ELEMENTOS DE ESTATÍSTICA DESCRITIVA

Amílcar Oliveira
Teresa A. Oliveira

Lisboa
Janeiro de 2011

Resumo. Pretende-se com o presente texto uma abordagem aos principais tópicos desenvolvidos em Estatística Descritiva, para apoio ao módulo introdutório sobre Estatística Descritiva inserido na unidade curricular de Matemática do Curso de Qualificação para Estudos Superiores da Universidade Aberta. O principal objetivo é fornecer e/ou relembrar um conjunto de conceitos, habitualmente adquiridos pelos estudantes ao nível do ensino secundário, que permita uma melhor adaptação às unidades curriculares da área da Estatística lecionadas em licenciaturas das áreas da matemática, informática, ciências, economia e gestão.

Conteúdo

Capítulo 1 Introdução. 3
Capítulo 2 População e Amostra, obtenção de dados. 4
2.1 Amostra e população. 4
2.2 Recenseamento e sondagens. 4
2.3 Amostragem aleatória simples. 5
2.4 Estatística descritiva e Estatística indutiva. 6
Capítulo 3 Análise, representação e redução de dados. 8
3.1 Variável estatística discreta e variável estatística contínua 8
3.2 Tabelas de frequências 9
3.2.1 Frequência absoluta simples 9
3.2.2 Frequência relativa simples 9
3.2.3 Frequência absoluta acumulada 9
3.2.4 Frequência relativa acumulada 9
3.2.5 Dados agrupados em classes 10
3.3 Representações gráficas 11
3.3.1 Histograma 11
3.2.2 Diagrama circular 12
3.2.3 Diagrama de barras 13
3.2.4 Diagrama de caule e folhas 13
3.2.5 Diagrama de extremos e quartis 14
Capítulo 4 Medidas de localização 16
4.1 Média aritmética 16
4.2 Mediana 17
4.3 Moda 18
4.4 Quantis 20
Capítulo 5 Medidas de dispersão ou variabilidade 21
5.1 Amplitude total 21
5.2 Amplitude interquartil 21
5.3 Variância 22
5.4 Desvio padrão 23
5.5 Propriedades algébricas da média e do desvio padrão 25
5.5.1 Propriedade 1 26
5.5.2 Propriedade 2 26
Capítulo 6 Exercícios propostos 28

CAPITULO 1

Introdução

Iniciamos este texto de apoio com a procura de uma definição para o conceito de Estatística. E vamos rápidamente concluir que tal não é tarefa fácil, pois a designação de Estatística aparece muitas vezes na bibliografia com diferentes significados. Talvez a primeira questão a colocar relativamente a este assunto, tenha a haver com a forma como este assunto é tratado, por exemplo, pela comunicação social. Com efeito, a tendência sistemática para se referirem à Estatística como simples informação numérica, pode levar muitas vezes, o cidadão comum, a absorver uma ideia demasiado simplista daquilo que é efectivamente a Estatística.
Numa visão um pouco mais ampla, podemos pensar em Estatística como sendo um conjunto de técnicas para tratamento e análise de dados, mas ainda assim ficaremos longe do seu verdadeiro significado.
Atualmente, é de certo modo consensual falarmos em Estatística como sendo simultâneamente uma ciência e uma arte que permite obter conclusões sobre um conjunto de dados, e que na sua vertente mais abrangente realiza aquilo que se designa por Inferência Estatística.
Até final do século XIX, a Estatística era apenas o que nos dias de hoje designamos de Análise de Dados ou Estatística Descritiva, tendo-se revelado bastante importante no apoio a muitas àreas científicas.

CAPÍTULO 2

População e Amostra, obtenção de dados

2.1 Amostra e população

Os conceitos de amostra e de população aparecem em inúmeras situações onde se utilizam técnicas estatísticas envolvendo análises sobre conjuntos de indivíduos. É então conveniente que comecemos por apresentar uma definição destes conceitos.

População - é uma coleção de elementos individuais, por exemplo, animais, pessoas, objectos, ou resultados experimentais que tenham uma ou mais características em comum e que se pretendam analisar.

A dimensão duma População pode ser mais ou menos clara consoante o problema em estudo. Vejamos alguns exemplos:

- Estudantes da Universidade Aberta
- Pessoas com mais de 18 anos residentes em Portugal Continental
- Conjunto de escaravelhos na região de Trás os Montes
- Conjunto dos acidentes num determinado dia na autoestrada A1
- Conjunto de golfinhos presos diáriamente em redes de pesca em todo o mundo

Sobre as populações, poderemos analisar as mais variadas características, consoante a natureza das mesmas e o objectivo do estudo. São exemplos dessas características:

- idade(em cm) dos estudantes da Universidade Aberta
- peso (em Kg) das pessoas residentes em Portugal continental com mais de 18 anos
- número de escaravelhos por hectare nos concelhos da Região de Trás os Montes
- número de acidentes provocados por excesso de alcool na A1
- número de golfinhos presos em redes de pesca

2.2 Recenseamento e Sondagens

Em situações onde é necessário estudar todos os elementos duma população procede-se habitualmente a um recenseamento ou censo. Em Portugal, o Instituto Nacional de Estatística tem a responsabilidade dessa tarefa que é realizada periódicamente, ao longo do tempo. Num

recenseamento é possível obter informação variada, respeitante a condições económicas e sociais dos habitantes, e pode contribuir para tomada de decisões importantes por parte do poder político. Por outro lado é possível analisar a evolução das características observadas ao longo do tempo. A problemática deste tipo de processo, que envolve toda a população de um país, é em regra, envolver bastantes meios e custos elevados. Um recenseamento pode averiguar aspectos ligados à habitação, mas também pode envolver aspetos ligados ao funcionamento da agricultura e indústria. Podemos assim definir **Recenseamento** como sendo um processo em que para além da recolha de informação está também envolvida a análise de todos os elementos da população em estudo, tendo como objectivo não apenas a enumeração dos seus elementos mas também o estudo de características importantes.
Noutros estudos porém, não é viável proceder a um levantamento tão exaustivo da informação. Nesse caso recorre-se a outro tipo de abordagem, que se traduz no conceito de Sondagem, de que passaremos de seguida a descrever.

Sondagem é um estudo baseado numa parte de uma população com intuíto de inferir resultados para toda a população. Usualmente os estudos associados às sondagens têm como objetivo conhecer gostos ou preferências em relação a certos assuntos ou acontecimentos comuns a toda a população. São exemplo disso as bem conhecidas sondagens de opinião nos períodos que antecedem os atos eleitorais.
Históricamente, podemos afirmar que este tipo de estudo começou a ter maior destaque a partir da segunda metade do século XX, período a partir do qual foi possível implementar um conjunto de métodos e técnicas estatísticas que lhe deram definitivamente um caráter científico.

2.3 Amostragem aleatória simples

A problemática envolvendo a recolha de amostras é sem dúvida bastante importante e merece destaque especial. Atendendo a factores vários, torna-se muitas vezes inviável a análise de todos os elementos de uma população. Vejamos algumas situações:

Caso 1: A população é infinita ou pode ser considerada como tal. Ex: Temperaturas nos pontos da superfície da Terra, num dado instante.
Caso 2: A recolha da informação obrigaria à destruição total dos elementos em estudo. Ex: Tempo de vida de determinado tipo de lâmpada.
Caso 3: A recolha de toda a informação ser muito dispendiosa ou ser muito demorada. Ex: Consulta de opinião pública sobre um candidato

presidencial através de censo.

Nos exemplos apresentados e noutros casos onde situações semelhantes se verifiquem, é de todo aconselhável recorrer à seleção de uma amostra. Quando falamos em amostragem, referimo-nos a um processo que consiste em considerar um certo número de elementos - **amostra** - do conjunto de elementos a estudar - **população**. Esse processo deve ter em atenção critérios adequados, de tal forma que nos conduzam a conclusões válidas para toda a população. Diremos que os critérios foram seguidos quando a amostra obtida, é representativa de toda a população. Uma amostra que pelo contrário não seja representativa da população, diz-se *enviesada*. Podemos enumerar os critérios, que no essencial nos orientam para a realização de uma amostragem adequada:

- Todos os indivíduos da população devem ter igual probabilidade de ser seleccionados;
- A População deve estar bem definida logo à partida, ou seja desde o início do estudo;
- Dimensão da amostragem deve ser adequada, ou seja nela devem figurar toda a variedade de subgrupos existentes na população.

Para exemplificar a ideia de amostra enviesada, vejamos alguns exemplos de amostragens não representativas da população:

- Utilização de uma amostra, constituída por opiniões de 10 membros da Liga Protectora dos Animais, sobre a abolição das touradas;

- Utilização de uma amostra de pluviosidades diárias verificadas nos meses de Junho, Julho e Agosto para tirar conclusões sobre um ano inteiro.

A amostragem aleatória simples consiste num processo onde se extrai de uma população um número de elementos préviamente fixado. Essa extracção deve no essencial, ser feita:

- ao acaso;
- de forma a ter em conta a composição da população;
- de forma que a escolha feita por um indivíduo não influencie a escolha de outro.

2.4 Estatística descritiva e Estatística indutiva

Numa análise estatística podem estar envolvidas fases distintas: uma análise descritiva dos dados, correspondente à obtenção, organização e

tratamento dos dados de uma determinada amostra, e uma análise inferencial onde a partir do conhecimento descritivo dos dados se procura inferir conclusões para toda a população.
Podemos afirmar que a estatística descritiva se resume ao estudo de uma amostra, onde o principal objetivo é a obtenção de algumas características amostrais e construção de tabelas e gráficos onde possa constar toda a informação na forma resumida. Efetivamente, nesta fase procuram-se representações alternativas e sugestivas que substituam um conjunto de dados que se tenha.
Considere-se um exemplo, onde numa turma se questionou cada um dos alunos acerca do número de irmãos que cada um tem. As respostas foram as indicadas:

0 1 2 0 0 2 4 1 2 3 2 1 1 1 1 0 0 0 1 2 3 4 2 2 1 1 0 1 0 1 0 1 0 2 1 1 2 0 1 1

Na presença destes dados, não percebemos de imediato, por exemplo qual a situação que predomina, mas se construirmos uma simples representação gráfica, ficamos com ideia bastante mais clara.

0 | **********
1 | **********
2 | *********
3 | **
4 | **

Desta forma podemos de imediato concluir que a situação que mais ocorre é ter um irmão, e que apenas uma pequena minoria tem três ou mais irmãos.
Esta organização dos dados, permitiu que duma forma simples pudessemos tirar melhores informações sobre os dados.
Mas num estudo estatístico, os objetivos vão habitualmente mais além duma descrição dos dados, quer seja em tabelas ou gráficos. Muitas vezes queremos mesmo é estimar quantidades ou testar hipóteses utilizando técnicas estatísticas adequadas, que nos permitam tirar conclusões acerca de uma população. Aqui entramos na área da Estatística Indutiva ou Inferência Estatística, tema que não é âmbito deste texto.

CAPÍTULO 3

Análise, representação e redução de dados

Abordámos no capítulo anterior a importância de se efetuar uma representação dos dados, quer através de tabelas, quer através de gráficos, por forma a podermos obter uma realce da informação contida num conjunto de dados. Dessa forma será possível obter respostas, à partida não óbvias, de questões relevantes. Por exemplo, podemos constatar se os dados são parecidos entre si, se apresentam alguma tendência ou se existem agrupamentos.

Neste capítulo iremos estudar alguns processos utilizados na redução de dados. Começaremos por ver alguns aspetos relacionados com o tipo de dados que podem ocorrer e formas de classificação dos mesmos. Numa primeira classificação podemos ter dados de caráter qualitativo, por exemplo a cor dos olhos, ou dados de caráter quantitativo, por exemplo a temperatura em ^{o}C.

3.1 Variável estatística discreta e variável estatística contínua

Uma variável estatística que apenas assume valores numéricos isolados denomina-se de variável discreta.

Exemplos:

- Número de irmãos de um conjunto de alunos duma turma do 12^{o} ano;
- Número mensal de viagens de avião efetuadas por um executivo de uma empresa;
- Número de golos marcados por uma equipa de futebol em cada jogo das 30 jornadas de um campeonato.

Quando uma variável toma qualquer valor num certo intervalo, designa-se de variável contínua.

Exemplos:

- Peso dos alunos de uma escola do ensino básico;
- Temperatura diária registada numa estação metereológica;
- Pressão arterial medida num grupo de 100 hipertensos.

3.2 Tabelas de frequências

Uma das formas comuns de organizar os dados que estejamos a tratar, é através das tabelas de frequências, onde habitualmente figuram entre outros valores, aqueles que dizem respeito às frequências absolutas simples, frequências relativas simples, frequências absolutas acumuladas e frequências relativas acumuladas.

3.2.1 Frequência absoluta simples

Corresponde ao número de vezes que um determinado valor x_i é observado na população ou amostra estudada, e designa-se por n_i.

3.2.2 Frequência relativa simples

Sendo n_i a frequência absoluta do valor da variável x_i e N a dimensão da população ou amostra estudada, designa-se por frequência relativa simples de x_i o quociente:

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

É comum a utilização das frequências relativas sob a forma de percentagem, que se obtêm calculando $f_i \times 100$.

3.2.3 Frequência absoluta acumulada

Designa-se por frequência absoluta acumulada de índice i a soma das frequências absolutas correspondentes aos valores da variável desde o primeiro valor até ao de ordem i, e representa-se por N_i. Para k observações tem-se

$$\begin{aligned}
&N_1 = n_1 \\
&N_2 = n_1 + n_2 = N_1 + n_2 \\
&N_3 = n_1 + n_2 + n_3 = N_2 + n_3 \\
&\vdots \\
&N_k = n_1 + n_2 + \ldots + n_k = N_{k-1} + n_k
\end{aligned}$$

$$N = \sum_{i=1}^{k} n_i$$

3.2.4 Frequência relativa acumulada

De modo idêntico designa-se de frequência relativa acumulada de índice i à soma das frequências relativas correspondentes aos valores da variável

desde o primeiro valor até ao de ordem i, e representa-se por F_i.

$F_1 = f_1$
$F_2 = f_1 + f_2 = F_1 + f_2$
$F_3 = f_1 + f_2 + f_3 = F_2 + f_3$
$\vdots$
$F_k = f_1 + f_2 + f_3 + \ldots + f_k = F_{k-1} + f_k$

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

Em termos genéricos, uma tabela de frequências será constituída da seguinte forma:

x_i	n_i	$f_i = \frac{n_i}{n}$	N_i	F_i
x_1	n_1	$f_1 = \frac{n_1}{n}$	n_1	f_1
x_2	n_2	$f_2 = \frac{n_2}{n}$	$n_1 + n_2$	$f_1 + f_2$
x_3	n_3	$f_3 = \frac{n_3}{n}$	$n_1 + n_2 + n_3$	$f_1 + f_2 + f_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = \frac{n_k}{n}$	$n_1 + n_2 + n_3 + \ldots + n_k$	$f_1 + f_2 + f_3 + \ldots + f_k$
		Total	$N = \sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} f_i = 1$

3.2.5 Dados agrupados em classes

Frequência absoluta e frequências relativa

Numa distribuição com dados agrupados em classes, designa-se de frequência absoluta da classe (n_i) ao número de casos que pertencem a essa classe e frequência relativa da classe (f_i), ao quociente entre a frequência absoluta da classe e a dimensão da população (N).
As **frequências absolutas acumuladas (N_i)** e as **frequências relativas acumuladas (F_i)** obtêm-se fazendo a acumulação dos valores em cada classe de acordo com o exemplo seguinte:

Exemplo 3.1 Foram agrupados em classes e registados na tabela seguinte os dados relativos aos tempos (em minutos) da duração de uma viagem de automóvel, entre duas localidades A e B, durante 30 dias consecutivos.

i	Classes	n_i	$f_i(\%)$
1	$[15;20[$	8	26.67
2	$[20;25[$	6	20
3	$[25;30[$	7	23.33
4	$[30;35[$	2	6.67
5	$[35;40[$	7	23.33

a) Qual a percentagem de dias para os quais a duração da viagem é inferior a 30 minutos?
R: $F(x < 30)$=f_1+f_2+f_3=26.67+20+23.33=70, logo conclui-se que em 70% dos dias a viagem demora menos de 30 minutos.
b) Qual a percentagem de dias para os quais a duração da viagem é superior a 35 minutos?
R: $F(x > 35)$=f_5=23.33, pelo que em 23.33% dos dias a viagem demora mais do que 35 minutos.
c) Complete a tabela anterior obtendo as colunas com as frequências absolutas acumuladas (N_i) e as frequências relativas acumuladas (F_i).

i	Classes	n_i	$f_i(\%)$	N_i	$F_i(\%)$
1	$[15;20[$	8	26.67	8	26.67
2	$[20;25[$	6	20	14	46.67
3	$[25;30[$	7	23.33	21	70
4	$[30;35[$	2	6.67	23	76.67
5	$[35;40[$	7	23.33	30	100

Centro da Classe

Em certos casos, torna-se necessário efetuar cálculos onde o valor central de cada classe deve ser conhecido. Esse valor central ou centro duma classe $[l_i; l_{i+1}[$ é calculado através de

$$x_i = \frac{l_i + l_{i+1}}{2}$$

sendo l_i o extremo inferior da classe e l_{i+1} o extremo superior da classe.

3.3 Representações gráficas

3.3.1 Histogramas

Os dados agrupados em classes podem ser representados graficamente por histogramas. Há dois tipos de histogramas que podem ser desenhados para as frequências absolutas ou ordinárias: o histograma de frequências simples e o histograma de frequências acumuladas. Os histogramas como impacto visual transmitem uma importante informação relativamente à forma, à tendência central e à dispersão dos dados.

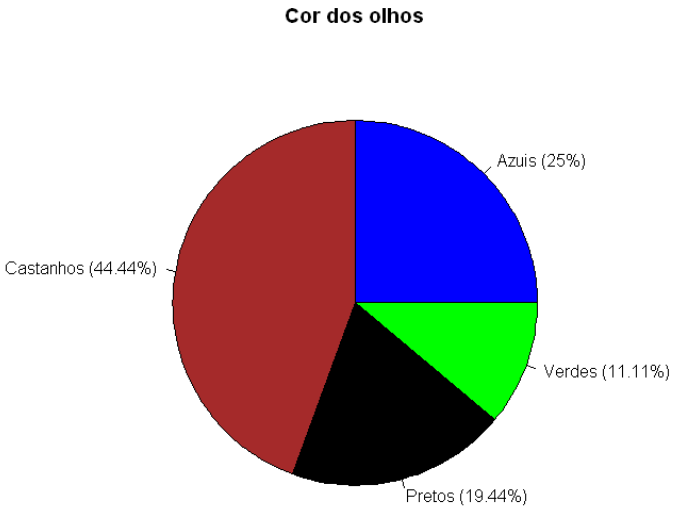
Para elaboração dos histogramas é conveniente a construção prévia de tabelas de frequências.

3.3.2 Diagrama circular

Para amostras de dados qualitativos, este é um tipo de representação bastante utilizado. Após a construção da tabela de frequências é bastante simples a obtenção deste diagrama, bastando associar a cada setor circular do diagrama, o valor da frequência de cada classe.
Podemos assim definir **diagrama circular** como sendo uma representação gráfica que tem base um circulo, que é dividido em tantos setores circulares, quantas as classes existentes para a variável em estudo e onde os ângulos dos setores são proporcionais às frequências das classes.

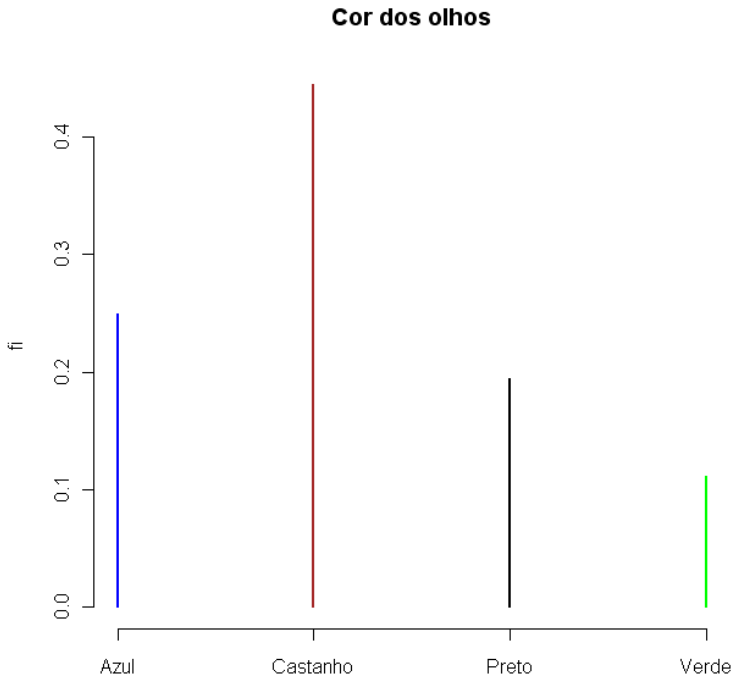
Vejamos o seguinte exemplo onde estão representados no diagrama circular, os resultados relativos às cores dos olhos dos alunos de uma turma.

Cor dos olhos	n_i	Graus
Castanhos	16	160º
Pretos	7	70º
Verdes	4	40º
Azuis	9	90º



3.3.3 Diagrama de barras

O diagrama de barras é uma representação gráfica, onde num sistema de eixos coordenados, associamos ao eixo das abcissas, marcas que representam as classes, e alinhadas com essas marcas colocamos barras verticais de altura igual ou proporcional à frequência da classe. Preferencialmente, deveremos usar as frequências relativas para construção do diagrama, pois nesse caso a soma das alturas das barras é unitária e torna-se mais simples a comparação de gráficos entre amostras diferentes.
Em regra estes diagramas devem ter as barras com a mesma largura, a não ser que se pretenda dar outra informação para essa característica. De seguida apresentamos um exemplo referente aos dados usados anteriormente para o diagrama circular.



3.3.4 Diagrama de caule e folhas

Este tipo de representação, pelo tipo de informação que contém, pode considerar-se como estando entre uma representação em tabela e um gráfico. Após um processo de contagem, organização e representação

dos dados, obtém-se um diagrama com barras na horizontal. Em primeiro lugar deve-se traçar uma linha vertical, colocando à esquerda dessa linha o dígito (ou dígitos) da classe com maior grandeza, seguido dos restantes. Este tipo de representação é de fácil construção, sendo possível fazer a ordenação dos dados e muitas vezes também fazer a reconstituição da amostra.

Exemplo 3.2 Num inquérito realizado a 50 estudantes universitários obteve-se a seguinte informação acerca do peso de cada um.

49 55 62 63 75 62 93 72 83 59 64 53 60 90 75 89 46 51 55 62 67 69 62 71 75 82 85 87 92 77 67 58 69 75 77 77 82 89 86 71 68 67 65 59 50 65 62 52 72 75

Para fazermos a representação de caule-e-folhas, começamos por traçar uma linha vertical, registar à esquerda os dígitos das dezenas e à direita os sucessivos digitos das unidades.

4 | 96
5 | 593158902
6 | 2324027927987552
7 | 525157577125
8 | 39257296
9 | 302

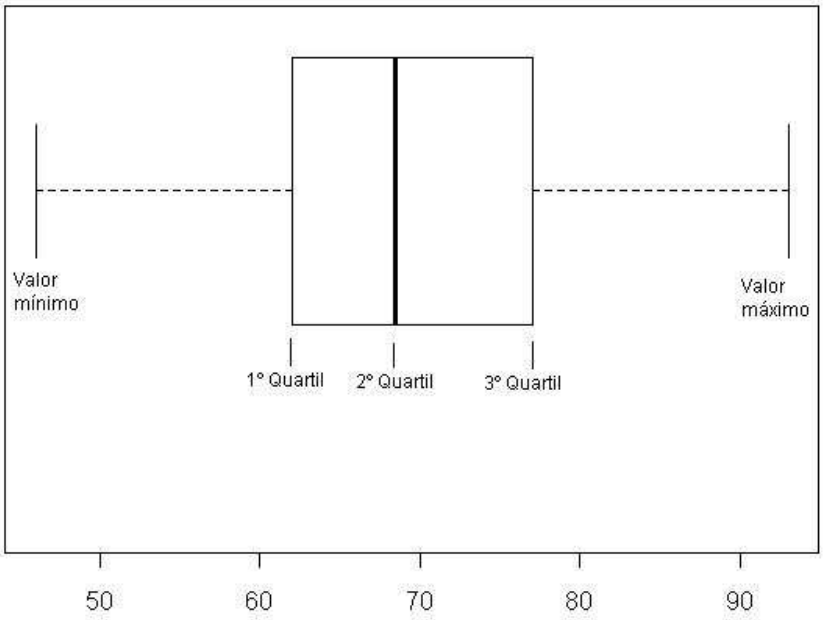
e ordenando, obtem-se
4 | 69
5 | 012355899
6 | 0222223455777899
7 | 112255555777
8 | 22356799
9 | 023

Este tipo de representação de dados permite duma forma simples, destacar alguns aspectos particulares dos mesmos, como por exemplo, existência ou não de simetria, maior ou menor dispersão.

3.3.5 Diagrama de extremos e quartis

O diagrama de extremos e quartis ou caixa de bigodes (box-plot) é um tipo de representação gráfica também de fácil construção e que realça muito bem a informação sobre os dados.
Para sua construção é necessária a obtenção de cinco medidas, valor mínimo, 1º Quartil(Q_1), 2º Quartil(Q_2), 3º Quartil(Q_3) e valor máximo.

Naturalmente verifica-se que $Q_1 < Q_2 < Q_3$, onde Q_2 representa a mediana. Estas medidas são apresentadas no capítulo 4.
Vejamos agora um exemplo do caso da representação dos dados do exemplo anterior em diagrama de extremos e quartis:



CAPÍTULO 4

Medidas de localização

No âmbito da estatística descritiva são apresentadas de seguida, algumas das medidas de localização mais usadas na análise de dados, nomeadamente a média aritmética, a mediana e a moda. São também apresentados os quantis ou amplitudes modificadas. As medidas de localização ou tendência central são muito importantes pois permitem caracterizar uma distribuição identificando um valor que a tipifíca.

4.1 Média aritmética

A média aritmética de uma amostra $\bar{x}$ é a medida de localização obtida através do quociente entre a soma de todos os valores observados nessa amostra e o número total de observações.
Se representarmos as n observações por $x_1, x_2, \ldots, x_n$ virá

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

Caso os dados estejam agrupados podem ocorrer duas situações distintas:

(1) Os dados são do tipo discreto e existem valores que se repetem na amostra ou população. Neste caso podemos obter a média com base na seguinte expressão:

$$\bar{x} = \frac{n_1x_1+n_2x_2+\ldots+n_kx_k}{n} \text{ ou abreviadamente } \bar{x} = \frac{\sum_{i=1}^{k} x_in_i}{n}$$

onde,
n_i é a frequência absoluta da classe i
k é o número de classes do agrupamento
x_i é o valor correspondente à classe i
$n = n_1 + n_2 + \ldots + n_k$

(2) Os dados são do tipo discreto ou contínuo e as classes são intervalos. Neste caso teremos um valor aproximado para a média dado pela expressão:

$$\bar{x} \approx \frac{n_1x_1+n_2x_2+\ldots+n_kx_k}{n}$$

onde,
n_i é a frequência absoluta da classe i
k é o número de classes do agrupamento
x_i é o ponto médio da classe i, valor considerado como representativo da classe i
$n = n_1 + n_2 + \ldots + n_k$

Para dados de carácter quantitativo, a média aritmética, é uma das medidas de localização ou tendência central mais importantes. A média aritmética tem no entanto a desvantagem de não poder ser usada para dados de carácter qualitativo, que traduzem qualidades, como por exemplo a cor dos olhos, e pelo facto de ser permeável a valores errados, uma vez que depende de todos os valores da série estatística, bastando um valor errado muito diferente de todos os outros para a afectar.

4.2 Mediana

A mediana $\tilde{x}$, é outra medida de localização cuja característica se traduz pela posição do meio que ocupa quando todos os dados estão ordenados por ordem crescente

$$x_1 < x_2 < \ldots < x_c < \ldots < x_{n-1} < x_n$$

Caso 1:

Se o número de dados é ímpar, existe um valor central que corresponde à mediana $\tilde{x}$

$$\underbrace{x_1 x_2 \ldots x_{c-1}}_{\frac{n-1}{2} valores} \quad \underbrace{x_c}_{mediana(\tilde{x})} \quad \underbrace{x_{c+1} \ldots x_{n-1} x_n}_{\frac{n-1}{2} valores}$$

Caso 2:

Se o número de dados é par, existem dois valores centrais cuja média aritmética é o valor da mediana ($\tilde{x}$):

$$x_1 x_2 \ldots \underbrace{x_c x_{c+1}}_{\tilde{x}=\frac{x_c+x_{c+1}}{2}} \ldots x_{n-1} x_n$$

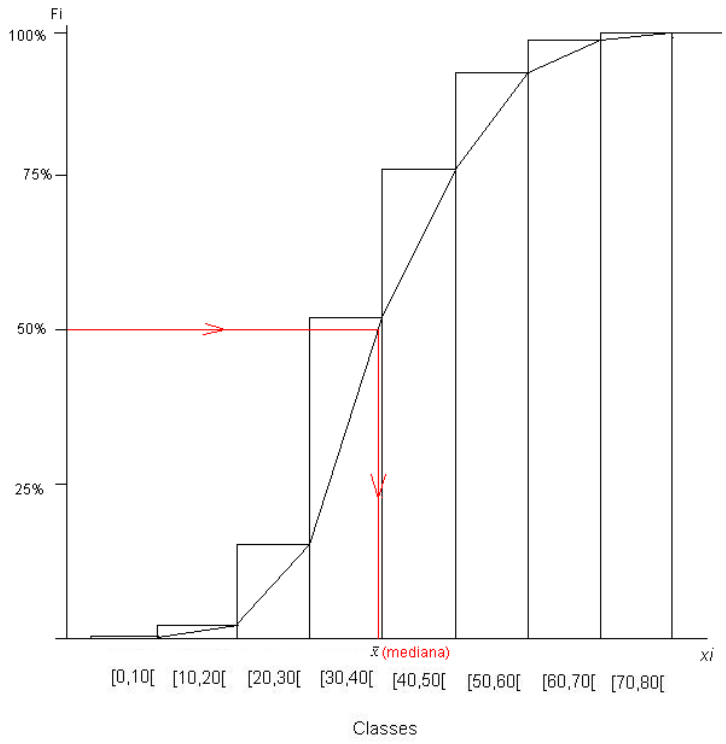
Caso 3:

Quando os dados se apresentam agrupados em classes, deve em primeiro lugar localizar-se a classe mediana, sendo posteriormente determinada a posição da mediana dentro dessa classe, por interpolação. A classe mediana é a primeira classe cuja frequência acumulada iguala ou excede metade do número total de observações. Uma vez identificada essa classe, determina-se o valor da mediana com base na seguinte expressão

$$\tilde{x} = l_i + \left(\frac{\frac{n}{2} - F'_{aci}}{F_{aci} - F'_{aci}}\right) \times a_i$$

onde l_i é o limite inferior da classe mediana; F'_{aci} frequência absoluta acumulada no limite inferior da classe mediana; F_{aci} frequência absoluta

acumulada no limite superior da classe mediana; a_i amplitude da classe mediana.
Graficamente, dado o histograma de frequências acumuladas e o respetivo poligono, tem-se:



Neste caso a classe mediana é a classe $[30, 40[$.

4.3 Moda

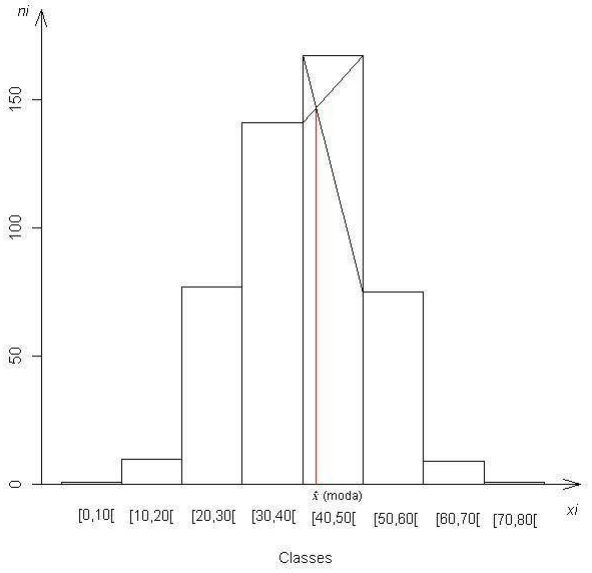
A moda $\hat{x}$, define-se como sendo o valor que ocorre com maior frequência num conjunto de dados, ou seja, uma vez conhecidas as frequências absolutas, a moda será o valor x_i ao qual corresponde o maior valor de n_i. Nas situações onde numa distribuição ocorrem dois valores apresentando a mesma e mais elevada frequência absoluta, diz-se que a distribuição é bimodal. Se uma distribuição estatística possui três ou mais modas, diz-se que a distribuição é plurimodal ou multimodal.
No caso de se dispor apenas de dados agrupados, deve em primeiro lugar localizar-se a classe modal, sendo posteriormente determinada a posição da moda dentro dessa classe. A classe modal é a classe com

maior frequência ordinária, quer absoluta quer relativa. O valor da moda será determinado por interpolação, usando a seguinte expressão

$$\hat{x} = l_i + \left[\frac{d_1}{d_1+d_2}\right] \times a_i$$

onde l_i é o limite inferior da classe modal; d_1 diferença entre a frequência da classe modal e a frequência da classe que lhe precede; d_2 diferença entre a frequência da classe modal e a frequência da classe que lhe sucede; a_i amplitude da classe modal.

Dado o histograma de frequências absolutas simples tem-se:



A classe modal é a classe $[40, 50[$.

4.4 Quantis

Os quantis são medidas de tendência central que permitem dividir uma amostra ou distribuição em partes iguais. Um caso particular dos quantis é a mediana, que permite dividir a distribuição em duas partes iguais.
Os quantis mais conhecidos são: os quartis, que dividem a distribuição em quatro partes iguais; os decis, que dividem a distribuição em dez partes iguais; os percentis, que dividem a distribuição em cem partes iguais.
O cálculo destes quantis é similar ao cálculo da mediana, uma vez que também subdividem a distribuição de medidas de acordo com a proporção das frequências observadas.

Para dados agrupados a expressão usada para a determinacão da mediana, mantem-se adaptando o quantil apropriado. Recordemos que

$$\tilde{x} = l_i + (\frac{\frac{n}{2} - F'_{aci}}{F_{aci} - F'_{aci}}) \times a_i$$

De um modo geral, para os restantes quantis, na expressão anterior em vez de $\frac{n}{2}$ aparece $\frac{n}{4}$ e $\frac{3n}{4}$ para o 1º e 3º quartis, $\frac{n}{10}$ e $\frac{2n}{4}$ para o 1º decil e 2º decil e assim sucessivamente, sendo os F_{aci} e F'_{aci} as repetivas frequências acumuladas no limite inferior da respetiva classe e no limite superior da respetiva classe.

CAPÍTULO 5

Medidas de dispersão ou variabilidade

As medidas de dispersão permitem avaliar o grau de variabilidade dos valores de uma distribuição. Estas medidas permitem complementar a informação obtida pelas medidas de localização abordadas anteriormente, dando um conhecimento mais efetivo acerca da variabilidade dos dados. Apresentamos de seguida algumas dessas medidas.

5.1 Amplitude total

A amplitude ou amplitude total, usualmente denotada por R (range) é o valor da diferença entre o maior e o menor valor observados. No caso de dados agrupados a amplitude total será dada pela diferença entre o limite superior da classe mais alta e o limite inferior da classe mais baixa. Obviamente quanto maior for a amplitude maior será a dispersão dos dados.

$$\text{R} = \text{valor máximo - valor mínimo}$$

Esta é a medida de dispersão mais simples em termos de cálculo, mas dá uma informação restrita pois considera apenas os valores extremos da distribuição. No que respeita aos valores intermédios não dá qualquer informação.

Exemplo 5.1 Num teste de Matemática realizado numa turma do 12º ano foram obtidas as seguintes classificações, numa escala de 0-20:

12.1 13.2 8.9 7.5 14.5 15.8 10.5 14.6 13.3 9.2 6.3 12.2 17.3 14.2 14.8 15.7 13.3 14.2 10.0 8.4 6.5 7.3 12.5 16.4 12.2

Neste caso, a amplitude total das classificações obtidas pelos alunos é 11, pois sendo a classificação máxima 17.3 e a classificação mínima 6.3, tem-se: R=17.3-6.3=11 .

Seguem-se algumas das medidas de dispersão mais utilizadas:

5.2 Amplitude inter-quartil

A amplitude inter-quartil é considerada uma medida resistente uma vez que é definida a partir de medidas resistentes, ou seja os quartis. Esta medida que é utilizada na construção do diagrama de extremos e

quartis, fornece-nos informação acerca da amplitude do intervalo onde se encontram 50% das observações centrais.

$$\text{Amplitude inter-quartil} = 3^{\text{o}} \text{ Quartil - } 1^{\text{o}} \text{ Quartil}$$

Relativamente a esta medida podemos fazer as seguintes observações:

- Quanto mais variabilidade houver entre os dados, maior será a amplitude inter-quartil.
- No caso de não haver variabilidade, ou seja todas as observações serem iguais, a amplitude inter-quartil terá valor nulo.

Exemplo 5.2 Considere-se a seguinte amostra obtida num inquérito estudantes do 12º ano acerca do número de horas de estudo semanais na disciplina de Matemática. Obtenha a mediana e a amplitude interquartil.

R:

$$2\ 2\ \underbrace{2}_{1^{\text{o}}quartil}\ 3\ \underbrace{45}_{mediana}\ 6\ \underbrace{6}_{3^{\text{o}}quartil}\ 6\ 6$$

Neste caso a mediana é dada pela média dos dois valores centrais, ou seja $\tilde{x} = \frac{4+5}{2} = 4.5$

A amplitude interquartil terá o valor

$$\text{Amplitude inter-quartil} = 3^{\text{o}} \text{ Quartil - } 1^{\text{o}} \text{ Quartil} = 6 - 2 = 4$$

5.3 Variância

Esta é uma importante medida de variabilidade que permite medir o grau de dispersão dos dados em relação à média.
Considere-se $x_1, x_2, \ldots, x_n$ as n observações de uma distribuição estatística, de média $\bar{x}$ e seja d_i o desvio entre o valor x_i e $\bar{x}$, ou seja $d_i = x_i - \bar{x}$. Então, dado que a soma dos desvios é igual a zero, considera-se os quadrados dos desvios.

$$d_i^2 = (x_i - \bar{x})^2$$

A variância s^2 é a média dos quadrados dos desvios e fornece-nos a informação de quão distantes se encontram os dados relativamente à média.

$$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

Em algumas situações nomeadamente quando se faz inferência da amostra para a população utiliza-se a fórmula

$$\hat{s}^{*2} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

que corresponde à variância corrigida, ou estimador não enviesado da variância.
Nesta fórmula o denominador n é substituído por $n-1$ sendo que para um número de dados suficientemente grande os resultados são muito próximos.

5.4 Desvio padrão

À semelhança do que acontece com a variância, o desvio padrão também é uma medida que permite avaliar o grau de dispersão dos valores da variável em relação à média, com a vantagem de que, contráriamente à variância, o desvio padrão aparece sempre nas mesmas unidades dos valores da variável. O desvio padrão s é simplesmente a raíz quadrada da variância s^2. Assim tem-se

$$s = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

Em muitos cálculos, nomeadamente envolvendo análises descritivas de dados, esta é a medida de dispersão mais utilizada, uma vez que oferece simultâneamente o uso de todos os valores da variável envolvida e ao mesmo tempo é expressa nas mesmas unidades.
A expressão anterior pode aparecer sob outras formas, por exemplo para dados agrupados

$$s = \sqrt{\frac{n_1(x_1-\bar{x})^2+n_2(x_2-\bar{x})^2+...+n_k(x_n-\bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{k} n_i(x_i-\bar{x})^2}{n}}$$

se recorremos às frequências absolutas dos k valores diferentes da variável.

Tal como vimos para o caso da variância, também poderemos ter necessidade de recorrer ao desvio padrão corrigido, nomeadamente no âmbito do Cálculo de Probabilidades ou em estudos de Inferência Estatística. Nessa situação recorremos a

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

Vejamos agora algumas propriedades do desvio padrão:

- O desvio padrão é sempre não negativo e será tanto maior quanto mais variabilidade houver entre os dados.
- Caso não haja variabilidade, o valor do desvio padrão é nulo.

- Se o desvio padrão for nulo, então não existe variabilidade, e nesse caso os dados são todos iguais.
- À semelhança da média, também o desvio padrão é uma medida *p*ouco resistente, pois é muito influenciável por valores que sejam muito diferentes dos restantes, quer seja por excesso quer por defeito.

Exemplo 5.3 Considere-se a seguinte tabela com indicação das frequências absolutas, onde se encontram registadas as classificações num teste de Matemática, obtidas por numa turma de 20 alunos. Pretende-se o cálculo do desvio padrão dessa distribuição.

x_i	n_i
8	3
10	4
12	5
14	2
16	4
18	2

Para tal, devemos em primeiro lugar proceder aos cálculos necessários, completando a tabela:

x_i	n_i	$d_i = x_i - \bar{x}$	$d_i^2 = (x_i - \bar{x})^2$	$n_i d_i^2 = n_i(x_i - \bar{x})^2$
8	3	$8 - 12.6 = -4.6$	21.16	63.48
10	4	$10 - 12.6 = -2.6$	6.76	27.04
12	5	$12 - 12.6 = -0.6$	0.36	1.8
14	2	$14 - 12.6 = 1.4$	1.96	3.92
16	4	$16 - 12.6 = 3.4$	11.56	46.24
18	2	$18 - 12.6 = 5.4$	29.16	58.32

A soma dos quadrados dos desvios e o valor $n = 20$ permite-nos calcular o valor do desvio padrão

$$s = \sqrt{\frac{200.8}{20}} = \sqrt{10.04} = 3.1686$$

Exemplo 5.4 A tabela seguinte foi obtida a partir do conhecimento das alturas (em metros) dos alunos de uma turma do 12º ano. Pretende-se o cálculo do desvio padrão desta distribuição, na qual os dados estão agrupados em classes.

Classes	n_i
$[1.50; 1.55[$	5
$[1.55; 1.60[$	4
$[1.60; 1.65[$	6
$[1.65; 1.70[$	4
$[1.70; 1.75[$	5
$[1.75; 1.80[$	6

Para responder à questão colocada, devemos começar por obter as médias de classe e a média da distribuição:

Classes	n_i	x_i	$x_i n_i$	$d_i = x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i d_i^2 = n_i(x_i - \bar{x})^2$
$[1.50; 1.55[$	5	1.525	7.625	-0.13	0.0169	0.0845
$[1.55; 1.60[$	4	1.575	6.3	-0.08	0.0064	0.0256
$[1.60; 1.65[$	6	1.625	9.75	-0.03	0.0009	0.0054
$[1.65; 1.70[$	4	1.675	6.7	0.02	0.0004	0.0016
$[1.70; 1.75[$	5	1.725	8.625	0.07	0.0049	0.0245
$[1.75; 1.80[$	6	1.775	10.65	0.12	0.0144	0.0864

$n = \sum_{i=1}^{6} n_i = 30 \qquad \sum_{i=1}^{6} x_i n_i = 49.65$

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i n_i}{n} = \frac{49.65}{30} = 1.655$$

Temos também $\sum_{i=1}^{6} n_i(x_i - \bar{x})^2 = 0.228$

No caso de distribuições com dados agrupados devemos utilizar a fórmula, onde x_i representa o ponto médio das classes e k o número de classes.
$s = \sqrt{\frac{n_1(x_1-\bar{x})^2+n_2(x_2-\bar{x})^2+...+n_k(x_n-\bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{k} n_i(x_i-\bar{x})^2}{n}}$

O desvio padrão da distribuição será então

$$s = \sqrt{\frac{0.228}{30}} = 0.2757$$

5.5 Propriedades algébricas da média e do desvio padrão

As propriedades que seguem pretendem ilustrar o que acontece se todas as observações de uma série estatística forem acrescidas na mesma quantidade. Ou seja, qual será a nova média? E qual o novo desvio padrão?
Por outro lado, e se todas as observações duplicarem ou triplicarem o seu valor, qual será a nova média? E o desvio padrão?

5.5.1 Propriedade 1

Dada uma distribuição da variável X de média $\bar{x}$ e desvio padrão s, se adicionarmos a cada valor da variável o mesmo valor c, obtemos uma distribuição X^* de média $\bar{x^*} = \bar{x}+c$ e com desvio padrão igual a $s^* = s$.

Dem.: Considerando a distribuição $X = \{x_1, x_2, \ldots, x_n\}$, temos o valor médio $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n}$. Se adicionarmos a cada valor da variável X a constante c, obtemos a variável X^* tal que

$$X^* = \{x_1 + c, x_2 + c, \ldots, x_n + c\}$$

cujo valor médio observado pode ser representado como se segue. Tem-se:

$\bar{x}^* = \frac{(x_1+c)+(x_2+c)+\ldots+(x_n+c)}{n} = \frac{x_1+x_2+\ldots+x_n}{n} + \frac{c+c+\ldots+c}{n} = \bar{x} + \frac{nc}{n} = \bar{x} + c$
como pretendiamos demonstrar.

Vejamos agora o que se passa com os desvios padrão s e s^*.
Sejam respetivamente d_i os desvios dos valores x_i em relação à média $\bar{x}$ e d_i^* os desvios dos valores x_i^* em relação à média $\bar{x}^*$.

$$d_1^* = (x_1 + c) - (\bar{x} + c) = x_1 - \bar{x} = d_1$$

$$d_2^* = (x_2 + c) - (\bar{x} + c) = x_2 - \bar{x} = d_2$$

$$\vdots$$

$$d_n^* = (x_n + c) - (\bar{x} + c) = x_n - \bar{x} = d_n$$

Tem-se $s^* = \sqrt{\frac{d_1^2+d_2^2+\ldots+d_n^2}{n}} = s$ como queriamos demonstrar.

5.5.2 Propriedade 2

Dada uma distribuição X de média $\bar{x}$ e desvio padrão s, multiplicando cada valor da variável pelo mesmo valor c, obtemos uma distribuição X^*, cuja média observada pode ser representada por $\bar{x^*} = c\bar{x}$ e o desvio padrão será $s^* = cs$.

Dem.: Consideramos a mesma variável $X = \{x_1, x_2, \ldots, x_n\}$. Se multiplicarmos cada valor da variável X pela constante c obteremos,

$$X^* = \{cx_1, cx_2, \ldots, cx_n\}$$

A média da distribuição X^* é agora

$\bar{x}^* = \frac{(cx_1)+(cx_2)+...+(cx_n)}{n} = \frac{c(x_1+x_2+...+x_n)}{n} = c\bar{x}$ como queriamos demonstrar

Por outro lado temos

$$d_1^* = cx_1 - c\bar{x} = c(x_1 - \bar{x}) = cd_1$$

$$d_2^* = cx_2 - c\bar{x} = c(x_2 - \bar{x}) = cd_2$$

$$\vdots$$
$$d_n^* = cx_n - c\bar{x} = c(x_n - \bar{x}) = cd_n$$

e neste caso a variância s^{2*} de X^* virá $s^{2*} = \frac{c^2d_1^2+c^2d_2^2+...+c^2d_n^2}{n} = \frac{c^2(d_1^2+d_2^2+...+d_n^2)}{n} = c^2s^2$, pelo que $s^* = \sqrt{c^2s^2} = cs$ como queriamos demonstrar.

CAPÍTULO 6

Exercícios Propostos

1. Calcule a média, a mediana e a moda para as séries A e B, seguintes:

a) Série A: Número de DVD's usados por mês, por um determinado aluno de um Curso de Estatística Computacional, ao longo de um ano lectivo:

7, 8, 12, 5, 8, 10, 3, 5, 8, 8, 8, 10

b) Série B: Classificações obtidas por 230 candidatos numa prova de selecção de um jogo de cálculo mental:

Classificação	nº de candidatos
$[0, 10[$	7
$[10, 20[$	11
$[20, 30[$	18
$[30, 40[$	9
$[40, 50[$	41
$[50, 60[$	37
$[60, 70[$	52
$[70, 80[$	25
$[80, 90[$	22
$[90, 100]$	8

2. Relativamente à série B anterior, determine:

a) O 3º quartil.
b) O 90º percentil.

3. Para os dados das séries A e B anteriores, calcule a amplitude total, amplitude inter-quartil, a variância e o desvio padrão.

4. Construa o histograma de frequências absolutas simples e o histograma de frequências absolutas acumuladas para os dados da série B, do exercício 1.

5. De acordo com o indicado na tabela que se segue, a distribuição dos salários (em Euros) dos trabalhadores de uma empresa é dada por:

x_i	n_i
1000	10
1100	8
1200	12
1300	8
1400	10
1500	12

a) Indique o número de trabalhadores da empresa.
b) Determine a média e o desvio padrão desta distribuição.
c) Face aos prejuízos da empresa, o dono decidiu reduzir todos os salários para metade. Nestas novas condições determine a média e o desvio padrão e compare-os com os obtidos na alínea anterior.

6. Num processo de produção de certos componentes para lâmpadas, foram registados numa hora os seguintes valores para a resistência dos componentes, de uma amostra de dimensão 20 (em unidades codificadas). Na tabela encontram-se os valores obtidos para os componentes retirados sequencialmente:

1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º	16º	17º	18º	19º	20º
23	45	24	23	54	43	46	57	42	21	31	65	32	34	35	32	33	37	41	30

Construa um diagrama de caule-e-folhas representativo destes dados.

7. Numa amostra de 12 telemóveis de uma determinada marca e modelo, verificaram-se os seguintes tempos operacionais, em horas, até descarregamento total da bateria:

40, 38, 41, 37, 39, 40, 35, 40, 42, 40, 41, 38

Com base nesta amostra:
a) Calcule a média, a moda e a mediana dos referidos tempos.
b) Calcule a variância , o desvio padrão e a amplitude total dos referidos tempos.
c) Represente os dados num diagrama de caule-e-folhas.

8. Foi realizado um estudo para análise da concentração de calcário na água da região do Algarve. Para tal, foram recolhidas 20 amostras de 1 litro de água em nascentes aleatoriamente seleccionadas nessa região, tendo-se verificado os resultados seguintes:

Concentração(mg/l)	Nº de amostras
$[1.0, 1.2[$	5
$[1.2, 1.4[$	8
$[1.4, 1.6[$	4
$[1.6, 1.8[$	2
$[1.8, 2.0[$	1

a) Elabore uma tabela de frequências para os dados.
b) Elabore o histograma das frequências absolutas simples.
c) Determine a média e o desvio padrão.