica o uso dos quartis e pode ser acompanhada das seguintes cinco etapas:

- ✓ Representação da variável em estudo no eixo horizontal.
- ✓ Desenho de uma caixa retangular no espaço acima do eixo horizontal de tal modo que o extremo esquerdo da caixa alinhe com o 1º quartil, e o extremo direito alinha com o terceiro quartil.
- ✓ Divisão da caixa em duas partes por uma linha vertical que alinha com a mediana.
- ✓ Desenho de uma linha horizontal desde o extremo esquerdo da caixa até ao ponto que alinha com o valor mínimo (Min (X_i)) do conjunto de observações.
- ✓ Desenho de outra linha horizontal desde o extremo direito da caixa até ao ponto que alinha com o valor máximo (Max (X_i)) do conjunto de observações.

Exemplo:

Dado um conjunto de dados:

55	50	80	60	70	75	45	45	80	70

Construir o diagrama de extremos e quartis.

1º) Ordenar os dados

45	45	50	55	60	70	70	75	80	80	(n = 10)

2º) Max(X_i)=80 e Min(X_i)=45

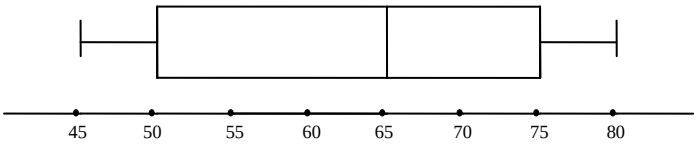
Como n é par:

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n+2}{2}}}{2} = \frac{X_5 + X_6}{2} = \frac{60+70}{2} = 65$$

$$Q_1 = X_{\frac{n+2}{4}} = X_{\frac{12}{4}} = X_3 = 50$$

$$Q_3 = X_{\frac{3n+2}{4}} = X_{\frac{32}{4}} = X_8 = 75$$

3º) Construção do diagrama de extremos e quartis



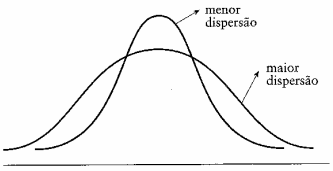
3. MEDIDAS DE DISPERSÃO

3.1 Introdução

A dispersão de um conjunto de observações refere-se à variabilidade que as observações exibem. Uma medida de dispersão dá a informação respeitante à quantidade dessa variabilidade.

Se todos os valores são iguais, não há dispersão. Se os valores não são todos iguais, então há dispersão.

A quantidade de dispersão pode ser baixa quando os valores embora diferentes se encontram muito próximos.



Exemplo:

Considerem-se os dois conjuntos de observações:

X_1	40	42	44	46	50	52	54	56
X_2	0	10	20	40	56	76	86	96

Os conjuntos têm a mesma média e a mesma mediana, mas o 2º conjunto é mais disperso.

As medidas de dispersão podem ser medidas de distância, cujos valores se apresentam na mesma unidade de medida dos dados originais e não recorrem ao cálculo prévio de uma medida de localização. São exemplos de medidas de distância o *intervalo de variação* e o *intervalo interquartis*.
Existem outras medidas de dispersão que utilizam uma medida de localização como termo de comparação. É o caso do *desvio absoluto médio*, da *variância* e do *desvio-padrão*.
Quando se pretende comparar a dispersão entre dois ou mais conjuntos de dados pode ainda utilizar-se uma medida de dispersão relativa como, por exemplo, o *coeficiente de variação*.

Como os próprios termos dispersão e variação o indicam, este tipo de medidas de estatística descritiva identifica o modo como os valores observados se afastam das medidas de tendência central.
Existem duas razões importantes que justificam o cálculo de uma medida de dispersão. A primeira consiste na necessidade de se obter um instrumento que permita avaliar a representatividade do valor médio num conjunto de dados. Se existir uma grande variação dos valores observados, uma medida de tendência central, por si só, pouca informação dá. A segunda razão prende-se com a necessidade de conhecer a dispersão dos valores com o objetivo de serem tomadas medidas concretas para a controlar.

3.2 Intervalo de variação

O *intervalo de variação* (R) é talvez a medida de dispersão mais simples de calcular. É a diferença entre os valores máximos e mínimo da variável:

$$R = Max(X_i) - Min(X_i)$$

A sua utilização tem como principal desvantagem o facto de ter em conta apenas os dois valores extremos que a variável toma e, portanto, não ser sensível aos valores intermédios podendo dar uma ideia errada da verdadeira dispersão da variável.

3.3 Intervalo interquartis

Uma outra medida de dispersão é o *intervalo interquartis*. É definido como a diferença entre o terceiro e o primeiro quartil e corresponde a um intervalo que engloba 50% das observações centrais.

$$IQ = Q_3 - Q_1$$

Tem a desvantagem de não ser influenciado por 50% dos valores observados que correspondem aos valores extremos.
Por não ser afetado por valores extremos, a sua utilização é aconselhada para distribuições altamente assimétricas.

3.4 Desvio absoluto médio

Permite estudar o comportamento dos desvios de cada valor relativamente à média, ou seja, $X_i - \mu$.

- Para dados desagregados:

$$DM = \frac{\sum \left| X_i - \overline{X} \right|}{n}$$

- Para dados agregados:

$$DM = \sum f_i \left| X_i - \overline{X} \right|$$

O desvio absoluto médio é uma medida de dispersão não negativa e quanto maior o seu valor, maior a dispersão da variável.

3.5 Variância e desvio-padrão

A *variância* permite-nos medir a dispersão dos valores da observação relativamente ao centro.

- Se os dados estão desagregados:

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

- Se os dados estão agregados:

$$S^2 = \frac{\sum F_i (X_i - \overline{X})^2}{n}$$

- Se os dados estão agregados em classes:

$$S^2 = \frac{\sum F_i (C_i - \overline{X})^2}{n}$$

Quando se utilizam valores retirados de uma amostra de pequena dimensão é mais correto dividir a soma dos quadrados por n – 1, em vez de n. O estimador assim encontrado constitui um estimador não enviesado ou centrado para o verdadeiro parâmetro da população, propriedade muito importante quando o objetivo é generalizar a toda a população a informação recolhida na amostra.

A variância tem a desvantagem de se traduzir no quadrado das unidades em que está definida a variável X. Por essa razão, é mais utilizado, como medida de dispersão, o *desvio-padrão* S que se define como sendo a raiz quadrada da variância:

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

Por exemplo, se se estiver a tratar uma variável cujas unidades são metros, a variância estará definida em metros^2 e o desvio-padrão em metros. O desvio-padrão é a medida de dispersão mais importante e mais utilizada.

Exemplos:

1. Calcular a média, a variância e o desvio-padrão da seguinte distribuição de frequências:

X_i	5	7	8	9	11
F_i	2	3	5	4	2

Considere-se o quadro seguinte:

X_i	F_i	$X_i\ F_i$	$F_i(X_i - \overline{X})^2$
5	2	10	18,764
7	3	21	3,390
8	5	40	0,020
9	4	36	3,512
11	2	22	17,252
	$\sum = 16$	$\sum = 129$	$\sum = 42{,}938$

Média:

$$\overline{X} = \frac{129}{16} \approx 8{,}063$$

Variância:

$$S^2 = \frac{\sum F_i \left(X_i - \overline{X}\right)^2}{n-1} = \frac{42{,}938}{15} \approx 2{,}863 \text{ (pequena amostra)}$$

Desvio-Padrão:

$$S = \sqrt{S^2} = \sqrt{2{,}863} \approx 1{,}692$$

2. Calcular a média, a variância e o desvio-padrão da seguinte distribuição de frequências:

Classes	F_i	C_i	$C_i\ F_i$	$F_i\left(C_i - \overline{X}\right)^2$
2 – 4	2	3	6	35,280
4 – 6	4	5	20	19,360
6 – 8	7	7	49	0,280
8 – 10	4	9	36	12,960
10 – 12	3	11	33	43,320
	$\sum = 20$		$\sum = 144$	$\sum = 111{,}200$

Média:

$$\overline{X} = \frac{144}{20} = 7{,}2$$

Variância:

$$S^2 = \frac{\sum F_i \left(C_i - \overline{X}\right)^2}{n-1} = \frac{111{,}200}{19} = 5{,}853$$

Desvio-Padrão:

$$S = \sqrt{S^2} = \sqrt{5{,}853} \approx 2{,}419$$

Quanto maior for o desvio-padrão maior será a dispersão dos valores relativamente à média.

Interpretação do desvio padrão

Quanto maior for o desvio padrão maior será a dispersão dos valores relativamente à média.

Combinando o conhecimento da média e do desvio-padrão, pode-se, em muitas situações, caraterizar a localização e a dispersão dos valores.

Caraterísticas do desvio padrão

1. É a medida de dispersão mais utilizada pois tem propriedades matemáticas que o tornam particularmente apropriado em situações de inferência estatística.

2. O desvio padrão é afetado por todos os valores observados e, portanto, qualquer alteração nestes provoca uma alteração do primeiro.

3. O seu valor pode ser fortemente influenciado por apenas alguns valores extremos, o que faz com que a sua utilização seja menos aconselhada em distribuições altamente assimétricas.

Quando a distribuição é normal, a média, a mediana e a moda coincidem e, entre outras características, tem-se, aproximadamente:

- no intervalo $]\overline{X}-S,\overline{X}+S[$, 68,3% das observações.
- no intervalo $]\overline{X}-2S,\overline{X}+2S[$, 95,4% das observações.
- no intervalo $]\overline{X}-3S,\overline{X}+3S[$, 99,7% das observações.

3.6 Coeficiente de variação

Quando é necessário comparar a dispersão em dois conjuntos dados, ao comparar os desvios-padrão respetivos, pode ser-se conduzido a resultados falsos se as duas variáveis envolvidas forem medidas em unidades diferentes. Por exemplo, se se pretender comparar os níveis de colesterol de uma certa população, em miligramas, com o peso do corpo, em quilos.
Nestes casos é preciso uma medida de variação relativa: *coeficiente de variação*, que se expressa em percentagem.

$$CV = \frac{S}{\overline{X}} \cdot 100$$

O coeficiente de variação é uma medida independente da unidade de medição. Em termos práticos é usual considerar-se que um coeficiente de variação superior a 50% indica alto grau de dispersão relativa.

Exemplo:

Numa empresa o salário médio dos homens é de 400€ com desvio padrão de 150€ e o das mulheres é em média 350€ com desvio-padrão de 100€. Então:

$$CV = \frac{S}{\overline{X}} \cdot 100 = \frac{150}{400} \cdot 100 = 37{,}5\% \text{ para os homens}$$

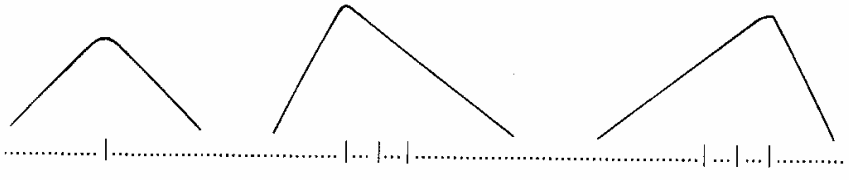
$$CV = \frac{S}{\overline{X}} \cdot 100 = \frac{100}{350} \cdot 100 = 28{,}6\% \text{ para as mulheres}$$

Conclui-se que o salário dos homens apresenta uma maior dispersão relativa que os salários das mulheres.

4. MEDIDAS DE ASSIMETRIA E CURTOSE

4.1 Medidas de Assimetria

O método mais simples para se medir o grau de assimetria de uma distribuição consiste na comparação de três medidas de tendência central: a média, a mediana e a moda.
Numa distribuição simétrica a média, a mediana e a moda são iguais (figura 1).
Se a média $\geq$ mediana $\geq$ moda, a distribuição é assimétrica positiva ou enviesada à esquerda (figura 2).
Se a média $\leq$ mediana $\leq$ moda, a distribuição é assimétrica negativa ou enviesada à direita (figura 3).



$\mu = M_0 = M_e$ | $M_0 \leq M_e \leq \mu$ | $\mu \leq M_e \leq M_0$

Figura 1: Distribuição simétrica | *Figura 2: Assimetria positiva* | *Figura 3: Assimetria negativa*

Quanto mais pronunciada for a assimetria da distribuição maior será a distância entre a moda e a mediana.

Através da comparação da média com a moda é possível medir com mais precisão a assimetria, calculando o coeficiente de Pearson:

$$G_1 = \frac{\overline{X} - Mo}{S}$$

Observe-se que se:

- $G_1 = 0$ a distribuição é simétrica
- $G_1 > 0$ a distribuição é assimétrica positiva;
- $G_1 < 0$ a distribuição é assimétrica negativa.

Observação:

Caso não se tenha a moda ou ocorram distribuições bimodais, trimodais ou multimodais; deve ser utilizado o 2º coeficiente de Pearson:

$$G_2 = \frac{3\left(\overline{X} - Me\right)}{S}$$

Da mesma forma:

- $G_2 = 0$ a distribuição é simétrica
- $G_2 > 0$ a distribuição é assimétrica positiva;
- $G_2 < 0$ a distribuição é assimétrica negativa.

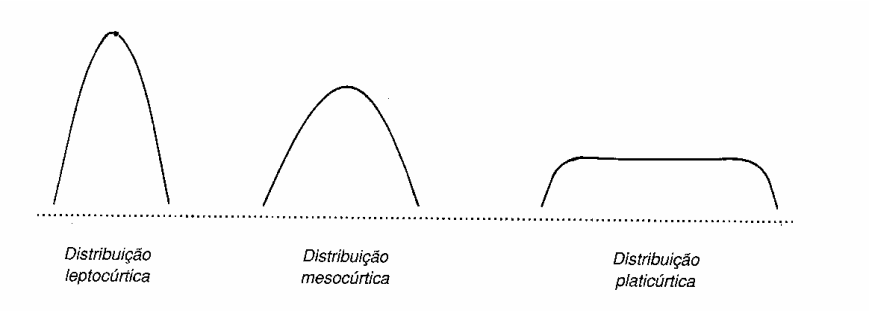
O 1º e 2º coeficiente de assimetria de Pearson permitem a comparação entre duas ou mais distribuições e a avaliação de qual delas é mais simétrica. O seu valor em módulo indica a intensidade da assimetria… quanto maior o coeficiente de assimetria de Pearson, mais assimétrica é a curva.

- Se $0 < \left| G_1 \right| < 0{,}15$, a assimetria é fraca.
- Se $0{,}15 \leq \left| G_1 \right| \leq 1$, a assimetria é moderada.
- Se $\left| G_1 \right| > 1$, a assimetria é forte.

4.2 Medidas de Achatamento ou Curtose

O grau de achatamento ou curtose de uma distribuição indica a intensidade das frequências na vizinhança dos valores centrais.

Grau de curtose de uma distribuição



O grau de curtose pode ser calculado da seguinte forma:

$$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

Observe-se que:

- $K = 0{,}263$: distribuição *mesocúrtica*.
- $K > 0{,}263$: distribuição *platicúrtica*.
- $K < 0{,}263$: distribuição *leptocúrtica*.

Exercícios:

1. A administração do condomínio de um prédio com 25 apartamentos, todos habitados, fez o levantamento do número de pessoas que habitam em cada um deles.
Os resultados foram os seguintes:

2	5	3	2	4
3	6	4	4	3
2	5	2	5	3
4	3	3	4	6
3	5	3	3	2

1.1. Construa uma tabela de frequências.

1.2. Construa um gráfico para as frequências absolutas.

1.3. Construa um gráfico para as frequências acumuladas.

2. Com o objetivo de fazer um estudo sobre o número de jornais e/ou revistas comprados ao fim de semana, fez-se uma sondagem, tendo-se organizado os resultados na seguinte tabela:

Nº de jornais e/ou revistas comprados	Frequência absoluta
0	15
1	70
2	54
3	42
4	15
5	4

2.1. Construa um gráfico para as frequências absolutas.

2.2. Construa um gráfico para as frequências acumuladas.

3. Uma empresa pretende efetuar o estudo estatístico das idades dos seus funcionários. Para tal seleciona aleatoriamente um grupo de 36 trabalhadores de diferentes países, e obtém os valores abaixo apresentados. Construa o diagrama de caule-e-folhas respetivo.

47	19	37	37	28	37	61	34	48	32	36	36
34	29	31	27	35	35	34	22	31	36	45	31
34	40	42	23	35	36	41	32	69	58	26	28

4. Uma amostra de 25 caixas de bombons foi selecionada de um stock de 1000 caixas. Os pesos em gramas das caixas foram os seguintes:

93	100	106	104	98
97	98	104	92	94
101	103	96	100	108
100	108	97	103	100
94	104	95	101	102

4.1. Construa uma tabela de frequências, agrupando o peso das caixas em intervalos de amplitude 5 gramas.

4.2. Construa o histograma e o polígono de frequências absolutas.

4.3. Construa o histograma e o polígono de frequências absolutas acumuladas.

5. As horas de sol por dia, registadas numa praia durante um período de 60 dias, foram as seguintes:

Horas de sol	5	6	7	8	9	10	11
Frequência (dias)	6	12	10	9	8	8	7

5.1. Indique o número modal de horas de sol por dia.

5.2. Indique o número mediano de horas de sol por dia.

5.3. Determine o número médio de horas de sol por dia.

5.4. Construa o diagrama diferencial e o diagrama integral.

6. Durante uma hora, na travessia de uma ponte, foi feito o registo do número de passageiros transportados em cada automóvel que atravessava a ponte em ambos os sentidos:

Nº de passageiros	Nº de automóveis
1	37
2	48
3	28
4	20
5	12

6.1. Construa uma tabela de frequências.

6.2. Construa o diagrama integral e o diagrama diferencial.

6.3. Determine as medidas de localização de tendência central.

6.4. Calcule o coeficiente de variação.

6.5. Construa o diagrama de extremos e quartis.

7. As alturas de 100 pessoas estão registadas na tabela seguinte:

Alturas (cm)	Frequência
[152,158[	5
[158,164[	18
[164,170[	42
[170,176[	27
[176,182[	8
Total	100

7.1. Determine a média, a moda, a mediana, o 3º quartil, o 7º decil e 90º percentil.

7.2. Construa o histograma e o polígono de frequências simples.

8. Suponha que, numa dada região, o Rendimento Social de Inserção por família obedece à seguinte distribuição:

Valor mensal (€)	Nº de famílias
[0,100[	6
[100,200[	19
[200,300[	34
[300,400[	6

8.1. Determine a média dos valores mensais do Rendimento Social de Inserção atribuídos na região considerada.

8.2. Determine a moda geometricamente e analiticamente.

8.3. Construa um histograma de frequências relativas acumuladas e um diagrama de extremos e quartis.

9. Numa inspeção a 34 aviões verificou-se o número de rebites que faltavam. Estes dados estão registados na tabela.

Número de rebites em falta	Frequência
0 – 4	4
4 – 8	9
8 – 12	11
12 – 16	6
16 – 20	2
20 – 24	1
24 – 28	0
28 – 32	1

9.1. Determine a média e a moda.

9.2. Desenhe o polígono de frequências relativas acumuladas e localize graficamente o 1º, o 2º e o 3º quartis.

9.3. Faça uma estimativa do valor destas medidas.

9.4. Determine o 9º decil e 80º percentil.

9.5. Classifique a distribuição quanto à simetria e ao achatamento.

10. Para a análise da estrutura salarial da indústria têxtil em Portugal, inquiriram-se 500 indivíduos sobre a remuneração mensal que auferem, tendo-se obtido a informação, que depois de classificada, deu origem ao seguinte quadro:

Classes salariais (€)	Número de trabalhadores
350 – 400	80
400 – 450	150
450 – 500	230
500 – 550	30
550 – 600	10

10.1. Determine o intervalo interquartis.

10.2. Calcule o desvio-padrão e a variância.

10.3. Calcule o coeficiente de variação.

10.4. Classifique a distribuição quanto à simetria e ao achatamento.

11. Em certa região do país a distribuição dos rendimentos semanais auferidos (em euros) pelos seus habitantes é a seguinte:

Classes de rendimentos	Número de habitantes
50 – 60	8
60 – 70	10
70 – 80	16
80 – 90	14
90 – 100	10
100 – 110	5
110 – 120	2

11.1. Determine as medidas de tendência central.

11.2. Calcule o coeficiente de variação.

11.3. Classifique a distribuição quanto à simetria e ao achatamento.

12. A seguinte tabela representa a distribuição do número de acidentes semanais ocorridos numa fábrica num período de dois anos.

Número de acidentes	Frequência
0	33
1	42
2	19
3	8
4	1
5	1

12.1. Calcule as medidas de tendência central.

12.2. Determine o coeficiente de variação.

12.3. Classifique a distribuição quanto à simetria e ao achatamento.