CAPÍTULO II

ESTATÍSTICA DESCRITIVA

2.1 INTRODUÇÃO

Pode dizer-se que a Estatística é a ciência de obter informação a partir dos dados. Os dados são simplesmente números, mas não só, são números *com contexto*. Por exemplo, o número 52 não nos dá alguma informação a não ser que esteja inserido num contexto. Se nos disserem que o bebé de uma pessoa que conhecemos nasceu com 52 cm, significa para nós que é um bebé comprido. O contexto em que o número está inserido permite-nos fazer um julgamento ou juízo de valor, o número tornou-se informativo.

Argumentos provenientes dos dados são cada vez mais utilizados em várias profissões e em política. Um espectador atento do noticiário questionar-se-á quando houve afirmações como estas:

O *ministério do Trabalho e da Solidariedade Social* comunicou que a taxa de desemprego no mês passado era de 4%. Como é que o governo obteve esta informação?
Numa outra notícia podemos ouvir dizer que há evidência estatística de que o consumo de tabaco provoca cancro. Que tipo de evidência é esta?

Os métodos estatísticos são hoje amplamente utilizados em praticamente todas as áreas da vida quotidiana.
Os consumidores são inquiridos (telefonicamente, directamente no seu domicílio, etc.) para fornecer informação sobre as suas preferências relativamente a certos produtos que se encontram no mercado, são feitas sondagens para prever o resultado de determinadas eleições, os médicos planeiam experiências para determinar o efeito de medicamentos ou ainda estudam o possível efeito de certas condições ambientais, hábitos alimentares ou nível de vida no aparecimento de determinadas doenças, engenheiros utilizam a estatística para efectuar controlo de qualidade de certos produtos, etc. O objectivo da recolha de dados é fazer inferências sobre os elementos da população em estudo. Assim, suponhamos que procuramos informação acerca de uma colecção de elementos com uma característica comum a que chamamos **população**. Por razões de tempo, dinheiro ou outras, pode não ser possível observar cada elemento da população sendo o nosso objectivo fazer inferências sobre a população com base na informação contida na **amostra**.
À característica comum dos indivíduos da população que assume valores diferentes de indivíduo para indivíduo chama-se **variável.** As variáveis podem ser de dois tipos: **qualitativas e quantitativas.** Para as observações destas variáveis, os dados, também se usa a mesma terminologia.

Dados qualitativos-representam a informação que identifica alguma qualidade, característica ou categoria, não susceptível de medida, mas de classificação.
Por exemplo, a cor de cabelo de uma pessoa: louro, castanho claro, castanho escuro, preto, ruivo, etc..

Dados quantitativos- resultam da atribuição a cada unidade do valor ou intensidade observada.

Uma variável quantitativa pode ser **discreta** se só pode tomar um nº finito ou infinito numerável de valores distintos. Por exemplo, nº de habitantes de uma cidade, nº de divisões de uma habitação, nº de alunos inscritos numa disciplina, nº de acidentes por dia numa determinada via, número de palavras contidas numa página de um livro, etc..
A variável é **contínua** se pode tomar qualquer valor dentro de um intervalo de variação.Estão neste caso o tempo de vida (horas) de uma peça, a duração de uma chamada telefónica, o caudal de um rio num determinado instante, o peso e a altura de uma pessoa, etc.

A colecção dos dados a estudar é em geral numerosa, havendo portanto uma certa dificuldade em tirar conclusões enquanto os dados se mantiverem nesta forma. Há pois necessidade de classificar os dados e indicar de forma sumária as características apresentadas por aquela amálgama de números, assim como efectuar a redução dos dados o que permite concentrar num pequeno número de valores característicos a informação contida nos dados. É destes aspectos que trata a **Estatística Descritiva.**

2.2 TABELAS DE FREQUÊNCIAS

Vamos começar por dar uma possível classificação aos dados e descrever processos adequados para a sua representação.

A **tabela de frequências** é uma organização dos dados que pode ser utilizada tanto para dados discretos como contínuos havendo no entanto diferenças a considerar.

Tomemos o seguinte exemplo relacionado com a autoria de textos, dados recolhidos de uma tese de mestrado realizada no DEIO e intitulada "Quem foi que? - Um desafio à Estatística: Questões de autoria em *Novas Cartas Portuguesas*"

EXEMPLO 1: Recolheu-se uma amostra de 165 palavras contextuais utilizadas na obra "Os Outros Legítimos Superiores" de Isabel Barreno, uma das autoras das referidas cartas.

Na tabela que se segue apresentam-se os dados já classificados da seguinte maneira: a coluna da esquerda indica as categorias (palavras escolhidas) e na coluna da direita está registado o número de vezes que cada palavra aparece na amostra, isto é, a **frequência absoluta** (ou frequência) de cada categoria ou **classe**.

Classes	Freq.abs.
Admirar	7
algodão	1
banido	1
Bocejar	2
capital	2
cerrados	1
Demonstrar	1
destroços	1
engenheiros	1
esboçam-se	1
Farsa	1
Géneros	3
Haver	69
imediato	1
inscrito	1
juvenis	1
lançam	1
Modificar	4
natureza	1
obrigatórias	1
palpando	1
pêlo	1
Querer	37
Raíz(es)	4
respeitosa	1
sacrifícios	1
semelhante	3
tapando	1
toga	4
universalmente	1
Valor(es)	9
zelar	1

Tabela 1

Numa tabela de frequências, além das frequências absolutas, também se podem apresentar as **frequências relativas.**

$$\textbf{frequência relativa} = \frac{\text{frequência absoluta}}{\text{dimensão da amostra}}$$

Dimensão da amostra é o número de elementos que constituem a amostra.

Classes	Freq. abs.	Freq. relativa
Admirar	7	.04242
algodão	1	.00606
banido	1	.00606
Bocejar	2	.01212
capital	2	.01212
cerrados	1	.00606
Demonstrar	1	.00606
destroços	1	.00606
engenheiros	1	.00606
esboçam-se	1	.00606
Farsa	1	.00606
Géneros	3	.01818
Haver	69	.41818
imediato	1	.00606
inscrito	1	.00606
juvenis	1	.00606
lançam	1	.00606
Modificar	4	.02424
natureza	1	.00606
obrigatórias	1	.00606
palpando	1	.00606
pêlo	1	.00606
Querer	37	.22424
Raíz(es)	4	.02424
respeitosa	1	.00606
sacrifícios	1	.00606
semelhante	3	.01818
tapando	1	.00606
toga	4	.02424
universalmente	1	.00606
Valor(es)	9	.05454
zelar	1	.00606

Tabela 2

Se estivermos interessados em estudar a variável *número de letras de cada palavra*, surge a tabela seguinte que resulta da anterior agrupando numa mesma classe as palavras com o mesmo nº de letras, assim:

Nº de letras	freq. absolutas	freq. relativas
4	5	0.03030303
5	71	0.43030303
6	43	0.26060606
7	26	0.15757576
8	5	0.03030303
9	6	0.03636364
≥10	9	0.05454545

Tabela 3

Note-se que $\sum_{i=1}^{k} f_i = n; \quad \sum_{i=1}^{k} \frac{f_i}{n} = \sum_{i=1}^{k} fr_i = 1$ onde *k* é o número de classes e *n* é a dimensão da amostra.

EXEMPLO 2: Os valores da tabela 4 representam o tempo de vida de n=88 rádio emissores e receptores.

Tempos de vida

16	448	552	256	40
224	716	72	246	12
16	304	184	328	112
80	16	240	464	288
96	72	438	156	168
536	8	120	216	352
400	80	308	168	56
80	72	32	184	72
392	56	272	168	64
576	608	152	40	40
128	108	328	152	184
56	194	480	360	264
656	136	60	96	176
224	224	208	224	160
40	80	340	168	208
32	16	104	168	152
358	424	72	114	
384	264	232	280	

Tabela 4

Trata-se de dados contínuos e temos de definir as classes da tabela de frequências, tarefa que não é tão evidente como no caso discreto.

Não tem sentido considerar para classes os diferentes valores que ocorrem na amostra pois eventualmente eles são todos diferentes, dado que a variável em estudo varia num intervalo. Como formar então as classes?

PASSOS:

1º Definição das classes

i) determinar a amplitude da amostra ou range

amplitude= R= valor máximo-valor mínimo

ii) dividir a amplitude pelo número de classe **k**, a determinar posteriormente. Tomar como amplitude de classe **h**, um valor aproximado por excesso do valor obtido anteriormente.

Nem sempre se consegue que a amplitude de classe **h** seja constante.

iii) representem-se as classe por I_j, j=1,...,k . Para não haver ambiguidades as classes devem ser disjuntas, isto é,

$$I_i \cap I_j = \varnothing, \quad i \neq j \qquad (1)$$

Por outro lado, para não ficarem elementos por classificar a união de todas as classes deve conter todos os elementos da amostra.

Seguindo estas duas regras podemos definir os k intervalos da seguinte maneira

$$I_1 = [l_1, l_2[, \quad I_2 = [l_2, l_3[, \ldots, \quad I_k = [l_k, l_{k+1}] \text{ com} \tag{2}$$

$l_1 < l_2 < \ldots < l_{k+1}$ e $l_1 \leq$ mínimo valor observado, $l_{k+1} \geq$ máximo valor observado

2º Contagem do número de elementos de cada classe:

Conta-se o número de elementos da amostra que pertence a cada classe, obtendo-se a frequência absoluta da classe.

Se a amostra é constituída pelos elementos $x_1, x_2, \ldots, x_n$

$$x \in I_j \Leftrightarrow l_j \leq x < l_{j+1} \tag{3}$$

À excepção da última classe para a qual se tem

$$x \in I_k \Leftrightarrow l_k \leq x \leq l_{k+1} \tag{4}$$

Quantas classes vamos considerar? Como regra prática podemos utilizar a regra de *Sturges* segundo a qual

$$K = [log_2\, n] + 1 \tag{5}$$

EXEMPLO 3: Voltemos aos dados do exemplo 2 e ordenemos a amostra, obtendo-se

T. de vida (amostra ordenada)

8	72	152	224	358
12	72	152	224	360
16	72	156	224	384
16	72	160	232	392
16	80	168	240	400
16	80	168	246	424
32	80	168	256	438
32	80	168	264	448
40	96	168	264	464
40	96	176	272	480
40	104	184	280	536
40	108	184	288	552
56	112	184	304	576
56	114	194	308	608
56	120	208	328	656
60	128	208	328	716
64	136	216	340	
72	152	224	352	

Tabela 5

A amplitude desta amostra é **R**= 716-8= 708, e porque n=88, $7 < K < 8 \ (2^7 = 128)$. Se tomarmos K=7 vem, $h = \frac{708}{7} = 101{,}14$, e arredondondando este valor para 102

Tempos de vida

classe	freq.absoluta	freq.relativa
[8,110[	30	0,34
[110,212[	22	0,25
[212,314[	16	0,18
[314,416[	9	0,10
[416,518[	5	0,06
[518,620[	4	0,05
[620,722]	2	0,02

Tabela 6

Nem sempre se pode trabalhar com classes de amplitude constante, é exemplo disso a distribuição dos rendimentos. Enquanto que as classes referentes aos rendimentos mais baixos devem ter uma amplitude pequena, as classes para os rendimentos mais elevados deverão possuir uma maior amplitude. De facto, interessa saber quantas famílias auferem rendimentos entre os 150 euros e os 200 euros, mas será informação desnecessária saber quantas têm rendimentos entre 2000 euros e 2050 euros. Isto é, basta saber quantas auferem rendimentos entre 2000 euros e 2250 euros, aumentando ainda mais a amplitude para rendimentos mais elevados. Se fosse adoptada uma amplitude constante de 50 euros teríamos um nº de classes incomportável.

2.3 REPRESENTAÇÃO GRÁFICA DOS DADOS

Numa primeira fase de estudo de uma população é importante a habilidade de descrever um conjunto de valores observados. Os gráficos de caule-e-folhas e os gráficos de frequências (histogramas ou polígonos de frequências) são um precioso auxiliar para esta análise preliminar.

2.3.1 Caule e Folhas

Os dados são usualmente apresentados como uma colecção de números, mas esta estrutura simples pode esconder características importantes num estudo mais superficial. A representação gráfica de caule-e-folhas permite-nos organizar os números graficamente, de forma a dirigir a nossa atenção para várias características dos dados. No gráfico de caule-e-folhas podemos ver a colecção como um todo e notar certas características, tais como:

A que ponto se pode considerar a colecção aproximadamente simétrica.
A que ponto os números estão dispersos.
Se existem alguns valores muito distantes de todos os outros.
Se existem concentrações de dados.
Se existem lacunas entre os dados.

Para explicar esta técnica de representação gráfica dos dados comecemos com o seguinte

EXEMPLO 1: TEMPERATURA MÉDIA DO AR (ºC) EM 1990

Tabela 1

Estações Meteorológicas	Temp. Média(ºC)	Estações Meteorológicas	Temp. Média(ºC)
Bragança	13.2	Portalegre	15.8
Viana	15.6	Santarém	16.7
Braga	15.3	Lisboa	17.5
Vila Real	14.1	Évora	16.2
Miranda	13.0	Setúbal	16.7
porto	15.5	Beja	17.1
Viseu	14.4	P.Rocha	17.6
Guarda	11.4	Faro	17.7
Coimbra	16.6	Ponta Delgada	17.4
C. Branco	16.0	Funchal	19.4

Anuário Estatístico de Portugal, I.N.E., Lisboa, 1991

Para fazer a representação de caule-e-folhas comecemos por traçar uma linha vertical e do lado esquerdo os dígitos dominantes, que neste caso são os das unidades e das dezenas.

No primeiro passo limitamo-nos a colocar os dígitos dominantes que são os caules, e depois teremos de pendurar em cada caule as folhas respectivas. O primeiro número da amostra é o 13.2, pelo que vamos pendurar o 2 no caule 13 (2º passo) e o processo repete-se até termos esgotado todos os elementos da amostra (passo 3).

1º passo

11
12
13
14
15
16
17
18
19

2º passo

11	
12	
13	2
14	
15	
16	
17	
18	
19	

3º passo

11	4
12	
13	2 0
14	1 4
15	6 3 5 8
16	6 0 7 2 7
17	5 1 6 7 4
18	
19	4

É usual apresentar as folhas de cada caule ordenadas, isto é,

Gráfico de caule-e-folhas para a temperatura média (ºC)

11	4
12	
13	0 2
14	1 4
15	3 5 6 8
16	0 2 6 7 7
17	1 4 5 6 7
18	
19	4

Figura 1

Nesta representação observa-se imediatamente que há duas situações extremas, a "fria" Guarda e o "quente" Funchal.

Consideremos os dados da tabela 1 por ordem crescente de valor, aliás basta escrever por ordem todos os valores da representação gráfica anterior. Genericamente estes valores representam-se por

$$x_{(1)},x_{(2)},..,x_{(n)} \quad \text{com} \quad x_{(1)} \le x_{(2)} \le ... \le x_{(n)} \qquad (1)$$

Assim a cada uma das observações pode associar-se uma ordem contando a sua posição a partir do menor ou do maior dos valores observados.

Por exemplo, $x_{(i)}$ tem ordem *i* se a contagem for crescente e iniciada em $x_{(1)}$ e tem ordem *n-i+1* se a contagem for decrescente e iniciada em $x_{(n)}$; a **profundidade** de uma observação é a menor das ordens *i* e *n-i+1*.

Na figura seguinte representa-se na coluna da esquerda a profundidade máxima das observações em cada caule

Profundidade	n=20		unidade=0.1ºC
1	11	4	(Guarda)
	12		
3	13	0 2	
5	14	1 4	
9	15	3 5 6 8	
(5)	16	0 2 6 7 7	
6	17	1 4 5 6 7	
	18		
1	19	4	(Funchal)

Figura 2

A excepção verifica-se na "linha central" em que se encontra a observação mediana, onde em vez da profundidade se reproduz entre parêntesis o número de observações deste caule. Esta representação permite saber qual o número total de observações de uma forma rápida, n=9+(5)+6.

EXEMPLO 2: DUREZA DE MOLDES DE ALUMÍNIO

Shewhart (1931) fornece a dureza de 60 moldes de alumínio. Na tabela seguinte estão representados os 30 primeiros valores desta colecção.

DADOS: Dureza de moldes de alumínio

53.0	70.2	84.3	55.3	78.5	63.5	71.4	53.4
82.5	67.3	69.5	73.0	55.7	85.8	95.4	51.1
74.4	54.1	77.8	52.4	69.1	53.5	64.3	82.7
55.7	70.5	87.5	50.7	72.3	59.5		

Tabela 2

Comecemos por ordenar os dados por ordem crescente do seu valor

DADOS ordenados

50.7	51.1	52.4	53.0	53.4	53.5	54.1	55.3
55.7	55.7	59.5	63.5	64.3	67.3	69.1	69.5
70.2	70.5	71.4	72.3	73.0	74.4	77.8	78.5
82.5	82.7	84.3	85.8	87.5	95.4		

Tabela 3

Qual vai ser o comprimento dos intervalos que vamos considerar para agrupar os dados?

A amplitude (range) destes valores é **R**=máximo valor observado-mínimo valor observado=95.4-50.7=44.7, um valor considerável. Podemos começar por escolher o número de linhas do gráfico usando a seguinte regra de *Sturges*, o número máximo de linhas

$$L = \lfloor log_2\, n \rfloor + 1$$

onde *n* é o número de observações e $\lfloor x \rfloor$ denota o maior inteiro que não excede x (ou parte inteira de x). Esta regra parece fornecer valores de L convenientes para 20≤n≤300, dimensões usuais em estatística. Neste caso concreto tem-se

$$2^{L-1} \approx 30$$

Logo 5< L< 6 como limite grosseiro do número de linhas na representação, temos agora que determinar os intervalos correspondentes a cada linha. A maneira mais simples de o fazer é usar uma potência de 10 como comprimento do intervalo. Calcule-se $7 < R/L < 9$ e arredondemos para cima até à potência de 10 mais próxima, isto é, neste caso 10. Este comprimento é usado na linha da esquerda da representação, assim o valor 50.7 aparece nesta representação como $5|0$, os valores 55.3, 55.7 e 55.7 aparecem todos como $5|5$. Isto é, os valores são truncados até ao ponto decimal e não arredondados.

Representação 1:

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

Esta representação tem poucas linhas, assim vamos separar as linhas e repetir os caules, obtendo-se a representação:

Representação 2:

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

Acontece por vezes que mesmo com duas linhas por caule continuamos a ter uma representação demasiado pesada, se o número de observações é considerável.

2.3.2 Diagrama de Barras

Comecemos por considerar uma característica de indivíduos de uma população que é uma variável discreta, o tipo de representação gráfica apropriada para estas variáveis é o diagrama de barras. Retomemos o exemplo 1 da secção 2, relacionado com a autoria de textos. Os dados estão agrupados na tabela de frequências (tabela 3). No diagrama de barras a altura de cada barra representa a frequência absoluta de cada valor numérico da variável- *nº de letras da palavra.*

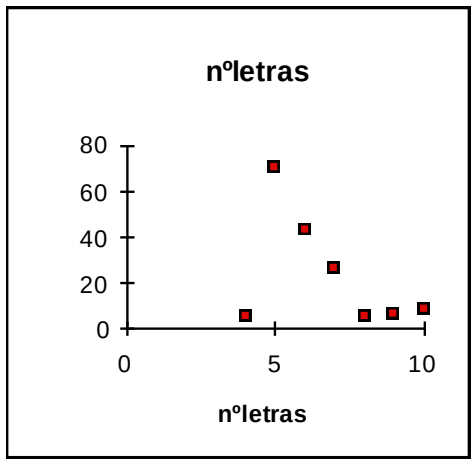

Figura 3

Podemos também obter o diagrama de barras representando em ordenadas as frequências relativas, isto é, $fr_i = \frac{f_i}{N}$, $N = \sum_{i=1}^{7} f_i - \text{nº total de observações}$. Por se ter $\sum_{i=1}^{7} fr_i = 1$, a representação em termos de frequências relativas permite a comparação de diagramas com um número diferente de observações.

Nº de letras	fr. relativas
4	0.03030303
5	0.43030303
6	0.26060606
7	0.15757576
8	0.03030303
9	0.03636364
≥10	0.05454545

Tabela 4

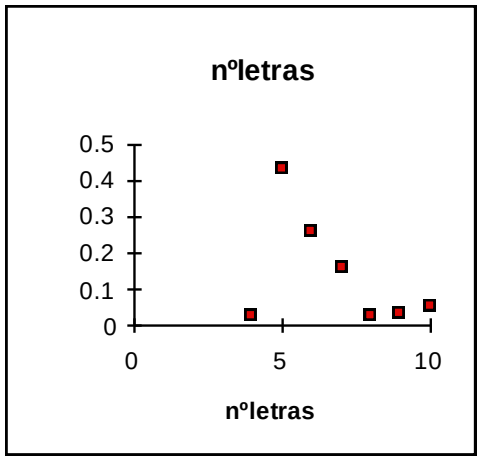

Figura 4

2.3.3 Histograma

Quando a variável que é objecto do nosso estudo é *contínua* a representação gráfica da distribuição de frequências é obtida por meio de um diagrama de áreas, **o histograma.** Este gráfico é formado por uma sucessão de rectângulos adjacentes tendo cada um por base um intervalo de classe e por altura a frequência absoluta ou relativa.Voltemos ao dados do exemplo, tempo de vida de rádios, que se encontram na tabela 5 da secção 2. De acordo com as classes e respectivas frequências absolutas registadas na tabela 6 da mesma secção, podemos construir um histograma para esta amostra.

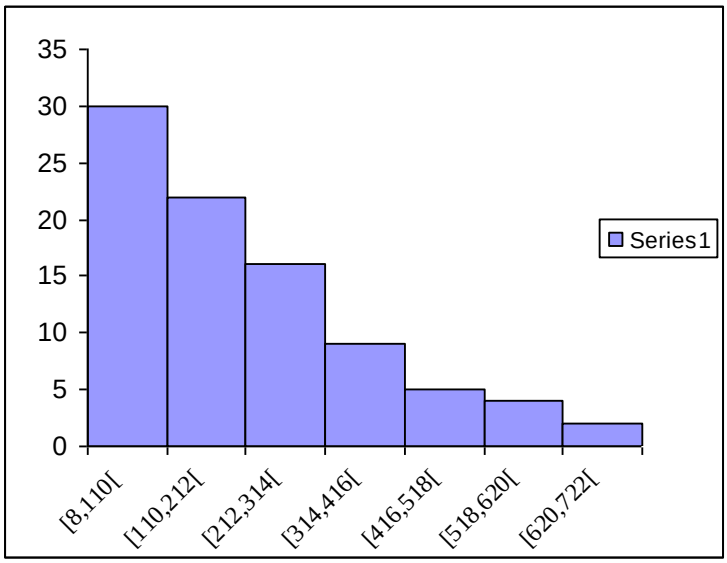

Figura-5

Como interpretar um histograma?

Como dissemos anteriormente, um histograma é somente uma análise preliminar das observações e o seu objectivo é ajudar-nos a interpretar os dados. Depois de fazermos o gráfico podemos perguntar "O que vemos"? Há características importantes para as quais devemos olhar, como por exemplo,

Num gráfico devemos olhar para o seu ***aspecto geral*** *e para os* ***desvios*** *relativamente a este aspecto. Podemos descrever o aspecto geral do histograma pela* ***forma, localização*** *e* ***dispersão.*** *Há medidas numéricas para quantificar a localização e a dispersão. Uma espécie importante de desvio é um* ***outlier****, um indivíduo que não segue o padrão geral.*

Ao observarmos o histograma anterior notamos ainda outras características como por exemplo, o facto da distribuição não apresentar **simetria**, mas ter uma **assimetria direita** dado que o lado direito da distribuição se prolonga mais do que o esquerdo (o lado direito contém as grandes observações). Haverá uma **assimetria esquerda** se for o lado esquerdo a prolongar-se mais.

2.3.4 Função de distribuição empírica

O que é? É uma função que nos dá informação sobre a percentagem de valores da amostra que são superiores ou inferiores a um determinado valor. A função de distribuição empírica designa-se por $F_N(x)$ e para cada valor real do argumento x, dá a frequência relativa das observações que são inferiores ou iguais a x. Esta função tem domínio |R e toma valores no intervalo [0, 1]. Como construir esta função?

CASO DISCRETO: Recolhamos uma amostra constituída pela 5ª e 10ª palavras de cada linha do parágrafo anterior

uma, informação, inferiores, a, para, argumento, são, esta, como

A variável que nos interessa estudar é o nº de letras de cada palavra. Temos então, a amostra que vamos analisar constituída pelos números:

3, 10, 10, 1, 4, 9, 3, 4, 4

A amostra ordenada é: 1, 3, 3, 4, 4, 4, 9, 10, 10. Trata-se de uma variável discreta e como era de esperar ocorrem valores repetidos. A função de distribuição empírica é neste caso

$$F_N(x)=\begin{cases} 0 & se \quad x<1 \\ \frac{1}{9} & se \quad 1\le x<3 \\ \frac{3}{9} & se \quad 3\le x<4 \\ \frac{6}{9} & se \quad 4\le x<9 \\ \frac{7}{9} & se \quad 9\le x<10 \\ 1 & se \quad x\ge 10 \end{cases}$$

O gráfico desta função é uma função em escada, que apresenta descontinuidades nos pontos: 1, 3, 4, 9 e 10. Note-se que quando há valores repetidos o valor do salto é igual a $\frac{r}{n}$, sendo r o nº de repetições desse valor.

Retomemos o exemplo 1 da secção 2, cuja variável em estudo é também o número de letras das palavras contextuais de um livro de Isabel Barreno, e reproduzamos a tabela 3.

Nº de letras	**Frequências acumuladas**
4	0.03030303
5	0.46060606
6	0.72121212
7	0.87878788
8	0.90909091
9	0.94545455
10	1

O gráfico de frequências acumuladas encontra-se representado na figura seguinte.

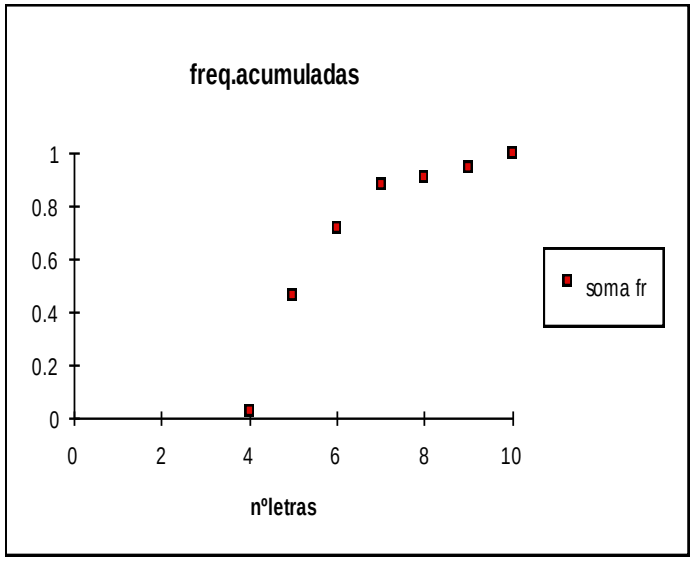

Figura 6

CASO CONTÍNUO:

Retomemos os dados do exemplo 2 e a partir da tabela 6 (secção 2), construa-se a tabela de frequências acumuladas

classe	freq.relativa	Freq. acumulada
[8,110[	0,34	0,34
[110,212[	0,25	0,59
[212,314[	0,18	0,77
[314,416[	0,10	0,87
[416,518[	0,06	0,93
[518,620[	0,05	0,98
[620,722[	0,02	1,00

Tabela 8

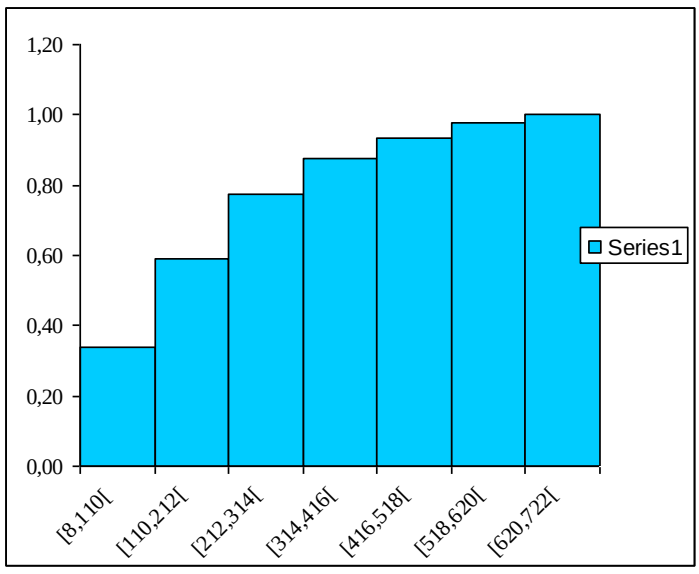

Figura 7

Por definição a função de distribuição empírica dá-nos a percentagem ou proporção de valores da amostra que são menores ou iguais a um determinado valor x. Assim, dado um valor **p** qualquer, entre 0 e 1, podemos obter o valor $\mathbf{Q_p}$, que divide a amostra em duas partes:

100p% dos elementos são menores ou iguais a $\mathbf{Q_p}$ e os restantes são maiores ou iguais a $\mathbf{Q_p}$.
A este valor $\mathbf{Q_p}$ dá-se o nome de **percentil** ou **quantil** de ordem p ou percentagem 100p%.
Alguns percentis têm nome especial como por exemplo:

Mediana. É o percentil correspondente à percentagem 50% e representa-se pela letra *M*.

Quartis: O **1º quartil** é o percentil correspondente à percentagem 25%, isto é, 25% dos elementos da amostra são inferiores ou iguais a ele e os restantes são maiores ou iguais. O **3º quartil** é o quantil correspondente à percentagem 75%.

Mais adiante veremos como caculá-los sem ter de recorrer à função de distribuição empírica.

2.4 CARACTERÍSTICAS NUMÉRICAS

2.4.1 -.Medidas de Localização: média, mediana e moda

Seja $x_1,\ldots,x_n$ um conjunto de n observações, ao valor $\bar{x} = \frac{x_1+\ldots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ chama-se **média** das observações.

Note-se que a média é de fácil cálculo e utiliza todas as observações efectuadas, está definida rigorosamente e pode interpretar-se sem ambiguidades. O facto da média ser calculada usando todas as observações tem vantagens e desvantagens, uma destas é tornar-se sensível a valores aberrantes.
Tomemos uma amostra constituída pelos seguintes salários diários em euros

10.5; 10.4; 10.7; 10.0; 11.0; 11.5; 20

$$\bar{x} = \frac{10.5+10.4+10.7+10.0+11.0+11.5+20}{7} = 12{,}01$$

que não é muito representativa dos salários apresentados, uma vez que é fortemente influenciada pelo salário máximo de 20 euros. Para que a média seja um indicador adequado da localização é necessário que não seja empurrada para os valores extremos do intervalo de variação da amostra por valores muito altos ou baixos.

Ao interpretar uma média devemos ter em conta que ela não tem de ser um valor que pertença à sucessão original de valores observados, e portanto, pode não ter existência real.

Outra medida de localização importante e que é definida pela sua posição na sucessão das observações ou na distribuição de frequências (percentil 50%) é

A **mediana** de um conjunto de observações é o valor em relação ao qual metade das observações são menores do que ele e a outra metade são superiores.

Como encontrar a mediana?

1) Ordenar as observações por ordem crescente de grandeza.

2) Se o número n de observações é *impar*, a mediana **M** é a observação de ordem $\frac{n+1}{2}$.

3) Se o número de observações é *par*, a mediana **M** é a média aritmética das duas observações (ordenadas) centrais.

Ao ordenar as observações devemos escrever todos os valores observados, mesmo que haja valores que se repetem.

EXEMPLO: Os dados ordenados dos tempos de vida de rádios são

T. de vida (amostra ordenada)

8	72	152	224	358
12	72	152	224	360
16	72	156	224	384
16	72	160	232	392
16	80	168	240	400
16	80	168	246	424
32	80	168	256	438
32	80	168	264	448
40	96	168	264	464
40	96	176	272	480
40	104	184	280	536
40	108	184	288	552
56	112	184	304	576
56	114	194	308	608
56	120	208	328	656
60	128	208	328	716
64	136	216	340	
72	152	224	352	

Tabela 1

Uma vez que n é par a mediana **M** é a média aritmética das 44ª e 45ª observações, i e., $\mathbf{M} = \frac{168+168}{2}$.

Como terceira medida de localização temos a

Moda: Define-se no caso discreto como sendo o valor mais frequente. Se os dados são contínuos define-se *classe modal*- classe com maior frequência.

EXEMPLO: Retomemos os exemplos da secção 2. No exemplo 1, a variável *nº de letras de cada palavra* é discreta. Analisando a tabela 3, verificamos que as palavras observadas mais frequentemente são as de 5 letras, logo **moda**=5.
No caso do exemplo 2, a tabela 6 indica-nos que a classe modal é [8, 110[.

2.4.2 - Medidas de dispersão

As medidas de localização só por si podem ser pouco elucidativas da distribuição da característica em estudo.
Consideremos as seguintes classificações obtidas por 3 alunos todos com média de 15, em 5 cadeiras de estatística do seu curso:

I	15	15	15	15	15
II	14	16	15	14	16
III	13	17	14	12	19

Estes conjuntos de dados têm todos a mesma localização, mas dispersões diferentes. Nos primeiros casos temos um aluno muito regular e no 3º caso um aluno que tanto pode ter notas relativamente baixas como notas muito altas, provavelmente só estuda para as cadeiras de que mais gosta.

Retomando ainda os dados relativos aos tempos de vida de rádios, sabemos que a média e a mediana das observações são respectivamente 209.66 e 168, isto é, metade dos rádios tem vida inferior a 168 (u. t.) e outra metade tem duração superior. No entanto se estivéssemos interessados em comprar um destes aparelhos este valor de pouco nos serviria. Duas amostras com a mesma mediana ou média podem ser muito diferentes quanto ao maior e menor valores observados. A tabela das observações ordenadas informa-nos que para esta amostra o rádio com maior duração teve um tempo de vida de 716 u.t. e o rádio que menos durou teve 8 u.t. de duração.

Podemos melhorar a nossa descrição calculando *os quartis*. Como proceder?

De uma forma geral, define-se um quantil empírico de ordem **p** da seguinte maneira:

DEFINIÇÃO: Se 0 < p< 1, o *quantil empírico de ordem p*, Q_p , *é dado por*

$$Q_p = \begin{cases} X_{([np+1])} & \text{se np não é inteiro} \\ \dfrac{X_{(np)} + X_{(np+1)}}{2} & \text{se np é inteiro} \end{cases} \qquad (1)$$

Assim, se $p=\frac{1}{4}$ *ou* $p=\frac{3}{4}$, temos o *1º e 3º quartil* respectivamente. Isto é, como os nomes indicam, o primeiro quartil tem um quarto das observações ordenadas à sua esquerda e três quartos à sua direita e tem-se a situação simétrica desta para o terceiro quartil. Quando $p=\frac{1}{2}$ temos a mediana:

$$M=\begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{se n é ímpar} \\ \dfrac{X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}}{2} & \text{se n é par} \end{cases} \qquad (2)$$

Note-se que se n é ímpar $\left[\frac{n}{2}+1\right]=\frac{n+1}{2}$. (Conferir pág. 18)

Para o exemplo dos tempos de vida de rádios vem:

$$Q_{\frac{1}{4}}=\frac{X_{(22)}+X_{(23)}}{2}=\frac{72+80}{2}=76 \quad \text{e} \quad Q_{\frac{3}{4}}=\frac{X_{(66)}+X_{(67)}}{2}=\frac{288+304}{2}=296.$$

Dispomos agora de 5 características numéricas (estatísticas de ordem) que nos dão uma ideia da dispersão da distribuição dos tempos de vida, são elas:

$$\textit{mínimo},\ Q_{\frac{1}{4}},\ M,\ Q_{\frac{3}{4}},\ \textit{máximo}$$

Para este exemplo temos: 8, 76, 168, 296, 716.

Podemos calcular ainda os quantis (percentis) 10 e o 90. Como os nomes indicam estes percentis têm à sua esquerda 10% e 90% das observações respectivamente e o restante (diferença para 100%) à sua direita.

O output seguinte dá indicação destes valores.

X_1: t.vida

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
209.659	163.22	17.399	26640.756	77.85	88
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**#Missing:**
8	716	708	18450	6185956	0
# < 10th %:	**10th %:**	**25th %:**	**50th %:**	**75th %:**	**90th %:**
8	40	76	168	296	445
# > 90th %:					
9					

Tabela 2

Podemos representar graficamente os 5 números anteriores obtendo-se o seguinte gráfico denominado "caixa de bigodes" (box-plot), onde também estão marcadas as observações à esquerda (8) do percentil 10 e as observações à direita (9) do percentil 90. Estas 17 observações por serem aberrantes, são candidatas a *outlier*.

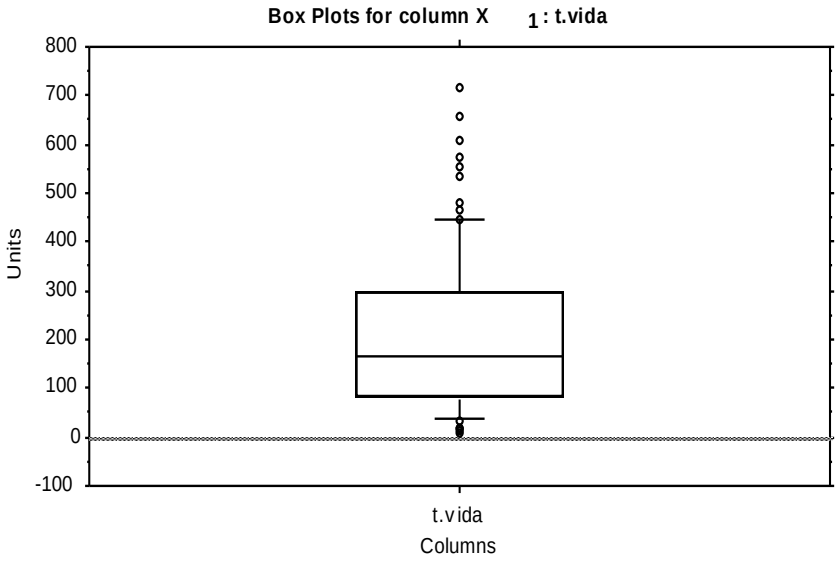

Figura 1

Podemos ainda definir medidas de dispersão a partir destas estatísticas de ordem. As mais comuns são a **Amplitude**

R=Máximo-Mínimo (3)

e a **Amplitude inter-quartil**

$$q = Q_{\frac{3}{4}} - Q_{\frac{1}{4}} \quad (4)$$

A medida de localização *mediana* aparece associada aos quartis da distribuição ou a outros percentis para descrever a localização e a dispersão da distribuição em estudo, no entanto a

mediana não é a única nem a mais comum das medidas de localização. A mais usual é a média e associada a ela a *variância empírica* ou a sua raiz quadrada positiva-o *desvio-padrão*. Estas medidas aparecem também no output anterior com os valores 26640.756 e 163.22 respectivamente (Variance e Standard-deviation) e calculam-se através das seguintes expressões:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 26640756,\ s = \sqrt{s^2} = 163.22$$

Isto é, a variância empírica é uma média ponderada dos desvios quadráticos relativamente à média da amostra $\bar{x}$, é pois uma medida de dispersão quando se utiliza a média como medida de localização.

Observação: Retiremos à nossa amostra ordenada as últimas 8 observações, isto é, os 8 maiores valores observados. Então, se recalcularmos a média, a variância e o desvio-padrão deste novo conjunto de observações obtemos:

X_2: 80 obser

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
173.275	118.102	13.204	13947.974	68.158	80
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**
8	448	440	13862	3503828	8
# < 10th %:	**10th %:**	**25th %:**	**50th %:**	**75th %:**	**90th %:**
8	36	72	164	251	355
# > 90th %:					
8					

Tabela 3

$\bar{x} = 173.275,\ s^2 = 13947.974,\ s = 118.102$, o que mostra que estas medidas são muito afectadas pelas existência de algumas observações extremas.

Se compararmos com o "output" anterior para a amostra completa

X_1: t.vida

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
209.659	163.22	17.399	26640.756	77.85	88
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum Squared:**	**# Missing:**
8	716	708	18450	6185956	0
# < 10th %:	**10th %:**	**25th %:**	**50th %:**	**75th %:**	**90th %:**
8	40	76	168	296	445
# > 90th %:					
9					

Podemos verificar que o mesmo não acontece relativamente à mediana e aos quartis que pouca alteração sofreram. Aliás a amplitude inter-quartis é, para a amostra completa, igual a

$$\mathbf{q_1} = 296\text{-}76 = 220$$

e para a amostra sem as 8 maiores observações

$$\mathbf{q_2} = 251\text{-}72 = 179$$

o que mostra que a amplitude inter-quartis é menos afectada pela presença de valores muito grandes ou muito pequenos do que o desvio-padrão, é pois uma medida de dispersão mais robusta do que este (note-se a semelhança de comportamento com a mediana e a média como medidas de localização). Por outro lado, a amplitude inter-quartis não utiliza todos os dados para o seu cálculo como o desvio-padrão.

Assim, podemos concluir que a mediana e os quartis podem descrever melhor uma distribuição quando esta não é simétrica, isto é, apresenta uma assimetria esquerda ou direita ou quando existem observações aberrantes que podem ser consideradas "outliers". No entanto para distribuições simétricas (ou com pequena assimetria) devemos utilizar a média e o desvio-padrão. É o que acontece, por exemplo, com a distribuição *Gaussiana.* Baseado no estudo dessa distribuição, uma regra empírica para descrever uma distribuição simétrica faz uso dos intervalos da forma $\left(\bar{x} - ks, \bar{x} + ks\right)$, com k inteiro positivo. Embora a distribuição dos tempos de vida apresente uma assimetria direita, ao aplicar a regra para esta amostra obtivemos o seguinte resultado:

	$\left(\bar{x} - ks, \bar{x} + ks\right)$	nº de observações	regra empírica
k=1	(46.439, 372.879)	74	88x0.68=59.84
k=2	(-116.781, 536.099)	82	88x0.950=83
k=3	(-280.001, 699.319)	87	quase todos os valores

Tabela 4

Note-se que apesar da assimetria a regra empírica fornece uma descrição bastante razoável dos dados.

Usando a regra empírica podíamos obter uma aproximação de s, da seguinte maneira:

$$s \cong \frac{amplitude}{6} = 118$$

Nota: Quando se têm dados "normais" uma maneira de construir histogramas pode basear-se na regra do desvio-padrão. Nesta regra a classe central é centrada em $\bar{x}$ e de amplitude $h \in \left(\frac{s}{3}, \frac{s}{2}\right)$, fechada à esquerda e aberta à direita, $\left[\bar{x} - \frac{h}{2}, \bar{x} + \frac{h}{2}\right[$. As classes laterais para a

direita são $\left[\bar{x}+\frac{h}{2},\bar{x}+\frac{3h}{2}\right[,\left[\bar{x}+\frac{3h}{2},\bar{x}+\frac{5h}{2}\right[,...,$ até ultrapassar $x_{(n)}$. Para a esquerda constroem-se classes $\left[\bar{x}-\frac{3h}{2},\bar{x}-\frac{h}{2}\right[,\left[\bar{x}-\frac{5h}{2},\bar{x}-\frac{3h}{2}\right[,...,$ até ultrapassar o valor $x_{(1)}$.

Qual das regras utilizar? A regra de Sturges ou a regra do desvio-padrão? Não há resposta concludente para esta pergunta, ambas as regras são empíricas. A regra de desvio-padrão é mais trabalhosa e sempre que se aumentar a amostra as classes têm de ser construídas de novo pois temos uma nova média e desvio-padrão. Por outro lado tem a vantagem de salientar a simetria da distribuição e o aspecto em forma de sino, se os dados forem "normais".