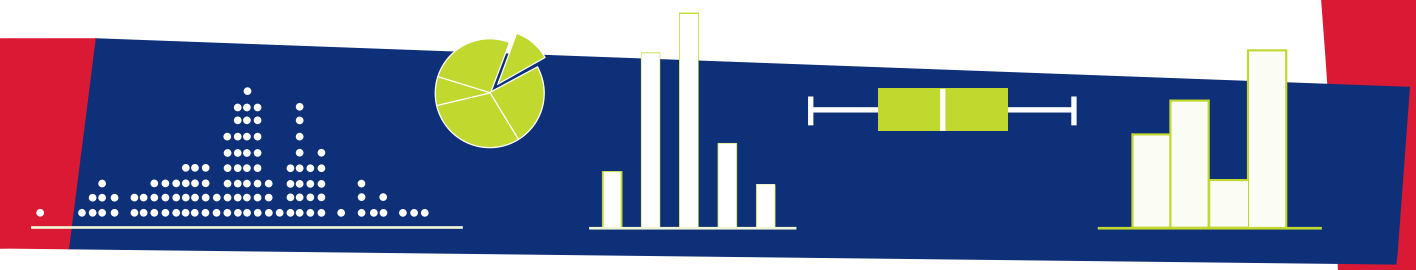


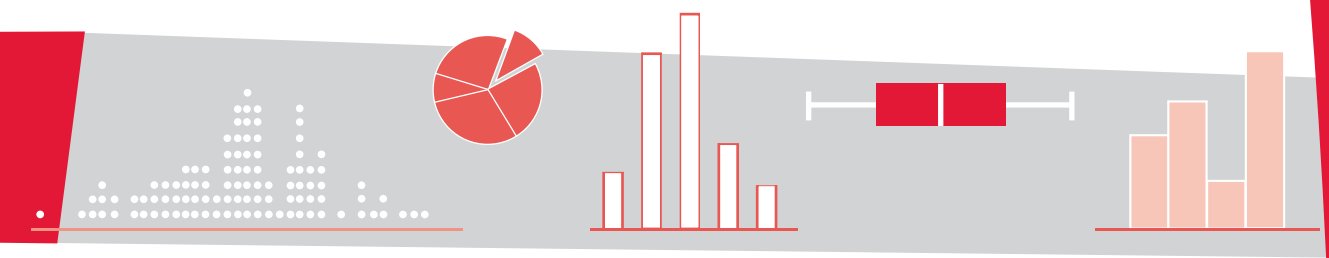
ANÁLISE de DADOS

Texto de Apoio para os Professores do 1.º ciclo

Maria Eugénia Graça Martins

Luísa Canto e Castro Loura

Maria de Fátima Mendes



ANÁLISE de DADOS

Texto de Apoio para os Professores do 1.º ciclo

Ministério da Educação

dgidc
Direcção-Geral de Inovação e de Desenvolvimento Curricular

Maria Eugénia Graça Martins

Luísa Canto e Castro Loura

Maria de Fátima Mendes

Biblioteca Nacional – Catalogação Nacional

MARTINS, Maria Eugénia Graça, 1947- , e outros

Análise de Dados: texto de apoio para os professores do 1.º ciclo/Maria Eugénia Graça Martins, Luísa Canto e Castro Loura, Maria de Fátima Mendes
ISBN 978-972-742-261-6

I – LOURA, Luísa Canto e Castro, 1954-
II – MENDES, Maria de Fátima, 1963-

CDU 371
51

Ficha Técnica

Análise de Dados
Texto de Apoio para os Professores do 1.º ciclo

Editor
Ministério da Educação
Direcção-Geral de Inovação e de Desenvolvimento Curricular

Autores
Maria Eugénia Graça Martins, Luísa Canto e Castro Loura,
Maria de Fátima Mendes

Design
Manuela Lourenço

Execução Gráfica
Editorial do Ministério da Educação

Tiragem
7500 Exemplares

Depósito Legal
262 674/07

ISBN
978-972-742-261-6

Nota de Apresentação

No âmbito do Programa de Formação Contínua em Matemática iniciado em 2005 para os professores do 1.º ciclo e que se alargou no ano seguinte aos professores do 2.º ciclo, foram identificados aspectos e temas relevantes para a formação em Matemática dos professores do Ensino Básico. Uma das vertentes que se destacou foi a importância de ter disponíveis documentos científicos que incidam nas temáticas abordadas nos primeiros anos de escolaridade.

A publicação desta brochura sobre *Análise de Dados* concretiza a iniciativa de organizar publicações de matemática focadas nas temáticas centrais do currículo do Ensino Básico.

A partir de uma proposta da Comissão de Acompanhamento do Programa de Formação Contínua em Matemática, o Ministério da Educação, através da Direcção-Geral de Inovação e de Desenvolvimento Curricular, convidou Maria Eugénia Graça Martins, Luísa Canto e Castro Loura e Maria de Fátima Mendes a elaborar uma brochura que apoiasse, do ponto de vista científico, os professores do Ensino Básico no domínio da organização, análise e interpretação de dados.

Esta publicação constitui-se como um importante recurso posto à disposição dos professores numa temática que assume cada vez maior relevância no mundo de hoje. Paralelamente, marca a afirmação da importância da temática da *Análise de Dados* desde os primeiros anos de escolaridade apoiando o professor no desenvolvimento do seu conhecimento matemático.

Lisboa, 20 de Julho de 2007

O Director da Direcção-Geral de Inovação e de Desenvolvimento Curricular

Luís Capucha

Prefácio

Esta brochura foi organizada no âmbito do Programa Nacional de Formação Contínua em Matemática para professores do 1.º ciclo do Ensino Básico. A sua finalidade é constituir um instrumento de apoio, científico e didáctico, no domínio da organização, análise e interpretação de dados.

A publicação foi organizada de modo a incluir duas vertentes, a primeira das quais relacionada com os conhecimentos científicos associados à Estatística, onde se procurou transmitir, de forma clara e simples, os conceitos e procedimentos que consideramos fundamentais serem do conhecimento de um professor do Ensino Básico. À medida que esses conceitos e procedimentos são desenvolvidos, vão sendo apresentados exemplos ilustrativos a partir de contextos do dia-a-dia. Para além dos exemplos são ainda propostas diversas tarefas, que possibilitam ao professor uma melhor apropriação dos conceitos envolvidos.
A outra vertente, de âmbito mais didáctico, pretende constituir um recurso para o trabalho a ser desenvolvido na sala de aula no âmbito da educação estatística. Assim, e ao longo de todos os capítulos, são apresentadas e exploradas tarefas que podem ser propostas a alunos do Ensino Básico. Foi ainda preocupação das autoras, dar exemplos, para além dos relacionados com a vida de todos os dias, de contextos provenientes de outras áreas curriculares.

Considerando que hoje em dia o computador faz parte, cada vez mais, do nosso quotidiano, sugerimos, a propósito da construção de diferentes modos de organização de dados, o recurso ao *Excel*, uma ferramenta informática de utilização acessível e que facilita muitos dos procedimentos propostos.

A exploração feita ao nível dos conceitos e processos de organização, análise e interpretação de dados, vai um pouco para além de todo o trabalho a desenvolver na sala de aula. No entanto, cremos que um professor não deve esgotar o seu conhecimento no que explora com os seus alunos, é necessário que tenha um conhecimento sólido e mais aprofundado sobre os mesmos assuntos.

Por outro lado é fundamental que a actividade na sala de aula, em torno da análise de dados, seja realizada de forma integrada no desenvolvimento de projectos que partam do interesse dos alunos e contribuam para o desenvolvimento das competências estatísticas.

Acreditamos que a publicação desta brochura possa contribuir para considerar a literacia estatística como uma vertente fundamental para o desenvolvimento de cidadãos críticos e intervenientes, apesar de, até agora, no currículo do ensino básico dos primeiros anos, o papel que lhe tem sido atribuído ter sido pouco relevante.

As autoras

Índice

DADOS e VARIÁVEIS

A Estatística é uma Ciência que se aplica em todos os campos do conhecimento. Costuma-se dizer que é a ciência que trata dos **dados**. Os dados têm sido, desde há muitos séculos, instrumentos essenciais à compreensão do mundo que nos rodeia. Neste capítulo procedemos à classificação dos dados, processo este que condiciona, de um modo geral, a ferramenta estatística a utilizar na sua organização e no seu tratamento.

Introdução

O registo e análise de dados têm sido, desde há muitos séculos, instrumentos essenciais à compreensão do mundo que nos rodeia. Os físicos, por exemplo, registavam os dados resultantes das suas experiências e, posteriormente, analisavam-nos em busca de uma lei que explicasse os resultados obtidos. Com o avanço das técnicas estatísticas de análise de dados, é possível encontrar padrões e tendências em colecções de dados provenientes de muitas outras fontes que não, somente, as resultantes de experiências físicas. Na verdade, são poucas as áreas do saber onde não se recorre à análise de dados para confirmar teorias e propor novas interpretações para os fenómenos que são o seu objecto de estudo.

Perante uma colecção de dados, há duas formas possíveis de abordar a sua análise consoante interesse:

- apenas explorá-los, e encontrar padrões na colecção de dados – que é, por assim dizer, a **população** em estudo.
- extrapolar para um universo mais vasto os padrões encontrados nessa colecção de dados, a qual é parte (ou **amostra**) desse universo (ou **população**).

Para dar dois exemplos da nossa vida corrente, pense-se nos resultados obtidos num teste que um professor deu à sua turma e nos resultados obtidos numa sondagem à boca da urna nas eleições presidenciais. No primeiro caso, a população é a turma e os dados que se têm referem-se a toda a população enquanto que, no segundo caso, os dados referem-se a uma pequena parte da população de interesse. A grande maioria das situações onde é necessária a utilização de metodologias estatísticas, enquadra-se neste segundo caso.

População – colecção de unidades individuais, que podem ser pessoas, animais, resultados experimentais, com uma ou mais características em comum, que se pretendem analisar.

Amostra – subconjunto da população, que se observa com o objectivo de tirar conclusões para a população de onde foi retirada.

Dimensão da amostra – número de elementos da amostra.

Ao longo deste texto iremos incidir, fundamentalmente, nas técnicas estatísticas destinadas a descrever, explorar e encontar padrões numa colecção de dados. Aliás, mesmo quando o objectivo é inferir para uma população mais vasta, é usual iniciar o estudo de uma colecção de dados com aquilo a que se chama **análise exploratória** ou **estatística descritiva**: fase da análise de dados onde estes são organizados em tabelas e gráficos e onde se calculam algumas características sumativas como a moda, a mediana, a média, o desvio padrão, entre outras. De notar que, quando a colecção de dados coincide com a população, o estudo desses dados resume-se à estatística descritiva.

A fase seguinte do estudo de uma colecção de dados (que não será, aqui, objecto de estudo) designa-se por **análise inferencial** ou **inferência**: fase da análise de dados onde se propõem possíveis modelos probabilísticos para a forma como os dados referentes a toda a população se distribuem e se interligam. É com base nesses modelos que se infere da amostra para a população (da parte para o todo).

1.2 Dados e Variáveis

Os três primeiros capítulos desta brochura têm por objectivo ilustrar as diferentes etapas por que passa uma análise descritiva dos dados. A primeira dessas etapas consiste na identificação do **tipo** de dados que temos para analisar.

Observe-se a seguinte tabela – **Dados sobre casas** – (fictícia):

Ident.	N.º assoalhadas	Área (m^2)	Estado	Garagem	Zona	Preço (10^3 €)
1	3	99,0	0	0	C	138,50
2	3	90,5	0	0	B	190,30
3	3	109,0	0	0	B	179,26
4	3	104,8	0	0	B	162,74
5	5	138,7	1	1	A	357,32
6	2	87,3	0	0	B	157,39
7	2	93,7	0	0	B	138,34
8	4	118,5	0	0	B	209,46
9	2	88,9	0	1	A	169,60
10	2	95,6	0	0	B	153,56
11	3	104,3	0	0	C	149,00
12	3	126,5	1	0	A	299,33
13	4	118,5	0	0	B	207,66
14	3	98,9	0	1	B	182,86
15	3	100,3	1	1	A	236,27
16	3	94,7	0	0	B	188,17
17	2	88,0	0	0	C	122,84
18	2	92,4	0	1	B	149,20
19	2	101,1	0	0	A	160,13
20	1	66,3	0	1	A	147,89
21	2	96,8	1	0	A	202,63
22	3	103,8	0	0	A	205,92
23	2	109,0	0	1	A	185,66
24	3	119,0	0	1	A	210,21
25	2	100,8	0	1	A	208,88
26	1	79,5	1	0	A	186,09
27	3	114,6	0	0	B	183,49
28	2	91,1	0	0	C	126,80
29	2	94,9	0	0	A	165,69
30	2	98,1	1	1	A	290,00
31	3	94,9	0	1	B	170,18
32	3	103,0	0	1	B	189,22
33	2	104,4	1	0	A	255,90
34	3	112,9	1	0	A	281,25
35	2	87,6	0	0	C	121,47
36	2	76,7	1	1	A	210,24
37	5	163,3	0	0	B	295,98
38	3	154,2	0	0	A	255,03
39	1	75,9	0	0	A	135,69
40	2	90,2	0	0	B	151,26

Tabela com algumas características de 40 casas.

Trata-se de um registo com informação referente a 40 casas que estão à venda, nomeadamente, número de assoalhadas, área, estado (0-usada, 1-nova), ter ou não ter garagem (0-não tem, 1-tem), zona (A, B ou C) e preço (em milhares de euros). Na tabela surge ainda uma coluna com o número de identificação de cada casa.

Olhando com um pouco mais de detalhe para as quatro primeiras casas, verificamos que todas são usadas, têm 3 assoalhadas e não têm garagem. No entanto, diferem na área e no preço – uma característica dos dados estatísticos é a **variabilidade**. Os dados variam e é essa variabilidade que é objecto de estudo da estatística.

Uma **variável** é qualquer característica de um indivíduo ou objecto à qual se possa atribuir um número ou uma categoria. O indivíduo ou coisa relativamente ao qual se recolhe a informação é designado por **unidade observacional** ou caso.
Uma variável diz-se **quantitativa** (ou numérica) se se referir a uma característica que se possa contar ou medir. Por exemplo, o número de irmãos de um aluno escolhido ao acaso, na turma, é uma variável quantitativa de contagem, enquanto que a sua altura é uma variável quantitativa de medição.
Uma variável diz-se **qualitativa** (ou categórica) se não for susceptível de medição ou contagem, mas unicamente de uma classificação, podendo assumir várias modalidades ou categorias. Por exemplo, a cor dos olhos do aluno referido anteriormente, é uma variável qualitativa. Se só assumir duas categorias, diz-se **binária**. É o caso da variável sexo, que assume as categorias Feminino e Masculino.

As variáveis quantitativas de contagem, isto é, que se referem a características que só se podem contar e não se podem medir, designam-se também por variáveis quantitativas **discretas**; por sua vez, as variáveis quantitativas de medição, isto é, que se podem medir, também se designam por variáveis quantitativas **contínuas**.

Estas designações são bastante importantes, pois a ferramenta estatística a utilizar, no estudo das variáveis, depende do tipo de variável em estudo.

O resultado da observação da variável, sobre o indivíduo, é o **dado estatístico** ou simplesmente **dado**.

Algumas variáveis qualitativas apresentam uma ordem subjacente – são designadas por **qualitativas ordinais.** São exemplos de variáveis qualitativas ordinais: o *nível social* (com as categorias "baixo", "médio" e "elevado"), o *grau de satisfação* com um produto (com as categorias "nada satisfeito", "pouco satisfeito", "satisfeito", "bastante satisfeito" e "muito satisfeito") e grande parte das variáveis utilizadas em inquéritos na área das ciências sociais onde se avalia o nível atingido em cada variável solicitando ao respondente que coloque uma cruz numa grelha numerada de 1 a 5 (escala de Lickert).

No nosso exemplo, cujos dados estão apresentados na tabela, as unidades observacionais são as "casas" e as variáveis são cada uma das características observadas para cada casa:

- ***Número de assoalhadas*** – variável quantitativa discreta (ou de contagem).
- ***Área*** – variável quantitativa contínua (ou de medição).
- ***Estado*** – variável qualitativa binária.
- ***Garagem*** – variável qualitativa binária.
- ***Zona*** – variável qualitativa.

De notar que a primeira coluna da tabela não se pode classificar como uma variável, uma vez que se trata de um mero identificador não se reportando a qualquer característica da "unidade observacional".

Dissemos anteriormente que o objectivo da Estatística é o estudo de **Populações**, isto é, conjuntos de indivíduos (não necessariamente pessoas) com características comuns, que se pretendam estudar. A uma característica comum, que assume valores diferentes de indivíduo para indivíduo, chamámos **variável**. Sendo então o nosso objectivo o estudo de uma (ou mais) característica(s) da População, vamos *identificar População com a variável que se está a estudar*, dizendo que a População é constituída por todos os valores que a variável pode assumir. Por exemplo, relativamente à população portuguesa, se o objectivo do nosso estudo for a característica altura, diremos que a população é constituída por todos os valortes possíveis para a variável altura. Do mesmo modo identificaremos **amostra** com os valores observados para a variável em estudo, sobre alguns elementos da População. Assim, na continuação do exemplo referido, os valores 156 cm, 171 cm, 163 cm, 168 cm, 166 cm, obtidos ao medir a altura de 5 portugueses, constituem uma amostra da população a estudar.

Na Sala de Aula

Tarefa

Vamos conhecer a turma!...

Conhecermo-nos uns aos outros faz parte do nosso dia a dia de vida em sociedade. Fazer ressaltar as semelhanças e diferenças do grupo de alunos da turma pode ser uma boa forma de sensibilizar os alunos para a importância de organizar e analisar dados e para os confrontar com os diversos tipos de dados.

Uma vez que interessa considerar e distinguir variáveis qualitativas e quantitativas (discretas e contínuas), eis alguns exemplos:

- ***Qualitativas*** – cor dos olhos, mês em que nasceu, transporte que usa para vir para a escola, cor de que mais gosta, animal de estimação,...
- ***Quantitativas discretas*** – número de irmãos, número de letras do nome, número de vogais no nome,...
- ***Quantitativas contínuas*** – comprimento do palmo, tempo que demora a ir de casa para a escola, peso da mochila,...

Destas variáveis escolhemos algumas para ilustrar de que modo poderão ser abordados diversos conceitos estatísticos muito simples.

A propósito de se conhecer melhor os alunos da turma, e da forma de organizar as diferentes características, o professor pode propor que se preencha uma tabela, como a que a seguir se apresenta, que reúne algumas características de cada aluno:

Nome	Número de letras no nome	Tempo que demora de casa à escola	Cor dos olhos	Comprimento do palmo	Número de irmãos

O professor pode ainda dar alguns esclarecimentos e fazer algumas recomendações, tais como:

- Se os alunos não souberem muito bem quanto tempo demoram no caminho entre a sua casa e a escola, basta darem um número aproximado.
- Os alunos deverão, no dia seguinte, ter o cuidado de escrever num papel a hora a que saem de casa e a hora a que chegam à escola.
- Para medir o comprimento do palmo, deve ser colocado o polegar da mão direita junto ao zero da régua e depois ver até quantos centímetros chega o dedo mindinho.

Eis o exemplo de uma tabela preenchida com as variáveis sugeridas anteriormente.

Nome	Número de letras no nome	Tempo que demora de casa à escola (minutos)	Cor dos olhos	Comprim. do palmo (cm)	Número de irmãos
Ana Patrícia Santos	17	3	Azuis	14,7	3
Ana Rita Pereira	14	32	Castanhos	15,6	1
Bruno Martins	12	25	Castanhos	15,9	1
Cátia Reis	9	20	Castanhos	14,2	1
Cláudia Rodrigues	16	17	Azuis	16,3	1
David Amaral	11	15	Azuis	13,5	2
Elisabete Soares	15	33	Pretos	14,4	1
José Manuel Rocha	15	22	Azuis	15,1	1
José Augusto Silva	16	9	Castanhos	15,2	1
Liliana Morais	13	35	Castanhos	16,2	1
Maria Isabel Antunes	18	25	Castanhos	15,9	2
Miguel Correia	13	28	Verdes	13,6	0
Patrícia Mendes	14	10	Castanhos	17,3	1
Pedro Mendes	11	21	Castanhos	14,7	2
Ricardo Freitas	14	20	Castanhos	15,0	0
Rui Eduardo Pires	15	6	Pretos	13,8	4
Sónia Gonçalves	14	5	Castanhos	14,3	1
Susana Alves	11	19	Castanhos	15,4	0
Tatiana Medeiros	15	13	Castanhos	14,8	1
Vasco Fernandes	14	5	Castanhos	13,2	3

Completada a tabela, chamar a atenção para os procedimentos que caracterizam a natureza dos dados, realçando as diferenças, mas sem insistir nas designações:

- Para preencherem a coluna do *número de letras no nome* os alunos têm de contar. Os dados que estão nessa coluna são, por isso, chamados **dados discretos** ou de contagem.
- Para preencherem a coluna do *comprimento do palmo* é necessário usar uma régua. Teve de se medir o palmo. Os dados que resultam de medições dizem-se **dados contínuos** ou de medição.
- A *cor dos olhos* não se mede, nem se conta!... Os dados que estão nessa coluna são chamados **qualitativos** ou categóricos.
- O *número de irmãos* conta-se, o *comprimento do palmo* mede-se usando uma régua ou uma fita métrica. O *tempo* também se mede mas usando um relógio ou um cronómetro.

Tarefa

Vamos conhecer os animais I

Uma outra proposta interessante para os alunos e que lhes permite distinguir diferentes tipos de variáveis, é a construção de um ficheiro com informação relativa a alguns animais. Por exemplo, numa turma cada aluno recolhe informação sobre um animal, nomeadamente no que diz respeito às seguintes características:

- Ter asas
- Ter penas
- Ter escamas
- Número de pernas
- Por ovos
- Viver na água

Nome	Tem asas	N.º de Pernas	Vive na água	Tem penas	Tem pêlo	Tem escamas	Põe ovos
Cão	Não	4	Não	Não	Sim	Não	Não
Gato	Não	4	Não	Não	Sim	Não	Não
Andorinha	Sim	2	Não	Sim	Não	Não	Sim
Elefante	Não	4	Não	Não	Sim	Não	Não
Burro	Não	4	Não	Não	Sim	Não	Não
Sardinha	Não	0	Sim	Não	Não	Sim	Sim
Melro	Sim	2	Não	Sim	Não	Não	Sim
Girafa	Não	4	Não	Não	Sim	Não	Não
Urso	Não	4	Não	Não	Sim	Não	Não
Rã	Não	2	Sim	Não	Não	Não	Sim
Pintassilgo	Sim	2	Não	Sim	Não	Não	Sim
Carapau	Não	0	Sim	Não	Não	Sim	Sim
Pescada	Não	0	Sim	Não	Não	Sim	Sim
Rato	Não	4	Não	Não	Sim	Não	Não
Piriquito	Sim	2	Não	Sim	Não	Não	Sim
Galinha	Sim	2	Não	Sim	Não	Não	Sim
Baleia	Não	0	Sim	Não	Sim	Não	Não
Mosca	Sim	6	Não	Não	Não	Não	Sim
Barata	Sim	6	Não	Não	Não	Não	Sim
Aranha	Não	8	Sim	Não	Não	Não	Sim

Depois da tabela construída, podem ser feitas perguntas do tipo:

- Todos os animais que vivem na água, são peixes? Consegues encontrar, na tabela anterior um animal que viva na água e não seja peixe?
- Recorda o que é um mamífero. Conheces algum mamífero que viva na água?
- Dá exemplo de uma característica que não se possa medir ou contar.
- Dá exemplo de uma característica que possa ser objecto de contagem e outra que possa ser medida, se as houver na tabela.

Uma característica que não se possa medir nem contar é, por exemplo, *ter asas*. Na verdade, um animal ou tem, ou não tem asas. Outra característica relacionada com as asas, seria *número de asas* de um animal. Neste caso já poderíamos contar o número de asas e por isso esta característica já não poderia ser dada como resposta a esta pergunta.
Uma característica que se possa contar é, por exemplo *número de pernas*. Na tabela não existe nenhuma característica que possa ser medida.

Tarefa proposta

Conhecer os hábitos de lazer

Outro exemplo de tarefa que pode ser proposta aos alunos na sala de aula, é a seguinte: Pretende-se conhecer os hábitos de lazer dos alunos da escola. Na turma, os alunos, com a ajuda da professora, preparam as perguntas convenientes para obter a informação desejada e classificam o tipo de variáveis utilizadas, num estudo análogo ao feito na tarefa anterior.

ORGANIZAÇÃO dos DADOS em TABELAS e GRÁFICOS

Neste capítulo são apresentados alguns processos, nomeadamente *tabelas* e *gráficos*, adequados para organizar e resumir a informação contida nos dados, de forma a realçar as características mais importantes.

2.1 Introdução

O objectivo de organizar dados em tabelas e de os representar graficamente é fornecer uma informação visual rápida de padrões e tendências. A forma como se estruturam as tabelas e as representações gráficas mais adequadas, depende do tipo de dados que temos para analisar e dos aspectos que se pretendem evidenciar.

Esta análise inicial de dados, que é feita utilizando tabelas e gráficos, vai-nos permitir responder rapidamente a algumas questões, tais como:

- Serão os dados quase todos iguais?
- Serão muito diferentes uns dos outros?
- Existe algum padrão subjacente ou alguma tendência?
- Existem alguns agrupamentos especiais?
- Existem alguns dados muito diferentes da maior parte?

Estas questões, de um modo geral, não podem ser respondidas facilmente a partir dos dados em bruto, com aspecto desorganizado.

2.2 Tabelas e gráficos para dados qualitativos

Os dados qualitativos ou categóricos são os que resultam da análise de variáveis qualitativas. Relembre-se que cada unidade observacional assume, no que respeita a este tipo de variáveis, a designação de uma categoria e não de uma grandeza quantitativa. Por vezes, escolhe-se como designação de cada categoria um número mas isso em nada altera a natureza da variável. A análise estatística deste tipo de dados resume-se, por isso, à contagem do número de indivíduos em cada categoria e ao cálculo das respectivas percentagens.

Tomemos o exemplo das casas, apresentado no capítulo anterior. Há três variáveis qualitativas – *Garagem, Estado* e *Zona*. Para as duas primeiras optou-se por utilizar designações numéricas (0 - sem garagem, 1 - com garagem e 0 - usada, 1 - nova, respectivamente). Antes de se passar à representação gráfica é, de um modo geral, necessário registar a informação numa tabela de frequências.

2.2.1 Tabela de frequências para dados qualitativos

Numa **tabela de frequências para dados qualitativos** ou categóricos a informação é organizada, de um modo geral, em 3 colunas: coluna das *categorias ou classes* – onde se indicam todas as categorias da variável em estudo; coluna das *frequências absolutas* – onde se regista o total de elementos da amostra que pertencem a cada categoria e coluna das *frequências relativas* (ou percentagens) – onde se coloca, para cada categoria, o valor que se obtém dividindo a respectiva frequência absoluta pela dimensão da amostra.

Uma tabela de frequências representa, portanto, a *distribuição* da variável, na amostra em estudo, isto é, quais as categorias ou modalidades que assume, assim como a frequência (absoluta ou relativa) com que assume essas modalidades.

Garagem	Frequência Absoluta **(n_i)**	Frequência Relativa **(f_i)**
Sem garagem	27	0,675
Com garagem	13	0,325
Total	40	1,000

Estado	Frequência Absoluta **(n_i)**	Frequência Relativa **(f_i)**
Usada	31	0,775
Nova	9	0,225
Total	40	1,000

Zona	Frequência Absoluta **(n_i)**	Frequência Relativa **(f_i)**
A	19	0,475
B	16	0,400
C	5	0,125
Total	40	1,000

Tabelas de frequências correspondentes às variáveis qualitativas *Garagem*, *Estado* e *Zona*

Quando se organizam os dados de uma amostra numa tabela de frequências, um processo de fácil verificação de que as frequências devem estar bem calculadas, consiste em somá-las para todas as classes e verificar que:

- A soma das frequências absolutas é igual à dimensão da amostra;
- A soma das frequências relativas é igual a 1.

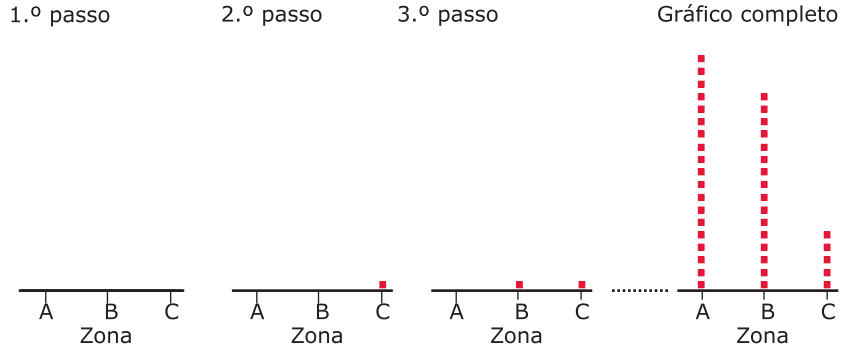
Observação:

Em muitas situações as frequências relativas são dízimas infinitas obrigando, por isso, a arredondamentos. Estes têm de ser feitos com algum cuidado, de modo a que o total seja igual a 1.

2.2.2 Gráfico de pontos e gráfico de barras para dados qualitativos

2.2.2.1 Gráfico de pontos

A representação gráfica mais simples que se pode construir é o gráfico (ou diagrama) de pontos (*dotplot*). Para obter esta representação basta desenhar um eixo horizontal (ou vertical), onde se assinalam as diferentes modalidades ou categorias da variável em estudo e, por cima de cada modalidade (ou ao lado), se representa um ponto, sempre que ao percorrer o conjunto de dados se encontrar a respectiva modalidade. Por exemplo, vejamos como obter o gráfico de pontos para a variável *Zona*, da tabela com os Dados sobre casas, do Capítulo 1. Num primeiro passo desenhamos um eixo, por exemplo horizontal, onde assinalamos as 3 modalidades diferentes da variável *Zona*: A, B e C. Depois, vamos nos passos seguintes colocando pontos, uns em cima dos outros, conforme formos percorrendo o conjunto dos dados C, B, B, B, A, ..., B relativos à variável *Zona*:



Algumas fases de construção de um gráfico de pontos

Esta representação é muito simples de fazer num papel quadriculado, em que se coloca um ponto em cada quadrícula:

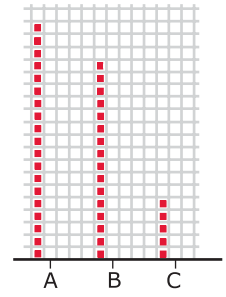
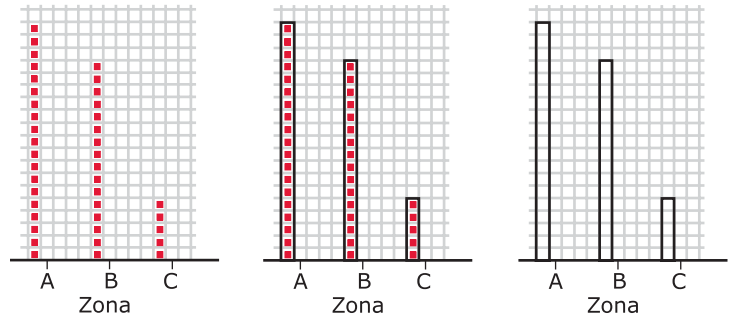


Gráfico de pontos construído em papel quadriculado

Podemos supor que, na representação gráfica anterior, se envolvem os pontos com um rectângulo e a seguir se retiram os pontos. O gráfico de pontos evolui para um outro gráfico, com aspecto semelhante ao gráfico de pontos, mas com barras:



Passagem de um gráfico de pontos a um gráfico de barras

Este tipo de gráfico (ou diagrama) de barras será objecto de estudo na secção seguinte.

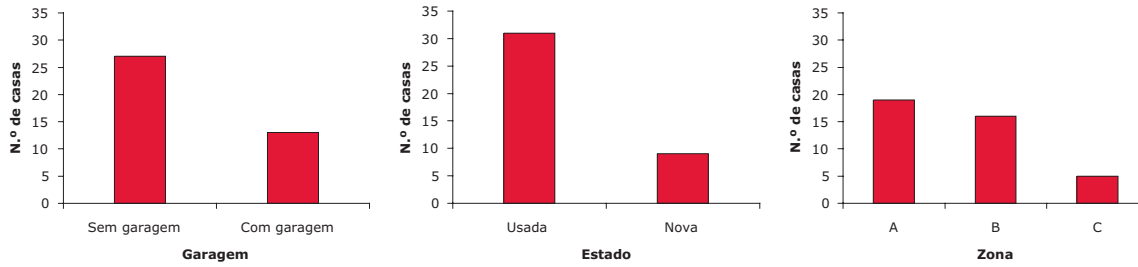
2.2.2.2 Gráfico de barras

Uma das representações gráficas mais utilizadas é o gráfico (ou diagrama) de barras. Neste tipo de gráfico desenha-se uma barra para cada categoria, sendo a altura da barra proporcional ao número de casos observados nessa categoria (frequência absoluta). Estas barras podem dispor-se ao longo de um eixo horizontal ou vertical. A ordem por que se colocam as barras é qualquer, salvo se existir alguma ordem subjacente, como nos dados qualitativos ordinais. Neste caso, deve-se respeitar a ordem colocando, da esquerda para a direita as diversas categorias, partindo da de menor nível para a de maior nível.

Não existem regras para a largura das barras nem para qualquer forma de acabamento gráfico – cor, textura, grossura dos traços, etc. No entanto, deve ter-se em atenção que as barras, no mesmo gráfico, devem ter a mesma largura, pois a mensagem que transmitem é a que está contida nas alturas, e umas barras mais largas do que outras poderiam chamar mais a atenção, induzindo em erro. Mais uma vez se frisa o cuidado a ter com as alturas das barras, que têm de ser iguais ou proporcionais à frequência observada em cada categoria.

Há ainda um cuidado suplementar a ter quando se representa, num mesmo gráfico, a informação contida em duas, ou mais, amostras de dimensão diferente. Nesse caso as alturas das barras têm de ser iguais à frequência relativa de cada categoria, pois só assim a soma das alturas das barras correspondentes a qualquer das amostras é idêntica (a soma dá sempre 1), permitindo a comparação. Se usássemos as frequências absolutas para alturas das barras dos gráficos, correspondentes às várias amostras, a comparação poderia induzir em erro, pois como a dimensão das amostras não é a mesma, estaríamos a comparar coisas diferentes.

Os gráficos de barras que correspondem às tabelas da secção 2.2.1 são, respectivamente, os seguintes:



Gráficos de barras correspondentes às variáveis qualitativas *Garagem*, *Estado* e *Zona*

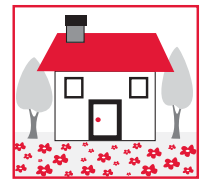
A principal vantagem dos gráficos relativamente às tabelas de frequências está na rapidez da leitura!... Não só há uma percepção imediata de qual a categoria de maior frequência, como também se fica com uma noção bastante precisa de qual a ordem de grandeza de cada categoria relativamente às restantes. Por isso se diz que "um gráfico vale mais que mil palavras!..."

Assim, observando os gráficos anteriores podemos afirmar, rapidamente, que, no que respeita às casas que constituem a nossa amostra, predominam as que não têm garagem (numa relação próxima de 2:1), a grande maioria das casas já teve algum dono (há cerca de três vezes mais casas usadas do que novas) e a distribuição do número de casas por cada zona é muito pouco uniforme, observando-se um número muito reduzido de casas na zona C, quando comparado com o das zonas A e B.

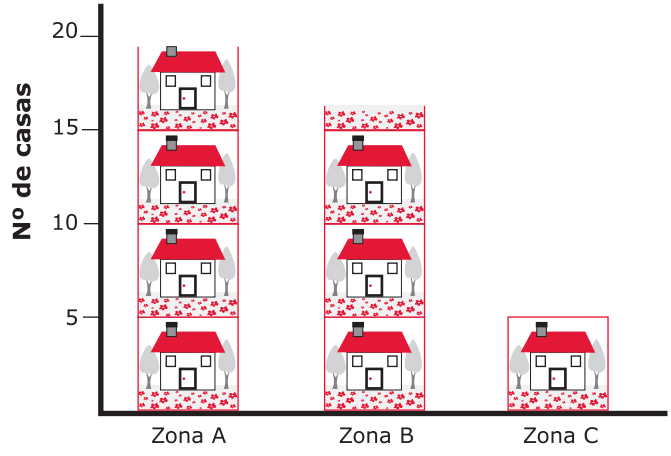
2.2.3 Pictograma

Uma representação gráfica que resulta especialmente atraente é o pictograma. Começa-se por escolher uma figura ilustrativa da unidade observacional. Cada figura pode representar uma ou mais unidades observacionais. De seguida procede-se como na construção do gráfico de barras mas, em vez de rectângulos, empilham-se as figuras que representam as unidades observacionais até perfazer a frequência absoluta observada em cada categoria. Esta representação só pode ser utilizada quando a variável em estudo é qualitativa.

As unidades observacionais no exemplo que temos vindo a tratar são "casas":



Admita-se que cada uma destas figuras representa 5 casas. O pictograma da variável qualitativa *Zona* terá 3 destas "casinhas" e mais uma quarta a que se lhe tira uma quinta parte, na categoria correspondente à zona A (onde a frequência absoluta é 19). Na categoria correspondente à zona B (onde a frequência absoluta é 16), terá 3 "casinhas" e mais um quinto de uma terceira "casinha" e a zona C (onde a frequência absoluta é 5) terá apenas uma "casinha".



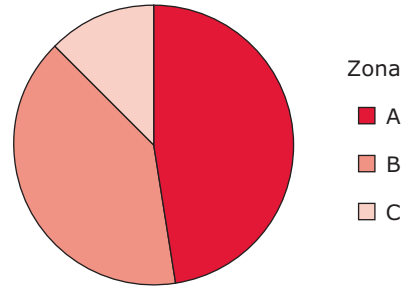
Pictograma correspondente à variável *Zona*

Embora seja uma representação gráfica muito sugestiva, é necessário ter os devidos cuidados com as figuras utilizadas e com a forma como são utilizadas, já que, com alguma frequência, dão origem a representações erradas, como veremos na secção 2.8.

2.2.4 Diagrama circular

Como o nome sugere, esta representação é constituída por um círculo, em que se apresentam vários sectores circulares, tantos quantas as categorias consideradas na tabela de frequências da amostra em estudo. O ângulo de cada sector circular é proporcional à frequência observada na classe que lhe corresponde.

Tomemos como exemplo a variável *Zona*. Tem 3 categorias: A, B e C com frequências relativas, respectivamente, iguais a 0,475, 0,400 e 0,125.



O sector circular correspondente à Zona A terá um ângulo de 360ºx0,475=171º, o da Zona B terá um ângulo de 360ºx0.400=144º, enquanto que o da Zona C terá 45º. A soma dos três ângulos é igual a 360º (171+144+45=360). É usual indicar os valores das frequências relativas junto dos respectivos sectores circulares, como se apresenta a seguir, sob a forma de percentagens:

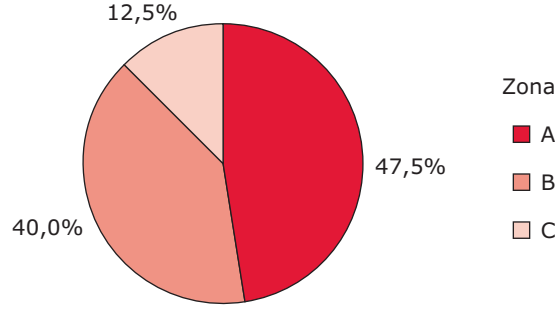


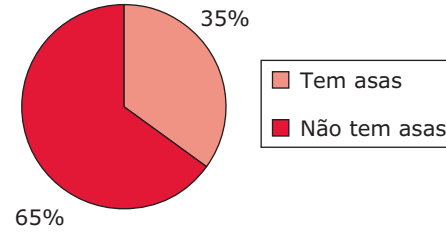
Diagrama circular correspondente à variável *Zona*

Tarefa

Vamos conhecer os animais II

Considere-se de novo a tarefa – Vamos conhecer os animais, e os dados da tabela associada. Pode-se escolher uma característica qualitativa e organizar os dados correspondentes na forma de uma tabela de frequências. Pode-se ainda construir uma representação gráfica conveniente.
Por exemplo, se for considerada a característica *ter asas*, que assume as modalidades "Tem asas" e "Não tem asas", a tabela de frequências permite concluir que, dos animais em estudo, predominam largamente os que não têm asas, relativamente aos que têm asas. Uma representação gráfica possível é o diagrama circular, que se apresenta a seguir:

	Freq. Abs.	**Freq. Rel.**
Tem asas	7	0,35
Não tem asas	13	0,65
Total	20	1.00



Utilização do *Excel* para construir uma tabela de frequências, um gráfico de barras e um diagrama circular para dados qualitativos

Tabela de frequências

Para construir uma tabela de frequências, para um conjunto de dados qualitativos, basta utilizar o seguinte procedimento:

- Inserir numa coluna do *Excel* os dados;
- Seleccionar as diferentes categorias que irão constituir as classes e inseri-las numa outra coluna a que chamamos Classes;
- Utilizar a função COUNTIF (CONTAR.SE) para obter as frequências absolutas para cada uma das classes;
- A partir das frequências absolutas, construir as frequências relativas.

Exemplificamos esta metodologia com uma das tabelas construídas anteriormente:

	F	H	Q	R	S
	Zona				
	C		Classes	Freq.abs.	Freq.rel
	B		A	=COUNTIF($F$2:$F$41;Q3)	=R3/$R$6
	B		B	=COUNTIF($F$2:$F$41;Q4)	=R4/$R$6
	B		C	=COUNTIF($F$2:$F$41;Q5)	=R5/$R$6
	A		Total	=SUM(R3:R5)	=SUM(S3:S5

Sheet1 Sheet2 Sheet3

Casas

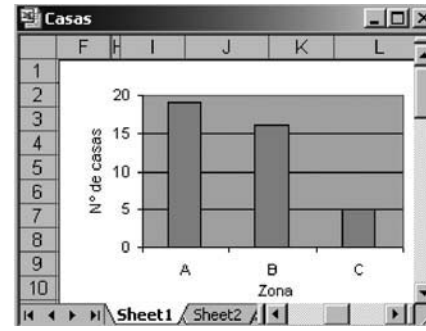
	F	H	Q	R	S
1	Zona				
2	C		Classes	Freq.abs.	Freq.rel
3	B		A	19	0,475
4	B		B	16	0,400
5	B		C	5	0,125
6	A		Total	40	1

Sheet1 Shee

Gráfico de barras

Para construir o gráfico de barras, a partir de uma tabela de frequências, se as classes são categorias, basta utilizar o seguinte procedimento:

- Seleccionar as células que contêm as classes e as frequências absolutas (ou frequências relativas), incluindo os cabeçalhos, ou seja Q2 a Q5 e R2 a R5 (se a coluna que contém as frequências relativas, não for adjacente à que contém as classes, então seleccione as classes e com a tecla CTRL pressionada seleccione as células que contêm as frequências relativas);
- Seleccionar, no menu, o ícone Chart  ;
- Na caixa de diálogo que aparece, seleccionar a opção *Column*;
- Clicar no botão *Next*, duas vezes, para passar dois passos, até aparecer uma caixa de diálogo, que apresenta várias opções: Em *Legend*, desactivar a legenda e em *Titles*, acrescentar o título no eixo dos Y's e no eixo dos X's.



Uma alternativa ao gráfico anterior, menos usual, é considerar as barras horizontais. Para obter a representação gráfica correspondente, basta seguir os passos anteriores, para a construção do gráfico de barras, com a única excepção de onde diz para seleccionar *Column*, seleccionar *Bar*:

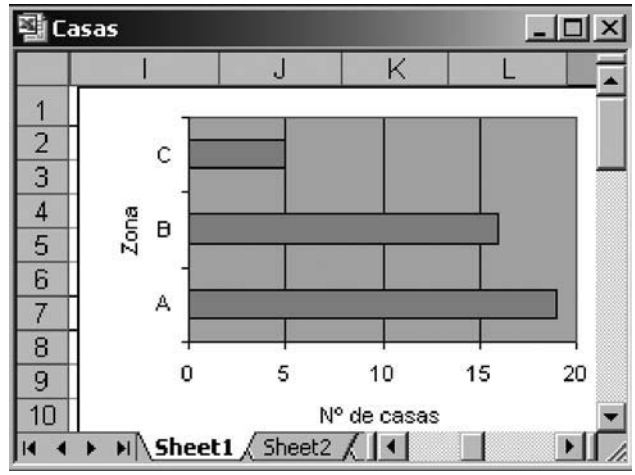
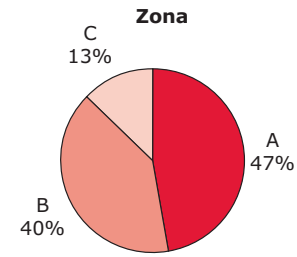


Diagrama Circular

A representação do diagrama circular, em *Excel*, é imediata, utilizando-se o seguinte procedimento:

- Seleccionar as células que contêm as classes e as frequências absolutas (ou frequências relativas), ou seja I3 a I5 e J3 a J5 (se a coluna que contém as frequências relativas, não for adjacente à que contém as classes, então seleccione as classes e com a tecla CTRL pressionada seleccione as células que contêm as frequências relativas);
- Seleccionar, no menu, o ícone Chart;
- Na caixa de diálogo que aparece, seleccionar a opção *Pie*; Escolher o subtipo pretendido (neste exemplo foi seleccionado o primeiro);



- Clicar no botão *Next*, duas vezes, para passar dois passos, até aparecer uma caixa de diálogo, que apresenta várias opções: Em *Legend*, desactivar a legenda; em *Titles* acrescentar o título, e em *Data Labels* seleccionar as opções pretendidas (nós seleccionámos *Category name* e *Percentage*).

Tabelas e gráficos para dados quantitativos discretos

Na sua definição formal, uma variável de natureza quantitativa diz-se **discreta** se o conjunto de valores que pode assumir for finito ou infinito numerável (isto é, pode-se estabelecer uma correspondência com os números naturais). Na prática, as variáveis discretas resultam sempre de contagens: número de filhos de cada família, número de carros que passam numa ponte por unidade de tempo, número de gralhas numa página dactilografada, número de chamadas telefónicas registadas por minuto numa central, etc.

A análise exploratória de dados quantitativos discretos tem duas abordagens possíveis: uma abordagem específica para dados discretos – quando o número de valores distintos na amostra for reduzido (por comparação com a dimensão da amostra) ou uma abordagem idêntica à utilizada para dados quantitativos contínuos – quando o número de valores distintos na amostra for muito elevado (quando comparado com a dimensão da amostra). Por exemplo, o tratamento de uma amostra constituída pelo número de chamadas telefónicas que um indivíduo recebe por dia, está na primeira situação, enquanto que a amostra do número de chamadas telefónicas recebidas por dia numa central, está na segunda situação.

Neste parágrafo vamos dar algumas indicações sobre a construção de tabelas e gráficos, específicos para dados discretos.

2.3.1 Tabela de frequências para dados quantitativos discretos

A construção da tabela de frequências para dados quantitativos discretos é idêntica à construída para dados qualitativos. Do mesmo modo que para os dados qualitativos, o primeiro passo é a escolha das classes, que aqui serão os diferentes valores que surgem na amostra:

Na **tabela de frequências para dados quantitativos discretos** a informação é organizada, no mínimo, em 3 colunas: coluna das *classes* – onde se indicam todos os valores distintos que surgem na amostra, que representamos por $\mathbf{x_i^*}$; coluna das *frequências absolutas* $\mathbf{n_i}$ – onde se regista o total de elementos da amostra que pertencem a cada classe (ou número de vezes que cada valor x_i^* surge na amostra) e coluna das *frequências relativas* (ou percentagens) $\mathbf{f_i}$ – onde se coloca, para cada classe, o valor que se obtém dividindo a respectiva frequência absoluta pela dimensão da amostra.

A tabela de frequências pode ainda incluir mais 2 colunas: a coluna das *frequências absolutas acumuladas* – onde, para cada classe, se coloca a soma da frequência absoluta observada nessa classe com as frequências absolutas observadas nas classes anteriores e a coluna das *frequências relativas acumuladas* – onde, para cada classe, se coloca a soma da frequência relativa observada nessa classe com as frequências relativas observadas nas classes anteriores. Como veremos mais à frente, esta coluna é bastante útil para o cálculo de algumas medidas, como a mediana e os quartis.

No exemplo das casas, temos uma variável quantitativa discreta que é o *Número de assoalhadas*. Após contagem do total de casas com cada número de assoalhadas obtém-se a seguinte tabela de frequências:

N.º de Assoalhadas x_i^*	Freq. Abs. n_i	Freq. Rel. f_i	Freq. Abs. Acum.	Freq. Rel. Acum.
1	3	0,075	3	0,075
2	17	0,425	20	0,500
3	16	0,400	36	0,900
4	2	0,050	38	0,950
5	2	0,050	40	1,000
Total	40	1,000		

Tabela de frequências para a variável *Número de assoalhadas*

Observe-se que, na coluna das frequências absolutas acumuladas, cada um dos valores é obtido fazendo a soma do valor que está na célula imediatamente acima, com o valor que está na célula das frequências absolutas. Assim, na linha correspondente a 3 assoalhadas, o valor 36, que surge como frequência absoluta acumulada, resulta da soma de 20 (que lhe está imediatamente acima) com 16. A excepção é o primeiro valor que coincide com a frequência absoluta. Para as frequências relativas acumuladas, processa-se de igual modo, usando a coluna das frequências relativas.

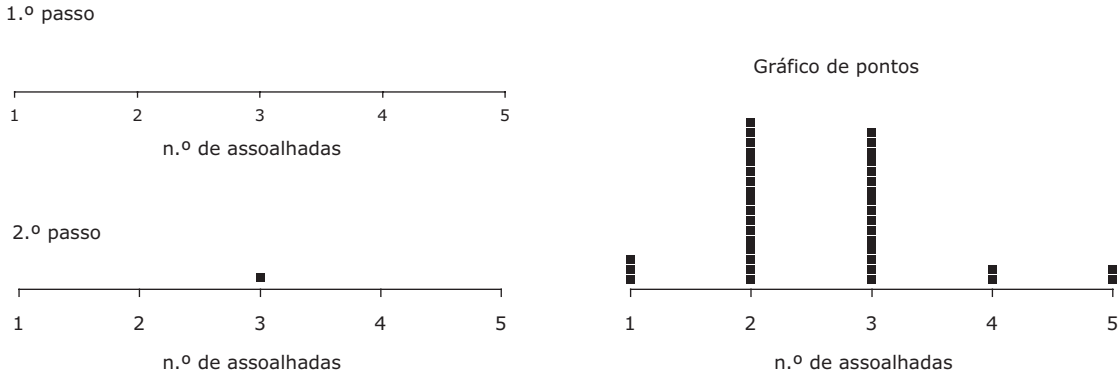
Esta tabela, para além de nos indicar a distribuição do número de assoalhadas na amostra, permite ainda fazer outro tipo de leituras: verificamos, por exemplo, que 90% das casas têm até um máximo de 3 assoalhadas (obtém-se a percentagem multiplicando 0,9 por 100); que a grande maioria das casas tem 2 ou 3 assoalhadas; que, na amostra, não há casas com mais de 5 assoalhadas, etc.

Convém salientar que as colunas referentes a frequências acumuladas só fazem sentido em tabelas de frequências onde a variável em estudo se possa ordenar.

2.3.2 Gráfico de pontos e gráfico de barras para dados quantitativos discretos

2.3.2.1 Gráfico de pontos

Tal como no caso de dados qualitativos ou categóricos, a representação gráfica mais simples é o gráfico ou diagrama de pontos. Para obter essa representação, basta traçar um eixo horizontal (ou vertical), onde se assinalam os diferentes valores que surgem na amostra ou mais correctamente, todos os valores entre o mínimo e o máximo, incluindo estes. Por cima de cada valor marca-se um ponto, sempre que se encontar um valor igual, ao percorrer a amostra. Por exemplo, vejamos como obter o gráfico de pontos para a variável *Número de assoalhadas*, da tabela com os Dados sobre casas, do Capítulo 1. Num primeiro passo desenhámos um eixo, onde assinalámos os diferentes valores que a variável pode assumir, ou sejam 1, 2, 3, 4 e 5. Depois, tal como fizemos para as variáveis qualitativas, fomos colocando pontos, uns em cima dos outros, à medida que percorremos o conjunto de dados 3, 3, 3, 3, 5,..., 2, correspondentes à variável *Número de assoalhadas*:



Algumas fases de construção de um gráfico de pontos

Da representação anterior, imediatamente se conclui que predominam as casas com 2 ou 3 assolhadas, sendo bastante inferior o número de casas com 1, 4 ou 5 assoalhadas.

Sugere-se que, para mais fácil execução, este gráfico seja feito em papel quadriculado, inserindo os pontos nas quadrículas convenientes.

Chamamos ainda a atenção, tal como fizemos para as variáveis qualitativas, que esta representação nos dá uma informação muito semelhante à dada pelo gráfico de barras, que apresentamos a seguir.

2.3.2.2 Gráfico de barras

O gráfico ou diagrama de barras é uma representação gráfica que consiste em marcar num sistema de eixos coordenados, no eixo horizontal, o valor correspondente a cada classe $\mathbf{x_i^*}$ e, nesses pontos, barras verticais de altura igual (ou proporcional) à respectiva frequência absoluta ou relativa. Devem-se utilizar as frequências relativas sempre que se pretenda comparar amostras de diferente dimensão (já que a soma das alturas das barras será, necessariamente, igual a 1 ou 100%, tornando possível a comparação de amostras de diferente dimensão).

Ilustramos esta representação gráfica com o gráfico de barras referente à variável *Número de assoalhadas*:

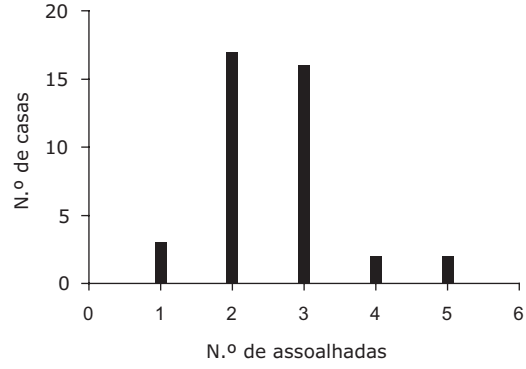


Gráfico de barras da variável *Número de assoalhadas*

Como se verifica a partir da representação gráfica anterior, predominam as casas com 2 ou 3 assoalhadas, havendo um número muito reduzido de casas com 4 ou 5 assoalhadas. Estas conclusões já tinham sido evidenciadas pela leitura da tabela de frequências e do gráfico de pontos.

Observação:

No eixo horizontal, deve ser marcada a sequência completa dos valores, entre o mínimo observado e o máximo observado, mesmo que algum esteja em falta na amostra. Nesse caso não haverá qualquer barra vertical nesse ponto.

Tarefa

Vamos conhecer os animais III

Consideremos ainda a tarefa – Vamos conhecer os animais.

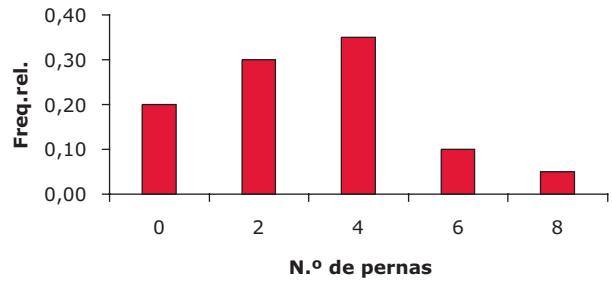
Pode ser sugerido aos alunos que, a partir dos dados da tabela associada:

- Organizem os dados dessa tabela, no que diz respeito ao *Número de pernas*, numa tabela de frequências.
- Construam uma representação gráfica adequada, tendo em conta a tabela de frequências, obtida anteriormente.

Para construir a tabela de frequências, deve-se começar por considerar os valores distintos que surgem no conjunto de dados e dispor estes valores por ordem crescente, numa coluna de uma tabela. Para ser mais fácil referirmo-nos a esses valores, vamos chamar-lhes classes. Depois contamos quantos dados são iguais a cada um dos valores seleccionados para as classes. Os valores obtidos são as frequências absolutas – indicam o número total de animais com 0, 2, 4, 6 e 8 pernas. Acrescentamos à tabela uma outra coluna, com as frequências relativas:

Classes	Freq. Abs.	Freq. Rel.
0	4	0,20 = 4/20
2	6	0,30 = 6/20
4	7	0,35 = 7/20
6	2	0,10 = 2/20
8	1	0,05 = 1/20
Total	20	1

Uma representação gráfica adequada é o gráfico de barras



Da tabela e gráfico anterior concluímos que predominam os animais de 4 pernas, seguidos dos de 2 pernas. De referir ainda a existência de um animal com 8 pernas, que ao consultar a tabela se verifica ser a aranha (Repare-se que no gráfico anterior não inserimos os números ímpares, entre o 0 e o 8, uma vez que eles não podem

fazer parte da população. Uma situação diferente seria a que se consideraria se no estudo da variável *Número de assoalhadas*, de uma amostra de casas, não tivéssemos obtido, por exemplo, o 2, que teria de ser incluído entre o 1 e o 3).

Quando inserido num contexto de sala de aula, pode pedir-se aos alunos para escreverem algumas frases a partir da observação do gráfico. O objectivo é irem desenvolvendo competências associadas à interpretação de dados organizados sob a forma de tabelas e gráficos.

2.3.3 Exemplos de tabelas e gráficos para dados quantitativos discretos

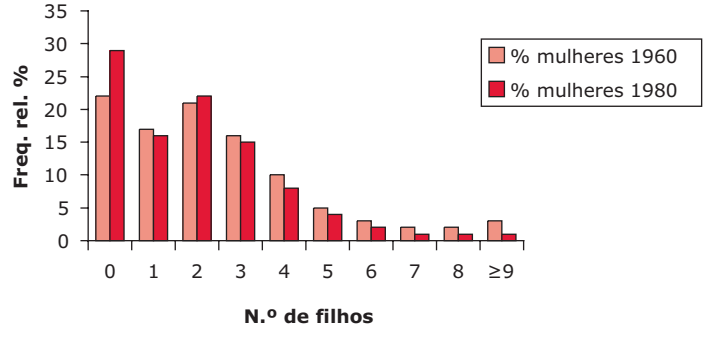
Vamos apresentar alguns exemplos relacionados com variáveis quantitativas discretas, onde se procura fazer uma interpretação dos dados a partir de tabelas ou gráficos.

Exemplo:

Número de filhos das mulheres americanas (Adaptado de Freedman *et al.,* 1991) – Em 1960 e novamente em 1980 foi feito um inquérito às mulheres americanas sobre o número de filhos. Os resultados obtidos foram os seguintes:

Número de filhos	*% mulheres 1960*	*% mulheres 1980*
0	22	29
1	17	16
2	21	22
3	16	15
4	10	8
5	5	4
6	3	2
7	2	1
8	2	1
≥9	3	1

Uma representação gráfica adequada, é o gráfico de barras, em que se apresenta lado a lado a distribuição das frequências para os anos de 1960 e 1980:



Da representação gráfica anterior ressalta o facto de a natalidade ter diminuído de 1960 para 1980. De facto, aumentou bastante a percentagem de mulheres sem filhos e diminuiu a percentagem de mulheres com 1 ou mais de 2 filhos. Esta diminuição só foi contrabalançada com um ligeiro aumento da percentagem de mulheres com 2 filhos.

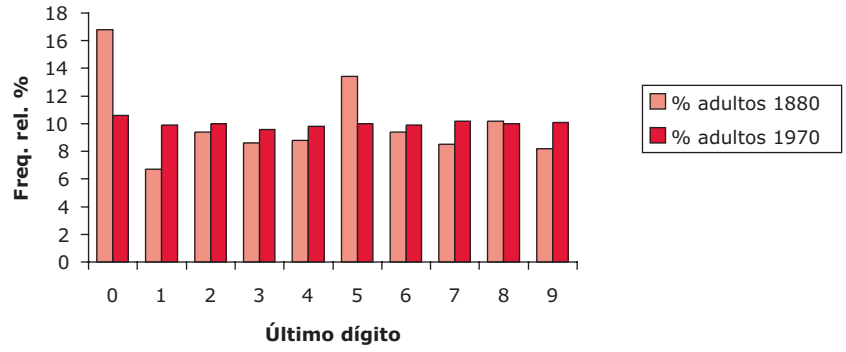
Exemplo:

Idade de indivíduos adultos (Adaptado de Freedman, 1991) – A tabela seguinte mostra a distribuição das frequências relativas do último dígito das idades dos indivíduos adultos. Esta informação foi recolhida relativamente a dois censos diferentes: o Censo de 1880 e o de 1970.

Último dígito da idade	% de indivíduos 1880	% de indivíduos 1970
0	16,8	10,6
1	6,7	9,9
2	9,4	10,0
3	8,6	9,6
4	8,8	9,8
5	13,4	10,0
6	9,4	9,9
7	8,5	10,2
8	10,2	10,0
9	8,2	10,1

Pode ser construído um diagrama de barras relativamente aos dois censos. Da consulta da tabela e do gráfico, verifica alguma anomalia?
Em 1880 havia uma nítida preferência pelos dígitos 0 e 5. Existe alguma explicação para este facto? Em 1970 essa preferência é muito mais fraca. Como se pode explicar esse facto?

Tal como se fez no exemplo anterior, construímos no mesmo gráfico de barras a distribuição das frequências para os anos de 1880 e 1970:

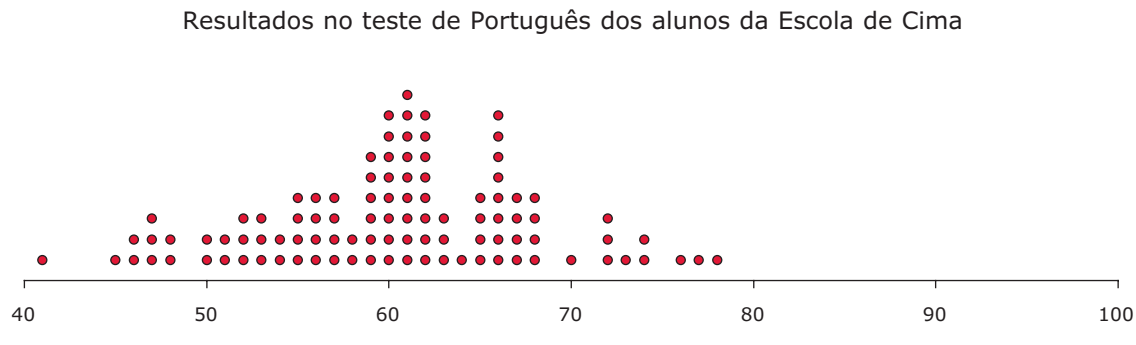


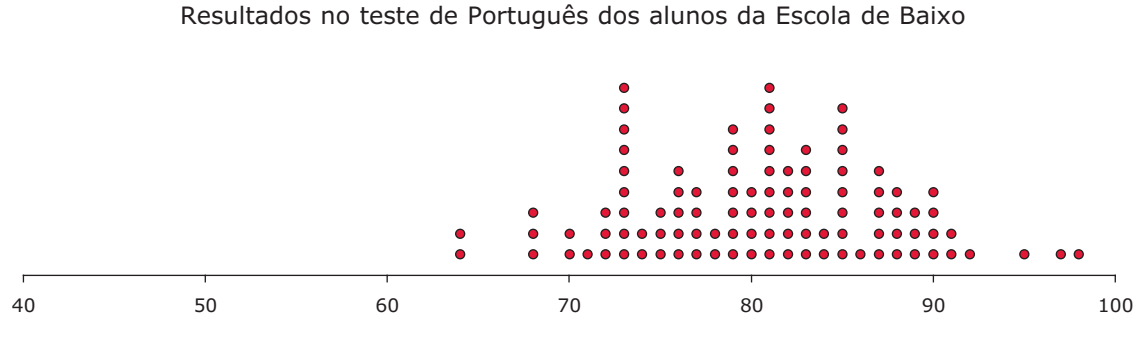
Também através do gráfico anterior ressalta o facto de haver, em 1880, uma predominância excessiva dos dígitos 0 e 5, em detrimento dos outros dígitos.

Uma explicação possível para, em 1880, as pessoas indicarem a idade a terminar em 0 ou 5, é não saberem ao certo a sua idade. Esta situação era vulgar, sobretudo nas pessoas mais idosas. Em 1970 esta situação já não se verifica, com a informação mais acessível a todos, verificando-se uma distribuição idêntica pelos 10 dígitos.

Exemplo:

Notas de duas escolas – A seguir apresentam-se dois gráficos de pontos com os resultados, numa escala de 0 a 100, dos alunos de duas escolas, num teste de Português:





Como se verifica, os alunos das duas escolas comportaram-se de forma muito diferente no teste. Os resultados da Escola de Baixo são nitidamente superiores aos resultados da Escola de Cima. Enquanto que a maior parte das notas dos alunos da Escola de Cima estão entre 50 e 70, a maior parte dos alunos da Escola de Baixo tiveram notas entre 70 e 90. Como seria um gráfico possível para as notas dos alunos de uma escola, cujos resultados estivessem entre os das duas escolas consideradas?

Exemplo:

Candidatos a algumas vagas (Adaptado de Freedman, 1991)- No Distrito Sanitário de Chicago, a escolha dos técnicos é feita mediante um exame. Em 1966, havia 223 candidatos para 15 vagas. O exame teve lugar no dia 12 de Março e os resultados dos testes (inteiros numa escala de 0 a 100) apresentam-se a seguir:

26	27	27	27	27	29	30	30	30	30	31	31	31	32	32
33	33	33	33	33	34	34	34	35	35	36	36	36	37	37
37	37	37	37	37	39	39	39	39	39	39	39	40	41	42
42	42	42	42	43	43	43	43	43	43	43	43	44	44	44
44	44	44	45	45	45	45	45	45	45	46	46	46	46	46
46	47	47	47	47	47	47	48	48	48	48	48	48	48	48
49	49	49	49	50	50	51	51	51	51	51	52	52	52	52
52	53	53	53	53	53	54	54	54	54	54	55	55	55	56
56	56	56	56	57	57	57	57	58	58	58	58	58	58	58
58	59	59	59	59	60	60	60	60	60	60	61	61	61	61
61	61	62	62	62	63	63	64	65	66	66	66	67	67	67
67	68	68	68	69	69	69	69	69	69	69	71	71	72	73
74	74	74	75	75	76	76	78	80	80	80	80	81	81	81
82	82	83	83	83	83	84	84	84	84	84	84	84	90	90
90	91	91	91	92	92	92	93	93	93	93	95	95		

Neste caso, a construção da tabela de frequências, segundo a metodologia descrita para dados discretos, conduziria a uma tabela com demasiadas classes. Assim, resolvemos tomar como classes uma partição natural, para os dados considerados, que é a seguinte: considerar como classes os intervalos 20 a 29, 30 a 39, 40 a 49, 50 a 59, 60 a 69, 70 a 79, 80 a 89, 90 a 99.

Classes	Freq. abs.	Freq. rel.
20 a 29	6	0,027
30 a 39	36	0,161
40 a 49	52	0,233
50 a 59	46	0,206
60 a 69	36	0,161
70 a 79	12	0,054
80 a 89	20	0,090
90 a 99	15	0,067
Total	223	1,000

Tabela de frequências para os resultados dos testes

A representação gráfica para os dados organizados desta forma já não pode ser um diagrama de barras, pois não existe um ponto onde colocar a barra, uma vez que as classes são intervalos. Veremos, mais à frente, que a representação gráfica adequada é o histograma.

A organização dos dados na forma da tabela anterior permite realçar o facto de predominarem as classificações entre 40 e 49, diminuindo progressivamente para baixo e para cima desses valores. Temos, no entanto de estar conscientes de que ao fazer a redução de dados há informação que sobressai, como a estrutura subjacente aos dados, embora haja outra informação que se possa perder. Vejamos qual o aspecto da tabela se tivéssemos considerado como classes todos os valores distintos da amostra, sem os agrupar:

Classe		Classe		Classe		Classe		Classe	
26	1	40	1	52	5	64	1	78	1
27	4	41	1	53	5	65	1	80	4
29	1	42	5	54	5	66	3	81	3
30	4	43	8	55	3	67	4	82	2
31	3	44	6	56	5	68	3	83	4
32	2	45	7	57	4	69	7	84	7
33	5	46	6	58	8	71	2	90	3
34	3	47	6	59	4	72	1	91	3
35	2	48	8	60	6	73	1	92	3
36	3	49	4	61	6	74	3	93	4
37	7	50	2	62	3	75	2	95	2
39	7	51	5	63	2	76	2		

Tabela de frequências para os dados sem estarem agrupados

O diagrama de barras correspondente tem o seguinte aspecto

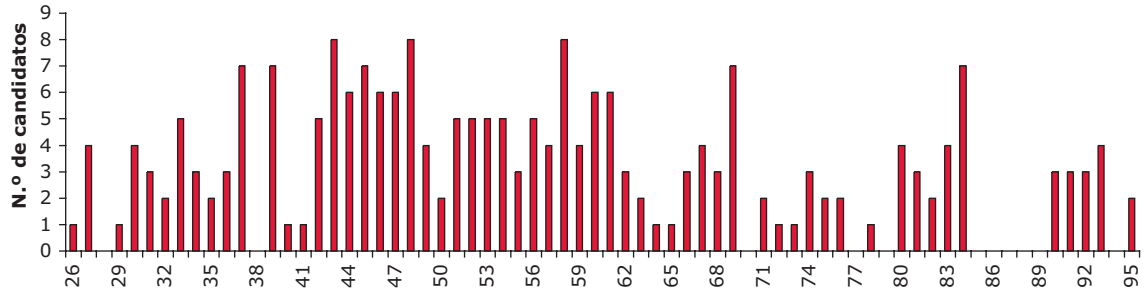


Diagrama de barras dos resultados nos testes

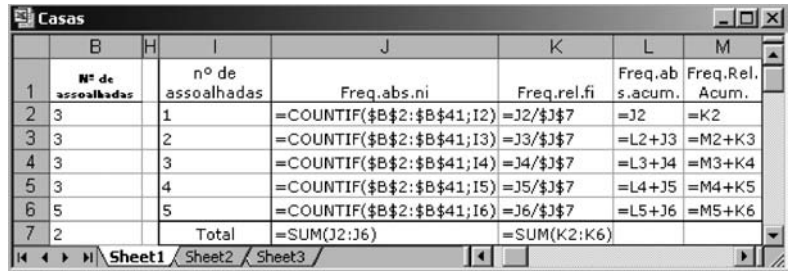
Da análise da tabela e do gráfico anterior verifica-se a existência de uma lacuna, não havendo classificações iguais a 85, 86, 87, 88 e 89 e o número de classificações iguais ou superiores a 90 ser de 15, precisamente igual ao número de vagas, para os 223 candidatos. Não terá havido batota da parte dos examinadores?

Chamamos a atenção para que esta representação, com tantas classes, não permite sobressair o padrão subjacente à distribuição dos dados, já que apresenta toda a variabilidade existente nesses dados. Como dissemos anteriormente, embora estejamos perante um conjunto de dados discretos, o tratamento adequado será o mesmo dos dados contínuos, apresentado na próxima secção.

Utilização do *Excel* para construir uma tabela de frequências e um gráfico de barras para dados quantitativos discretos

Tabela de frequências

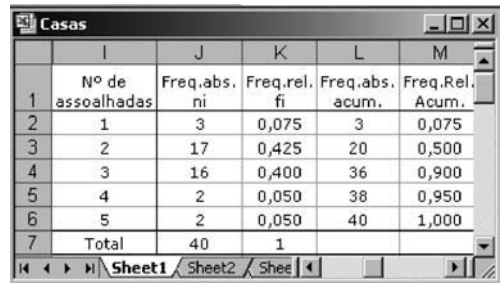
Para construir uma tabela de frequências, para um conjunto de dados quantitativos discretos, basta utilizar um procedimento idêntico ao utilizado para dados qualitativos. Assim, para a variável *Número de assoalhadas*, vem:



Casas

	B	H	I	J	K	L	M
1	Nº de assoalhadas		nº de assoalhadas	Freq.abs.ni	Freq.rel.fi	Freq.abs.acum.	Freq.Rel. Acum.
2	3		1	=COUNTIF($B$2:$B$41;I2)	=J2/$J$7	=J2	=K2
3	3		2	=COUNTIF($B$2:$B$41;I3)	=J3/$J$7	=L2+J3	=M2+K3
4	3		3	=COUNTIF($B$2:$B$41;I4)	=J4/$J$7	=L3+J4	=M3+K4
5	3		4	=COUNTIF($B$2:$B$41;I5)	=J5/$J$7	=L4+J5	=M4+K5
6	5		5	=COUNTIF($B$2:$B$41;I6)	=J6/$J$7	=L5+J6	=M5+K6
7	2		Total	=SUM(J2:J6)	=SUM(K2:K6)		

Sheet1 Sheet2 Sheet3



Casas

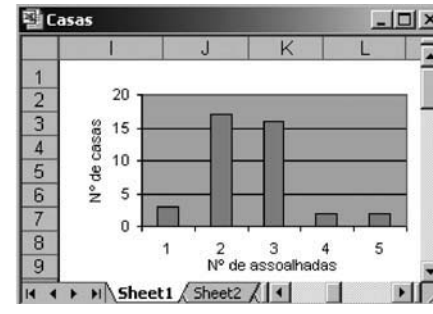
	I	J	K	L	M
1	Nº de assoalhadas	Freq.abs. ni	Freq.rel. fi	Freq.abs. acum.	Freq.Rel. Acum.
2	1	3	0,075	3	0,075
3	2	17	0,425	20	0,500
4	3	16	0,400	36	0,900
5	4	2	0,050	38	0,950
6	5	2	0,050	40	1,000
7	Total	40	1		

Sheet1 Sheet2 Shee

Gráfico de barras

Para construir o gráfico de barras, a partir de uma tabela frequências, que agrupa dados discretos, basta utilizar um procedimento idêntico ao utilizado para as variáveis qualitativas, em que as classes eram categorias, mas tendo em atenção o seguinte artifício:

- Apagar o título da coluna que contém as classes, No caso do exemplo apagar o conteúdo da célula I1, ou seja, "Número de assoalhadas";
- Seleccionar as células I1 a I6 e J1 a J6, caso pretenda construir o gráfico de barras com as frequências absolutas, ou K1 a K6, se desejar as frequências relativas;
- Proceder como se indicou na construção do gráfico de barras para variáveis qualitativas.



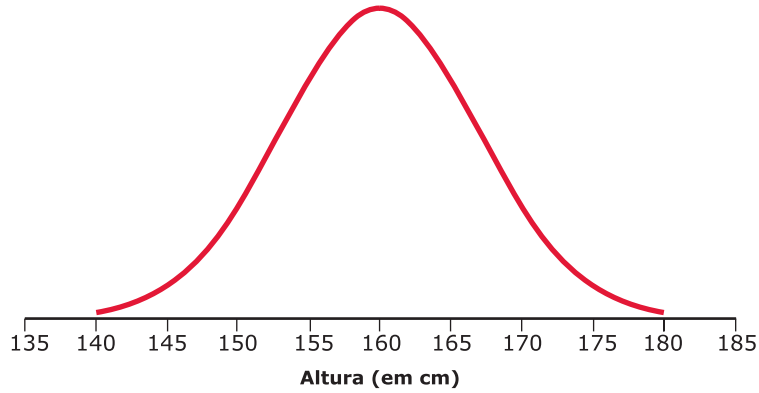
2.4 Tabelas e gráficos para dados quantitativos contínuos

Dados quantitativos contínuos são todos os que resultam de "medições", tal como foi dito anteriormente. Por outras palavras, a variável em estudo é passível de ser "medida" com algum "instrumento" (régua, balança, relógio, termómetro, etc.) e os dados são constituídos pelos valores resultantes das medições efectuadas. Para estas variáveis, qualquer valor num certo intervalo é um potencial candidato a aparecer na amostra. Por isso se chamam variáveis contínuas.

No nosso exemplo inicial das casas, há uma variável que se enquadra perfeitamente nesta definição: a *Área*. A área da casa resulta de uma medição e, embora seja apresentada com um arredondamento ao metro quadrado, sabemos que o verdadeiro valor pode ser qualquer número real num certo intervalo. Outra variável que também se pode considerar de natureza contínua é o *Preço*. O "instrumento" de medida é aqui menos preciso porque resulta de leis de mercado, mas não deixa, por isso, de "medir" o valor da casa. É de alguma forma semelhante à classificação em percentagem, obtida num teste pelos alunos de uma turma – o professor pretente "medir" o nível de conhecimentos de cada aluno e constrói o seu próprio instrumento de medida que é o teste. Como resultado das "medições" obtém uma amostra constituída pelas classificações dos alunos nesse teste.

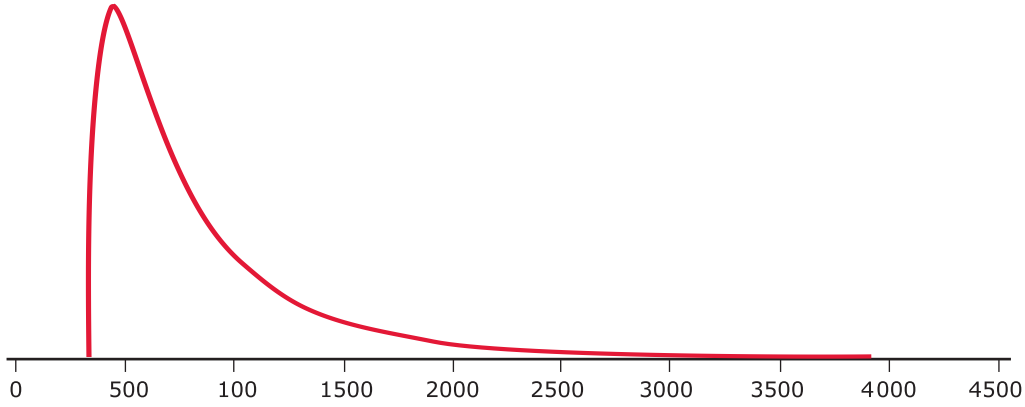
Uma característica comum a qualquer amostra cujos dados são de natureza contínua, é a grande diversidade de valores que a constituem. São poucos os valores repetidos. Como tal, para podermos visualizar a forma como os dados se distribuem, de nada nos serve fazer uma tabela onde se registe a frequência de cada valor distinto (como se fez para os dados quantitativos discretos). A alternativa aqui é organizar os dados num número conveniente de classes (intervalos) que permita condensar a informação sem esconder o padrão subjacente.

Não há regras rígidas para a forma como se constróem as classes, pois tal depende bastante da maior ou menor simetria na maneira como os dados se distribuem. Por exemplo, a subdivisão em classes de uma amostra de alturas de mulheres portuguesas processa-se de modo distinto da subdivisão em classes da amostra dos vencimentos auferidos por essas mesmas mulheres (onde é quase certo que a maior concentração seja em torno dos pequenos valores, podendo, no entanto surgir alguns valores extremamente elevados). Mais precisamente, é natural que a forma genérica da distribuição das alturas das mulheres portuguesas tenha um aspecto simétrico, como ilustrado na seguinte figura,



já que se espera que haja uma grande concentração em torno de 1,60m, com uma rarefacção gradual na direcção dos valores menores, ou maiores, que este valor central.

Por outro lado, no que diz respeito à distribuição dos vencimentos, o nosso conhecimento empírico leva-nos a supor que a sua forma genérica seja muito mais enviesada, como se apresenta na figura seguinte,



com a grande maioria dos vencimentos a não ultrapassar os 800 euros, dispersando-se os restantes ao longo de um intervalo, que pode atingir alguns milhares de euros.

Perante uma amostra de dados de tipo contínuo, o que se pretende com a subdivisão em classes é, exactamente, tornar patente a forma como esses dados se distribuem. Em muitos casos o bom senso preside à escolha das classes (principalmente em amostras muito enviesadas). No entanto, para dados que se distribuem de forma aproximadamente simétrica, é usual construir classes de igual comprimento (amplitude) e há uma regra relativamente simples para a determinação do número de classes, inspirada no Binómio de Newton*. Chama-se *regra de Sturges*, e consiste em determinar o menor inteiro k tal que $2^k>n$ (onde n é a dimensão da amostra):

- Regra de Sturges – Para organizar uma amostra, de dados contínuos, de dimensão n, pode considerar-se para número de classes o valor k, onde k é o menor inteiro tal que $2^k>n$.

Para a formação das classes pode-se escolher uma de duas estratégias:

Passo 1 – Subdividir um intervalo onde se encontrem todos os valores da amostra em k subintervalos de igual amplitude, h. O bom senso preside à escolha do referido intervalo. Assim, pode-se escolher como extremo esquerdo do intervalo o mínimo da amostra ou um valor que lhe seja um pouco inferior e escolher como extremo direito o máximo da amostra ou um valor que lhe seja um pouco superior.

* Tomemos uma potência de 2, por exemplo, 64 (que é igual a 2^6). Pelo Binómio de Newton sabemos que $2^6=1+6+15+20+15+6+1$, onde cada parcela da soma é cada uma das combinações do número 6 "j a j" com j a variar de 0 a 6. Na soma, o número de maior valor é o central e os restantes decrescem gradualmente à medida que se caminha para a direita e para a esquerda. Como 2^6 é igual a 64, se tivermos uma amostra de dimensão n=64, cujos dados se distribuam de forma aproximadamente simétrica, uma subdivisão em 7 classes (tantas quantas as parcelas que surgem na decomposição de 2^6) deverá conduzir a uma distribuição de frequências que capte bem a simetria da distribuição.

Passo 2 – Formar as classes como intervalos semiabertos (fechados à esquerda e abertos à direita, ou vice-versa), sendo o extremo esquerdo do primeiro intervalo coincidente com o extremo esquerdo do intervalo que se utilizou no passo 1.

ou

Passo 1' – Escolher como amplitude h, de cada intervalo, um valor arredondado por excesso daquele que se obtém dividindo a amplitude da amostra (máximo – mínimo) pelo número de classes, k.

Passo 2' – Formar as classes como intervalos semiabertos, fechados à esquerda e abertos à direita (ou vice-versa), sendo o extremo esquerdo do primeiro intervalo o mínimo da amostra.

Exemplo:

Subdivisão em classes dos dados referentes à variável Área

Uma vez que a nossa amostra tem dimensão n=40, o menor inteiro k tal que $2^k>40$ vem igual a 6. De acordo com a regra de Sturges, vamos então subdividir a amostra em 6 classes de igual amplitude. Para escolher as classes temos de começar por escolher um intervalo onde estejam todos os valores da amostra. Ora, ao ordenar a amostra verificamos que a área mínima é 66,3 m^2 e a área máxima é 163,3 m^2. Uma possibilidade razoável para o intervalo a subdividir será então o que vai de 65 m^2 a 165 m^2, com uma amplitude de 100 m^2 (165 m^2- 65 m^2). Dividindo 100 por 6, obtém-se a amplitude h=16,6(6) para cada um dos intervalos de classe. Em alternativa, também se pode escolher um intervalo com uma amplitude múltipla de 6 (de 64 m^2 a 166 m^2, por exemplo) o que conduz a um valor inteiro para h (h=17) e, consequentemente, a intervalos de classe cujos extremos são também números inteiros. Vamos optar por esta segunda hipótese, por ser a de mais fácil leitura. Antes de apresentar a tabela convém ainda estabelecer uma convenção quanto à inclusão ou não de cada extremo dos intervalos de classe. Assim, vamos convencionar que todos os intervalos são fechados à esquerda e abertos à direita, isto é, da forma [a, b[, onde o número que surge no extremo esquerdo (a) pertence ao intervalo, mas o número que surge no extremo direito (b) já não pertence. Esta metodologia é utilizada em algum *software* estatístico, mas não necessariamte em todo o *software*, pois há situações em que os intervalos considerados para as classes são abertos à esquerda e fechados à direita. O Excel, que não é um *software* estatístico, mas que permite construir tabelas de frequência, utiliza esta última metodologia, isto é, considera como elementos pertencentes à classe, os que são iguais ao limite superior.

Como optámos por subdividir o intervalo que vai de 64 m^2 a 166 m^2, com uma amplitude de classe igual a 17, o primeiro intervalo de classe será então [64, 81[, porque 64+17=81, o segundo [81, 98[e assim por diante até ao sexto e último intervalo que é [149 ,166[. Após a subdivisão em classes, o passo seguinte será construir a respectiva tabela de frequências.

2.4.1 Tabela de frequências para dados contínuos

Uma vez escolhidas as classes, a construção da tabela de frequências é idêntica à considerada para dados discretos:

Na **tabela de frequências para dados quantitativos contínuos** a informação é organizada, no mínimo, em 3 colunas: coluna das *classes* – onde se identificam os intervalos (classes) em que se subdividiu a amostra; coluna das *frequências absolutas* $\mathbf{n_i}$ – onde se regista o total de elementos da amostra, que pertencem a cada classe e coluna das *frequências relativas* $\mathbf{f_i}$ – onde se coloca, para cada classe, o valor que se obtém dividindo a respectiva frequência absoluta pela dimensão da amostra.

A tabela de frequências pode ainda incluir mais 3 colunas: coluna do representante da classe – onde se indica o ponto médio $\mathbf{x'_i}$ de cada intervalo de classe (usualmente escolhido para representante da classe); coluna das *frequências absolutas acumuladas* – onde, para cada classe, se coloca a soma da frequência absoluta observada nessa classe com as frequências absolutas observadas nas classes anteriores e coluna das *frequências relativas acumuladas* – onde, para cada classe, se coloca a soma da frequência relativa observada nessa classe com as frequências relativas observadas nas classes anteriores.

Apresentamos a seguir a tabela de frequências para a variável *Área*, do exemplo que tem vindo a ser tratado ao longo deste texto. Como sugerido pela regra de Sturges, considerámos 6 classes. Optámos por considerar classes fechadas à esquerda e abertas à direita e de amplitude 17 m^2. Como representante das classes considerámos os pontos médios, apresentados na 2.ª coluna da tabela. Por exemplo, o ponto médio da classe [64, 81[é (64+81)/2 = 72,5. Para obter as frequências absolutas percorre-se o conjunto de dados e contam-se os que caem dentro de cada classe (intervalo):

Classes	Rep. classe $\mathbf{x'_i}$	Freq. Abs. $\mathbf{n_i}$	Freq. Rel. $\mathbf{f_i}$	Freq. Abs. Acum	Freq. Abs. Acum
[64, 81[	72,5	4	0,100	4	0,100
[81, 98[	89,5	14	0,350	18	0,450
[98, 115[	106,5	15	0,375	33	0,825
[115, 132[	123,5	4	0,100	37	0,925
[132, 149[	140,5	1	0,025	38	0,950
[149, 166[	157,5	2	0,050	40	1,000
Total		40	1,000		

Tabela de frequências da variável *Área*

Por exemplo a frequência absoluta da classe [64, 81[é 4, porque só existem na amostra 4 valores maiores ou iguais a 64 e menores que 81, e assim sucessivamente, para as outras classes.

Como se verifica a partir da tabela, predominam as casas com áreas entre 81 e 115 m^2. Há uma assimetria no sentido de haver algumas casas, embora poucas, com áreas razoavelmente grandes, nomeadamente superiores a 149 m^2.

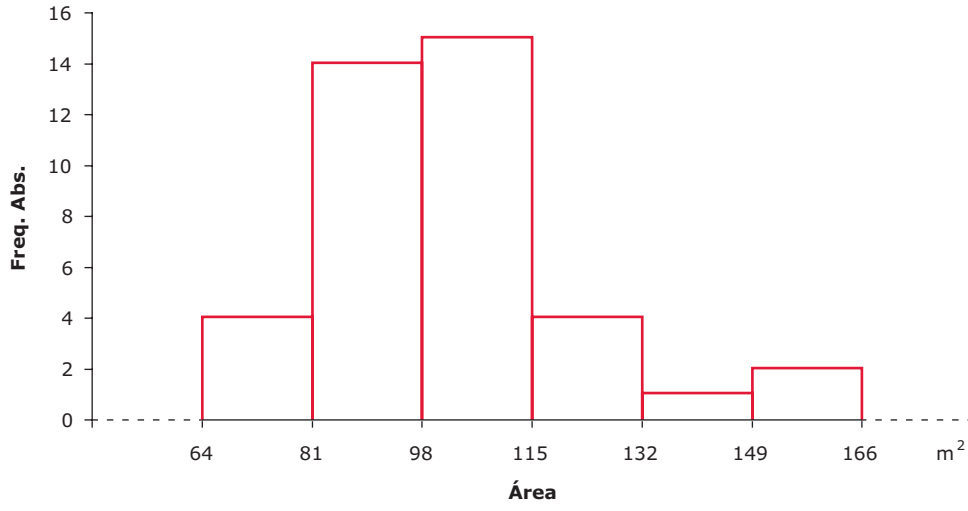
2.4.2 Histograma

O histograma é um tipo de representação usado para dados quantitativos contínuos. É um diagrama de áreas, formado por uma sucessão de rectângulos adjacentes, tendo cada um por base um intervalo de classe e por área a frequência relativa (ou absoluta) dessa classe.

Deste modo a área total coberta pelo histograma é igual a 1 (ou igual à dimensão da amostra) e a área determinada por dois pontos *a* e *b* dá-nos a percentagem de elementos da amostra que apresentam valores entre *a* e *b*. Observe-se que, para que a área de cada rectângulo seja igual à frequência relativa, é necessário que a altura seja o quociente entre a frequência relativa (f_i) e a amplitude da classe (h_i). Quando as classes têm todas a mesma amplitude (h), o aspecto gráfico não se altera se se considerar como altura a frequência relativa ou absoluta, uma vez que tal corresponde a uma simples mudança de escala no eixo vertical. Chama-se, no entanto, a atenção para o facto de a área total do histograma deixar de ser unitária passando a ser igual, respectivamente, à amplitude de classe h, ou ao produto da dimensão da amostra pela amplitude de classe (área total=n x h), caso se utilizem para alturas dos rectângulos as frequências relativas ou as frequências absolutas.

Nota 1: Se se pretender comparar várias amostras através de histogramas deve-se ter o cuidado de os construir de modo a que a área total seja unitária, para ser possível a comparação.
Nota 2: Um erro que se costuma cometer com frequência é construir o histograma com os rectângulos separados! Este procedimento não é correcto, pois os rectângulos devem ser adjacentes, dando no seu conjunto uma informação em termos de área.

Um histograma correspondente à tabela de frequências que construímos para a variável *Área* tem o seguinte aspecto (com alturas dos rectângulos iguais às frequências absolutas):



Histograma para a variável *Área*

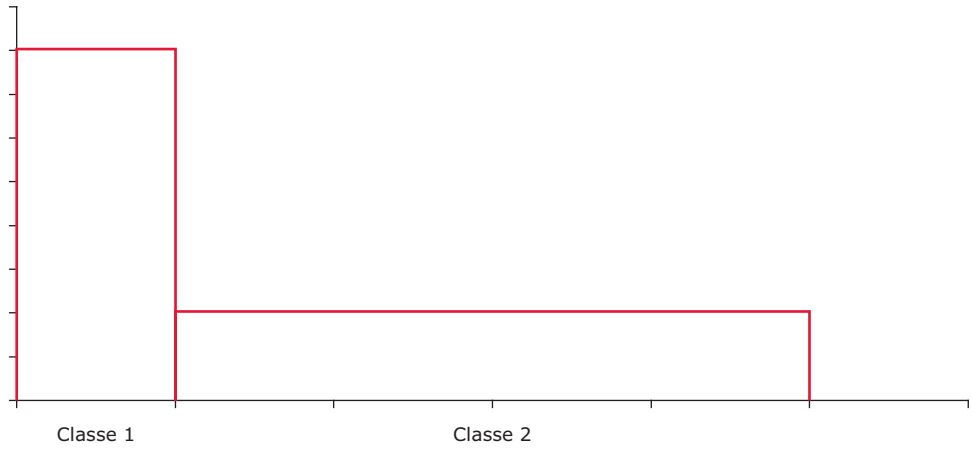
Mais uma vez, consegue-se com a representação gráfica uma percepção rápida e clara da forma como os dados se distribuem!

Assim, podemos fazer, por exemplo, as seguintes observações:

- há uma grande concentração de valores entre os 81 m^2 e os 115 m^2, indicando que é neste intervalo que se encontra a maioria das áreas das casas que constituem a amostra;
- são poucas as casas com áreas inferiores a 81 m^2;
- há uma ligeira assimetria no sentido das maiores áreas, pois surgem nesta zona alguns valores mais distantes dos valores centrais, que na zona das menores áreas. Em terminologia estatística diz-se que a distribuição apresenta uma cauda direita mais longa do que a cauda esquerda, havendo, por isso, uma assimetria positiva ou um enviesamento positivo.

Construção de histogramas com classes com amplitudes diferentes

Quando as classes em que os dados estão organizados não têm a mesma amplitude, tem que se ter o devido cuidado na construção das barras do histograma, pois a área de cada uma deve ser igual (ou proporcional) à frequência relativa. Se tivermos uma tabela de frequências em que, por exemplo, duas das classes tenham amplitudes diferentes, mas a que corresponda a mesma frequência, a relação entre as alturas dos rectângulos correspondentes a essas classes, deve ser a inversa da relação entre as suas amplitudes, como se apresenta a seguir:



Como a amplitude da classe 2 é 4 vezes maior que a amplitude da classe 1, então a altura do rectângulo correspondente à classe 2 deverá ser 4 vezes menor que a altura do rectângulo correspondente à classe 1.

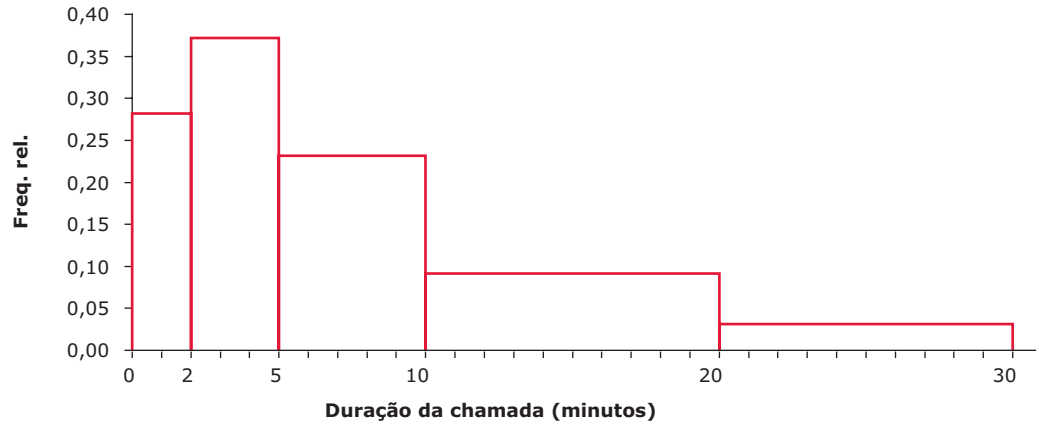
Exemplo:

Duração de chamadas telefónicas – Uma empresa, preocupada com os gastos em telefone, decidiu fazer um estudo sobre a duração (em minutos) das chamadas telefónicas. Assim, o departamento de controlo de qualidade recolheu uma amostra de dimensão 100, tendo construído a seguinte tabela de frequências, com os dados recolhidos:

Classes	Freq. absoluta	Freq. relativa
[0, 2[	28	0,28
[2, 5[	37	0,37
[5, 10[	23	0,23
[10, 20[	9	0,09
[20, 30[	3	0,03
Total	100	1,00

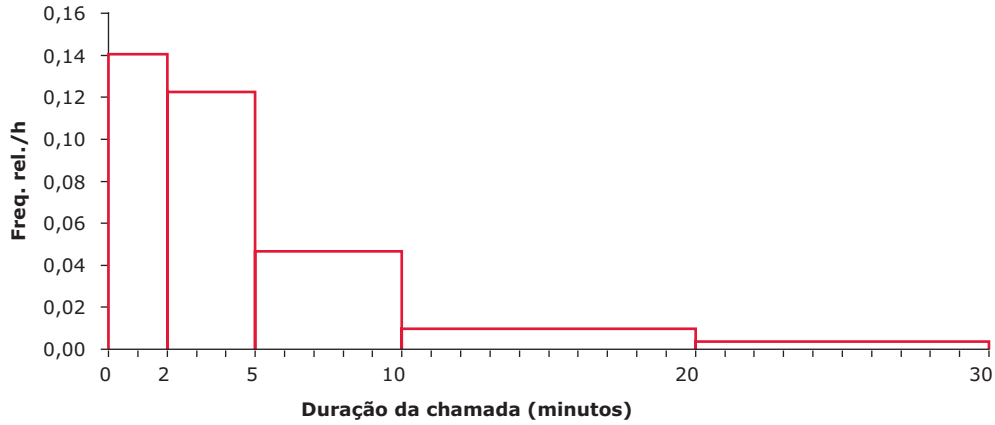
Duração da chamada (em minutos)

Construíram depois o seguinte histograma, que apresentaram à gerência (costuma-se dizer que um gráfico vale mais que mil palavras!):



Um dos gerentes, que sabia o que era um histograma, manifestou-se bastante preocupado com a percentagem de chamadas razoavelmente longas, já que a percentagem de chamadas com duração entre 5 e 10 minutos era um pouco superior às de duração entre 2 e 5 minutos e só um pouco inferior às de duração de 10 a 20 minutos, como se depreende pelas áreas dos rectângulos correspondentes às classes respectivas. Pediu para consultar a tabela de frequências e concluiu que aquela representação gráfica não estava correcta, pois as áreas dos rectângulos não eram proporcionais às frequências, induzindo em erro. Ele próprio acrescentou mais uma coluna à tabela de frequências, com as alturas correctas dos rectângulos e construiu o histograma correspondente:

Classes	Freq. absoluta	Freq. relativa	Freq.relativa/amplitude classe
[0, 2[	28	0,28	0,140
[2, 5[	37	0,37	0,122
[5, 10[	23	0,23	0,046
[10, 20[	9	0,09	0,009
[20, 30[	3	0,03	0,003
Total	100	1,00	



Repare-se que as duas representações são completamente diferentes.

2.4.3 Histograma acumulado

O histograma acumulado ou gráfico de frequências relativas acumuladas, tal como o nome indica, apresenta a evolução das frequências relativas acumuladas ao longo das classes, em que se subdividiu a amostra. Utiliza-se principalmente na determinação gráfica da mediana, dos quartis e de outros percentis quando os dados estão agrupados em classes. Estas medidas serão estudadas com mais pormenor no capítulo 3, mas devido à sua simplicidade e à sua importância na construção de uma representação gráfica muito simples, mas muito útil, vamos indicar a forma de as obter.

Como veremos, a mediana (Me) é um valor que divide a amostra, ordenada, ao meio, isto é, 50% dos elementos da amostra são menores ou iguais à mediana e os outros 50% são maiores ou iguais à mediana. Ficando a amostra dividida em duas partes, com igual número de elementos, cada uma destas partes ainda pode ser dividida ao meio. À mediana da parte inferior dos dados, chamamos 1.º quartil (Q_1), enquanto que à mediana da parte superior dos dados, chamamos 3.º quartil (Q_3). Repare-se que, deste modo, o 1.º quartil, a mediana e o 3.º quartil dividem os dados em 4 partes iguais: o 1.º quartil é tal que 25% dos dados são inferiores a ele; entre o 1.º quartil e a mediana estão outros 25% dos dados; entre a mediana e o 3.º quartil estão 25% dos dados, fazendo com que abaixo do 3.º quartil estejam 75% dos dados, enquanto que acima dele estão os restantes 25% dos dados.

Para obter graficamente estas medidas, tomemos de novo a seguinte tabela de frequências, obtida em 2.4.1, mas em que agora consideramos as percentagens para as frequências relativas (multiplicamos as frequências relativas por 100):

Classes	Rep. classe x'_i	Freq. Abs. n_i	Freq. Rel. (%) f_i	Freq. Abs. Acum.	Freq. Rel. Acum. (%)
[64, 81[	72,5	4	10,0	4	10,0
[81, 98[	89,5	14	35,0	18	45,0
[98, 115[	106,5	15	37,5	33	82,5
[115, 132[	123,5	4	10,0	37	92,5
[132, 149[	140,5	1	2,5	38	95,0
[149, 166[	157,5	2	5,0	40	100,0
Total		40	100,0		

O gráfico de frequências relativas acumuladas correspondente é

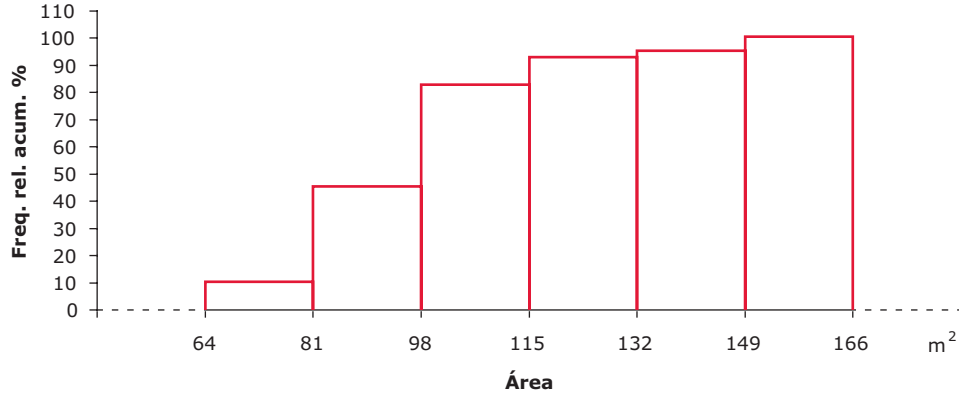
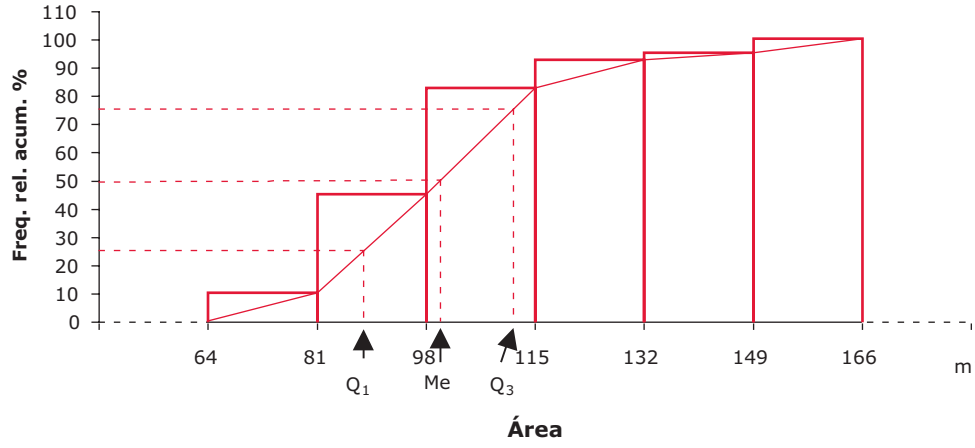


Gráfico das frequências relativas acumuladas

Para se obter graficamente a mediana (Me) e os quartis (Q_1 e Q_3), começa-se por traçar uma linha poligonal que une, em cada um dos rectângulos, o vértice inferior esquerdo com o vértice superior direito (ver figura). De seguida, toma-se no eixo vertical uma percentagem conveniente (50% para a mediana, 25% para o 1.º quartil e 75% para o 3.º quartil). Traça-se uma linha paralela ao eixo horizontal passando pelo ponto correspondente à percentagem de interesse e prolonga-se até encontrar a linha poligonal. Finalmente projecta-se sobre o eixo horizontal e obtém-se o respectivo quartil (repare-se que, à mediana, também podemos chamar 2.º quartil):



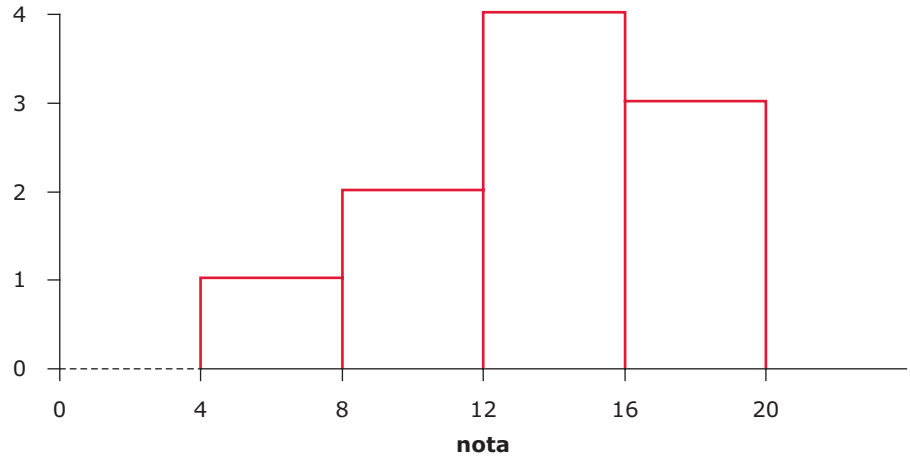
Como se verifica a partir da representação gráfica anterior, a mediana deve estar próxima de 100, enquanto o primeiro quartil deve estar próximo de 87 e o terceiro quartil andará à volta de 113. Salientamos que este procedimento, utilizado para dados agrupados, só dá valores aproximados.

2.4.4 Exemplos de tabelas e gráficos para dados quantitativos contínuos

Vamos apresentar alguns exemplos relacionados com variáveis quantitativas contínuas, onde se procura fazer uma interpretação dos dados a partir de tabelas ou gráficos.

Exemplo:

Notas finais a Matemática – O histograma seguinte mostra a distribuição das notas finais de Matemática (numa escala de 0 a 20) de uma determinada turma.



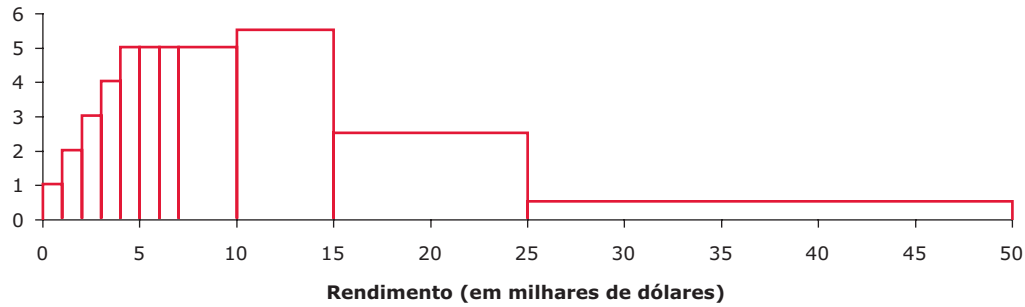
A partir do histograma anterior pode-se verificar que não houve nenhum aluno com nota inferior a 4.

Podem-se ainda colocar questões do género: Admitindo que 10% dos alunos da turma tiveram nota entre 4 e 8, qual a percentagem de alunos com nota entre 8 e 12?
Para responder a esta questão é fundamental ter presente que o histograma é um diagrama de áreas, pelo que se se está a admitir que 10% dos alunos tiveram nota entre 4 e 8, significa que a uma área de 4 unidades, que é a área do rectângulo mais à esquerda, corresponde uma frequência relativa de 10%. Então a percentagem de alunos com nota entre 8 e 12 será 20%, pois a área do rectângulo que corresponde a este intervalo é o dobro da área do rectângulo da classe anterior. De forma idêntica pode-se concluir que a percentagem de alunos que tiveram nota maior ou igual a 12 é 70%.

Neste exemplo convém fazer a seguinte observação: os valores assinalados no eixo vertical não correspondem necessariamente a frequências absolutas. Servem como orientação para calcular as áreas dos rectângulos correspondentes às classes. Assim, não sabemos quantos alunos fizeram o teste de Matemática.

Exemplo:

Rendimento familiar (Adaptado de Freedman *et al.,* 1991) – O histograma seguinte representa o rendimento familiar, em milhares de dólares de famílias americanas.



Tendo em conta que cerca de 1% das famílias têm rendimentos entre 0 e 1000 USD, estime a percentagem de famílias com rendimentos:

i) a) Entre 1000 USD e 2000 USD
b) Entre 2000 USD e 3000 USD
c) Entre 3000 USD e 4000 USD
d) Entre 4000 USD e 5000 USD
e) Entre 4000 USD e 7000 USD
f) Entre 7000 USD e 10000 USD

ii) a) Haverá mais famílias com rendimentos entre 6000 USD e 7000 USD ou entre 7000 USD e 8000 USD? Ou será aproximadamente o mesmo?
b) Haverá mais famílias com rendimentos entre 10000 USD e 11000 USD ou entre 15000 USD e 16000 USD? Ou será aproximadamente o mesmo?
c) Haverá mais famílias com rendimentos entre 10000 USD e 12000 USD ou entre 15000 USD e 20000 USD?

i) a) Se se diz que 1% das famílias têm rendimentos entre 0 e 1000 USD, então a área do rectângulo assente na classe [0, 1[, é igual a 1%. Repare-se que a escala do eixo vertical é tal que se se multiplicar a base do rectângulo pela sua altura, se obtém precisamente 1. Assim, para as outras classes, para obter as frequências relativas, basta calcular as áreas dos rectângulos respectivos. A resposta a esta alínea é então 2%; b) 3%; c) 4%; d) 5%; e) 15%; f) 15%.

ii) a) O mesmo, já que as áreas dos rectângulos correspondentes a essas classes são idênticas.

b) Mais entre 10000 USD e 11000 USD, pois a área do rectângulo correspondente a essa classe é superior ao da outra classe.

c) Mais entre 15000 USD e 20000 USD, pela mesma razão da alínea anterior.

Utilização do *Excel*, na construção da tabela de frequências e do histograma para dados quantitativos contínuos

Tabela de frequências

No caso de dados contínuos, o processo de construção das classes é um pouco mais elaborado do que no caso de dados discretos, já que a definição das classes não é tão imediata. De um modo geral as classes são intervalos com a mesma amplitude, fechados à esquerda e abertos à direita ou abertos à esquerda e fechados à direita. Em certos casos não é conveniente que as classes tenham a mesma amplitude, o que em si não é um problema para a construção da tabela de frequências, mas que implica alguma complicação na construção do histograma associado, quando pretendemos utilizar o *Excel*. Limitar-nos-emos a utilizar o *Excel* para a construção de histogramas associados a tabelas com as classes com igual amplitude.

Vamos exemplificar a construção de uma tabela de frequências com a variável *Área* do conjunto de dados, que temos vindo a estudar.

Definição das classes:

- Determinar a amplitude da amostra, subtraindo o mínimo do máximo;
- Dividir essa amplitude pelo número **k** de classes pretendido. Existe uma regra empírica que nos dá um valor aproximado para o número **k** de classes e que consiste no seguinte: para uma amostra de dimensão n, considerar para **k** o menor inteiro tal que $2^{\mathbf{k}}>n$. Uma expressão equivalente para obter **k**, consiste em considerar **k**=*INT*(LOG(n;2))+1 ou **k**=*ROUNDUP*(LOG(n;2);0), em que a função *ROUNDUP*(x;m), devolve um valor de x, arredondado por excesso, com m casas decimais;
- Calcular a amplitude de classe **h**, dividindo a amplitude da amostra por **k** e tomando para h um valor aproximado por excesso, do quociente anteriormente obtido;
- Construir as classes **C1, C2, ..., Ck**. Vamos considerar como classes os intervalos [mínimo, mínimo + **h**[,[mínimo + **h**, mínimo + 2**h**[, ..., [mínimo + (**k**-1)**h**, mínimo + **kh**[.Uma alternativa a este procedimento seria considerar as classes abertas à esquerda e fechadas à direita, da seguinte forma:]max – **kh**, max – (**k**-1)**h**],]max – (**k**-1)**h**, max – (**k**-2)**h**], **...,**]max – **h**, max].

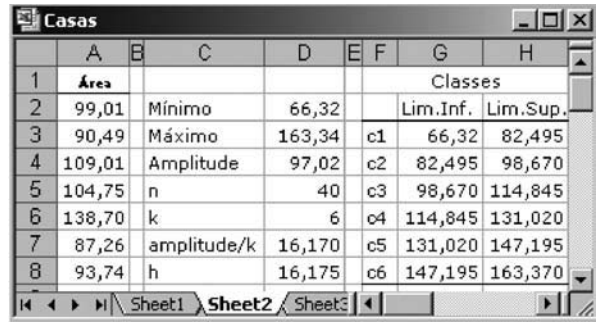
Estes passos são representados na figura seguinte:

Casas

	A	B	C	D	E	F	G	H
1	Área						Classes	
2	99,01		Mínimo	=MIN(A2:A41)			Lim.Inf.	Lim.Sup.
3	90,49		Máximo	=MAX(A2:A41)		c1	=D2	=G3+$D$8
4	109,01		Amplitude	=D3-D2		c2	=H3	=G4+$D$8
5	104,75		n	=COUNT(A2:A41		c3	=H4	=G5+$D$8
6	138,7		k	=INT(LOG(D5;2)		c4	=H5	=G6+$D$8
7	87,26		amplitude/k	=D4/D6		c5	=H6	=G7+$D$8
8	93,74		h	16,175		c6	=H7	=G8+$D$8

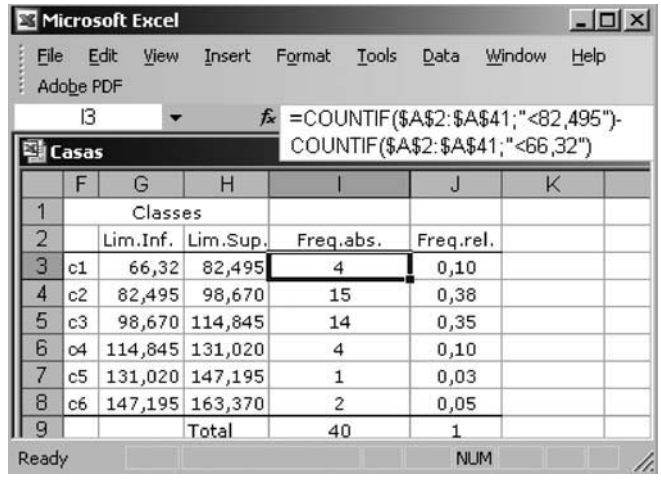
Sheet1 Sheet2 Sheet3

com os seguintes resultados:



Cálculo das frequências

Para obter as frequências absolutas, vamos utilizar a função COUNTIF, como se exemplifica para a classe c1:



As frequências das classes c2, c3, c4, c5 e c6, são obtidas de forma idêntica à de c1, mudando os limites das classes.

A construção de uma tabela de frequências pode ser feita utilizando um item chamado *Histogram*, disponível no *Excel*, em *Tools* → *Data Analysis*. Chama-se a atenção para que o nome deste item é enganador, pois na realidade, esta "função" limita-se a construir uma tabela de frequências. Para proceder ao agrupamento em k classes, utilizando o *Histogram*, é necessário começar por construir um conjunto de separadores de classes, b_1, b_2, ..., b_{k-1}, e as frequências absolutas obtidas com a "função" *Histogram*, são as correspondentes às seguintes classes:

1.ª classe – conterá todos os elementos $\leq b_1$;
2.ª classe – conterá todos os elementos $\leq b_2$ e $> b_1$;
3.ª classe – conterá todos os elementos $\leq b_3$ e $> b_2$;
....
k-ésima classe – conterá todos os elementos $> b_{k-1}$.

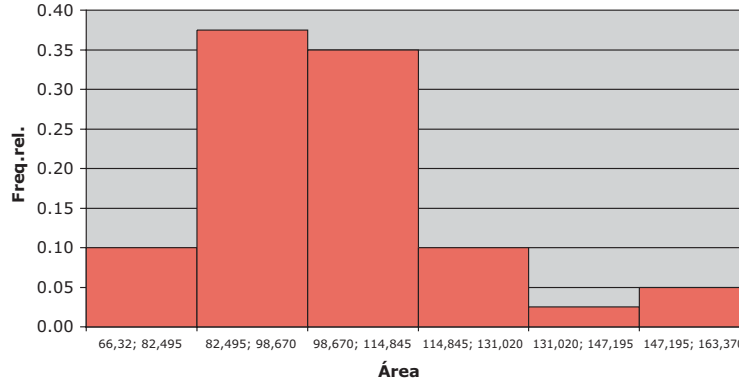
Construção do histograma

Para construir o histograma, a partir da tabela de frequências, pode-se utilizar o seguinte procedimento:

- Seleccionar as células que contêm as classes e as que contêm as frequências relativas (se pretender seleccionar células não adjacentes, basta seleccionar as células da primeira coluna e se a coluna seguinte não for adjacente, começar por carregar a tecla *CTRL* e com ela pressionada seleccionar, então, as células pretendidas, da coluna não adjacente);
- Proceder como se fosse construir um gráfico de barras;
- Clicar duas vezes sobre as barras, de forma a que apareça o menu *Format Data Series* ou *Format data Points*.; Seleccionar *Options* e em *Gap Width* seleccionar 0; *OK*:

Casas

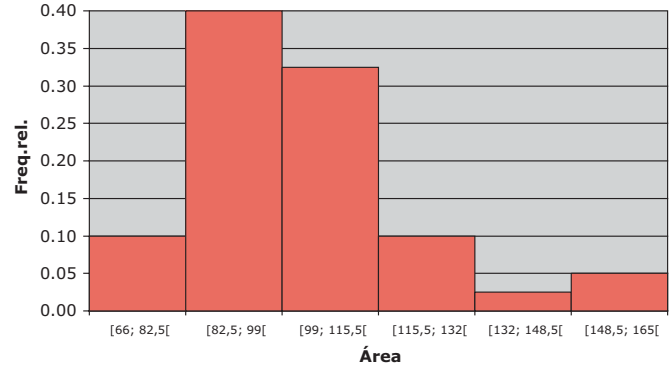
	L	N
2	Classes	Freq.rel.
3	66,32; 82,495	0,100
4	82,495; 98,670	0,375
5	98,670; 114,845	0,350
6	114,845; 131,020	0,100
7	131,020; 147,195	0,025
8	147,195; 163,370	0,050



Fica visualmente mais elucidativo se considerarmos as classes com outros limites, como por exemplo [66; 82,5[, [82,5; 99[, [99; 115,5[, [115,5; 132[, [132; 148,5[, [148,5; 165[, que não se afastam muito dos anteriores. Construindo a nova tabela de frequências e o correspondente histograma, vem:

Casas

	P	Q
2	Classes	Freq.rel.
3	[66; 82,5[	0,100
4	[82,5; 99[	0,400
5	[99; 115,5[	0,325
6	[115,5; 132[	0,100
7	[132; 148,5[	0,025
8	[148,5; 165[	0,050



Repare-se que a modificação processada nas classes, provocou uma alteração no histograma. Efectivamente, o histograma é uma representação que depende muito da amplitude que se considera para as classes e do ponto onde se inicia a construção das classes.

Nota: A observação anterior é importante, pois chama a atenção para o facto de, para o mesmo conjunto de dados, se poderem construir vários histogramas, nem todos com aspecto semelhante. Este facto faz com que se diga que o histograma não é uma representação *resistente*, pois pode mudar drasticamente de aspecto, quando se altera a amplitude da classe ou o valor em que se inicia a construção destas.

2.5 Outras representações gráficas

Além das representações gráficas consideradas anteriormente, de que se destacam o diagrama de barras e o histograma, utilizados especialmente para variáveis quantitativas discretas e contínuas, respectivamente, existem outras representações gráficas que se usam tanto para dados discretos, como contínuos. Passamos a apresentar as mais usuais.

2.5.1 Diagrama de extremos e quartis

Uma forma simples de evidenciar a forma como os dados se distribuem é através de uma representação gráfica que envolve apenas 5 valores retirados ou calculados a partir da amostra. Esses valores são: o mínimo, o máximo, a mediana, o 1.º quartil e o 3.º quartil. O diagrama de extremos e quartis é constituído por um rectângulo e por dois segmentos de recta que partem de cada um de dois lados opostos do rectângulo. Pode ser colocado na vertical ou na horizontal. O que mostramos na figura seguinte, do lado esquerdo, diz respeito à variável *Preço* e foi obtido através do *software* estatístico SPSS que utiliza a representação vertical:

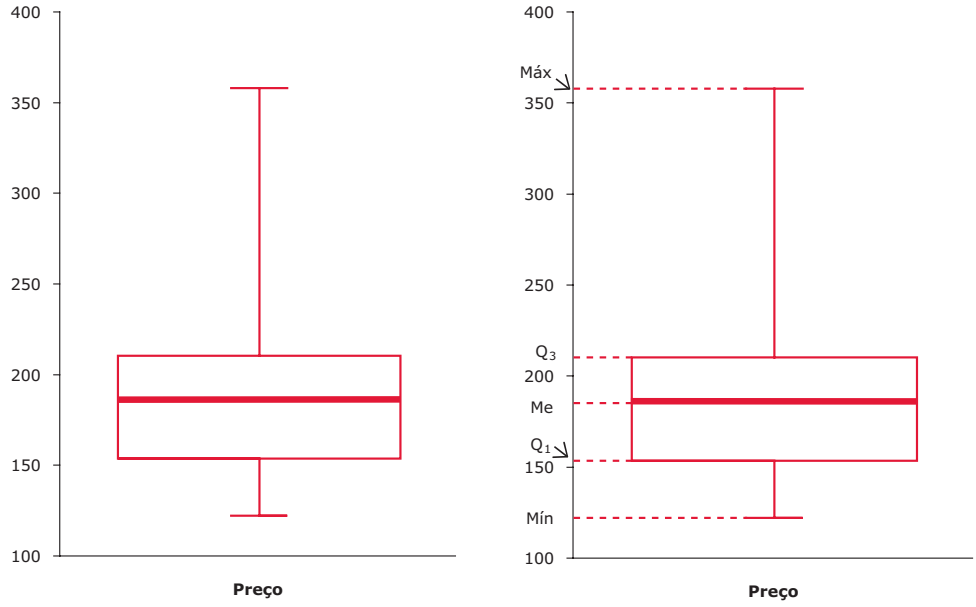


Diagrama de extremos e quartis para a variável *Preço*

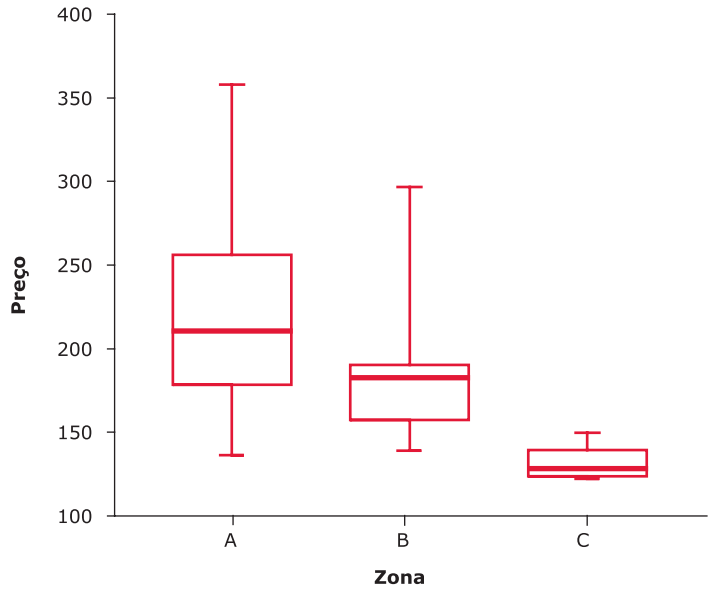
Como se pode ver na figura anterior, no gráfico do lado direito, o rectângulo é desenhado desde o 1.º quartil (que é 151,83 mil euros) até ao 3.º quartil (que é 210,02 mil euros). Dentro do rectângulo coloca-se um traço para assinalar a posição da mediana (que é 184,575 mil euros). Os dois segmentos de recta que completam esta representação gráfica estendem-se, um desde o mínimo da amostra (que é 121,47 mil euros) até ao lado do rectângulo determinado pelo 1.º quartil e o outro desde o lado do rectângulo determinado pelo 3.º quartil até ao máximo (que é 357,32 mil euros). Os diagramas de extremos e quartis permitem tirar conclusões importantes

acerca da forma como os dados se distribuem dentro da amostra. A partir do gráfico anterior podemos desde logo dizer que os preços das casas se distribuem de forma enviesada, com uma cauda mais longa no sentido dos grandes valores. Os 50% de valores centrais para os preços das casas situam-se entre cerca de 150 mil euros e 210 mil euros; o preço mínimo é pouco abaixo dos 150 mil euros, mas o preço máximo é bastante superior aos 210 mil euros, atingindo cerca de 350 mil euros; verifica-se ainda que metade das casas têm preços que não excedem o valor indicado pelo traço da mediana que, apenas pela leitura do gráfico, se verifica ser próxima dos 180 mil euros.

Utilização do diagrama de extremos e quartis para comparar várias amostras

Quando colocados em paralelo, os diagramas de extremos e quartis, permitem estabelecer comparações entre amostras, evidenciando as principais semelhanças e diferenças entre os padrões de distribuição, nomeadamente no que diz respeito à localização de algumas das medidas características dos dados, assim como à maior ou menor dispersão dos dados.

Que pode dizer acerca dos preços das casas nas zonas A, B e C?



A representação anterior torna evidente que os preços das casas da zona C são os mais baixos das 3 zonas consideradas, apresentando ainda uma pequena variabilidade entre o preço mais baixo e o preço mais alto. Pelo contrário, as casas da zona A são, de um modo geral, mais caras.

2.5.1.1 Construção do diagrama de extremos e quartis para dados agrupados

Como vimos anteriormente, na secção 2.4.3, o histograma acumulado permite obter valores aproximados para a mediana e quartis, quando os dados estão agrupados. Vamos então aproveitar essa facilidade para obter, neste caso, o diagrama de extremos e quartis. Para isso basta completar a representação gráfica com um diagrama que se desenha por baixo do gráfico de frequências relativas acumuladas, como se apresenta a seguir:

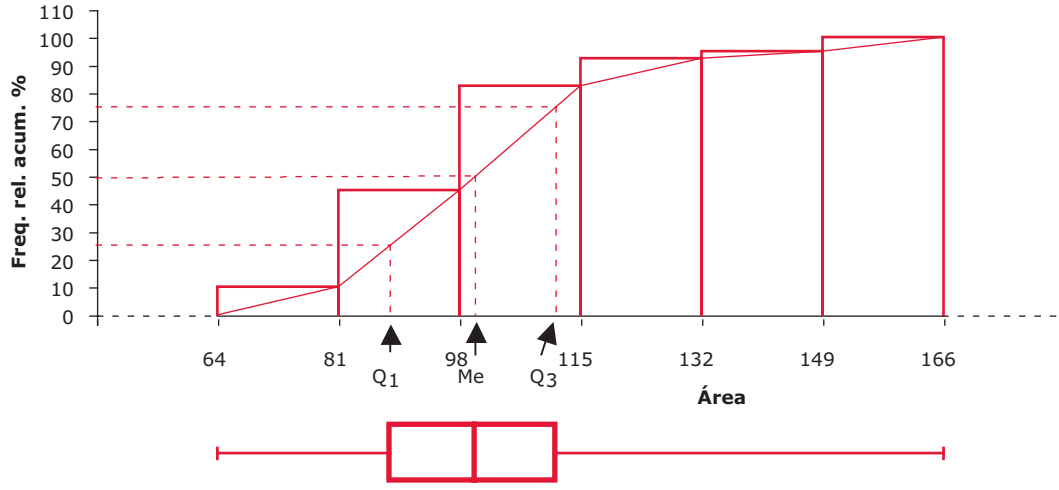


Diagrama de extremos e quartis horizontal

Mesmo sem ter explicitamente os valores da mediana e dos quartis, o histograma acumulado permite, de forma fácil, obter o diagrama de extremos e quartis.

2.5.2 Gráfico de caule-e-folhas

É um tipo de representação que se pode considerar entre a tabela e o gráfico. É com os próprios números que constituem a amostra que se vai construindo a representação gráfica. Cada dado é separado em duas partes: o "caule" e a "folha". Tomando por base a ordem de grandeza dos valores da amostra, escolhe-se o(s) dígito(s) dominante(s) (ver mais à frente) que se coloca(m) ao longo de um eixo vertical, do lado esquerdo. Os dígitos dominantes constituem os caules. Para cada valor da amostra toma-se o dígito que se segue imediatamente ao(s) dígito(s) dominante(s) e coloca-se do lado direito do eixo, em frente ao respectivo caule. Colocam-se assim as folhas. Após colocadas todas as folhas, é usual ordená-las por ordem crescente, dentro de cada caule. Se os dados são constituídos por dois dígitos, então é natural escolher o algarismo das dezenas para caule e o das unidades para folha.

Para ilustrar este procedimento, vamos usar o gráfico de caule-e-folhas como uma forma de organizar os dados resultantes de uma tarefa que facilmente se pode realizar numa turma do 1.º ciclo do ensino básico.

Tarefa

Quantos segundos se consegue estar sem respirar?

Gostaríamos de ter uma ideia de quantos segundos conseguimos estar sem respirar. Suponha que um grupo de alunos fez esta experiência na turma e obteve os seguintes valores: 59, 38, 47, 23, 48, 55, 37, 48, 53, 37, 52, 39, 54, 57, 38, 46, 40, 41, 62, 63, 38, 65, 44, 68, 27, 35, 46, 60.

Podem ser feitas perguntas do tipo:

- Quantos segundos esteve sem respirar o aluno que aguentou menos tempo? E o aluno que aguentou mais tempo?

O professor pode, com a ajuda dos alunos, organizar os dados num diagrama de caule-e-folhas.

Como o menor e o maior dos valores anteriores são, respectivamente, 23 e 68, para organizar os dados num gráfico de caule-e-folhas, vamos começar por considerar os seguintes caules (algarismos das dezenas dos valores iniciais):

2
3
4
5
6

Depois de considerar um segmento de linha vertical, ao lado dos caules, vamos pendurar as folhas, nos caules respectivos. Exemplificamos a seguir, um gráfico com a 1.ª folha, um outro com a 1.ª e a 2.ª folha e finalmente o gráfico com as folhas todas:

2	
3	
4	
5	9
6	

2	
3	8
4	
5	9
6	

2	37
3	8779885
4	78860146
5	953247
6	23580

É costume ordenar as folhas correspondentes a cada caule, de modo que o gráfico final é o seguinte:

2	37
3	5778889
4	01466788
5	234579
6	02358

Repare-se que agora é muito fácil ordenar o conjunto de dados inicial, pois basta percorrer o gráfico de caule-e-folhas:

23, 27, 35, 37, 37, 38, 38, 38, 39, 40, 51, 44, 46, 46, 47, 48, 48, 52, 53, 54, 55, 57, 59, 60, 62, 63, 65 e 68.

Sugestão

Pode ser repetida a tarefa anterior, mas depois de ter aspirado e expirado, profundamente, 3 vezes. Compare os resultados agora obtidos, com os obtidos anteriormente.

Algumas considerações sobre o gráfico de caule-e-folhas

A representação em gráfico de caule-e-folhas tem muitas vantagens:

- É, em geral, muito simples de fazer e torna-se, por isso, acessível, até a alunos do 1.º ciclo do ensino básico (é necessário ter algum cuidado na escolha do exemplo, para que não haja problemas na escolha do(s) dígito(s) dominante(s), mas é esse o único cuidado a ter).
- Dá uma informação visual sobre a forma como os dados estão distribuídos.
- Permite ordenar rapidamente a amostra.
- Facilita o cálculo da mediana e dos quartis.

Escolha dos dígitos dominantes

Na construção de um gráfico de caule-e-folhas nem sempre é imediata a escolha dos dígitos dominantes. Se essa escolha conduzir a muitos caules o resultado final tem pouco de representação gráfica, pois será muito disperso. Se conduzir a poucos caules, para além de poder esconder padrões nos dados, de pouca utilidade se torna na tarefa de ordenação da amostra. Vamos ver o que acontece, por exemplo, com os dados da variável *Preço* do exemplo das casas que temos vindo a tratar. Os preços das casas variam entre 121,47 mil euros e 357,32 mil euros. Se tomarmos como dígito dominante o das centenas, ficaremos apenas com 3 caules. Se tomarmos os dois primeiros dígitos (até à classe das dezenas), ficaremos com 24 caules, o que é demasiado tendo em conta que a dimensão da amostra é n=40. Este problema pode ser resolvido subdividindo em dois cada um dos 3 caules que se obtêm no primeiro caso. No primeiro desses dois caules, identificado com um asterisco (*), colocam-se as folhas de dígitos 0,1,2,3, e 4 e no outro, identificado com um ponto (.), as folhas de dígitos 5,6,7,8, e 9. Deste modo ficamos ao todo com 6 caules que é um número razoável para a dimensão de amostra que temos. Há ainda a possibilidade de subdividir cada caule em 5, um para as folhas 0 e 1, outro para as folhas 2 e 3, e assim por diante até ao último que terá as folhas 8 e 9, mas iríamos obter 15 caules que já é excessivo.

Um gráfico de caule-e-folhas para a variável *Preço* (onde a unidade de cada caule é a centena de milhares de euros) é, então:

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

Note-se que se pendurou como folhas unicamente os algarismos que figuram na classe das dezenas. Neste caso não se consegue recuperar exactamente os valores da amostra, mas apenas uma aproximação. Pode-se observar, por exemplo, que o mínimo da amostra é próximo dos 120 mil euros e que o máximo é próximo dos 350 mil euros.

Utilização do caule-e-folhas para comparar duas amostras

A representação em caule-e-folhas é muito sugestiva para comparar duas amostras, como se apresenta no exemplo seguinte:

Exemplo:

O tempo de sono do Pedro e do David – Apresentam-se, a seguir, os tempos de sono, em horas, medidos durante 30 noites seguidas, do Pedro e do David.

Pedro			David		
8,7	9,3	8,7	7,1	9,5	7,1
9,4	5,3	7,4	8,3	7,1	7,4
6,6	7,3	6,3	7,1	7,5	7,4
6,0	6,7	5,9	7,9	7,9	7,8
6,9	5,8	10,0	7,5	6,4	6,2
9,9	4,7	6,5	6,2	6,2	8,6
6,3	5,6	8,6	8,2	7,5	8,4
8,9	5,9	7,7	8,7	7,7	6,6
10,1	9,4	9,0	8,5	7,6	8,1
9,6	7,6	7,9	7,6	8,8	7,1

Para comparar os tempos de sono dos dois jovens, vamos representar os caule-e--folhas paralelos, isto é, determinamos os caules (comuns) a partir da amostra de maior amplitude, ou seja, neste caso, dos dados correspondentes ao David, e depois colocamos as folhas correspondentes às observações do Pedro para um lado e as correspondentes às do David para o outro:

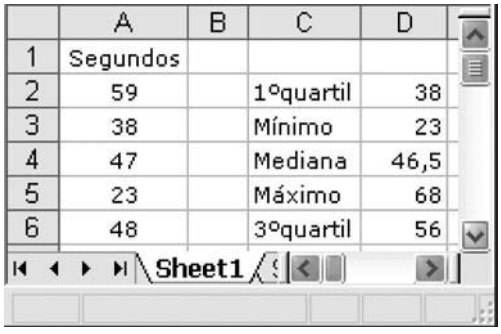
Pedro					David								
			7	4.									
			3	5*									
9	9	8	6	5.									
	3	3	0	6*	2	2	2	4					
9	7	6	5	6.	6								
		4	3	7*	1	1	1	1	1	4	4		
	9	7	6	7.	5	5	5	6	6	7	8	9	9
				8*	1	2	3	4					
9	7	7	6	8.	5	6	7	8					
4	4	3	0	9*									
		9	6	9.	5								
		1	0	10*									

A representação anterior permite realçar a maior dispersão do sono do Pedro, enquanto que o David é mais regular, com uma duração de sono de um modo geral entre as 7 e as 8 horas.

Utilização do *Excel*, na construção do diagrama de extremos e quartis e do caule-e-folhas

Construção do diagrama de extremos e quartis

Utilizando o *Excel*, começam por se calcular os 5 valores necessários para a construção do diagrama de extremos e quartis, que se apresentam da seguinte forma, e pela ordem indicada:



	A	B	C	D
1	Segundos			
2	59		1ºquartil	38
3	38		Mínimo	23
4	47		Mediana	46,5
5	23		Máximo	68
6	48		3ºquartil	56

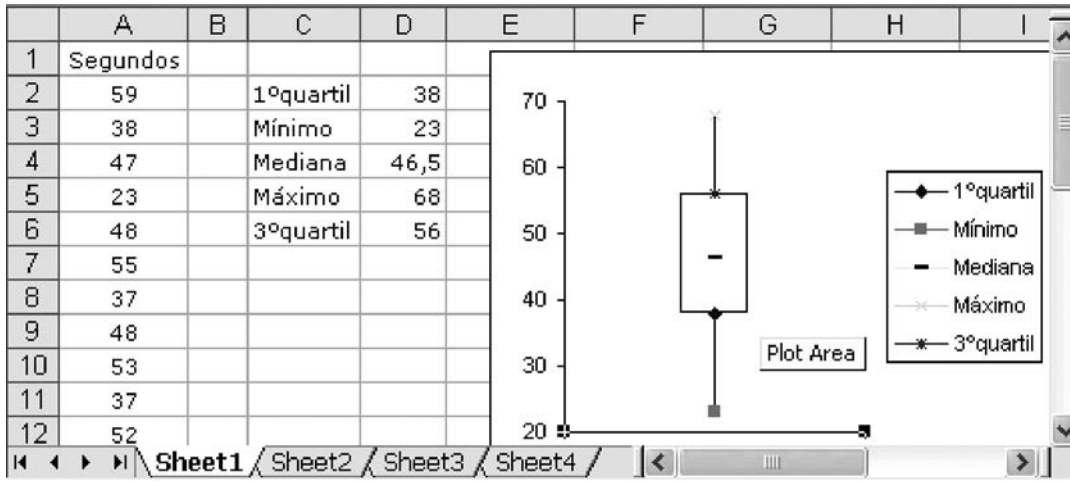
Sheet1

- Seleccionar as células que contêm as estatísticas, assim como as suas etiquetas: C2 a D6;
- No módulo Chart Wizard seleccionar:

  Line

  Seleccionar *Line with markers displayed at each data value*
  Clicar *Next*
  Seleccionar *Series in Rows*
  Clicar *Finish*
- Clicar com o botão direito do rato num dos pontos. Seleccionar:

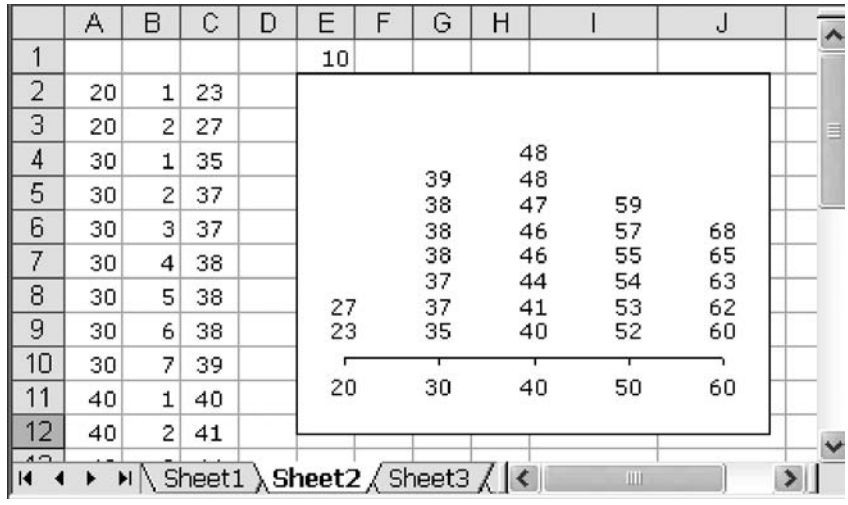
  Format Data Series
  Seleccionar *Options*
  Escolher *High-low lines* e *Up-down bars*;
  Ajustar à sua escolha *Gap width*;
  OK
- Arranjar "esteticamente" o gráfico:



Construção do caule-e-folhas

Não existe no *Excel* uma representação imediata para a construção de um caule-e--folhas, pelo que vamos utilizar um processo desenvolvido por Neville Hunt (Hunt, 2006), para o Excel:

- 1.º passo – Insira os dados na coluna C, começando na célula C2; se não estiverem ordenados, ordene-os por ordem crescente;
- 2.º passo – Insira na célula E1 o valor que deseja para o comprimento de linha: 10, 5 ou 2 ou uma potência de 10, destes valores;
- 3.º passo – Na célula A2 escreva a seguinte fórmula = *INT*(C2/E$1)*E$1 e replique-a tantas vezes quantos os dados inseridos no 1.º passo, na coluna C;
- 4.º passo – Na célula B2 escreva o valor 1. Na célula B3 escreva a fórmula = *IF* (A3=A2; B2+1; 1) e replique a fórmula, tantas vezes quantos os dados inseridos no 1.º passo, na coluna C;
- 5.º passo – Seleccione as células das colunas A, B e C com os resultados obtidos nos passos anteriores e no módulo *Chart Wizard* (Assistente de Gráficos) escolha *Bubble*;
- 6.º passo – Faça um duplo clique numa das bolas representadas e na janela *Format data Series* (ou clique com o botão direito do rato e seleccione *Format data Series*) seleccione *Patterns*:
  Border: None
  Area: None
  Data Labels: Show bubbles sizes
  OK;
- 7.º passo – Faça um duplo clique numa das "*Data labels*" (ou clique com o botão direito do rato e seleccione *Format Data Labels*), e na janela Format Data Labels, em Alignment:
  Label Position: Centre
  OK;
- 8.º passo – Clique numa das linhas horizontais que atravessam o gráfico e apague-as com a tecla Delete. Faça o mesmo ao fundo cinzento, seleccionando-o e carregando na tecla Delete. Apague também a legenda.
- 9.º passo – Formate convenientemente os eixos.



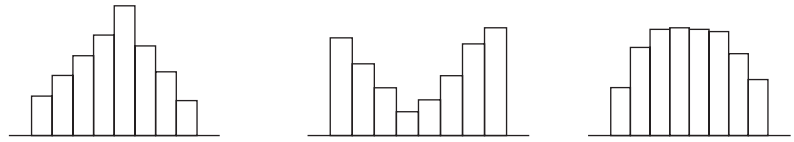
Como se verifica, a "mancha" obtida é idêntica à representação anteriormente feita à mão, mas aqui não existe o mesmo conceito para o caule e a folha.

2.6 Algumas formas básicas de distribuição de dados

Numa fase mais avançada da análise dos dados, o histograma pode ser utilizado como uma ajuda na escolha de um modelo teórico para a distribuição subjacente à população de onde os dados foram retirados.

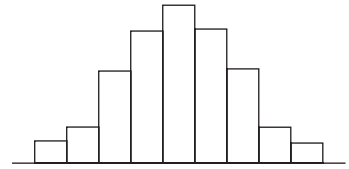
Alguns histogramas apresentam formas que, pela frequência com que surgem, merecem referência especial. Assim, as distribuições mais comuns, apresentadas pelos dados, são:

Distribuições simétricas – A distribuição das frequências faz-se de forma aproximadamente simétrica, relativamente a uma classe média:

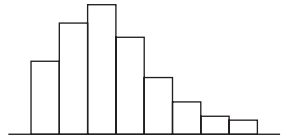


Caso especial de uma distribuição simétrica

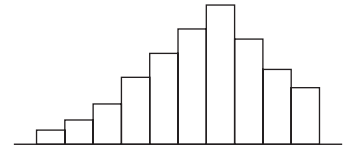
Um caso especial de uma distribuição simétrica é aquele que sugere a forma de um "sino" e que é apresentada por amostras provenientes de Populações *Normais*:



Distribuições enviesadas – A distribuição das frequências faz-se de forma acentuadamente assimétrica, apresentando valores substancialmente mais pequenos num dos lados, relativamente ao outro:

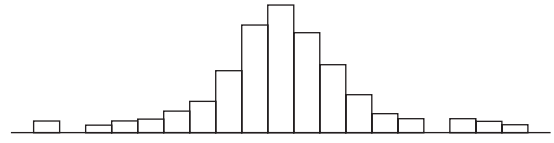


Enviesada para a direita

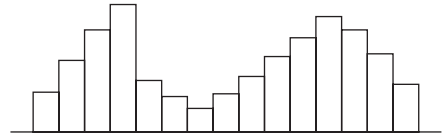


Enviesada para a esquerda

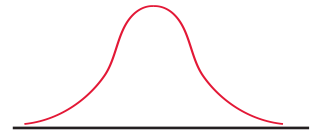
Distribuições com caudas longas - A distribuição das frequências faz-se de tal forma que existe um grande número de classes nos extremos, cujas frequências são pequenas, relativamente às classes centrais:



Distribuições com vários "picos" ou modas – A distribuição das frequências apresenta 2 ou mais "picos" a que chamamos modas, sugerindo que os dados são provenientes de vários grupos distintos:



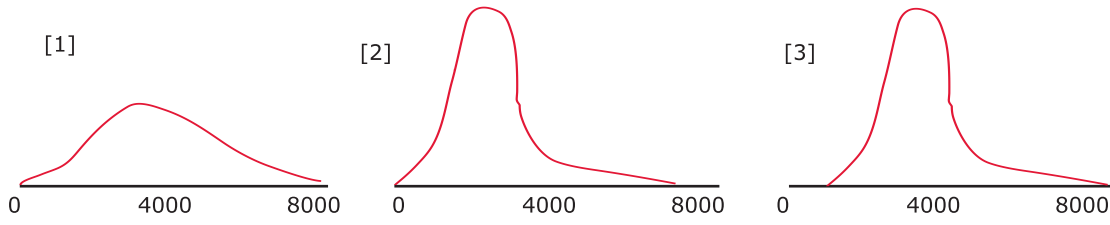
No caso das variáveis contínuas, os modelos teóricos são caracterizados pelas chamadas curvas de densidade. Estas são funções não negativas, que têm a particularidade de terem uma área unitária entre o eixo dos xx e o gráfico que as representa. Por exemplo, o seguinte gráfico



pode ser considerado a função densidade do modelo Normal, e a sua aplicação pode ser sugerida por um histograma como o que apresentámos anteriormente, como caso especial de uma distribuição simétrica. A seguir apresentamos alguns exemplos em que são apresentados diversos esquemas de histogramas estilizados, que procuram traduzir a distribuição subjacente a várias variáveis quantitativas contínuas.

Exemplo:

Salários de trabalhadores (Adaptado de Freedman *et al.,* 1991) – Recolheram-se os preços dos salários mensais de 3 tipos de trabalhadores. Os trabalhadores do grupo B ganham cerca de duas vezes mais do que os trabalhadores do grupo A; os trabalhadores do grupo C ganham mais 1500 euros por mês do que os do grupo A. Qual das "manchas" seguintes, de histogramas, pertence a cada um dos grupos?



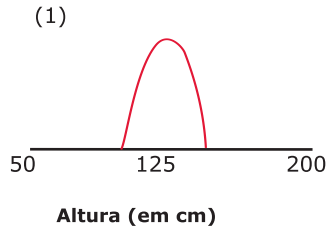
Para tentarmos resolver esta questão, podemos pensar que se se diz que os trabalhadores do grupo B ganham o dobro dos trabalhadores do grupo A, isto significa, por exemplo, que enquanto a maior parte dos trabalhadores do grupo B aufere um salário à volta de 4000 euros, os do grupo A auferem um salário à volta de 2000 euros. Então é natural esperar que a figura (1) corresponda aos salários dos trabalhadores do grupo B, enquanto a figura (2) corresponde aos trabalhadores do grupo A. Por outro lado, se os trabalhadores do grupo C ganham 1500 euros a mais do que os do grupo A, isto significa que a distribuição dos salários dos trabalhadores do grupo C terá um aspecto idêntico ao dos trabalhadores do grupo A, mas deslocada para a direita de 1500 euros. Então a figura (3) corresponderá aos salários dos trabalhadores do grupo C.

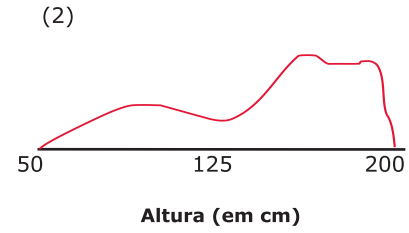
A distribuição com o aspecto (1) não é muito usual para representar salários, sendo mais usuais as distribuições com o aspecto (2) ou (3). Efectivamente, em geral, a distribuição dos salários tem um aspecto assimétrico, com um enviesamento para a direita. Isto deve-se ao facto de a maior parte dos salários se concentrarem numa determinada região, havendo alguns (poucos) salários que são substancialmente maiores que os restantes, provocando uma cauda da distribuição, alongada para a direita.

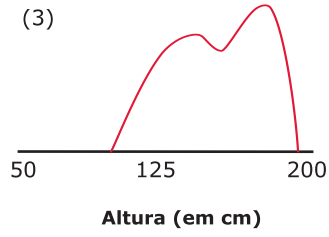
Exemplo:

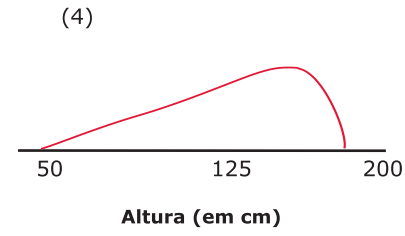
Qual o aspecto da distribuição? (Adaptado de Freedman *et al.,* 1991) – Seguidamente apresentam-se 6 "manchas" de histogramas, 4 dos quais apresentam os resultados do estudo, numa pequena cidade, das 4 características seguintes:

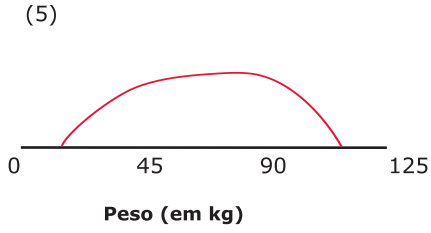
- a) Alturas de todos os elementos das famílias, em que os pais tenham idade inferior a 24 anos.
- b) Alturas dos casais (marido e mulher).
- c) Alturas de todos os indivíduos da cidade.
- d) Alturas de todos os automóveis.

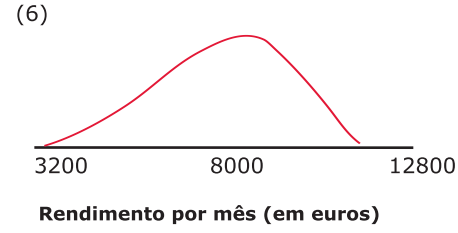








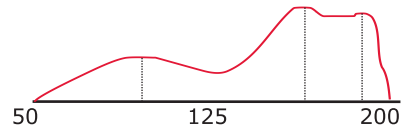




Quais dos histogramas podem representar cada uma das variáveis anteriores?

Pensando na variável que representa a altura de um elemento, escolhido ao acaso, de uma família, em que os pais tenham idade inferior a 24 anos, esperamos obter um histograma com uma mancha idêntica à (2), onde se vislumbram 3 pontos, à volta dos quais se nota uma maior frequência, e que corresponderão à altura dos filhos – entre 80 e 90 cm, que para casais com idades inferiores a 24 anos, ainda devem ser muito pequenos, e à altura dos membros do casal, da mulher ou do marido, respectivamente à volta de 165 cm e 190 cm, aproximadamente:

(2)



Altura (em cm)

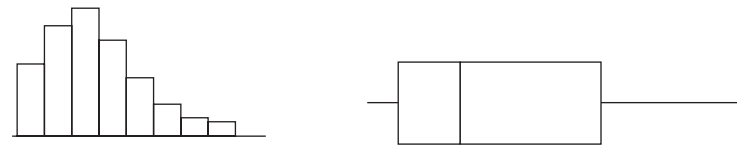
Quando consideramos a distribuição das alturas dos elementos de um casal, é natural esperar um esquema idêntico ao da figura (3), com duas modas, reflectindo que, de um modo geral, as alturas dos homens se concentram em torno de um valor um pouco superior ao valor em torno do qual se concentram as alturas das mulheres.

Ao escolher um indivíduo ao acaso, na cidade, esperamos que a distribuição das alturas seja descrita pela figura (4) que apresenta um enviesamento para a esquerda, correspondente às alturas das crianças, que estão em minoria.

Finalmente, quando se pretende estudar a variável que representa a altura de um carro, o histograma adequado é o que corresponde à mancha 1) que traduz o facto de os carros terem quase todos a mesma altura, andando à volta de 125 cm.

A informação transmitida pelo histograma, sobre o padrão da distribuição da população subjacente aos dados, também é igualmente transmitida pelo gráfico de caule-e-folhas e pelo diagrama de extremos e quartis. Por exemplo, as seguintes representações indicam o mesmo tipo de informação, sugerindo que a distribuição da população tem um enviesamento para a direita:

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



Quando se faz a redução dos dados, perde-se sempre alguma informação contida nesses dados, mas em contrapartida obtemos a estrutura da população que eles pretendem representar. Das representações gráficas anteriores, aquela em que se perdeu mais informação foi o diagrama de extremos e quartis, mas também foi a mais simples de ser construída – bastou recolher, a partir dos dados, informação sobre 5 números (mínimo, máximo, 1.º quartil, 3.º quartil e mediana).

2.7 Representações gráficas e tabelas de frequências para dados bivariados

Retomemos os Dados sobre casas, apresentados no Capítulo 1. Do nosso conhecimento do dia a dia, sabemos que, entre outras variáveis, a área de uma casa influencia directamente o seu preço de venda. Diz-se por isso que as variáveis *Área* e *Preço* estão correlacionadas. De igual modo estão correlacionadas as variáveis *Altura* e *Peso* em muitos seres vivos; a *Oferta/Procura* e o *Preço* de produtos, a *Cilindrada* e o *Consumo* de combustível nos carros, só para mencionar alguns exemplos. Nalguns casos o aumento de valor de uma das variáveis acarreta o aumento de valor na outra variável (correlação positiva) e noutros acarreta uma diminuição de valor na segunda variável (correlação negativa). À excepção do exemplo ligado à lei da oferta e da procura, em todos os outros é possível identificar uma das variáveis como sendo explicativa e a outra como sendo uma variável resposta. Por outras palavras, uma das variáveis é independente (ou explicativa) e a outra é dependente (ou resposta). Assim, o Preço da casa é dependente da Área da casa; o Peso é que depende da Altura e não a Altura que depende do Peso; o Consumo de combustível é directamente influenciado pela Cilindrada e não vice-versa. Em estatística, quando o objectivo do estudo é analisar a relação de dependência entre duas variáveis, o registo das observações tem de preservar o emparelhamento, obtendo-se assim uma amostra de **dados bivariados**.

2.7.1 Diagrama de dispersão

O diagrama de dispersão é uma representação gráfica de dados bivariados, utilizada quando qualquer das duas variáveis em estudo é de tipo quantitativo contínuo. Cada par de dados (x,y) é representado, num sistema de eixos ortogonais, por um ponto de coordenadas (x,y). Obtém-se assim uma nuvem de pontos que nos permite avaliar de imediato se há ou não uma forte relação entre as duas variáveis.

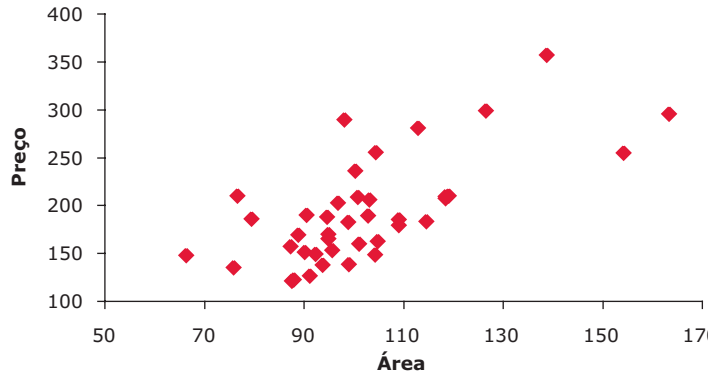


Diagrama de dispersão para os pares (*Área*, *Preço*)

No diagrama de dispersão anterior, estão representados os pares (*Área*, *Preço*) das 40 casas que constituem a nossa amostra. A nuvem de pontos apresenta-se um pouco dispersa, mas não deixa por isso de ser bem patente a sua forma alongada que se desenvolve em torno de uma recta com um declive positivo.

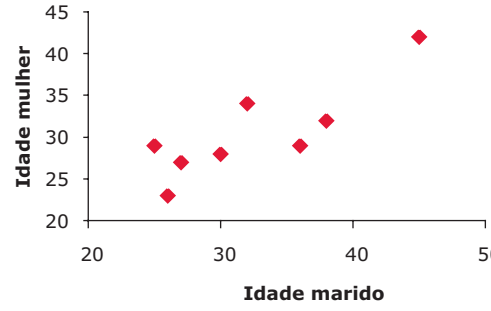
Diagrama de dispersão – É uma representação gráfica para os dados bivariados quantitativos, em que cada par de dados (x,y) é representado por um ponto de coordenadas (x,y), num sistema de eixos coordenados.

Este tipo de representação é muito útil, pois permite realçar algumas propriedades entre os dados, nomeadamente no que diz respeito ao tipo de associação entre as variáveis representadas por x e y. Quanto mais alongada for a nuvem de pontos ao longo de uma recta, isto é, quanto maior for o grau de proximidade dos pontos a uma linha recta, maior será o grau de associação entre as variáveis. Esta associação pode ser medida numericamente com um coeficiente a que se dá o nome de coeficiente de correlação, que será estudado no capítulo seguinte. No diagrama de dispersão para os pares (*Área, Preço*) verifica-se uma **tendência** para que casas de maior área tenham preços mais elevados. O facto de existir esta tendência não significa que se tenha necessariamente uma casa mais cara, quando tem maior área, mas, de um modo geral, as casas maiores tendem a ser mais caras.

Exemplo:

Idades do marido e da mulher – Considere os seguintes dados que representam as idades de 8 casais:

Casal	Marido	Mulher
1	26	23
2	25	29
3	45	42
4	27	27
5	38	32
6	30	28
7	32	34
8	36	29



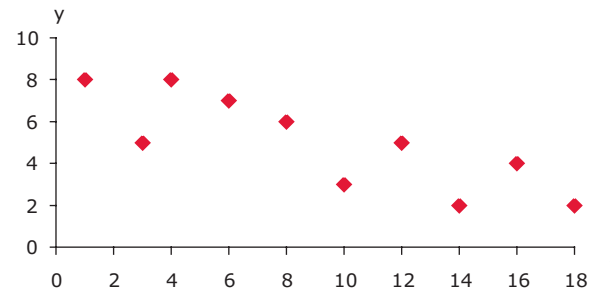
Verifica-se uma associação linear positiva entre a idade do marido e a idade da mulher, isto é, existe tendência a que mulheres mais velhas estejam casadas com homens mais velhos.

Exemplo:

Número de faltas – Considere os seguintes dados, que representam o número de faltas não autorizadas por ano e a distância (em km) a que os empregados de determinado armazém estão de casa.

Construa o diagrama de dispersão e comente-o.

Distância x	N.º faltas y
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
18	4
18	2



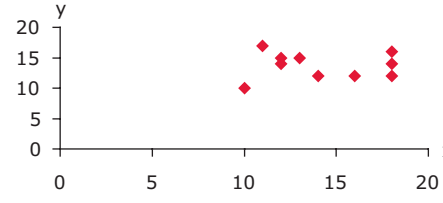
O gráfico mostra uma associação, de sentido contrário, entre o número de faltas e a distância. Assim, quanto maior é a distância de casa, menor é a tendência para faltar!

Exemplo:

Notas a Matemática e Educação Física – Considere os seguintes dados, que representam as notas obtidas por 10 alunos nas disciplinas de Matemática e Educação Física.

Construa o diagrama de dispersão e comente-o.

Matemática x	Ed. Física y
12	14
13	12
10	10
11	17
18	16
16	12
12	15
14	12
18	14
18	12



Aparentemente não existe nenhuma associação linear entre as notas obtidas nas duas disciplinas, uma vez que os pontos se encontram dispersos de forma "aleatória".

2.7.2 Tabelas de frequências para dados bivariados

Um outro processo de organizar a informação correspondente a dados bivariados, normalmente de tipo qualitativo, é utilizando uma tabela de frequências, a que damos o nome de **tabela de contingência**.

De uma maneira geral, uma tabela de contingência é uma representação dos dados, quer de tipo qualitativo, quer de tipo quantitativo, especialmente quando são de tipo bivariado, isto é, podem ser classificados segundo dois critérios. O aspecto de uma tabela de contingência é o de uma tabela com linhas, correspondentes a um dos critérios, e colunas correspondente ao outro critério. Seguidamente apresentamos um exemplo, para ilustrar o que acabámos de dizer.

Exemplo:

As casas – Considerando de novo o exemplo das casas, pretende-se organizar as variáveis *Zona* e *Estado* na forma de uma tabela de contingência. Para isso deve-se começar por construir uma tabela idêntica à que se segue:

Zona / Estado	A	B	C
Usada	10	16	5
Nova	9	0	0

que depois será preenchida com as frequências absolutas correspondentes a cada uma das células. Assim, na célula que corresponde às casas usadas da zona A, escrevemos 10, pois encontraram-se 10 casas nessas condições. As outras células são preenchidas de forma idêntica. Uma tabela destas ainda pode ser completada com mais uma linha e uma coluna, onde se colocam os totais de linhas e de colunas:

Zona / Estado	A	B	C	Total
Usada	10	16	5	31
Nova	9	0	0	9
Total	19	16	5	40

A leitura da tabela permite concluir que 31 das casas são usadas e 9 são novas. Também se pode concluir que 19 casas pertencem à zona A, 16 à zona B e 5 à zona C. A célula do canto inferior direito apresenta o número total de unidades observadas, que neste caso foram as casas.

Em vez das frequências absolutas, também se podem utilizar as frequências relativas, com um tipo variado de informação possível. Por exemplo, a tabela

Zona / Estado	A	B	C	Total
Usada	32%	52%	16%	100%
Nova	100%	0%	0%	100%

permite obter informação diferente da tabela que se apresenta a seguir:

Estado \ Zona	A	B	C	Total
Usada	25%	40%	13%	78%
Nova	23%	0%	0%	23%
Total	48%	40%	13%	100%

Da primeira das duas tabelas anteriores pode-se concluir, por exemplo, que das casas usadas, 32% pertencem à zona A, 52% à zona B e 16% à zona C. Repare-se que nessa tabela se calcularam, em separado, as percentagens relativamente ao número de casas usadas e relativamente ao número de casas novas.

Por outro lado, da segunda tabela pode-se concluir, por exemplo, que 25% das casas são usadas e pertencem à zona A; 23% das casas são novas e pertencem à zona A; etc. Nesta tabela, as percentagens foram calculadas relativamente ao número total de casas.

2.8 Um gráfico vale mais do que mil palavras?

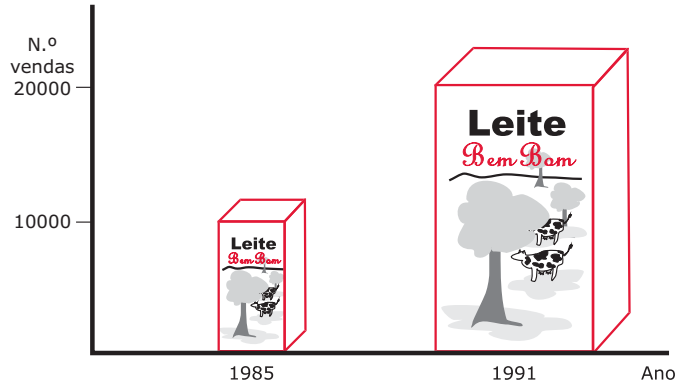
É costume dizer-se que um gráfico vale mais do que mil palavras. De facto, para que isso aconteça, é necessário tomar alguns cuidados na construção dessas representações gráficas. Damos de seguida alguns exemplos de representações gráficas incorrectas.

2.8.1 Utilização de pictogramas

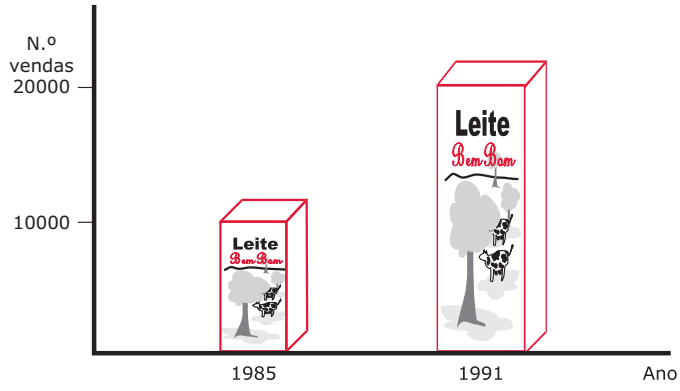
Os pictogramas são representações gráficas que utilizam figuras, o que faz com que essas representações se tornem bastante apelativas. No entanto, a utilização de pictogramas, nem sempre se faz de forma correcta.

Exemplo:

Aumento da quantidade de leite – Consideremos a seguinte representação, que pretende mostrar que a quantidade de leite, de uma determinada marca, vendida desde 1985 até 1991, duplicou:



Efectivamente a altura do pacote de leite, em 1991, é o dobro da de 1985, mas quando olhamos para as figuras, ficamos com a impressão que esse aumento foi muito superior ao verificado, induzindo o leitor em erro. Se pretendermos continuar a utilizar o pacote de leite como referência, então uma solução possível será a seguinte, em que os pacotes só diferem na altura. Deste modo, o volume da cada um é proporcional à frequência absoluta, sendo, neste caso, o volume do pacote referente a 1991, o dobro do referente a 1985:



Exemplo:

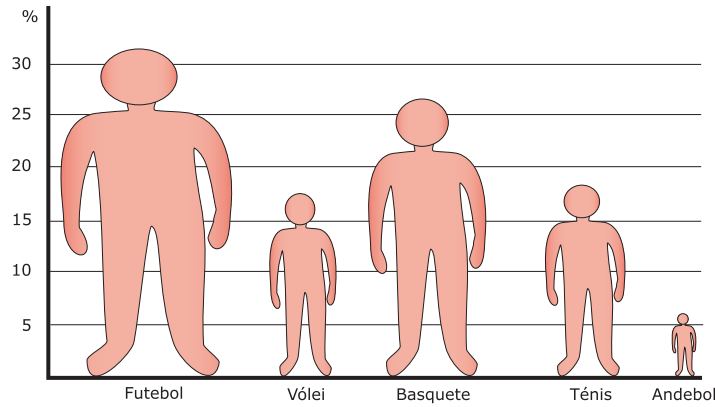
Os jogos preferidos pelos alunos (adaptado de Graça Martins et al. 1999) - Numa escola o Conselho Directivo pretende construir um campo de jogos, pelo que gostaria de ter uma ideia de quais os jogos preferidos dos alunos. Encarregou um aluno de recolher a informação necessária, o qual utilizou a seguinte metodologia: elaborou uma lista de jogos possíveis e percorreu todas as turmas da escola, em número de 20, perguntando dentro de cada turma qual a opinião dos alunos cujo número fosse um múltiplo de 5. Em três turmas foram seleccionados 6 alunos e nas restantes 5. O resultado da recolha da informação tinha o seguinte aspecto

Futebol	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
Vólei	x x x x x x x x x x x x x x x x x x
Basquete	x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
Ténis	x x x x x x x x x x x x x x x x x x x
Andebol	x x x x x x x

O aluno, ao perguntar a cada elemento da amostra a sua opinião, apontava o resultado com um x à frente da modalidade seleccionada. A forma como a informação foi recolhida permite imediatamente concluir que a modalidade preferida foi o futebol. A fim de transmitir verbalmente a informação ao Conselho Directivo, o aluno construiu a seguinte tabela de frequências

Classes	*Freq. abs.*	*Freq. rel(%)*
Futebol	32	31,1
Vólei	18	17,5
Basquete	27	26,2
Ténis	19	18,4
Andebol	7	6,8
Total	103	100,0

Então o Conselho Directivo foi informado que as preferências dos alunos vão para o futebol seguindo-se o basquete. Depende agora das disponibilidades financeiras contemplar as diferentes modalidades, tendo em conta as preferências dos alunos. Procurando transmitir a informação graficamente, construíram o seguinte pictograma, onde se substituiu a barra por uma figura humana:



Pictograma para a variável *Jogo preferido*

Na figura anterior a imagem correspondente à classe futebol é substancialmente maior que a que é utilizada para as outras modalidades ou classes. Daí dar uma ideia, errada, de que por exemplo a percentagem de alunos que preferem o futebol é várias vezes superior aos que preferem vólei, quando nem sequer chega a ser o dobro. Este problema foi ocasionado pelo facto de se pretender que a figura humana ficasse proporcional, pelo que à medida que se aumentou a altura, também se aumentou a largura. O gráfico de barras correspondente tem o seguinte aspecto:

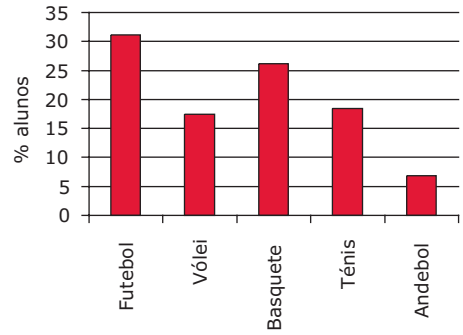
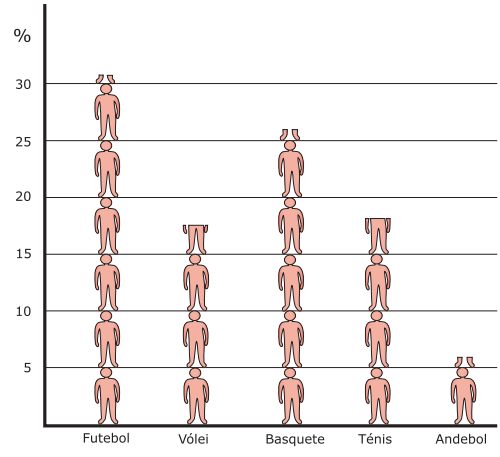


Gráfico de barras para a variável *Jogo preferido*

Na construção do gráfico de barras, como já dissemos nas indicações para a sua construção, deve ter-se em atenção que as barras devem ter a mesma largura, pois a mensagem que devem transmitir é a que está contida nas diferentes alturas das barras. Se umas barras forem mais largas do que outras, temos tendência a crer que as classes a que correspondem as barras mais largas têm maior frequência do que a que efectivamente têm. Este é um problema que não é tido em conta na construção de muitos *pictogramas*, em que as barras são substituídas por figuras, para tornar a representação gráfica mais atraente, como aconteceu no caso deste exemplo. Um pictograma possível, é o que se apresenta a seguir, em que a figura utilizada é uma figura humana, que corresponde a uma percentagem de 5%, que se replica o número de vezes que for necessário, sendo possível utilizar uma fracção da figura:



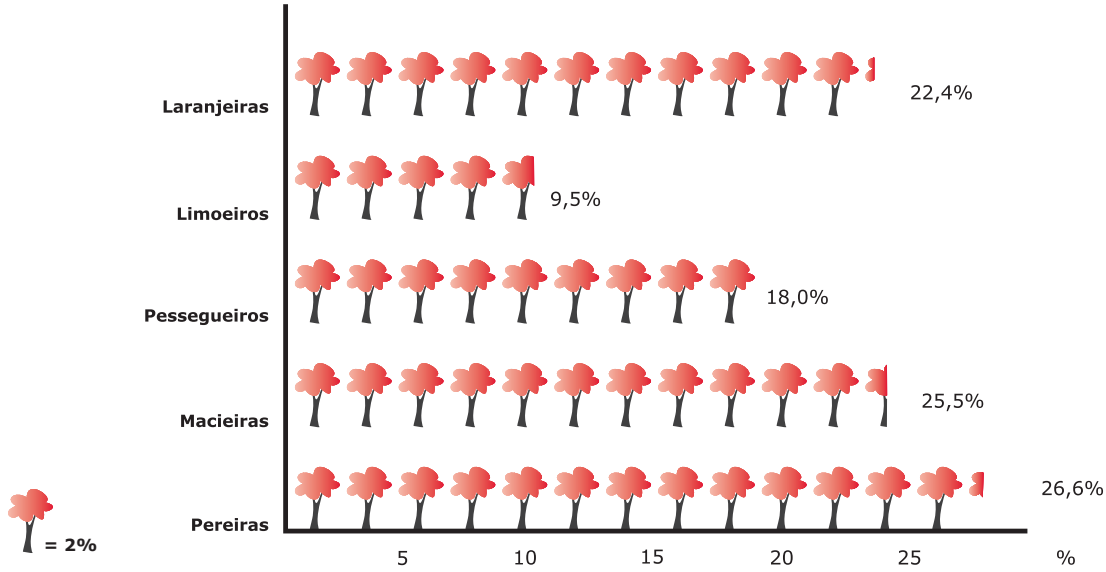
Pictograma para a variável *Jogo preferido*

Exemplo:

Seguro do agricultor (Graça Martins *et al.,* 1999) **-** Com o objectivo de fazer um seguro, um agricultor teve de fazer o levantamento do número e tipo de árvores de fruto existentes no seu pomar. O resultado apresenta-se na tabela seguinte:

Classes	*Freq. abs.*	*Freq. rel.(%)*
Laranjeiras	320	22,4
Limoeiros	135	9,5
Pessegueiros	257	18,0
Macieiras	335	23,5
Pereiras	379	26,6
Total	1426	100,0

Uma representação gráfica possível seria a seguinte, considerando uma figura sugestiva, mas sem incorrer no erro da representação do exemplo anterior, inicialmente apresentada:



Pictograma para a variável *Tipo de árvore*

Embora seja comum dizer que uma imagem vale mais do que mil palavras, não podemos deixar de chamar a atenção para que esta frase tem sentido se a informação transmitida pela imagem for correcta, o que nem sempre acontece, como vimos anteriormente.

2.8.2 Utilização do diagrama circular

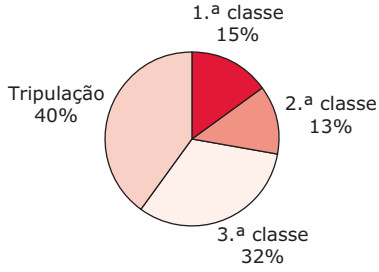
O diagrama circular é uma representação gráfica, utilizada para representar a distribuição de dados de tipo qualitativo. É das representações gráficas mais utilizadas pela comunicação social, em jornais, revistas ou televisão. No entanto, a sua utilização nem sempre se faz da forma mais correcta, nomeadamente quando se faz o diagrama circular a 3 dimensões, pois, neste caso, não transmite uma ideia clara das áreas que pretende representar, embora se tornem visualmente mais atractivas.

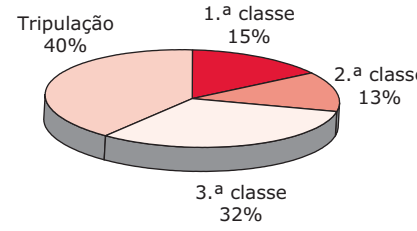
Exemplo:

Passageiros do Titanic (De Veaux *et al.,* 2004) – Considere a seguinte tabela com a distribuição dos 2201 passageiros do Titanic, na altura do naufrágio:

Classe	Freq. abs.	Freq. rel.(%)
1.ª classe	325	15%
2.ª classe	285	13%
3.ª classe	706	32%
Tripulação	885	40%

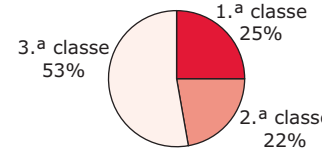
Para estes dados podemos construir algumas representações gráficas na forma de um diagrama circular, nomeadamente:





A representação a 3 dimensões torna difícil a comparação das frequências das diferentes classes, que é, afinal, o objectivo principal de uma construção destas. Esta situação verifica-se, sobretudo se não juntarmos as etiquetas com as percentagens respectivas, junto de cada sector. Uma regra básica é a de que as áreas ou volumes ocupadas pelas diferentes classes, devem reflectir, sem ambiguidade, o valor que representam, o que não é o caso da representação do lado direito.

Suponhamos, agora, que só desejávamos representar os passageiros que não faziam parte da tripulação:
Neste caso a representação correcta é a que se apresenta ao lado. As percentagens são diferentes das consideradas anteriormente, uma vez que passámos a representar um outro conjunto de dados. Uma outra regra básica é a de que, num diagrama circular, *a soma das percentagens tem que ser igual a 100%*, ou a soma dos efectivos tem que ser igual ao número de dados.

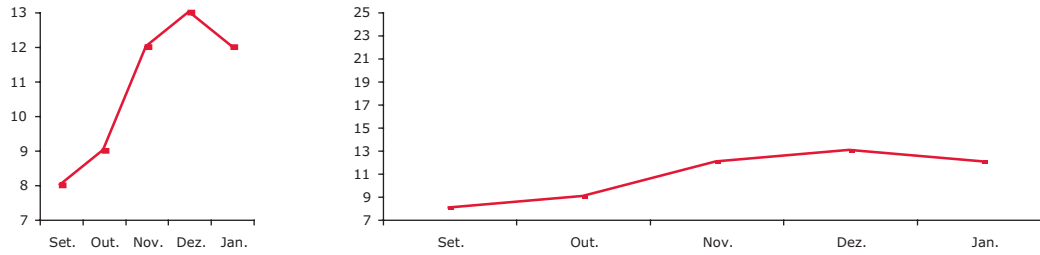


2.8.3 Escalas e escalas

A utilização e manipulação das escalas pode transmitir informação de acordo com a vontade do utilizador, o que se torna perigoso. Vejamos os três exemplos seguintes:

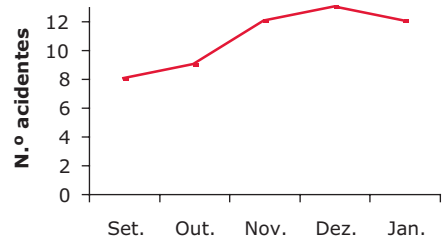
Exemplo:

Número de acidentes no IP5 (Hipotético) - Suponha que o número de acidentes no IP5 foi, no período de Setembro de 1997 a Janeiro de 1998, o seguinte: 8, 9, 12, 13 e 12. Dois jornais apresentaram as seguintes representações gráficas para transmitirem a informação anterior:



Número de acidentes no IP5

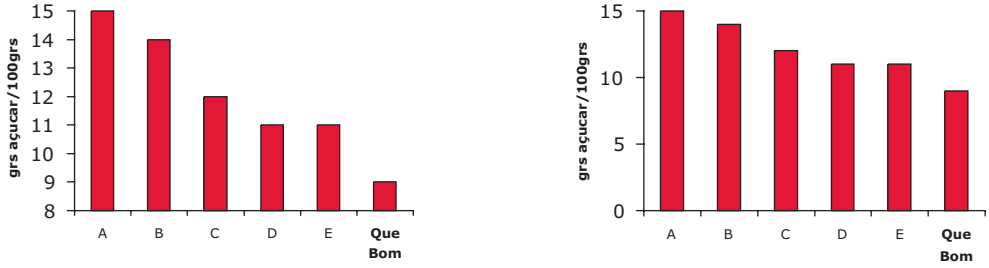
Repare que a representação gráfica da esquerda procura enfatizar o facto do número de acidentes ter aumentado substancialmente, enquanto que o do lado direito procura desvalorizar esse aumento. No primeiro caso não iniciámos a escala das frequências no ponto 0, enquanto que no 2.º caso diminuímos a distância entre os incrementos do eixo vertical, para diluir a variação da curva, ao mesmo tempo que aumentamos a distância entre as categorias no eixo horizontal. Uma representação correcta pode ser a seguinte:



Número de acidentes no IP5

Exemplo:

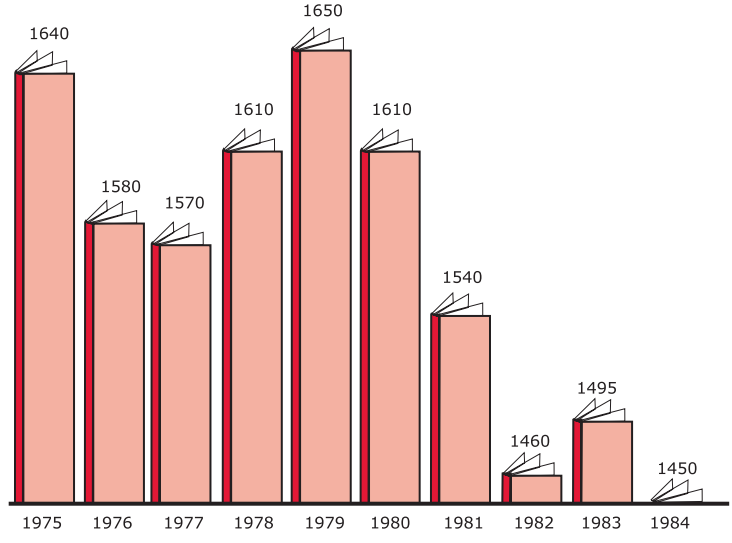
Quantidade de açúcar nos cereais para crianças - Uma empresa que vende cereais para crianças faz publicidade dos seus cereais da marca "Que Bom", dizendo que os seus cereais têm muito menos açúcar, por 100 gramas de cereal, do que os da concorrência. Para isso apresenta a representação gráfica do lado esquerdo da figura seguinte, onde compara os 9 gramas de açúcar dos cereais "Que Bom", com os 15, 14, 12, 11 e 11 gramas, respectivamente dos cereais A, B, C, D e E:



Nessa representação gráfica, a escala do eixo vertical não se inicia no ponto 0, como devia. Assim, uma representação correcta poderá ser a do lado direito da figura anterior, pois ao contrário dessa, já não induz o leitor em erro.

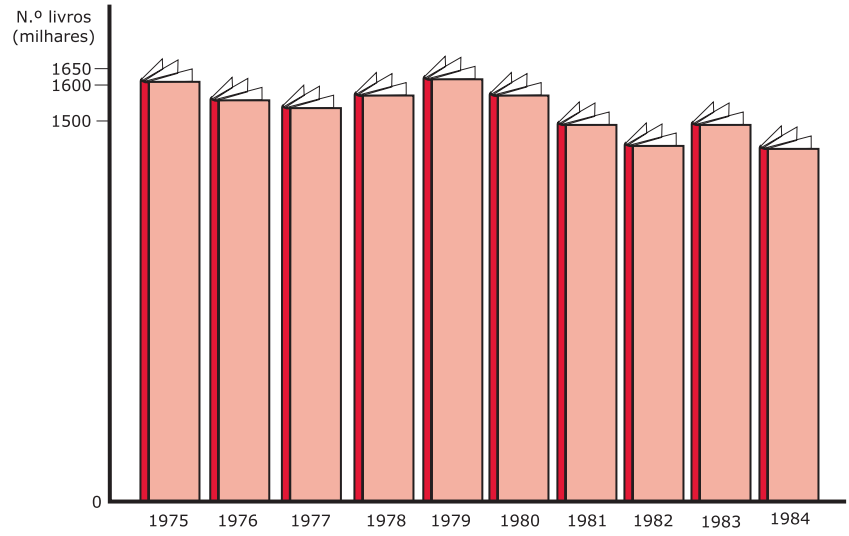
Exemplo:

Andamos a ler pouco – O gráfico a seguir apresentado pretende mostrar a diminuição na venda de livros de 1975 a 1984, num determinado país. Embora esteja indicado para cada ano o número de livros, em milhares, as alturas das barras transmitem-nos a ideia de que houve uma diminuição acentuada, sobretudo a partir de 1979:



Número de livros vendidos

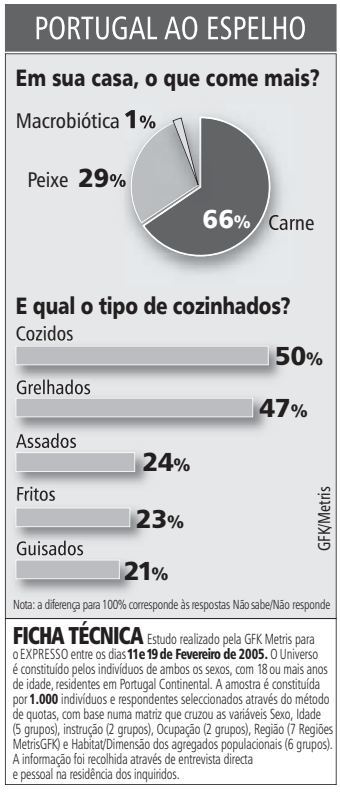
Temos, no entanto, de ter em atenção que o eixo horizontal não representa o zero. Temos uma escala que faz sobressair as pequenas diferenças de ano para ano. Com uma representação numa escala que inclua o zero obter-se-á o seguinte gráfico



Como se verifica, a variação não é tão grande, como o primeiro gráfico fazia supor.

2.8.4 Outras situações - Exemplo de um gráfico pouco elucidativo

O jornal Expresso do dia 9 de Abril de 2005 apresentava um artigo sobre a alimentação dos portugueses. Entre outras representações gráficas, apresentava a seguinte:



No que diz respeito ao diagrama circular, em que se apresenta o resultado da pergunta "Em sua casa, o que come mais?", ficamos a saber que praticamente 2/3 da população (e estamos a inferir para a população, os resultados verificados na amostra) come mais carne do que peixe, embora os especialistas não se cansem de referir os malefícios de comer carne a mais, nomeadamente para o colesterol.

Quanto à representação gráfica (?) que procura traduzir os resultados da questão "E qual o tipo de cozinhados?", como é que deve ser interpretada? O que é que significa a percentagem de 50% de respostas em "Cozidos"? E as percentagens nas outras modalidades de cozinhados?

A quantas modalidades é que as pessoas puderam responder? Evidentemente que não puderam responder só a 1, pois nesse caso a soma das percentagens teria de dar 100%!

Estamos perante uma representação gráfica para a qual faltam algumas palavras, de certeza menos que *mil palavras*.

2.9 Algumas "delicadezas" no tratamento estatístico dos dados

Vimos nas secções 2.3 e 2.4, tratamentos estatísticos utilizados para classificar a informação contida em dados discretos e contínuos. Apresentámos algumas representações gráficas especialmente adequadas para dados discretos – gráfico de barras, e para dados contínuos – histograma, além de outros gráficos utilizados indiferentemente para dados discretos ou contínuos.

Embora a classificação de uma variável quantitativa em discreta ou contínua possa não oferecer dúvidas, já a forma como os dados se apresentam pode causar alguma confusão. Por exemplo as variáveis *Peso*, *Altura*, *Idade*, são de natureza contínua, pois os dados são recolhidos procedendo a uma medição. No entanto, estes dados aparecem-nos discretizados. É comum o peso aparecer em Kg, a altura em cm e a idade em anos. Embora a diferença entre dois valores possa ser tão pequena quanto se queira, essa diferença é condicionada pelo instrumento de medida e pela necessidade de uma representação numérica simples.

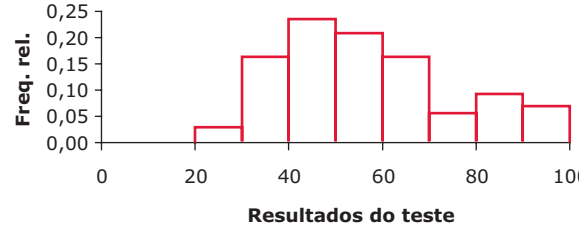
Por outro lado, algumas variáveis discretas, como por exemplo *Número de carros* que atravessam a portagem da ponte 25 de Abril num dia útil, escolhido ao acaso, *Salário* auferido por um trabalhador, são variáveis discretas, já que os dados são recolhidos procedendo a uma contagem. Por exemplo, no caso do salário, a diferença entre dois valores não pode ser inferior a um cêntimo.

Assim, embora não seja correcto utilizar o gráfico de barras para representar observações recolhidas de uma variável contínua, já o mesmo não se pode dizer da utilização do histograma para representar observações provenientes de variáveis discretas. Sempre que no estudo de uma variável discreta, o número de dados distintos seja muito grande, quando comparado com a dimensão da amostra, então deve-se utilizar o histograma, em vez do diagrama de barras. Voltemos ao exemplo **Candidatos a algumas vagas**, da página 41:

Exemplo:

Candidatos a algumas vagas (cont) – Uma vez que o número de valores distintos é muito grande, a construção de um gráfico de barras, conduziria a um gráfico com demasiadas classes, que não permitiria fazer sobressair o padrão da distribuição subjacente aos dados. Foi então sugerida a organização dos dados em classes, obtendo-se a seguinte tabela de frequências e o histograma correspondente:

Classes	Freq. absoluta	Freq. relativa
20 a 29	6	0,027
30 a 39	36	0,161
40 a 49	52	0,233
50 a 59	46	0,206
60 a 69	36	0,161
70 a 79	12	0,054
80 a 89	20	0,090
90 a 99	15	0,067
Total	223	1,000



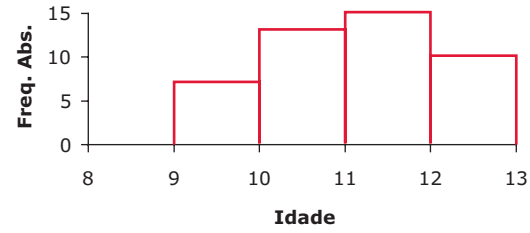
Quando os dados a classificar são provenientes de uma variável contínua, isto significa que poderemos obter, pelo menos teoricamente, um número infinito de valores distintos. Efectivamente, se a variável é de tipo contínuo, significa que não se pode passar de um valor a outro, sem passar por todos os valores intermédios. No entanto, estes dados, como dissemos anteriormente, podem-nos aparecer discretizados. Vejamos o seguinte exemplo:

Exemplo:

Idades dos alunos – Numa escola do 2.º ciclo recolheu-se informação sobre as idades de 45 alunos, tendo-se obtido os seguintes valores: 9, 11, 12, 10, 9, 10, 10, 10, 11, 12, 9, 9, 12, 12, 11, 11, 11, 11, 11, 12, 10, 10, 11, 9, 10, 9, 9, 10, 10, 10, 12, 12, 11, 10, 12, 11, 10, 11, 11, 10, 11, 11, 12, 11, 12
Note-se que quando se diz que um aluno tem 9 anos, por exemplo, este valor engloba todas as idades compreendidas entre os 9 e os 10 anos, exclusive. O mesmo se passa com as outras idades.

Ao estudar o histograma, considerámos uma metodologia que incluía uma regra que nos dá uma indicação sobre o número de classes que se podem considerar. Acontece que neste caso essa metodologia não deve ser aplicada, já que as classes, à partida estão bem definidas. Não teria qualquer sentido considerar 6 classes (valor sugerido pela regra de Sturges, cada uma com amplitude ligeiramente superior a 0.5 $\left(=\frac{12-9}{6}\right)$(recomendação feita na escolha da amplitude de classe). A organização dos dados pode ser feita de acordo com a seguinte tabela e respectivo histograma:

Classes	Freq. Absoluta	Freq. relativa
[9, 10[	7	0,16
[10, 11[	13	0,29
[11, 12[	15	0,33
[12, 13[	10	0,22
Total	45	1



Na Sala de Aula

Tarefa

Vamos conhecer a turma!...

Ao nível do 1.º ciclo do ensino básico, a forma como se introduz cada uma das técnicas de organização e representação gráfica de dados terá de ser muito alicerçada em actividades. Os alunos começam por recolher a informação e depois, naturalmente, terão curiosidade em "ver" um pouco mais para além daquele conjunto de valores que conseguiram obter.

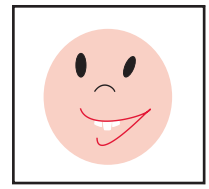
Neste texto vamos limitar-nos a apresentar algumas sugestões de como se poderão desenvolver um conjunto de actividades em que se faça tratamento estatístico de dados, nomeadamente a sua organização em tabelas e a construção de alguns gráficos.

O exemplo "Vamos conhecer a turma" irá servir de base a alguns procedimentos já referidos anteriormente:

Nome	Número de letras no nome	Tempo que demora de casa à escola (minutos)	Cor dos olhos	Comprimento do palmo (cm)	Número de irmãos
Ana Patrícia Santos	17	3	Azuis	14,7	3
Ana Rita Pereira	14	32	Castanhos	15,6	1
Bruno Martins	12	25	Castanhos	15,9	1
Cátia Reis	9	20	Pretos	14,2	1
Cláudia Rodrigues	16	17	Azuis	16,3	1
David Amaral	11	15	Azuis	13,5	2
Elisabete Soares	15	33	Pretos	14,4	1
José Manuel Rocha	15	22	Azuis	15,1	1
José Augusto Silva	16	9	Castanhos	15,2	1
Liliana Morais	13	35	Castanhos	16,2	1
Maria Isabel Antunes	18	25	Azuis	15,9	2
Miguel Correia	13	18	Verdes	13,6	0
Patrícia Mendes	14	10	Castanhos	17,3	1
Pedro Mendes	11	21	Castanhos	14,7	2
Ricardo Freitas	14	20	Verdes	15,0	0
Rui Eduardo Pires	15	6	Pretos	13,8	4
Sónia Gonçalves	14	5	Castanhos	14,3	1
Susana Alves	11	19	Azuis	15,4	0
Tatiana Medeiros	15	13	Castanhos	14,8	1
Vasco Fernandes	14	15	Castanhos	13,2	3

Indo por grau de dificuldade, deve-se começar por organizar os **dados de tipo qualitativo**. Para estes, a representação gráfica na forma de **pictograma** é especialmente atraente para os alunos e, por isso, vamos apresentar duas propostas de pictograma para a variável qualitativa *Cor dos olhos*.

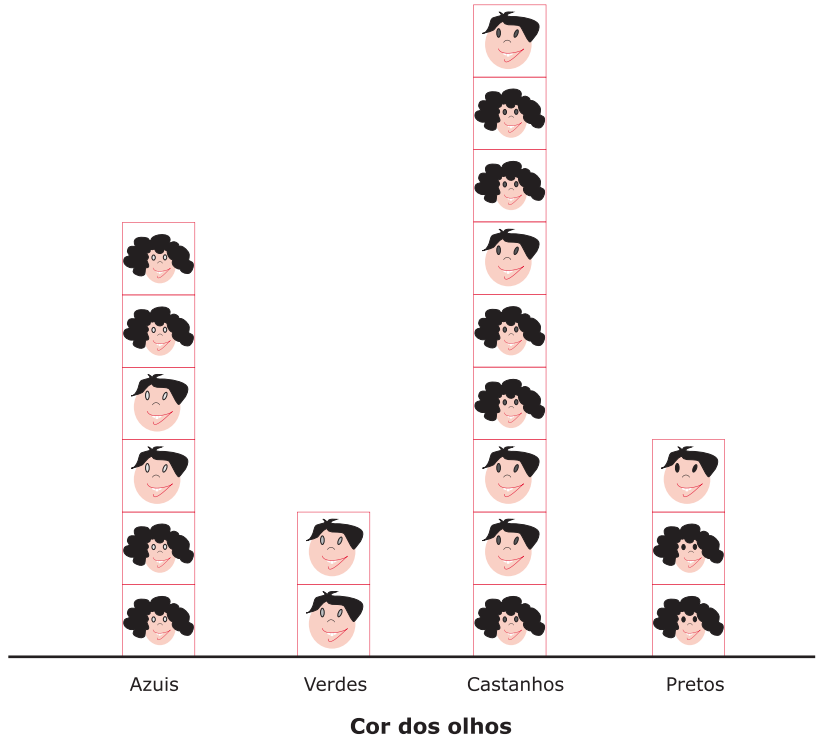
Entrega-se a cada aluno um pequeno quadrado de papel com uma cara desenhada. As caras devem ser todas iguais e o aluno terá de pintar os olhos da cor dos seus próprios olhos e desenhar os cabelos (para diferenciar entre rapaz e rapariga):



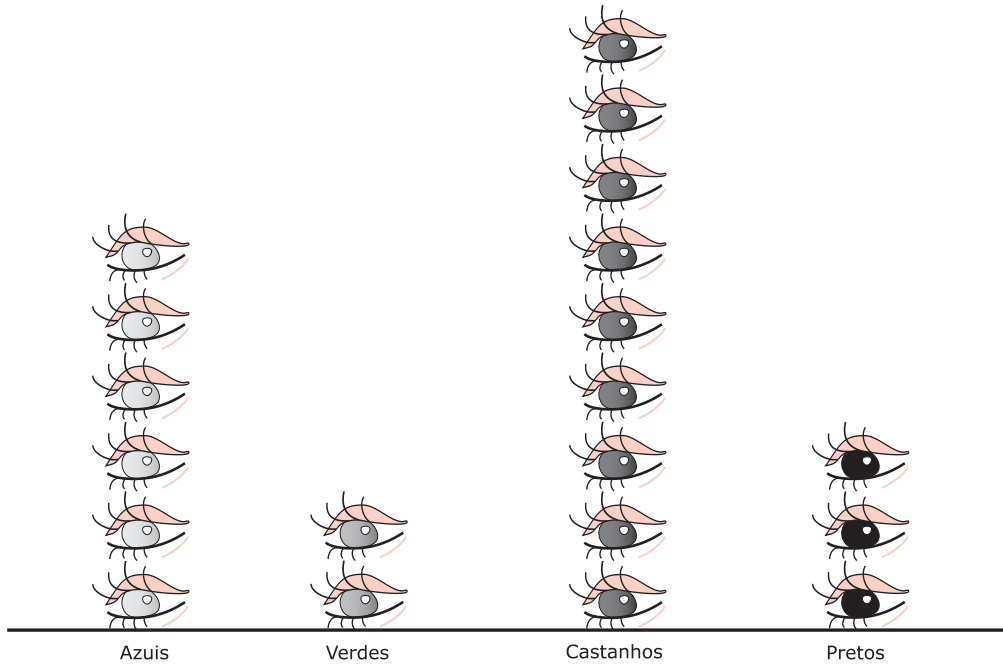
Numa folha de cartolina grande, traça-se uma linha horizontal e escreve-se sob essa linha as quatro cores de olhos que surgem na amostra. Coloca-se como legenda "Cor dos olhos":

Cada um dos alunos deverá colar a cara que pintou no local respectivo, imediatamente acima de alguma cara que já esteja lá colocada.

No final obtém-se um pictograma muito divertido!...



Em alternativa pode também utilizar-se o desenho de um "olho" como representante das unidades observacionais. Na amostra em estudo as unidades observacionais são os alunos mas, no que respeita à característica cor dos olhos, pode admitir-se que elas possam ser, simplesmentes, os "olhos":



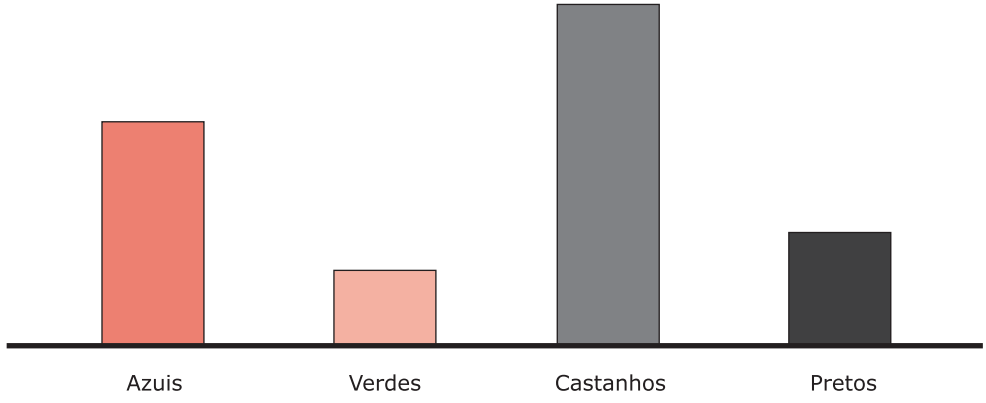
Nesta primeira abordagem à organização dos dados procedeu-se pela ordem contrária ao que é habitual. Fez-se a representação gráfica antes de fazer a tabela de frequências! Tal faz sentido tendo em conta a idade dos alunos, pois a representação gráfica é muito mais apelativa e, no caso das variáveis qualitativas, pode servir de base para a construção da tabela de frequências. Aliás, esta situação não é nova, pois quando falámos no gráfico de pontos, também o construímos antes da tabela de frequências.

Organizados os dados numa tabela de frequências, obtém-se

Cor dos olhos	Frequência Absoluta	Frequência Relativa
Azuis	6	0,30
Verdes	2	0,10
Castanhos	9	0,45
Pretos	3	0,15
Total	20	1,00

Nota: A coluna das frequências relativas é facultativa, deixando-se ao critério do professor apresentá-la ou não, pelo menos nesta fase.

Pode agora passar-se à construção de um **gráfico de barras**. Pede-se aos alunos que desenhem 4 rectângulos, todos com a mesma largura, em papel quadriculado, por exemplo, e cujas alturas sejam iguais às frequências absolutas. De seguida poderão recortar os rectângulos e colá-los numa folha de papel onde tenham desenhado um eixo e identificado as categorias da variável *Cor dos olhos*.

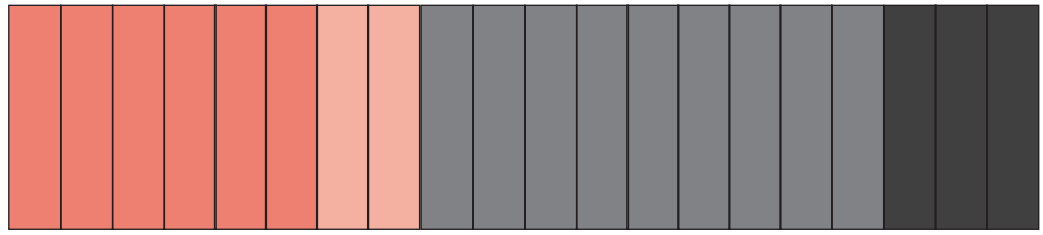


Cor dos olhos

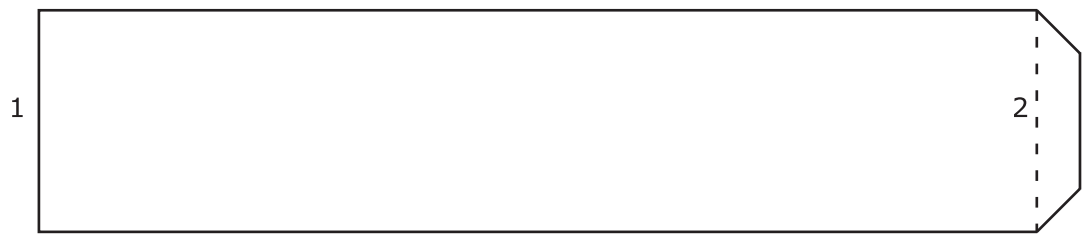
Numa fase posterior pode-se ainda pedir que desenhem o gráfico de barras numa folha de papel quadriculado.

Ainda utilizando o papel quadriculado, pode-se ensinar os alunos a desenharem um **diagrama circular**, para a variável *Cor dos olhos*, da seguinte forma:

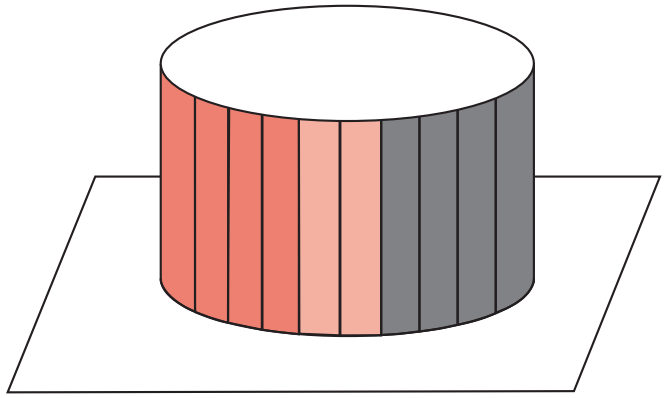
Numa folha desenha-se um rectângulo com largura igual a 20 unidades (pode-se considerar como unidade 1 ou 2 quadrículas) e uma altura qualquer. Divide-se essa largura em 4 partes de comprimentos 6, 2, 9 e 3 unidades, que se pintam de cores diferentes, conforme as classes a que dizem respeito:

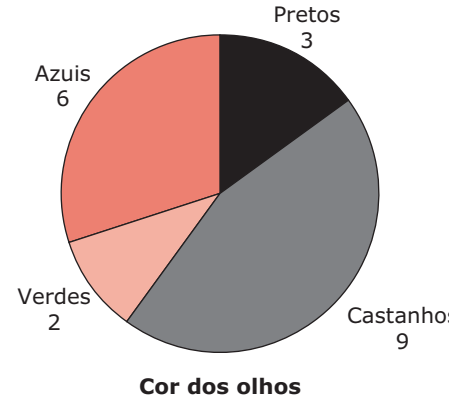


Cola-se o rectângulo anterior a uma cartolina com as mesmas dimensões, com uma margem para colar os lados 1 e 2 de modo a obter um cilindro:



Apoiam o cilindro numa folha de papel e desenham a circunferência assinalando os pontos onde muda a cor. Com a ajuda do professor procuram encontrar um ponto aproximado para o centro, que unem com os pontos da circunferência anteriormente assinalados:





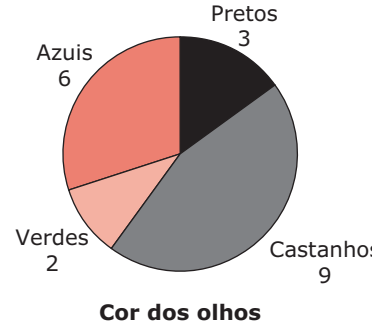
Completam a figura pintando as partes em que o círculo ficou dividido e colocando o nome das classes e as frequências absolutas respectivas.

Em turmas onde os alunos já conheçam as medidas das amplitudes de ângulo e saibam usar o transferidor para desenhar ângulos com uma amplitude que lhes é indicada, é também possível ensinar a construir o **diagrama circular** sem recorrer ao processo anterior.

Começa-se por dizer que se vai fazer uma representação gráfica na forma de um círculo e aproveita-se para recordar que a amplitude de um ângulo giro é igual a 360º. Os alunos têm então de desenhar sectores circulares, todos com o mesmo raio e amplitudes que se obtêm multiplicando a frequência relativa pelos 360º:

Cor dos olhos	Frequência Absoluta	Frequência Relativa (%)	Amplitude do ângulo
Azuis	6	30	108º
Verdes	2	10	36º
Castanhos	9	45	162º
Pretos	3	15	54º
Total	20	100	360º

Cada sector circular deverá ser pintado com uma cor diferente e o "puzzle" deverá no final ser montado de modo a formar um círculo completo. Não esquecer de colocar a legenda:

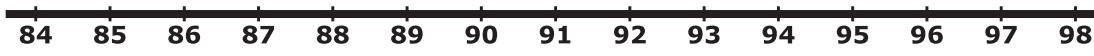


Para os **dados de tipo quantitativo** a representação gráfica mais fácil de ensinar, a alunos do 1.º ciclo do ensino básico, é o **gráfico de pontos**.

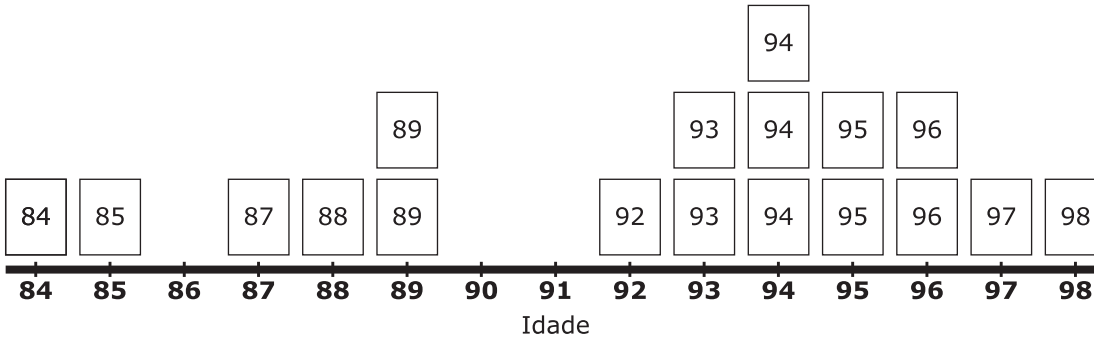
Vamos ver uma forma simples, de construir este gráfico considerando a variável *Idade*, medida em meses.

Pode começar-se por representar a idade de cada aluno em meses. De seguida o professor dá um quarto de uma folha A4 a cada aluno para registar o valor que obteve para a sua idade, que também é registada no quadro. Suponhamos que numa turma de 18 alunos se obtiveram os seguintes valores: 87, 88, 85, 84, 89, 92, 89, 94, 93, 98, 94, 97, 95, 95, 96, 96, 95, 96.

Numa cartolina grande desenha-se uma linha horizontal e, por baixo, igualmente espaçados, escrevem-se todos os números entre a menor e a maior das idades obtidas:



Depois cada aluno irá colocar o pedaço de folha com a sua idade, por cima do valor respectivo. Quando todos os alunos tiverem terminado, obter-se-á uma representação com o seguinte aspecto, em que os pontos foram substituídos por pedaços de papel:



A leitura e interpretação da representação gráfica obtida permite responder a algumas questões, como por exemplo:

Há algum aluno na turma cuja idade seja 90 meses?

Quantos colegas teus têm a tua idade?

Há mais alunos com idade inferior ou superior a 90 meses?

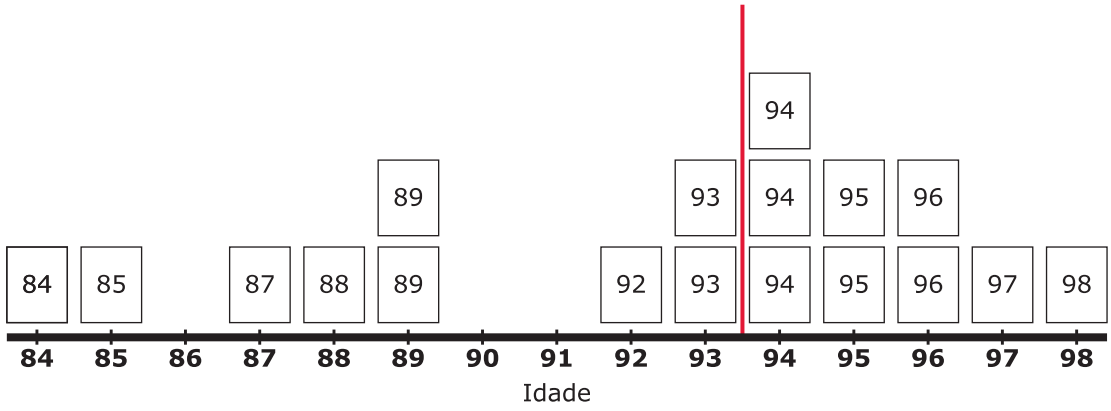
Quantos alunos têm idade menor ou igual a 93 meses? E maior ou igual que 94 meses?

Preenche a seguinte tabela:

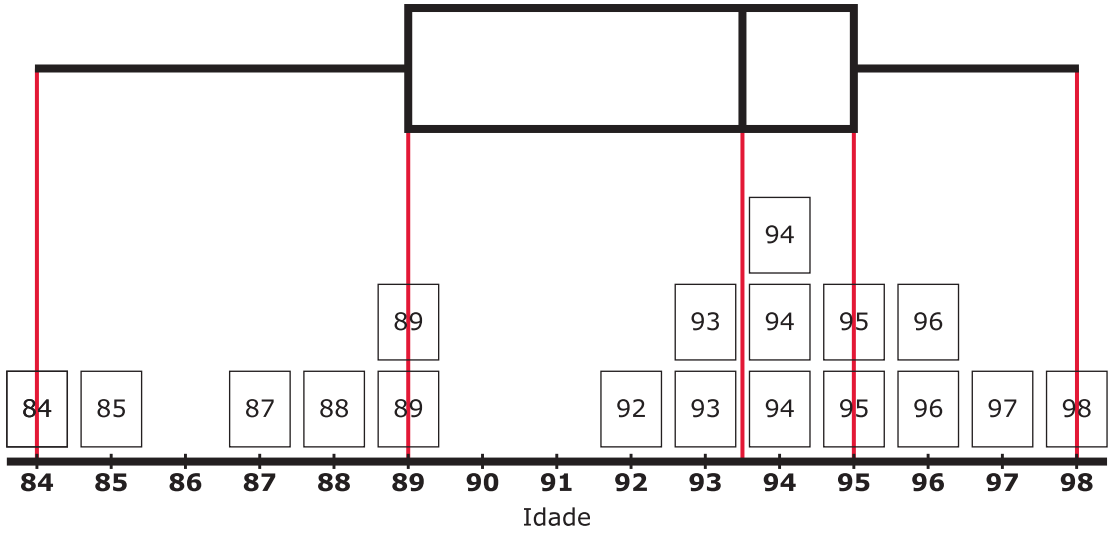
Idade (em meses)	Número de alunos
Menos de 85	
de 85 a 90	
de 90 a 95	
de 95 a 100	
Total	

Nota: Quando se escreve "de 85 a 90" entende-se que é maior ou igual que 85 e menor que 90. A convenção é idêntica para as outras classes.

A partir da representação gráfica anterior pode-se ainda calcular a mediana e os quartis para desenhar um diagrama de extremos e quartis. Assim, começa-se por identificar o "sítio" da mediana, que não será difícil se se tiver já concluído que o número de alunos com idade menor ou igual a 93 meses, é igual ao número de alunos com 94 ou mais meses de idade. Assinala-se a mediana com um traço:



A mediana dividiu o conjunto dos 18 papéis em duas partes, cada uma com 9 papéis. Agora os alunos com a ajuda do professor determinam as medianas de cada uma destas partes, que assinalam do mesmo modo que fizeram para a mediana. Uma vez estes 3 pontos determinados, pode construir-se o diagrama de extremos e quartis, como se apresenta na figura seguinte:



Podem fazer-se algumas perguntas que evidenciem a facilidade de leitura deste tipo de gráfico.

Sugestão: Pedir aos alunos para representarem graficamente os dados da variável *Número de letras do nome*, da tabela "Vamos conhecer a turma", utilizando um procedimento idêntico ao utilizado para a variável *Idade*.

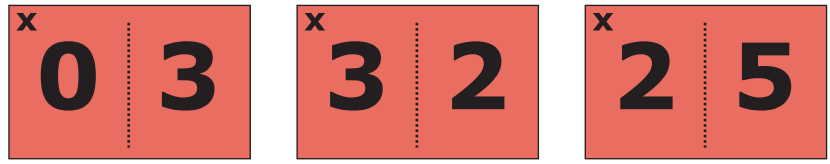
Ainda para **dados de tipo quantitativo** uma outra representação gráfica muito fácil de utilizar com os alunos do 1.º ciclo do ensino básico, é o **gráfico de caule-e--folhas**.

Vamos ilustrar uma forma simples de proceder usando como exemplo a variável *Tempo que demoras de casa à escola*, medido em minutos, da tabela "Vamos conhecer a turma" (ver página 87).

Começa-se por dar a cada aluno um rectângulo de cartolina (fina) com uma linha vertical tracejada a dividi-lo a meio e uma pequena cruz no canto superior esquerdo:

Do lado esquerdo do rectângulo o aluno terá de colocar o algarismo das dezenas do número que representa o tempo que ele demora de casa à escola. Do lado direito coloca o algarismo das unidades.

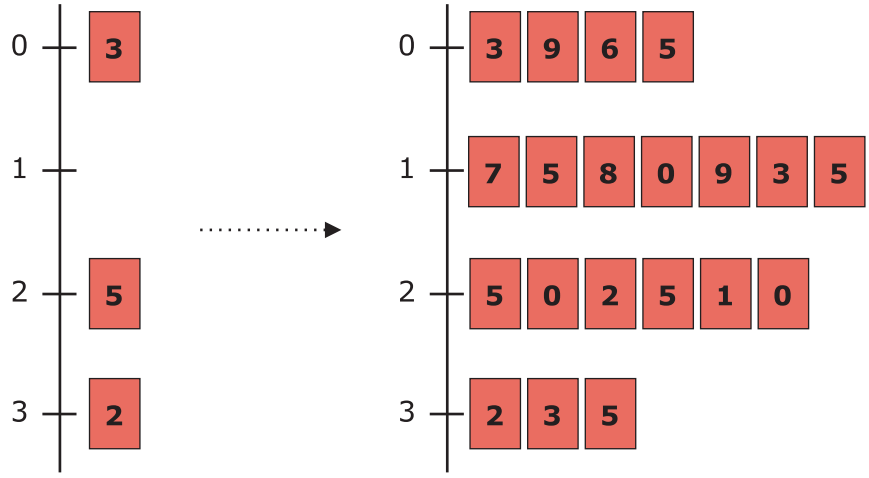
Os pequenos rectângulos de cartolina referentes aos 3 primeiros alunos da lista terão então o seguinte aspecto



De seguida, cada aluno dobra a cartolina pela linha tracejada, mantendo visíveis para o exterior os algarismos, e cola as duas metades pela parte de dentro.

Numa cartolina grande, desenha-se uma linha vertical e marcam-se de forma igualmente espaçada os algarismos dominantes (neste caso, das dezenas). Colocam-se todos, do mais pequeno ao maior, mesmo que na amostra haja algum que não apareça. No caso do exemplo que estamos a tratar os dígitos dominantes (os caules) são 0, 1, 2, e 3. Agora é só ir colocando cada cartão (folha) à frente do respectivo caule.

De notar que, em cada cartão, as folhas estão do lado que não tem "cruz" enquanto os caules se podem identificar virando o cartão e vendo o dígito que surge marcado com a dita "cruz". A evolução da representação gráfica entre a colocação dos 3 cartões acima e a fase final em que já estão colocados todos os cartões será então:



Para terminar basta agora ordenar, por ordem crescente, as folhas que estão em frente de cada um dos caules:

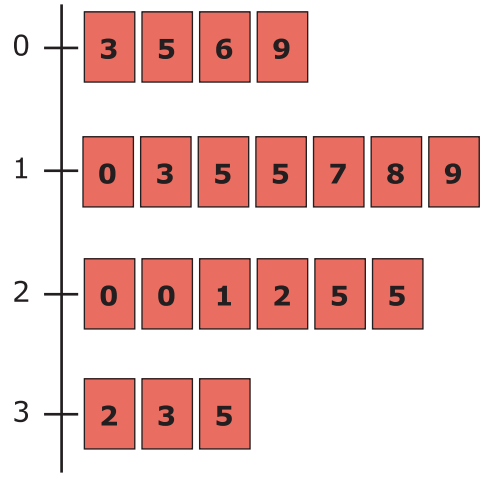


Gráfico de caule-e-folhas

A leitura e interpretação da representação gráfica é também muito importante. Eis algumas questões que podem ser colocadas a partir da leitura do gráfico de caule-e-folhas construído:

- Quantos alunos demoram mais do que 30 minutos a chegar à escola?
- Quantos alunos da turma demoram a chegar à escola entre 10 minutos (conta os que demoram 10 minutos) e 20 minutos (não consideres os que demoram 20 minutos)?
- Há mais alunos na turma a demorar mais tempo do que aquele que tu demoras ou há menos?
- Será verdadeira a frase "A maioria dos alunos da turma demora menos de 20 minutos a chegar à escola"? Justifica.
- Preenche a seguinte tabela de frequências

Tempo de casa à escola	Número de alunos
Até 10m	
de 10m a 20m	
de 20m a 30m	
de 30 a 40m	
Total	

Tarefa

Vamos conhecer algumas características dos alunos da escola

Será que predominam os olhos castanhos? Ou serão os pretos? E será que a cor dos olhos depende do sexo, isto é, se é rapaz ou rapariga? Para responder a esta questão, decidiu um professor nomear duas comissões de alunos, em que uma das comissões iria averiguar a cor dos olhos de 30 raparigas e a outra comissão iria averiguar a cor dos olhos de 25 rapazes. No dia escolhido para a recolha de dados, os alunos que pertenciam às comissões foram para a porta da escola e registaram a cor dos olhos das primeiras 30 alunas e dos primeiros 25 alunos a chegarem. Observe-se que as comissões acabaram a recolha da informação praticamente ao mesmo tempo, pois na escola havia mais raparigas que rapazes. Os resultados obtidos foram os seguintes:

Raparigas

pretos, castanhos, castanhos, azuis, pretos, castanhos, verdes, azuis, castanhos, castanhos, azuis, pretos, cinzentos, verdes, azuis, castanhos, castanhos, castanhos, castanhos, pretos, verdes, azuis, castanhos, pretos, pretos, castanhos, castanhos, pretos, castanhos, castanhos

Rapazes

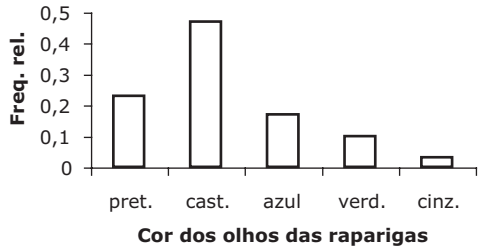
Castanhos, azuis, castanhos, pretos, castanhos, castanhos, pretos, castanhos, verdes, castanhos, pretos, castanhos, castanhos, pretos, azuis, azuis, verdes, castanhos, castanhos, verdes, castanhos, pretos, azuis, pretos, castanhos.

Para cada conjunto de dados construa uma tabela de frequências para organizar a informação recolhida e a seguir construa uma representação gráfica adequada. Tire conclusões.

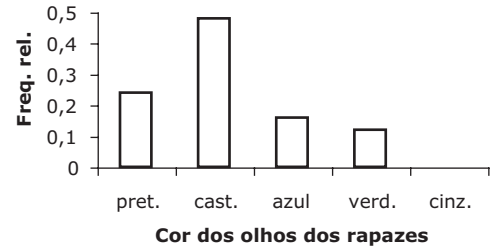
Resolução: Para construir a tabela de frequências, deve verificar-se quais as categorias ou modalidades que a variável (qualitativa) em estudo – *Cor dos olhos*, pode assumir. Embora no caso dos dados recolhidos para os rapazes, não se tivesse observado nenhum com olhos cinzentos, decidiu-se incluir essa categoria na tabela de frequências, para melhor se fazer a comparação com os dados recolhidos para as raparigas:

	Raparigas	
Classes	Freq.abs.	Freq.rel.
preto	7	0,23
castanho	14	0,47
azul	5	0,17
verde	3	0,10
cinzento	1	0,03
Total	30	1,00

	Rapazes	
Classes	Freq.abs.	Freq.rel.
preto	6	0,24
castanho	12	0,48
azul	4	0,16
verde	3	0,12
cinzento	0	0,00
Total	25	1,00



Cor dos olhos das raparigas



Cor dos olhos dos rapazes

Como se verifica a partir das frequências relativas ou dos gráficos de barras respectivos, construídos para estudar como se distribui a *Cor dos olhos* pelas raparigas e rapazes, podemos admitir que, na escola:

- Predominam os olhos castanhos.
- Em segundo lugar predominam os olhos pretos.
- Os olhos cinzentos são raros.
- A distribuição da variável C*or dos olhos*, é idêntica para as raparigas e rapazes.

Exercício:

Fazer um estudo análogo ao anterior, mas em que a variável a estudar seja P*rograma da televisão favorito*. Quais os programas favoritos? Haverá diferença entre os programas favoritos dos rapazes e das raparigas?

Tarefa

Vamos comparar a temperatura entre Lisboa e Porto

Durante 2 semanas, cada um dos 28 alunos de uma turma, ficou encarregue de registar a temperatura máxima observada num dos 14 dias e numa das 2 cidades. Essas temperaturas eram apontadas diariamente, numa tabela idêntica à seguinte:

Dia / Cidade	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lisboa	19	21	20	23	18	20	22	24	23	21	20	20	19	19
Porto	17	22	21	18	16	19	17	20	21	18	15	17	16	17

Utilizando uma representação gráfica adequada, vão-se comparar as temperaturas nas duas cidades.

Para comparar os 2 conjuntos de dados, pode-se utilizar a representação gráfica de caule-e-folha, considerando os mesmos caules para os dois conjuntos de dados:

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

Da representação gráfica anterior conclui-se, imediatamente, que, de um modo geral, as temperaturas em Lisboa são superiores às do Porto.

Exercício:

Será que a temperatura habitual do local onde a escola se situa, é muito diferente da temperatura de uma cidade afastada, por exemplo, 200 Km? Para planear este estudo:

- a) O professor tenta arranjar um termómetro para medir a temperatura exterior e, durante alguns dias, antes de iniciar as aulas, regista a temperatura exterior ou pede a um aluno que a registe.
- b) Por outro lado, antes de sair de casa, o professor toma atenção ao noticiário, e aponta qual a temperatura que faz na cidade escolhida para a comparação, ou combina com outra escola, com quem faça inter-câmbio.

Depois, para comparar os dois conjuntos de dados, procede de forma idêntica à da actividade anterior.

Tarefa

Quais são os nossos animais domésticos?

Na escola, um grupo de alunos decidiu averiguar se as famílias têm animais domésticos e no caso de os terem, que animais domésticos é que têm. Acompanhados do professor, foram para a porta da escola (ou para uma rua com algum movimento) e às primeiras 50 pessoas que passaram fizeram as seguintes perguntas:

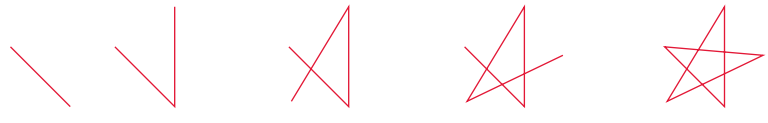
Tem algum animal doméstico? Se sim, qual o animal doméstico que tem há mais tempo?

Para anotar a informação que iam recebendo, tinham preparado uma folha de papel, idêntica à seguinte:

Não:	
Sim:	
Cão	
Gato	
Cágado	
Peixes	
Passarinho(s)	
Porquinho(s)-da-Índia	
Ratinho(s)	
Coelho(s)	
Galinha(s)	
Outros:	

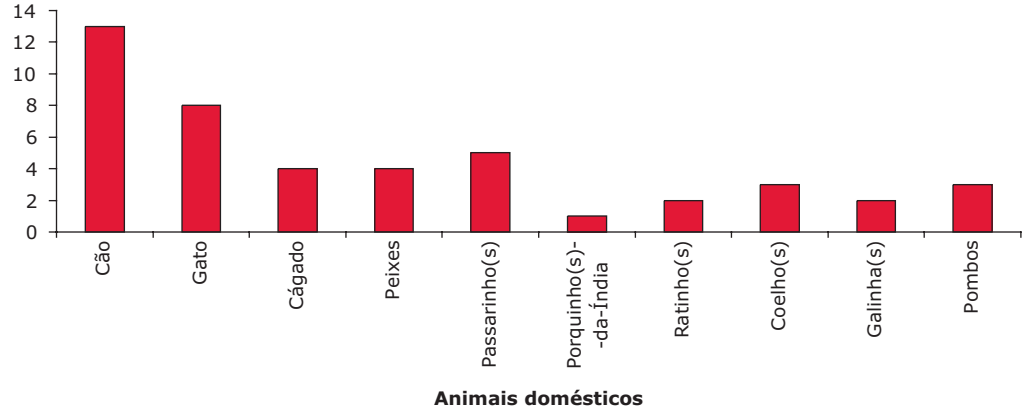
Não:	𝍸
Sim:	
Cão	𝍸 𝍸 III
Gato	𝍸 III
Cágado	IIII
Peixes	IIII
Passarinho(s)	𝍸
Porquinho(s)-da-Índia	I
Ratinho(s)	II
Coelho(s)	III
Galinha(s)	II
Outros:	
Pombos	III

À medida que as pessoas iam respondendo, anotavam com um traço. Faziam grupos de 5 traços, em que o quinto traço é oblíquo, por cima dos outros 4. Estes grupos tornam mais fácil a contagem posterior. Uma alternativa a estes montinhos, é o desenho de uma estrela, em que se representa sucessivamente:



Nota: Uma representação idêntica à anterior, recebe o nome de diagrama de marcas ou registos (*tally chart*).

Pode-se construir, com a ajuda do professor, em papel quadriculado, um gráfico semelhante ao da figura:



Algumas questões que podem ser feitas a partir da observação do gráfico:

- a) Houve mais pessoas a responderem que tinham cão ou gato?
- b) Das pessoas que responderam, qual o animal que as pessoas tinham menos em casa?
- c) Se outro grupo de alunos tivesse feito a mesma pergunta a outras 50 pessoas, o que é que se esperava que as pessoas respondessem mais vezes?
- d) Se no grupo das 50 pessoas considerado na alínea anterior, 14 pessoas respondessem que tinham cão, ficavas muito admirado ou achas que essa resposta é muito possível?
- e) Se, ainda neste novo grupo, 10 pessoas dissessem que tinham em casa galinhas, ficavas admirado? Porquê? Mais ou menos quantas pessoas esperarias que dissessem que tinham galinhas?

Algumas respostas:

c) Esperava-se que respondessem que tinham cão.
d) Não ficava admirado, porque se esperava obter um valor perto de 13, que foi o que se obteve como resposta nas primeiras 50 pessoas.
e) Sim, ficava admirado, porque esperava que houvesse poucas pessoas a responderem galinhas. Mais precisamente, esperávamos que o número de pessoas que respondessem galinhas andasse à volta de 2.

Tarefa

Qual o desporto favorito?

Para verificar se haveria evidência de que os desportos favoritos fossem diferentes para os rapazes e para as raparigas de uma determinada escola com 1567 alunos, um grupo de alunos dessa escola, resolveu fazer um estudo, baseado num inquérito feito a 160 alunos, dos quais 100 eram raparigas. As respostas ao inquérito foram organizadas nas seguintes tabelas, onde se apresenta o número de raparigas e o número de rapazes, cujo desporto favorito é o futebol, a natação, o atletismo, o ténis ou o ciclismo:

Raparigas	
Futebol	41
Natação	25
Atletismo	8
Ténis	23
Ciclismo	3

Rapazes	
Futebol	30
Natação	12
Atletismo	8
Ténis	7
Ciclismo	3

Tendo em consideração os resultados da tabela anterior, o grupo encarregue do estudo elaborou um relatório, onde se fazem as seguintes afirmações:

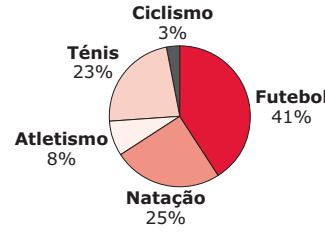
1. Ao contrário do que se pensava, há mais raparigas a preferirem o futebol, do que rapazes.
2. É interessante verificar que o atletismo e o ciclismo, é igualmente preferido por raparigas e rapazes.
3. O número de raparigas que prefere a natação, é mais do dobro do número de rapazes que prefere este desporto.

Concorda com as conclusões? Caso não concorde, apresente a sua versão das respostas que considera correctas.

Resolução:

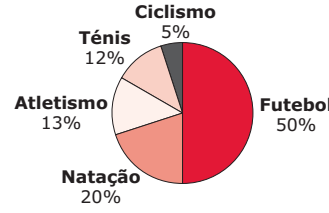
As conclusões estão erradas, pois estão baseadas nas frequências absolutas, quando se deveriam exprimir em termos das frequências relativas, uma vez que o número de raparigas inquiridas é diferente do número de rapazes inquiridos. Considerando as tabelas anteriores, onde adicionámos uma coluna com as frequências relativas, temos:

Classes	Raparigas Freq.abs.	Freq.rel.
Futebol	41	0,41
Natação	25	0,25
Atletismo	8	0,08
Ténis	23	0,23
Ciclismo	3	0,03



Desporto favorito (raparigas)

Classes	Rapazes Freq.abs.	Freq.rel.
Futebol	30	0,50
Natação	12	0,20
Atletismo	8	0,13
Ténis	7	0,12
Ciclismo	3	0,05



Desporto favorito (rapazes)

Como se verifica a partir dos resultados das tabelas e das representações gráficas:

1. Há uma maior percentagem de rapazes (50%), do que de raparigas (41%) a preferirem o futebol.
2. O atletismo e o ciclismo são desportos mais preferidos pelos rapazes.
3. A percentagem de raparigas que prefere a natação (25%), é um pouco superior à percentagem de rapazes que prefere esta modalidade (20%).

Tarefa

Vamos pesar laranjas

O(a) professor(a) pede a cada aluno da turma para, no dia seguinte, trazer uma ou duas laranjas (ou outro fruto, à escolha), pois vão fazer uma actividade, em que procurarão recolher informação sobre o peso desse fruto. No caso de não haver uma balança na escola, o professor providenciará para a arranjar. No dia escolhido para fazer pesagens, cada aluno vai pesar a(s) sua(s) laranja(s) e vai registar no quadro o peso (em gramas) observado. Suponha que os pesos obtidos foram os seguintes:

152	142	157	168	167	172	133	153	166	144	148	138	137	145
147	134	149	151	156	151	152	151	168	154	153	140	175	164
176	148	172	139	160	164	174	154	150	162	151	163	141	146

- a) O que é que se está a estudar?
- b) Estes dados resultam de uma contagem, ou de uma medição?
- c) Organiza os dados na forma de um caule-e-folhas
- d) A partir da representação gráfica, sabes dizer quantas laranjas pesam mais do que 170 gramas?
- e) E quantas laranjas têm um peso maior ou igual a 150 gramas, mas menor que 160 gramas?
- f) Alguém trouxe uma laranja com peso igual ou superior a 180 gramas?
- g) (Só para o professor) Organizar os dados na forma de um histograma, considerando como classes [130, 140[, [140, 150[, [150, 160[, [160, 170[e [170, 180[. Comparar a representação em caule-e-folhas obtida na alínea c) com o histograma.

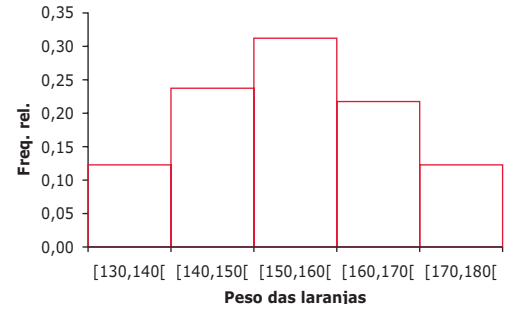
Resolução:

a) A variável a ser estudada é o "peso" de uma laranja.
b) Os dados foram obtidos através de uma medição. O objecto utilizado para a medição foi a balança.
c) Para obter a representação em caule-e-folhas, vamos considerar como caules 13, 14, 15, 16 e 17. Pendurando nestes caules as folhas respectivas e ordenando as folhas de cada caule, obtemos a seguinte representação:

13	3 4 7 8 9
14	0 1 2 4 5 6 7 8 8 9
15	0 1 1 1 1 2 2 3 3 4 4 6 7
16	0 2 3 4 4 6 7 8 8
17	2 2 4 5 6

d) As laranjas que têm peso igual ou superior a 170 gramas, são as que, na representação gráfica do caule-e-folhas, têm os pesos com caule 17. Assim, temos 5 laranjas com peso igual ou superior a 170 gramas.
e) As laranjas que têm peso maior ou igual a 150 gramas, mas menor que 160 gramas, são as que, na representação gráfica do caule-e-folhas, têm os pesos com caule 15. Assim, 13 laranjas estão nas condições pretendidas.
f) Ninguém trouxe laranjas com peso igual ou superior a 180 gramas. Para tirar esta conclusão, basta ver que na representação do caule-e-folhas, não existe o caule 18.
g) Para construir o histograma, começa-se por construir uma tabela de frequências em que se consideram como classes as seguintes: [130, 140[, [140, 150[, [150, 160[, [160, 170[, [170, 180[:

Classes	Freq. Abs.	Freq. Rel.
[130, 140[	5	0,12
[140, 150[	10	0,24
[150, 160[	13	0,31
[160, 170[	9	0,21
[170, 180[	5	0,12
Total	42	1,00



A escolha das classes anteriores para construir o histograma, foi feita com o objectivo de fazer sobressair a semelhança do histograma com a representação em caule-e-folhas. Se repararmos nos caules considerados para o caule-e-folhas, cada um tem penduradas as folhas correspondentes aos elementos dentro de cada uma das classes consideradas para o histograma.

Observemos que as duas representações gráficas consideradas, dão informação muito semelhante, no que diz respeito à distribuição dos pesos das laranjas.
Nomeadamente, realçamos a seguinte informação relevante, a retirar do gráfico:

- Predominam as laranjas com peso entre 150 e 160;
- O número de laranjas com peso inferior a 150, é sensivelmente igual ao número de laranjas com peso superior a 160;
- A média dos pesos observados deve andar à volta de 155 gramas.

Tarefa

Hábitos alimentares – comemos fruta suficiente?

Dizem os nutricionistas que, para uma alimentação saudável, além de outros requisitos, deveríamos comer 3 peças de fruta, por dia. Vamos investigar se os alunos comem fruta suficiente... Esta actividade vai ser realizada por duas turmas, pelo que num dia escolhido pelos professores para a realizar, começa-se por se debater:

- O que é que se vai perguntar a cada aluno;
- Como registar a informação recolhida.

Depois de alguma discussão, decide-se perguntar a cada aluno, quantas peças de fruta e que tipo de fruta, comeu no dia anterior. Convém explicar que, se por exemplo a fruta for cerejas, uma peça de fruta não será uma cereja! Pode ser, por exemplo, um copo cheio de cerejas. Analogamente, se se tratar de uvas, será um cacho de uvas. Depois de decidida a pergunta a fazer, começa-se a discutir sobre qual a melhor forma de registar a informação. Com a ajuda dos professores, pode chegar-se à conclusão que uma forma possível, seria construírem uma tabela, análoga à seguinte:

Quantas peças? / Quais	0	1	2	3	4	5	Mais de 5	Total
Ameixa								
Ananás								
Banana				|				
Cereja								
Figo								
Laranja								
Maçã				|				
Melancia								
Melão								
Meloa								
Morango								
Nêspera								
Papaia								
Pêra								
Pêssego								
Tângera								
Tangerina								
Uva				|				
Nenhuma								
Total								

Todos os frutos apresentados na tabela foram sugeridos pelos alunos. Para exemplificar o preenchimento da tabela, suponhamos que um aluno tinha no dia anterior comido uvas, uma maçã e uma banana. Então esse aluno ia ao quadro e na coluna com o número 3, colocava um risquinho (|) nas linhas que dizem respeito às Uvas, Maçãs e Bananas, como está assinalado na tabela. Um aluno que não tivesse comido fruta nenhuma, colocaria um risquinho na coluna com o 0 e na linha onde está escrito Nenhuma. Vamos admitir que os 35 alunos das turmas tinham ido ao quadro preencher a tabela com a informação que lhes dizia respeito e que a tabela obtida foi a seguinte:

Quantas peças? / Quais	0	1	2	3	4	5	Mais de 5	Total
Ameixa		|	|		|			3
Ananás			|					1
Banana		|	|||||	||||	|||||	|||||	||	23
Cereja			|	||	|			4
Figo			|					1
Laranja			||	|	|	||	|	7
Maçã		|	|||	|||||	|	|		12
Melancia			|	|				2
Melão			|	|				2
Meloa			|		|			2
Morango				|	|			2
Nêspera			|					1
Papaia		|						1
Pêra		|	|||	||||	||	||	|	13
Pêssego			|	|	|		|	4
Tângera			|					1
Tangerina			|	|	|			3
Uva		|	|	||	|		|	6
Nenhuma	||							2
Total	2	6	26	24	16	10	6	

- a) A partir da tabela pode-se concluir que há uma fruta que é preferida pelos alunos. Qual é essa fruta?
- b) Houve só um aluno a dizer que comeu figos. Poderemos concluir imediatamente que os alunos não gostam de figos? Ou poderemos, por exemplo, estar numa época em que só agora é que os figos começaram a amadurecer?
- c) Quantos alunos responderam que comeram 3 peças de fruta, no dia anterior?
- d) Com a ajuda do professor, constrói uma tabela de frequências onde se possa ver quantos alunos comeram 0, 1, 2, 3, 4, 5 ou 6 peças de fruta.

Tarefa proposta

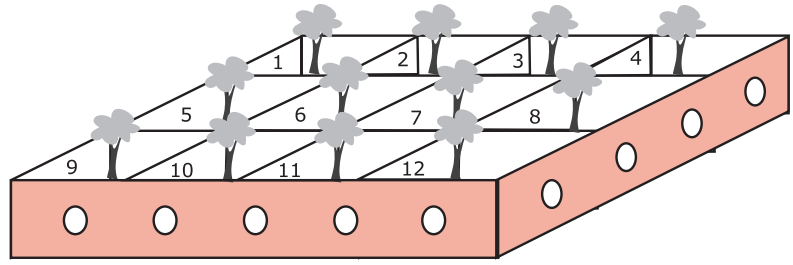
Vamos acompanhar o crescimento do milho

Será que os grãos de milho crescem o mesmo, durante um certo período de tempo? Num vaso rectangular, vamos plantar vários grãos de milho. É necessário saber qual a melhor época para plantar este cereal. Depois do milho começar a germinar, durante algumas semanas, os alunos terão como tarefa, acompanhar o seu crescimento, medindo os pezinhos do milho (esta medição deve ser feita, num dia fixo da semana).

- a) Considerando 3 semanas consecutivas, em que se registou a altura dos pés de milho, calcular o crescimento entre a 1.ª e a 2.ª semana e entre a 2.ª e a 3.ª semana.
- b) Comparar os crescimentos anteriores.

Resolução:

Para melhor identificarem os pés de milho, sugere-se que se faça uma quadrícula no vaso, com cordel ou fio de pesca, e em cada célula da quadrícula, semear um grão de milho. Constróem uma tabela com tantas células, quantos os grãos semeados, onde registarão o comprimento de cada pé de milho, em cada uma das semanas:



	Grão 1	Grão 2	Grão 3	Grão 4	Grão 5	Grão 6	Grão 7	Grão 8	Grão 9	Grão 10	Grão 11	Grão 12
1.ª semana												
2.ª semana												
3.ª semana												
2.ª – 1.ª												
3.ª – 2.ª												

Na tabela anterior já inserimos 2 linhas, onde serão calculados os crescimentos verificados para cada pé de milho, entre as 2.ª e 1.ª semanas e entre as 3.ª e 2.ª semana.

Tarefa proposta

Qual a dimensão do nosso salto em comprimento?

Os professores de 2 turmas da escola, de anos diferentes, decidiram levar a cabo uma experiência para averiguar se, como suspeitavam, a idade tinha influência no comprimento do salto de um jovem. Então, num dia em que as turmas tinham Educação Física, arranjaram um espaço no recreio da Escola, onde os alunos poderiam dar saltos em comprimento. Munidos de uma fita métrica, procederam à medição dos comprimentos dos saltos dos alunos de cada uma das turmas.

Organizar os dados em tabelas e construir os histogramas associados.

Tarefa proposta

Será que os autocarros que passam à frente da escola passam com a regularidade que está prevista no horário afixado na paragem?

Às vezes os alunos queixam-se de que os autocarros demoram muito a chegar e estão muito tempo na paragem, à espera que venha um! Então, um grupo de alunos decidiu realizar um projecto que consistia em estudar os tempos entre passagens consecutivas dos autocarros da carreira mais frequente. Escolheram alguns dias para recolher dados para esse estudo, e nesses dias o grupo de alunos (acompanhado do professor), foi mais cedo para a escola e instalou-se à porta, pronto a apontar as horas de passagem da dita carreira, no período das 8 às 9 horas da manhã (ou noutro período à escolha).

- a) De acordo com os dados registados, quantos autocarros passaram no período das 8 às 9 horas?
- b) Consulta o horário afixado na paragem dessa carreira, mais perto da escola. De acordo com esse horário, quantos autocarros deveriam passar no período em estudo?
- c) Como organizar os dados de forma a ser mais fácil a sua comparação?

Tarefa proposta

Vamos acompanhar o crescimento do milho

Na turma decidiram levar a cabo um estudo sobre qual será o supermercado mais barato, das redondezas. Como o preço de um determinado produto não é suficiente para avaliar qual dos supermercados é mais barato, começaram por definir um "cabaz de compras" que iria servir para fazer a avaliação pretendida. Então fixou-se que o "cabaz de compras" seria constituído pelos seguintes produtos:

- 1 kg de batatas para cozer
- 1 kg de cebolas
- 1 kg de açúcar
- 1 litro de azeite extra-virgem com 0,7º de acidez
- 1 litro de leite meio-gordo (o mais barato do supermercado)

Para recolher a informação sobre os preços dos produtos do "cabaz de compras" os alunos pediram aos pais para os acompanharem ao supermercado e levaram lápis e papel para apontarem os preços, que levaram no dia seguinte para a escola. Com a ajuda do professor, pode ser construída uma tabela (idêntica à que se apresenta a seguir) onde registam os preços dos produtos que constituem o cabaz de compras, para cada um dos supermercados visitados (que representamos pelas letras A, B, C..., enquanto não soubermos os nomes dos supermercados):

Produto	Sup A	Sup B	Sup C	...	...	...	...
1 kg batatas							
1 kg cebolas							
1 kg açúcar							
1 litro azeite							
1 litro leite							
...							

Organizar os dados de forma a tirar algumas conclusões.

CARACTERÍSTICAS AMOSTRAIS. MEDIDAS de LOCALIZAÇÃO e DISPERSÃO

No capítulo anterior foram apresentados alguns processos para organizar a informação contida nos dados, utilizando tabelas e gráficos. Neste capítulo veremos outro processo de resumir a informação, através de algumas medidas calculadas a partir desses dados, a que se dá o nome de *estatísticas*. Destas medidas distinguiremos as medidas de localização, nomeadamente as que localizam o centro da distribuição de dados, e as medidas de dispersão, que medem a variabilidade dos dados.

3.1 Introdução

As tabelas e, principalmente, as representações gráficas permitem-nos identificar e comparar padrões subjacentes à distribuição dos dados. No entanto, sente-se desde logo a necessidade de traduzir a informação visual em "números": um "número" que seja representativo da ordem de grandeza dos valores da amostra, outro que revele o maior ou menor grau de dispersão dos dados, outro que dê informação acerca do enviesamento, etc. Estes "números" são sempre calculados a partir dos valores da amostra e designam-se por **características amostrais**. Mais geralmente, às medidas que resumem, através de números, a informação contida nos dados, dá-se o nome de "**estatísticas**".

De entre as muitas características amostrais de interesse, destacam-se a *média*, a *mediana*, a *moda* e os *percentis*, que são **características** (ou medidas) **de localização**, o *desvio padrão* e a *amplitude interquartis*, que são **características** (ou medidas) **de dispersão**.

Antes de apresentar as fórmulas de cálculo e as propriedades das principais características amostrais, necessitamos de introduzir algumas notações.

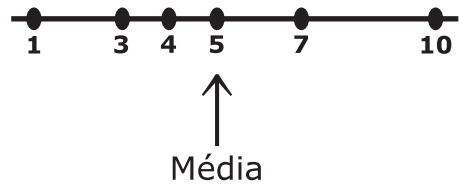
A dimensão da amostra será sempre representada pela letra **n**. A amostra será representada por uma lista, $(x_1,x_2,...,x_n)$, onde x_1 é o primeiro elemento da lista, x_2 é o segundo elemento da lista, e, assim por diante, até x_n, que é o último, ou n-ésimo, elemento da lista. Note-se que esta notação para representar a amostra não implica qualquer critério de ordenação.

Medidas de localização

Damos o nome de medida de localização a qualquer característica amostral que seja informativa da ordem de grandeza dos dados que surgem na amostra. Na maioria das vezes interessa-nos, unicamente, a localização da zona central da amostra, pois, em geral, é aí que se concentra a maior parte dos valores, mas pode também ser importante dar informação sobre a ordem de grandeza dos valores que surgem nas caudas. As medidas de localização central mais comuns são a média e a mediana.

3.2.1 Média

A média é a medida de localização central por excelência!... No seu cálculo intervêm todos os valores da amostra e não é mais do que o número que "equilibra" os grandes valores com os pequenos valores. É o centro de gravidade da distribuição dos dados. Se imaginarmos a recta real representada por uma vara sem peso e colocarmos massas unitárias nos pontos correspondentes aos valores que surgem na amostra, a média localiza-se no centro de gravidade deste objecto:



A **média** dos valores (1,3,4,5,7,10) é **5**, como facilmente se obtém

$$\frac{1+3+4+5+7+10}{6} = 5$$

e é esse o ponto onde o objecto físico representado se equilibra.

O maior óbice à utilização da média como um resumo indicador da localização da amostra, é o efeito de *contra-peso* que os valores extremos nela exercem. No exemplo anterior se, em vez de 10, tivéssemos 25, a média passaria de 5 para 7,5 (superior a todos os valores da amostra à excepção de um):

Se alguém nos disser que um conjunto de valores tem média 7,5, imaginamos que os valores se distribuem em volta do 7,5, aproximadamente metade de cada lado. Não pensamos num conjunto de valores em que todos, à excepção de um deles, são inferiores à média!

Efectivamente a média constitui um bom resumo dos dados nos casos em que estes se distribuem de forma aproximadamente simétrica, com uma zona central de maior concentração e caudas que não se alonguem demasiado. Quando a distribuição dos

dados não é aproximadamente simétrica, tem pouco interesse a utilização da média como centro da distribuição dos dados. Aliás, quando a distribuição dos dados não for aproximadamente simétrica é o próprio conceito de "centro da distribuição" que deixa de ter sentido.

De ora em diante, utilizaremos a notação $\overline{x}$ para representar a média da amostra $(x_1,x_2,\ldots,x_n)$:

$$\overline{x} = \frac{x_1 + x_2 \ldots + x_n}{n}$$

Voltemos de novo ao exemplo dos Dados sobre as casas, apresentado no Capítulo 1. Uma questão que pode ter interesse é saber quantas assoalhadas, em média, têm as casas que constam da tabela. Para obter esse valor basta somar o número de assoalhadas das 40 casas e dividir o resultado obtido por 40:

$$\overline{x} = \frac{3+3+3+3+5+2+2+4+2+2+3+3+4+\ldots+2+3+3+2+3+2+2+5+3+1+2}{40}$$

$$= 2{,}6.$$

E se além da média do número de assoalhadas, estivermos interessados na média das áreas, das casas observadas? O processo é precisamente o mesmo

$$\overline{x} = \frac{99 + 90{,}5 + 109 + 104{,}8 + \ldots + 154{,}2 + 75{,}9 + 90{,}2}{40}$$

$$= 102{,}19 \text{ m}^2$$

A média é uma medida muito importante na vida de um estudante. Durante os anos em que estiver a estudar será constantemente questionado sobre a sua média! Terá ainda que ter uma média de candidatura positiva (superior a 9,5) se pretender candidatar-se ao ensino superior... Convirá terminar um curso com uma média razoável, se pretender arranjar um emprego..., etc.

A média só pode ser calculada para dados quantitativos!

Quando a natureza da variável em estudo é qualitativa, acontece, por vezes, atribuir códigos numéricos às diferentes categorias. O cálculo da média desses códigos não tem, obviamente, qualquer sentido. Por exemplo, no caso dos Dados sobre casas, não tem qualquer sentido calcular a média das observações respeitantes à variável qualitativa *Estado*, que assume as categorias usada e nova, representadas respectivamente por 0 e 1.

Outro exemplo que surge com frequência é o seguinte: ao classificar um conjunto de pessoas, quanto ao sexo, é vulgar utilizar o número 1 para significar o sexo masculino e o número 2 para o sexo feminino. Assim, a amostra (2, 2, 1, 2, 2, 2, 2, 1, 2, 1, 1, 2, 1, 1, 2) representa um conjunto de 15 pessoas, classificadas quanto ao sexo, das quais 6 são do sexo masculino e 9 do sexo feminino. Obviamente que não tem qualquer sentido dizer que a média da amostra é 1,6, embora seja este o valor que se obtém fazendo a média daquelas valores! Neste exemplo, se utilizássemos o 0 para representar o sexo masculino e o 1 o sexo feminino já viria a média igual a 0,6!

Cálculo da média para dados discretos agrupados

Em amostras de **dados quantitativos discretos** aparecem muitos valores repetidos e, em vez de se somarem separadamente todos os valores da amostra, pode-se agrupar os valores que se repetem, vindo

$$\overline{x} = \frac{x_1^* n_1 + x_2^* n_2 + \ldots + x_k^* n_k}{n}$$

onde $x_1^*, x_2^*, \ldots, x_k^*$ representam os k valores distintos que surgem na amostra e n_i representa a frequência absoluta com que x_i^*, i=1, 2, ..., k, surge.

Por exemplo, para calcular a média do número de assoalhadas, podemos considerar a tabela de frequências com os dados agrupados, construída na secção 2.3.1,

N.º de Assoalhadas x_i^*	Freq. Abs. n_i	Freq. Rel. f_i
1	3	0,075
2	17	0,425
3	16	0,400
4	2	0,050
5	2	0,050
Total	40	1,000

e utilizá-la para calcular a média:

$$\overline{x} = \frac{1 \times 3 + 2 \times 17 + 3 \times 16 + 4 \times 2 + 5 \times 2}{40}$$

Sugestão – Verificar como é que se pode calcular a média, quando os dados estão agrupados, utilizando as frequências relativas, em vez de utilizar as frequências absolutas.

Cálculo da média para dados contínuos agrupados

Para dados quantitativos contínuos, já organizados em classes, utiliza-se a fórmula acima apresentada para calcular um valor aproximado para a média dos dados agrupados, sendo agora x_i^*, substituído por x'_i, o representante ou marca da i-ésima classe e n_i a respectiva frequência absoluta. O valor que se obtém para a média, quando os dados estão agrupados é, neste caso, um valor aproximado, já que não estamos a calcular a média com os verdadeiros valores. Assim, utilizando a tabela de frequências construída, na secção 2.4.1, para a variável *Área*

Classes	Rep. classe x'_i	Freq. Abs. n_i	Freq. Rel. f_i
[64, 81[	72,5	4	0,100
[81, 98[	89,5	14	0,350
[98, 115[	106,5	15	0,375
[115, 132[	123,5	4	0,100
[132, 149[	140,5	1	0,025
[149, 166[	157,5	2	0,050
Total		40	1,000

podemos obter um valor aproximado para a média das áreas:

$$\bar{x} \approx \frac{72{,}5 \times 4 + 89{,}5 \times 14 + 106{,}5 \times 15 + 123{,}5 \times 4 + 104{,}5 \times 1 + 157{,}5 \times 2}{40}$$

$$\approx 102{,}25 \text{ m}^2$$

O valor obtido para a média, considerando os dados agrupados, é uma boa aproximação do valor obtido quando se consideram todos os dados.

3.2.2 Mediana

A mediana é um valor que divide a amostra ao meio: metade dos valores da amostra são não superiores (menores ou iguais) à mediana e os restantes são não inferiores (maiores ou iguais) à mediana. Por outras palavras, até à mediana (inclusivé) está, pelo menos, 50% da amostra; para lá da mediana (inclusivé) está também, pelo menos, 50% da amostra.

Contrariamente com o que se passa com a média, o cálculo da mediana envolve um passo prévio de ordenação da amostra.

Como obter a mediana?

Para determinar a mediana é fundamental, como dissémos anteriormente, começar por ordenar os dados. Entretanto podem-se verificar duas situações, quanto à dimensão da amostra:

- Se a dimensão da amostra é ímpar, há um dos elementos da amostra ordenada que tem tantos elementos para a esquerda como para a direita. A título de exemplo, se a amostra tiver dimensão 11, o elemento na 6.ª posição tem 5 elementos da amostra para a sua esquerda e outros tantos para a sua direita. Esse elemento *central* da amostra será, neste caso, a mediana.
- Se a dimensão da amostra é par, não há nenhum elemento que tenha a propriedade de a dividir ao meio. Há dois valores *centrais* e define-se a mediana como sendo a média aritmética desses dois valores.

Repare-se que da forma como se calcula a mediana, quando a dimensão **n** da amostra é ímpar, a mediana é um elemento da amostra. Quando **n** é par, só será um elemento da amostra se os dois elementos centrais forem iguais.

Uma regra prática para obter a posição da mediana consiste em fazer o quociente $\frac{n+1}{2}$:

- Se este quociente for um número inteiro, o que se verifica quando n é ímpar, toma-se para mediana o elemento nessa posição;
- Se este quociente terminar em 0,5, o que se verifica quando n é par, considera-se a sua parte inteira e faz-se a semi-soma do elemento a que corresponde essa ordem, com o elemento da ordem seguinte.

Por exemplo, suponhamos que se pretende saber qual a mediana dos pesos (em kg) dos 15 alunos de uma turma do 2.º ano. Recolhida a informação sobre esses pesos, obtiveram-se os seguintes valores:

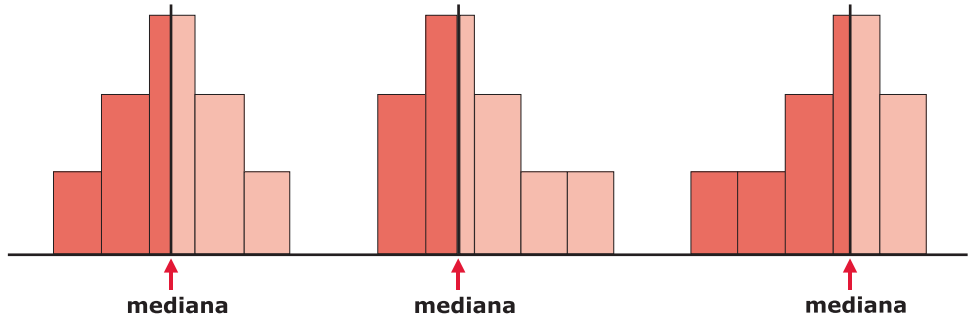
27 32 26 27 30 30 33 29 41 27 31 32 29 31 28

Para calcular a mediana é necessário começar por ordenar a amostra:

26 27 27 27 28 29 29 30 30 31 31 32 32 33 41

Então a mediana é o elemento na 8.ª posição $(\frac{15+1}{2})$, ou seja o 30. Se a amostra anterior tivesse só 14 elementos e o 41 não pertencesse à amostra, então a mediana seria a semi-soma dos elementos da 7.ª posição (parte inteira de $\frac{14+1}{2} = 7{,}5$) e da 8.ª posição, ou seja, 29,5 $(=\frac{29+30}{2})$.

Dado um histograma, é fácil obter a posição da mediana, pois esta está numa posição tal, que passando uma linha vertical por esse ponto, o histograma fica dividido em duas partes com áreas iguais, como se representa na figura seguinte:



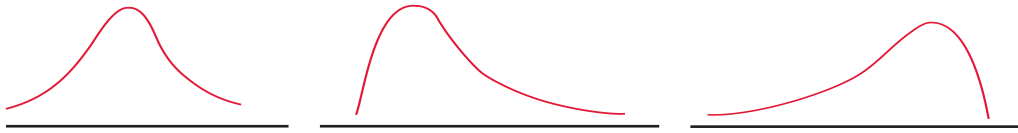
Ao contrário da mediana que "divide" o histograma em duas partes com áreas iguais, a média é o ponto de equilíbrio do histograma, em que se entra em linha de conta não só com a frequência das classes, mas também, com a distância a que estão do centro. Assim, na figura anterior, enquanto que no histograma do lado esquerdo, a média coincidirá com a mediana, no do centro, que apresenta um enviesamento para a direita, a média será "puxada" para a direita da mediana. Por outro lado, no histograma que apresenta o enviesamento para a esquerda, a média será "puxada" para a esquerda da mediana.

Como já referimos, a média, ao contrário da mediana, é uma medida muito pouco resistente, isto é, é muito influenciada por valores "muito grandes" ou "muito pequenos", mesmo que estes valores surjam em pequeno número na amostra. Estes valores, a que se dá o nome de *outliers*, são os responsáveis pela má utilização da média em muitas situações em que teria mais significado utilizar a mediana.

A mediana tem como principal desvantagem o facto de, no seu cálculo, só fazer intervir 1 ou 2 valores da amostra. No entanto, esta desvantagem transforma-se em vantagem, por comparação com a média, quando a distribuição da amostra é muito enviesada. A mediana é muito resistente e não é afectada pelos valores extremos.

Se tomarmos as duas amostras utilizadas na exemplificação das propriedades da média – (1,3,4,5,7,10) e (1,3,4,5,7,25) – facilmente se verifica que a mediana é igual a 4,5 para qualquer delas, enquanto que a média passou de 5 para 7,5!

Resumindo, como a média é influenciada quer por valores muito grandes, quer por valores muito pequenos, se a distribuição dos dados for enviesada para a direita (alguns valores grandes como *outliers*), a média tende a ser maior que a mediana; se for aproximadamente simétrica, a média aproxima-se da mediana e se for enviesada para a esquerda (alguns valores pequenos como *outliers*), a média tende a ser inferior à mediana. Representando as distribuições dos dados (esta observação é válida para as representações gráficas na forma de diagrama de barras ou de histograma) na forma de uma mancha, temos, de um modo geral (Graça Martins, 2005):



média ≈ mediana **média > mediana** **média < mediana**

Observe-se que o simples cálculo da média e da mediana nos pode dar informação sobre a forma da distribuição dos dados.

No estudo de dados qualitativos ordinais (isto é, onde se pode considerar uma ordem subjacente à categorias) faz sentido indicar a categoria mediana. A categoria mediana é aquela onde, pela primeira vez, a frequência relativa acumulada atinge ou ultrapassa os 50%. Esta mesma definição serve para identificar a classe mediana no caso de se estar perante dados agrupados.

Consideremos o exemplo apresentado para trabalhar na sala de aula, através da tabela da página 87, mais precisamente a variável *Número de irmãos*. Admitamos que os dados estavam organizados na forma de uma tabela de frequências, como se apresenta a seguir:

N.º de irmãos	Freq. Abs.	Freq. Rel.%	Freq. Rel. Acum.%
0	3	15	15
1	11	55	**70**
2	3	15	85
3	2	10	95
4	1	05	100
	20	100	

A mediana do *Número de irmãos* é igual a 1, já que a frequência relativa acumulada ultrapassa os 50% quando se soma a frequência correspondente à classe 1.

Consideremos ainda, da tabela da página 87, a variável *Tempo de casa à escola*. Organizados os dados na forma de uma tabela de frequências, como a sugerida na página 96,

Tempo de casa à escola (minutos)	Freq. Abs.	Freq. Rel. (%)	Freq. Rel. Acum.%
Até 10	4	20	20
de 10 a 20	7	35	55
de 20 a 30	6	30	85
de 30 a 40	3	15	100
Total	20	100	

pretende-se obter a mediana. Neste caso a classe mediana é a classe constituída pelos valores maiores ou iguais a 10 minutos e menores de 20 minutos, uma vez que a frequência relativa desta classe, adicionada à frequência acumulada que vem da classe anterior, faz com que se ultrapasse os 50%.

Suponhamos agora um outro exemplo, mas com uma situação especial, como a que se apresenta na tabela de frequências seguinte, correspondente à variável *Número de assoalhadas* do exemplo Dados sobre casas:

N.º de Assoalhadas x_i^*	Freq. Abs. n_i	Freq. Rel. f_i	Freq. Abs. Acum.	Freq. Rel. Acum. %
1	3	0,075	3	7,5
2	17	0,425	20	**50**
3	16	0,400	36	90
4	2	0,050	38	95
5	2	0,050	40	100
Total	40	1,000		

Temos uma classe a que corresponde exactamente 50% de frequência acumulada! Isto é, 50% dos elementos da amostra são menores ou iguais a 2 e os outros 50% são maiores ou iguais a 3. Esta situação só pode ocorrer quando o número de dados é par, e como vimos anteriormente, neste caso, a mediana é a semi-soma dos dois elementos centrais. Assim, neste caso, a mediana será 2,5.

Exemplo:

Salários dos trabalhadores - Os salários dos 160 trabalhadores de uma determinada empresa, com 6 níveis de salários, distribuem-se de acordo com a seguinte tabela de frequências:

Salário (euros)	400	450	600	700	1000	5000
N.º empregados	23	58	50	20	7	2

Calcule a média e a mediana e comente os resultados obtidos.

Resolução:

Cálculo da média: $\bar{x} = (400 \times 23 + 450 \times 58 + 600 \times 50 + 700 \times 20 + 1000 \times 7 + 5000 \times 2)/160$
≈ 602 euros

Cálculo da mediana: Considerando na tabela anterior as frequências relativas acumuladas, temos

Salário (euros)	400	450	600	700	1000	5000
N.º empregados	23	58	50	20	7	2
Freq. Rel. %	14,38%	36,25%	31,25%	12,50%	4,38%	1,25%
Freq. Rel. Acum. %	14,38%	**50,63%**	81,88%	94,38%	98,75%	100,00%

Então a mediana é igual a 450 euros.

Repare-se que a média é muito superior à mediana, o que acontece sobretudo devido aos 2 salários substancialmente superiores aos restantes, eventualmente dos administradores, que inflacionaram a média. Efectivamente, dos 160 trabalhadores, só 29 é que têm um salário superior à média.

A mediana dá-nos uma ideia mais correcta do nível dos salários, que são de um modo geral baixos. Assim, dá-nos a indicação de que 50% dos salários são menores ou iguais a 450 euros, enquanto que os restantes são maiores ou iguais àquele valor.

Suponha que no cálculo do salário médio dos trabalhadores da empresa, retirava os dois supostos administradores, com salários de 5000 euros, cada um. A média dos 158 trabalhadores restantes desce de 602 euros para 546 euros. Este exemplo é, mais uma vez, elucidativo do cuidado que é necessário ter com a interpretação da média. Esta é uma medida muito pouco resistente, isto é que "não resiste" a valores muito grandes ou muito pequenos, quando comparados com os restantes, sendo muito inflacionada por eles. Um valor grande provoca um "aumento" da média, assim como um valor pequeno provoca uma "diminuição" da média. Quando o nosso conjunto de dados tiver destes valores extremos, denominados de *outliers*, convém utilizar a mediana, como medida de localização do centro da distribuição dos dados. Vejamos ainda o seguinte exemplo.

Exemplo

Velocidade média – Em determinado dia e em determinado ponto da autoestrada, a polícia registou a velocidade (média) dos 5 primeiros carros que passaram após as 10 horas. Calculou a média das velocidades desses 5 carros e obteve 130 km (por hora). Embora a velocidade máxima permitida fosse 120 km (por hora), só autuou um dos carros! Na realidade as velocidades registadas foram 120 km, 115 km, 120 km, 110 km e 185 km, pelo que só um ultrapassou a velocidade máxima permitida.

3.2.3 Quartis

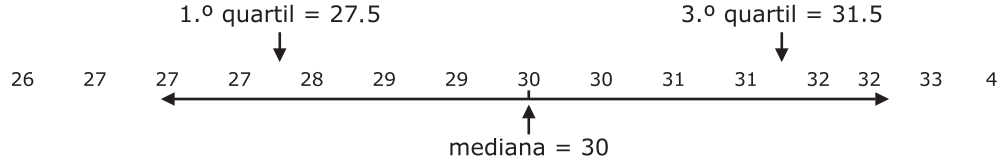
Os quartis, já utilizados anteriormente na construção do diagrama de extremos e quartis, são outras medidas de localização de alguns pontos de interesse, que não o centro da distribuição dos dados. Como vimos na definição da mediana, esta divide a amostra em duas partes com igual percentagem de elementos. Considerando cada uma destas partes e calculando a sua mediana, obteremos os quartis. Assim, a mediana e os quartis localizam pontos que dividem a distribuição dos dados em 4 partes com igual percentagem de elementos.

Há vários processos para calcular os quartis, nem todos conducentes aos mesmos valores, mas a valores aproximados. A metodologia que, a este nível, recomendamos para os obter é a seguinte:

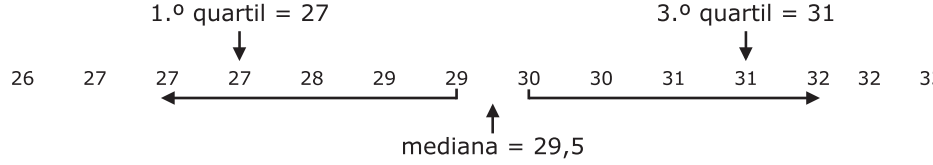
- Ordenar os dados e calcular a mediana Me;
- O 1.º quartil, Q_1, é a mediana dos dados que ficam para a esquerda de Me;
- O 3.º quartil, Q_3, é a mediana dos dados que ficam para a direita de Me.

Ao calcular os quartis pelo processo anterior, podem-se levantar algumas dúvidas, no caso em que a dimensão da amostra é ímpar. Efectivamente, neste caso a mediana coincide com um dos elementos da amostra e poderíamos optar por considerá-lo incluído nas duas metades em que fica dividida a amostra, ou não o considerar em nenhuma das metades. A nossa opção é considerá-lo pertencente às duas metades.

Consideremos de novo o exemplo utilizado para ilustrar o cálculo da mediana, dos pesos dos 15 alunos de uma turma do 2.º ano, já ordenados:



Como a mediana é um dos elementos da amostra, para o cálculo dos quartis, considerámos dois conjuntos de dados, cada um com 8 elementos, onde incluímos a mediana. Se a amostra inicial só tivesse 14 elementos, sem o valor 41, então teríamos:



Neste caso a mediana divide a amostra em duas partes de 7 elementos, cada uma, e, do mesmo modo que anteriormente, os quartis serão as medianas dessas partes.

Os quartis são medidas de localização com algum interesse prático, pois permitem localizar os 50% dos valores centrais dos dados e como veremos, são utilizados para definir uma medida de dispersão ou variabilidade desses dados.

3.2.4 Percentis

Os percentis de que a mediana e os quartis são casos particulares, são medidas de localização com grande interesse, nomeadamente para avaliar a posição relativa dos dados. Por exemplo, suponha que uma mãe vai, com o seu bebé de 6 meses, à consulta de rotina, do pediatra. Este, depois de pesar e medir a criança, consulta umas tabelas e só nessa altura comenta com a mãe, o estado de crescimento do seu filho. Pode acontecer que alguns dos seus comentários sejam desta forma:

- – Minha senhora, o seu filho, no que diz respeito ao peso, está no *percentil* 90. Vamos ter que ter algum cuidado!

Afinal o que significa o percentil 90? Significa que 90% das crianças com 6 meses têm um peso menor ou igual ao do bebé e só 10% têm um peso maior ou igual!

De um modo geral define-se **percentil P** de um conjunto de dados, como sendo o valor que tem **P%** dos dados menores ou iguais a ele, e os restantes maiores ou iguais. O 1.º e o 3.º quartis também são conhecidos como percentil 25 e 75, respectivamente. Analogamente, a mediana é o percentil 50.

Exemplo:

A obesidade é um problema – A comunicação social tem alertado a opinião pública para o problema da obesidade, nomeadamente nas crianças. Então, como é que poderemos saber se o nosso filho está obeso? Como é que o médico, além da sua experiência, sossega a mãe sobre a saúde e bem estar do seu filho? Existem tabelas, que apresentam, para cada idade, os valores dos percentis para as variáveis "peso" e "altura". A tabela seguinte, que se retirou da Internet, apresenta, para os vários meses de idade, valores adequados entre os quais deve estar o peso (em kg) da criança. Estes valores poderiam ser, por exemplo, os percentis 25 e 75, considerando-se um "peso normal" o que esteja nos 50% dos pesos centrais, quando se considera o conjunto dos pesos dos bebés (da população que se está a estudar, quer seja portuguesa, italiana, inglesa, alemã, etc.) com determinada idade:

	Ao nascer	1 mês	2 meses	3 meses	4 meses	5 meses	6 meses	7 meses	8 meses
Mínimo	2.750	3.500	4.000	4.750	5.500	6.000	6.500	7.000	7.500
Máximo	4.000	5.000	6.000	7.000	7.800	8.500	9.000	9.700	10.000
	9 meses	10 meses	11 meses	1 ano	1 ano 1m	1 ano 2m	1 ano 3m	1 ano 4m	1 ano 5m
Mínimo	7.900	8.300	8.500	8.800	9.000	9.250	9.500	9.700	9.800
Máximo	10.500	10.900	11.250	11.500	11.800	12.000	12.400	12.600	12.800
	1 ano 6m	1 ano 7m	1 ano 8m	1 ano 9m	1 ano 10m	1 ano 11m	2 anos	2 anos 1m	2 anos 2m
Mínimo	10.000	10.150	10.300	10.500	10.600	10.700	10.900	11.000	11.200
Máximo	13.000	13.300	13.600	13.800	14.000	14.200	14.500	14.650	14.800
	2 anos 3m	2 anos 4m	2 anos 5m	2 anos 6m	2 anos 7m	2 anos 8m	2 anos 9m	2 anos 10	2 anos 11m
Mínimo	11.300	11.500	11.600	11.750	11.900	12.000	12.100	12.250	12.400
Máximo	15.000	15.250	15.500	15.700	15.900	16.000	16.300	16.500	16.750
	3 anos	3 anos 4m	3 anos 8m	4 anos	4 anos 4m	4 anos 8m	5 anos		
Mínimo	12.600	13.200	13.750	14.300	15.000	15.500	16.000		
Máximo	17.000	17.700	18.500	19.300	20.200	21.000	21.800		

A partir da tabela anterior, concluímos que um peso razoável, nem muito magro, nem muito gordo, para um bebé de 2 anos e meio, será um peso compreendido no intervalo [11,750kg, 15,700kg].

Exemplo:

Conversa entre mãe e filho – Imagine a seguinte conversa entre uma mãe e o seu filho de 15 anos.
Filho - Mãe, tive 14 no teste de Biologia!
Mãe – E então isso é bom ou nem por isso?
Filho – Como assim? Digo que tive 14 e ainda me perguntas se isso é bom?
Mãe – Pois, pergunto. E até pergunto a que percentil é que corresponde essa nota?
Filho – Mas o que é isso de percentil? Não sei do que estás a falar!
Mãe – Quantos alunos na tua escola fizeram esse teste?
Filho – Foram 100, porquê?
Mãe – E quantos tiveram nota maior que 14?
Filho – Bom, não vi bem, mas parece-me que foram uns 80!
Mãe – Afinal, não tens razão para estar tão satisfeito! Ficaste no percentil 20. Só 20% dos teus colegas tiveram nota menor ou igual à tua. Esse exame foi mesmo muito fácil.

Exemplo:

Nota mínima de acesso – Uma Universidade pretendia estabelecer uma nota mínima de acesso para a prova específica de Matemática. Estava, no entanto, com o seguinte problema: se a prova fosse muito difícil, como tinha sido nos anos anteriores, corria o risco de não ter alunos, ou ter muito poucos, com nota maior ou igual a 95 (numa escala de 0 a 200) e ficar com as vagas por preencher. Então o Conselho Directivo tomou a seguinte decisão. Independentemente da distribuição que se vier a verificar para as notas no exame de Matemática, fixaram como nota mínima aquela que permita que 55% dos alunos que realizarem o exame, se possam candidatar. Com esta decisão, a nota mínima de acesso não é necessariamente positiva.
Nota – Este exemplo não é ficção e foi a metodologia seguida durante alguns anos pelo Conselho de Reitores das Universidades Portuguesas (CRUP) e outras instituições de Ensino Superior, na definição da nota mínima de acesso, como refere o Decreto-Lei que se trancreve, em parte, a seguir:

ENSINO SUPERIOR PÚBLICO

Decreto-Lei n.º 296-A/98, (alíneas a) e c) do artigo 24.º) de 25 de Setembro, alterado pelo Decreto-Lei n.º 99/99, de 30 de Março

CLASSIFICAÇÃO MÍNIMA NAS PROVAS DE INGRESSO

I. Recomendação do CRUP

1. Para candidatura aos pares estabelecimento/curso que adoptaram a Recomendação do CRUP no tocante à fixação da classificação mínima prevista na alínea a) do artigo 24.º do Decreto-Lei n.º 296-A/98, de 25 de Setembro, alterado pelo Decreto-Lei n.º 99/99, de 30 de Março, os candidatos devem obter no exame nacional de cada uma das provas de ingresso exigidas para o curso superior a que se candidatam, classificação não inferior a 95 pontos na escala de 0 a 200.

2. Se, excluídos os casos de classificação igual a zero pontos, o número de examinandos com classificação igual ou superior a 95 pontos em determinado exame nacional de prova de ingresso for inferior a 55% do número total, o valor da classificação mínima é aquele que permita a admissão ao concurso, por esta via, de 55% dos examinandos.

3. A regra é aplicada a cada chamada de cada exame.

II. Recomendação do CCISP 96

1. Para candidatura aos pares estabelecimento/curso que adoptaram a Recomendação do CCISP 96 no tocante à fixação da classificação mínima prevista na alínea a) do artigo 24.º do Decreto-Lei n.º 296-A/98, de 25 de Setembro, alterado pelo Decreto-Lei n.º 99/99, de 30 de Março, os candidatos devem obter no exame nacional de cada uma das provas de ingresso exigidas pelo curso superior a que se candidatam, classificação não inferior a 95 pontos na escala de 0 a 200.

2. Se, excluídos os casos de classificação igual a zero pontos, na escala de 0 a 200, o número de examinandos com classificação igual ou superior a 95 pontos for inferior a 75% do número total, o valor da classificação mínima é aquele que permita a admissão ao concurso, por esta via, de 75% dos examinandos.

3. A regra é aplicada a cada chamada de cada exame.

III. Recomendação do CCISP 97

1. Para candidatura aos pares estabelecimento/curso que adoptaram a Recomendação do CCISP 97 no tocante à fixação da classificação mínima prevista na alínea a) do artigo 24.º do Decreto-Lei n.º 296-A/98, de 25 de Setembro, alterado pelo Decreto-Lei n.º 99/99, de 30 de Março, os candidatos devem obter no exame nacional de cada uma das provas de ingresso exigidas pelo curso superior a que se candidatam, classificação não inferior a 95 pontos, na escala de 0 a 200.

2. Se, excluídos os casos de classificação menor ou igual a 14 pontos na escala de 0 a 200, o número de examinandos com classificação igual ou superior a 95 pontos for inferior a 75% do número total, a classificação mínima é aquela que permita a admissão ao concurso, por esta via, de 75% dos examinandos.

3. A regra é aplicada a cada chamada de cada exame.

A Rede do Ensino Superior

em Portugal / 2001
306

IV. Percentil 50

1. Para candidatura aos cursos da Faculdade de Ciências da Universidade do Porto, os candidatos devem obter no exame nacional de cada uma das provas de ingresso exigidas para o curso superior a que se candidatam, classificação não inferior a 100 pontos na escala de 0 a 200.

2. Se, excluídos os casos de classificação igual a zero pontos, na escala de 0 a 200, o número de examinandos com classificação igual ou superior a 100 pontos for inferior a 50% do número total, o valor da classificação mínima é aquele que permita a admissão ao concurso, por esta via, de 50% dos examinandos.

3. A regra é aplicada a cada chamada de cada exame.

3.2.5 Moda

Em amostras de dados qualitativos dá-se o nome de *categoria modal* ou, também, simplesmente, *moda*, à categoria de maior frequência. Esta mesma definição é usualmente utilizada também para dados quantitativos (discretos ou contínuos), ou seja, define-se moda como sendo a classe de maior frequência (dados agrupados). No entanto, neste caso, de dados quantitativos, tem vindo a ter cada vez maior interesse generalizar a definição de moda. Assim, vamos designar por moda qualquer classe que esteja ladeada por classes de menor frequência. As modas serão, pois, os "picos" na distribuição de frequências.

***Exemplo**:*

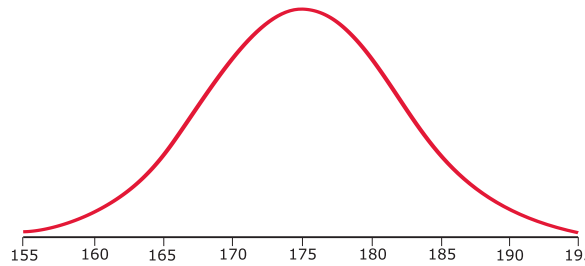
Virose desconhecida (Graça Martins, 1999) – Suponha que numa região começaram a aparecer pessoas com uma virose desconhecida. Os médicos do Centro de Saúde dessa região procuraram recolher alguma informação sobre as pessoas atacadas por essa virose. Foi recolhida uma amostra de 35 desses doentes a quem se perguntou, entre outras características, a idade. Depois de analisados os dados os médicos foram informados que a idade média dos doentes era de 32 anos. Um dos médicos, mais curioso que os outros, pediu que lhe mostrassem a distribuição dos dados, tendo-lhe sido apresentada a seguinte distribuição num gráfico de caule-e-folhas:

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

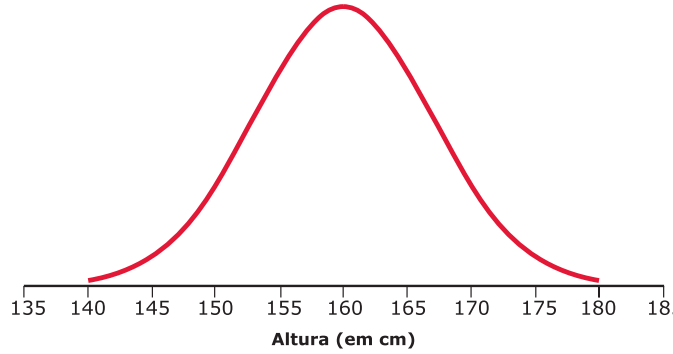
Perante a representação anterior, com duas modas, o médico não teve dúvidas em pôr de parte a média, assim como qualquer outra medida de localização do centro da amostra. Efectivamente, para dados deste tipo é enganador qualquer medida de localização do centro da distribuição. O que o médico concluiu imediatamente foi que a doença ataca crianças e pessoas na 3.ª idade.

Não sendo propriamente uma medida de localização, a moda deve a sua importância ao facto de ser a única medida que é susceptível de ser calculada para os dados qualitativos, em que não se possa estabelecer uma hierarquia entre as diferentes modalidades ou classes, que a variável possa assumir.

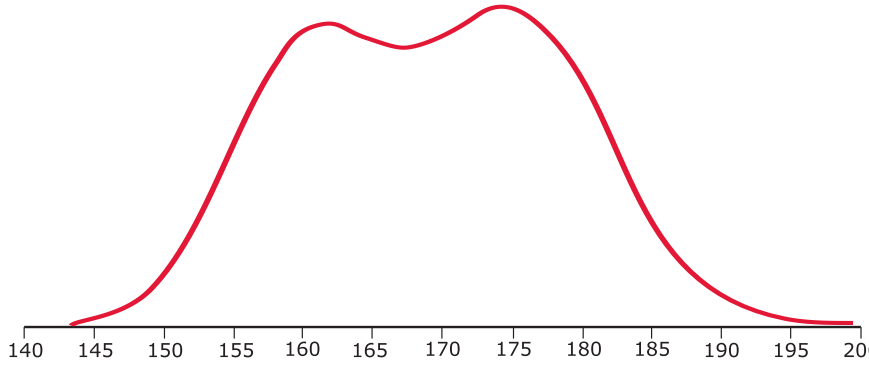
Em curvas que modelam muitas situações da vida real, dá-se o nome de moda a qualquer máximo relativo da curva de densidade. Os modelos teóricos de interesse têm uma única moda e é usual dizer que o aparecimento de várias modas pode evidenciar mistura de populações. Para ilustrar esta ideia, tome-se o exemplo das alturas na população portuguesa. Se considerarmos somente a subpopulação dos homens, a distribuição das suas alturas não deve afastar-se muito do seguinte padrão:



Note-se que a zona de maior concentração ou densidade, está entre 1,70m e 1,80m, sendo a moda (máximo relativo da curva) igual a 1,75m. A forma da distribuição das alturas das mulheres deverá ser idêntica, mas localizada em torno de 1,60m:

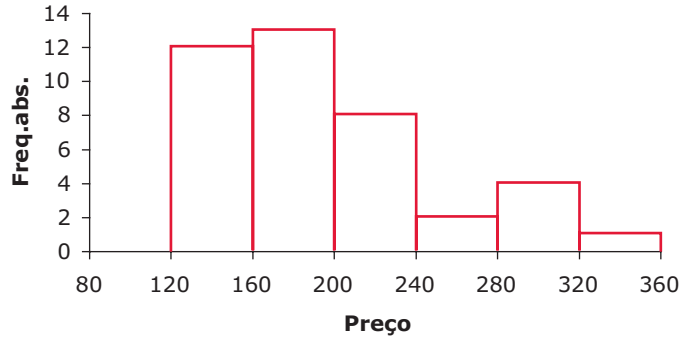


Que aconteceria se considerássemos as duas subpopulações em conjunto? Onde ficaria a moda? Em 1,75m, em 1,60m ou algures entre estes dois valores? Na verdade o que acontece é que surgem duas modas!... Uma, um pouco à direita de 1,60m e outra, um pouco à esquerda de 1,75m:



A bimodalidade torna-se ainda mais evidente se a zona central de uma das distribuições se encontrar muito afastada da zona central da outra e se a percentagem de observações pertencentes a cada uma das duas subpopulações for idêntica. Retomando o exemplo das alturas, se numa amostra de 100 indivíduos tivermos 10 mulheres e 90 homens é muito pouco provável que o histograma apresente bimodalidade, contrariamente ao que deverá ocorrer em amostras com 50 homens e 50 mulheres.

Considere-se o exemplo dos Dados sobre casas, do capítulo 1. No parágrafo 2.4.2 apresentámos um histograma construído para a variável *Preço*, que tinha o seguinte aspecto:



Histograma para a variável *Preço* das casas do ficheiro Dados sobre Casas

Este histograma apresenta duas classes modais! Uma delas é a classe dos 160 aos 200 mil euros, e a outra é a classe que vai de 280 a 320 mil euros. Olhando para as características das casas, podem apontar-se algumas possíveis causas para esta bimodalidade: há casas novas e casas usadas; há casas com garagem e casas sem garagem; as casas não são todas da mesma zona e pode haver alguma zona onde, em média, as casas são mais caras que nas outras duas zonas!... É claro que, como a frequência da segunda classe modal é relativamente baixa, pode-se ainda argumentar que a bimodalidade não é devida a uma mistura de populações mas sim "obra do acaso"!... Ainda a propósito deste exemplo, chamamos mais uma vez a atenção, para o facto de o histograma ser uma representação gráfica que, para alguns conjuntos de dados, pode mudar sensivelmente de aspecto, quando se altera a amplitude de classe ou o ponto onde se começam a construir as classes. Assim, para o mesmo conjunto de dados pode acontecer haver representações gráficas diferentes, nomeadamente em termos do número de modas.

Observação:

Quando se pretende saber qual o *centro de uma distribuição de dados*, a resposta a esta pergunta é fácil se a distribuição for aproximadamente simétrica e unimodal (só com uma moda). Se a distribuição dos dados apresentar outras formas, nomeadamente enviesamento ou várias modas, já o conceito de centro da distribuição dos dados pode não fazer qualquer sentido, como já referimos anteriormente ao tratarmos das medidas de localização.

Tarefa

Vamos pesar laranjas (cont.)

Considerando, de novo, a Tarefa - Vamos pesar laranjas, do capítulo 2, pretende-se agora obter a média, mediana e os quartis da distribuição dos dados e construir o diagrama de extremos e quartis.

A partir da representação em caule-e-folhas, que entretanto se fez, é fácil de obter os dados ordenados, pois basta percorrer os caules, de cima para baixo, juntando-lhe as folhas respectivas:

1.ª	2.ª	3.ª	4.ª	5.ª	6.ª	7.ª	8.ª	9.ª	10.ª	*11.ª*	12.ª	13.ª	14.ª
133	134	137	138	139	140	141	142	144	145	*146*	147	148	148
15.ª	16.ª	17.ª	18.ª	19.ª	20.ª	**21.ª**	**22.ª**	23.ª	24.ª	25.ª	26.ª	27.ª	28.ª
149	150	151	151	151	151	**152**	**152**	153	153	154	154	156	157
29.ª	30.ª	31.ª	*32.ª*	33.ª	34.ª	35.ª	36.ª	37.ª	38.ª	39.ª	40.ª	41.ª	42.ª
160	162	163	*164*	164	166	167	168	168	172	172	174	175	176

Como o número de dados é par, a mediana é a semi-soma dos dados que se encontram nas posições 21.ª e 22.ª, ou seja

$$\text{Mediana} = \frac{152 + 152}{2} = 152$$

Para determinar os quartis, vamos considerar as medianas de cada uma das partes em que ficaram divididos os dados, pela mediana: o 1.º quartil será a mediana dos dados que estão nas posições de 1 a 21, enquanto que o 3.º quartil será a mediana dos dados que estão nas posições de 22 a 42. Como agora temos um número ímpar de dados, a mediana será o elemento do meio. Assim, temos:

1.º quartil = 146 (elemento na 11.ª posição)

3.º quartil = 164 (elemento na 32.ª posição)

Para construir o diagrama de extremos e quartis, necessitamos de 5 números, obtidos a partir dos dados: mínimo, máximo, 1.º quartil, 3.º quartil e mediana:

Mínimo = 133
Máximo = 176
1.º quartil = 146
3.º quartil= 164
Mediana = 152

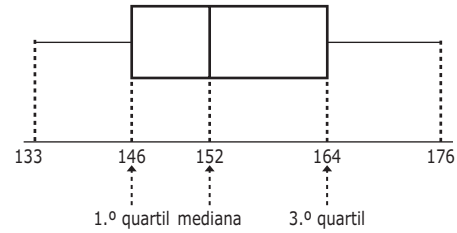


Diagrama de extremos e quartis para a variável *Peso* das laranjas

Desta representação gráfica, sobressai a simetria aproximada dos dados, como também já tínhamos visto com a representação em caule–e-folhas ou com o histograma. Calculando a média obtém-se o valor 154, um pouco superior à mediana, como se esperava pelo pequeno enviesamento para a direita, apresentado pelos dados.

Na Sala de Aula

Tarefa

O melhor é dar a cada um a média!

Os 3 netos da avó Maria, Huguinho, Zezinha e Luisinha, queriam ir à feira popular, mas não tinham dinheiro. Então decidiram ir ter com a avó, para ver se esta "subsidiava" os seus divertimentos! Até parece que não sabiam que a avó era um bocadinho "agarrada" ao dinheiro... Mas, por estranho que pareça, ela estava "muito" benevolente e disse aos netos para cada um procurar uns trocos nos bolsos dos 2 casacos e da saia que tinha vestido ultimamente. O Huguinho encontrou num casaco 8 euros, a Luisinha encontrou 2 euros no outro casaco e finalmente a Zezinha encontrou na saia 5 euros.

A avó, que embora fosse um bocadinho "agarrada" ao dinheiro, era justa, não achava bem que cada neto ficasse com a quantia que encontrou e gostaria de contentar todos de igual modo. Como fazer?

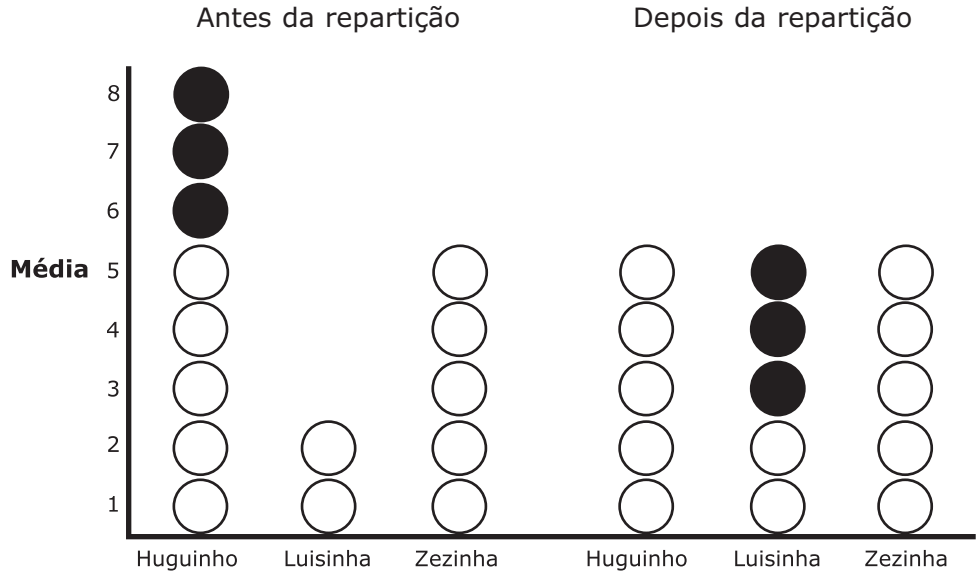
Como fazer, para cada neto ficar com igual quantia?

Uma proposta foi juntar o dinheiro todo e reparti-lo igualmente pelos 3, ou seja, calcular a **média** das quantias 8, 2 e 5. Assim, decidiu que cada um ficaria com a seguinte quantia:

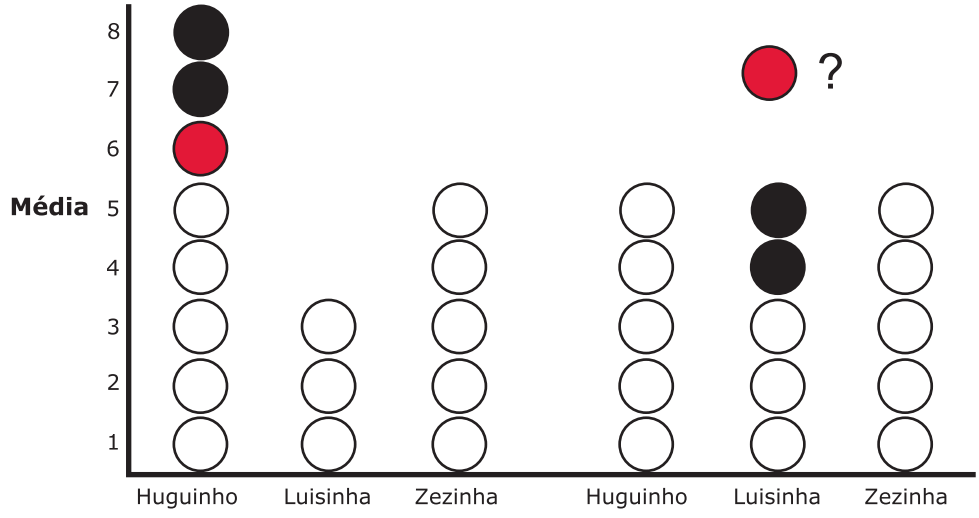
$$\frac{8 + 2 + 5}{2} = 5$$

Afinal basta o Huguinho dar 3 euros à Luisinha e cada um fica com 5 euros!

Esta situação pode ser apresentada graficamente, da seguinte forma, em que cada bola representa uma moeda de um euro:

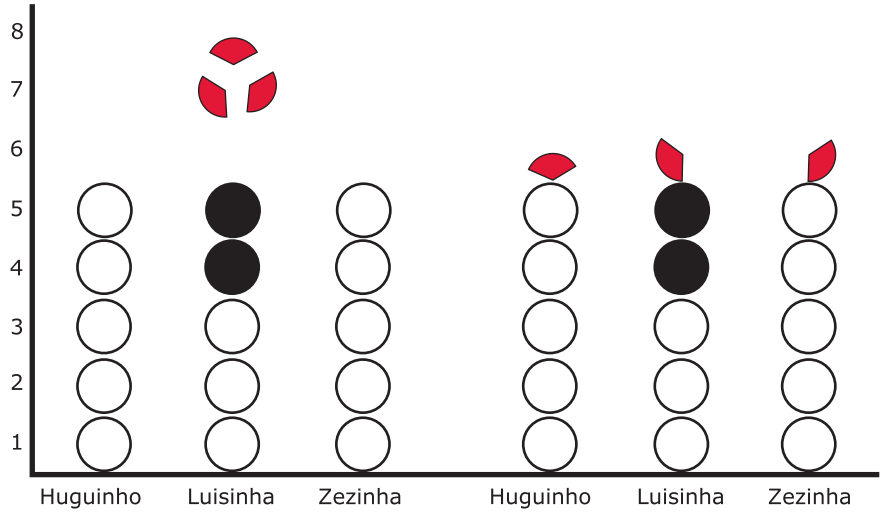


Uma questão que pode ser levantada por um aluno é, por exemplo, a seguinte: e se a Luisinha tivesse encontrado 3 euros em vez de 2 euros? Como é que resolvíamos a situação?



O Huguinho deu 2 euros à Luisinha, que ficou com a mesma quantia que a Zezinha, mas ainda sobrou 1 euro. Para ficarem os 3 com quantias iguais, teríamos de dividir o euro em 3 partes iguais e dar a cada um, uma dessas partes. Só assim é que cada um ficava com uma quantia igual, ou seja a média.

Se em vez de euros, tivessemos bolos, seria mais fácil dividir o bolo sobrante em 3 partes iguais e dar a cada um uma das partes:



Embora o conceito de média seja um conceito a desenvolver sobretudo ao nível do 2.º ciclo, este exemplo pode servir para o professor ter uma conversa com os alunos sobre o significado da média, que, em certas situações, pode não passar de um conceito abstracto, não possível de ser materializado.

Por exemplo, se na turma o professor perguntar a cada um dos alunos quantos irmãos tem e calcular a média dos valores registados, é natural que obtenha um valor não inteiro. Se obtiver o valor 1,6, como podemos interpretá-lo? O professor pode incentivar os alunos a registar os valores obtidos num diagrama de pontos e verificarem que a maior concentração de valores se regista à volta do 1 e do 2 (estamos a admitir que na turma nenhum aluno tem um número de irmãos substancialmente maior que os outros alunos, que provocasse uma inflação na média...). Pode-se dar ainda como exemplo a informação fornecida pelo Instituto Nacional de Estatística sobre o número médio de filhos das famílias portuguesas.

Tarefa

Vamos comer queijo, mas não exageremos...

O queijo, proveniente do leite, é um alimento rico em cálcio. No entanto, é necessário não abusar, já que, de um modo geral, é um alimento muito calórico e a maior parte das vezes rico em gordura. Na tabela seguinte apresentamos, para vários tipos de queijo, a quantidade de gordura e o número de calorias, por cada 100 gramas de queijo:

Alimento (100g)	Gordura *(g)*	Calorias
■ Queijo Brie	20	263
▲ Queijo Camembert	23	313
▲ Queijo da Ilha	26	357
▲ Queijo da Serra curado	32	385
▲ Queijo da Serra fresco	27	327
▲ Queijo de Azeitão	25	309
▲ Queijo de Évora	34	412
▲ Queijo de Serpa	26	330
▲ Queijo de Tomar	27	305
● Queijo flamengo 20%	8	185
■ Queijo flamengo 30%	14	246
▲ Queijo flamengo 45%	23	315
■ Queijo fresco	21	265
▲ Queijo Gorgonzola	37	407
■ Queijo Gruyère	20	315
▲ Queijo Parmesão	28	401
▲ Queijo Roquefort	32	371
▲ Queijo Suíço	29	357

- ● – Alimento com baixo teor em gordura mas podendo ter um elevado conteúdo em calorias.
- ■ – Alimento intermediário: consumir com moderação.
- ▲ – Alimento rico em gordura: comer pontualmente ou moderar o seu consumo.

A tabela anterior permite vários estudos no que diz respeito à quantidade de gordura e ao número de calorias dos diferentes tipos de queijo. Uma possível abordagem é começar por considerar os dados respeitantes à quantidade de gordura por cada 100 gramas de queijo e organizá-los na forma de um gráfico de caule-e-folhas. Uma pergunta que esta representação gráfica nos poderá imediatamente responder é a existência de algum possível enviesamento e, caso afirmativo, o que se espera para a relação de grandeza entre a média e a mediana?

Organizando os dados num gráfico de caule-e-folhas, obteve-se a seguinte representação:

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

A representação anterior apresenta algum enviesamento para a esquerda, pelo que se espera que a média seja inferior à mediana.

No que diz respeito à média, tem-se:

$$\text{Média} = \frac{8+14+20+20+21+23+23+25+26+26+27+27+28+29+32+32+34+37}{18}$$

$$= 25{,}1 \text{ gramas}$$

Para calcular a mediana é necessário considerar a amostra ordenada, que facilmente se obtém a partir da representação em caule-e-folhas. Como temos 18 dados, a mediana é a semi-soma dos dados que estão na posição 9 e 10, que ocupam os 2 pontos centrais dos dados ordenados:

$$\text{Mediana} = \frac{\text{dado na posição 9} + \text{dado na posição 10}}{2} = \frac{26+26}{2} = 26 \text{ gramas}$$

A média é um pouco inferior à mediana, como se esperava. Esta situação é possivelmente provocada pelo valor 8 que é bastante menor que os restantes, influenciando a média.

Uma questão que pode ainda ser respondida é a de saber quais os queijos menos gordos e os mais gordos. Podemos, por exemplo procurar saber quais os 25% dos queijos menos gordos e quais os 25% dos queijos mais gordos, calculando os quartis.

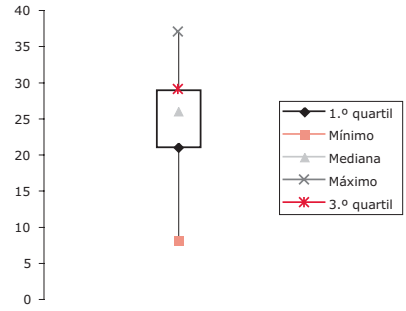
Para determinar os quartis, consideramos as duas partes em que os dados ficam divididos pela mediana e determina-se a mediana de cada uma dessas partes:

8 14 20 20 21 23 23 25 26 26 27 27 28 29 32 32 34 37

1.º quartil = 21 mediana = 26 3.º quartil = 29

Podemos dizer que os 25% dos queijos menos gordos, são os que têm uma quantidade de gordura menor ou igual a 21 g por cada 100 g de queijo, ou sejam: Brie, Flamengo 20%, Flamengo 30%, Fresco e Gruyère. Analogamente, os 25% dos queijos mais gordos, são que têm quantidade de gordura superior ou igual a 29 g, por cada 100 gramas de queijo e que são: Serra Curado, Évora, Gorgonzola, Roquefort e Suíço.

Uma vez que temos calculados os quartis e a mediana, vamos construir o diagrama de extremos e quartis:

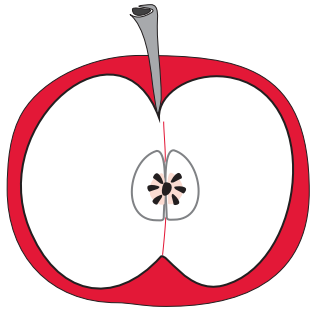


A representação anterior sugere algum enviesamento para a esquerda (embora o diagrama de extremos e quartis esteja ao alto, continuamos a falar no enviesamento para a esquerda, quando este for na direcção dos valores mais pequenos), tal como a representação em caule-e-folhas já havia sugerido.

Tarefa proposta

Vamos comparar vários tipos de maçãs

Será que os diferentes tipos de maçãs, têm características idênticas? Para preparar esta actividade, cada aluno pode ficar encarregue de levar uma maçã de um dos 3 tipos seguintes (ou outros): Red Delicious, Starking ou Golden.



As características que se decidiram estudar foram o peso, o perímetro e a altura de cada maçã.

Elaborar uma tabela, onde se regista o tipo do fruto e os valores observados das características anteriormente referidas.

Utilizando diagramas de extremos e quartis, comparar os três tipos de maçãs.

Tarefa proposta

Os frutos têm muitas calorias?

Dizem os nutricionistas que, para uma alimentação saudável, além de outros requisitos, deveríamos comer 3 peças de fruta, por dia. Apresentamos a seguir, para vários frutos, uma tabela com a quantidade de gordura e o número de calorias por cada 100 gramas de fruto:

Nome	Gordura	Calorias	Nome	Gordura	Calorias
Abacate	13	130	Limão	1	37
Ameixa	1	59	Maçã	1	64
Amêndoa	56	626	Manga	0	57
Amendoim	48	596	Maracujá	1	90
Amoras	1	59	Melancia	0	25
Ananás	1	49	Melão	0	31
Avelãs	65	676	Morango	1	34
Banana	0	90	Nêspera	1	54
Cajú	48	573	Noz	67	686
Castanha	1	182	Papaia	0	50
Cereja	0	63	Pêra	1	37
Coco	60	630	Pêssego	1	45
Figo	1	64	Pinhão	52	618
Framboesa	2	50	Pistácio	54	594
Ginja	2	70	Romã	0	54
Groselha	0	54	Tângera	0	41
Laranja	0	51	Tangerina	0	46
Lichias	0	58	Toranja	1	43
Lima	0	41	Uva	1	89

Analisando os dados anteriores, é nítido que os frutos se podem dividir em duas grandes categorias.

Tentar averiguar quais são essas categorias e calcular a quantidade média de calorias em cada uma dessas categorias.

Analisar com os alunos quais os frutos que se devem privilegiar, para uma alimentação saudável.

Medidas de dispersão

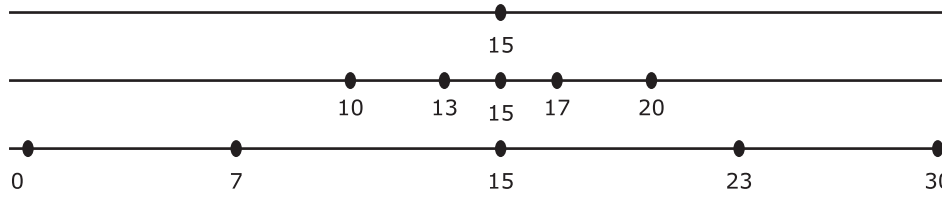
Na secção anterior estudámos algumas medidas que procuram transmitir alguma informação contida nos dados, em termos de localização de alguns pontos importantes, como por exemplo, o centro da distribuição dos dados. No entanto, uma distribuição não está completamente estudada enquanto não estudarmos a variabilidade associada aos dados. Algumas das questões a que as medidas de localização não dão resposta, são, por exemplo, as seguintes:

- Serão os dados quase todos iguais?
- Serão muito diferentes uns dos outros?
- De que modo é que são diferentes?
- ...

Por exemplo, consideremos os três conjuntos de dados:

Conjunto 1	15	15	15	15	15
Conjunto 2	10	13	15	17	20
Conjunto 3	0	7	15	23	30

Os conjuntos de dados anteriores, embora tenham a mesma média e a mesma mediana, nomeadamente igual a 15, têm um aspecto completamente diferente, no que diz respeito à variabilidade, como facilmente se vê, a partir da representação dos valores ao longo de segmentos de recta:



Enquanto que no Conjunto 1, os 5 dados são iguais, não havendo qualquer variabilidade, já no Conjunto 2 e no Conjunto 3 os valores são diferentes uns dos outros, e podemos mesmo avançar que a variabilidade ou dispersão verificada no Conjunto 3 é superior à verificada no Conjunto 2.

Existem algumas medidas para descrever a variabilidade presente num conjunto de dados, umas mais adequadas do que outras, dependendo a sua aplicação, por vezes, da forma da distribuição dos dados.

3.3.1 Amplitude

A medida mais simples para descrever a variabilidade ou dispersão dos dados, é a amplitude, que é a diferença entre o máximo e o mínimo do conjunto de dados:

Amplitude = máximo – mínimo

Esta medida, muito simples, é raramente usada como medida de variabilidade, pois tem a desvantagem de ser muito dependente dos valores extremos, que podem dar origem a uma amplitude muito grande, que não seja representativa do conjunto de dados. Uma alternativa é considerar só a parte central dos dados, obtendo-se uma outra medida a que damos o nome de amplitude interquartis.

3.3.2 Amplitude interquartis

Como o nome está a dizer, obtém-se a amplitude interquartis, fazendo a diferença entre o 3.º e o 1.º quartis. Esta medida, que já foi utilizada na construção do diagrama de extremos e quartis, dá-nos informação sobre a amplitude do intervalo em que se encontram 50% das observações centrais.

Algumas propriedades da *amplitude interquartis* são:

- A *amplitude interquartis* será tanto maior, quanto mais variabilidade houver entre os dados. Se não houver variabilidade, isto é, se as observações forem todas iguais, então a *amplitude interquartis* vem igual a zero.
- No entanto, uma *amplitude interquartis* nula, não significa necessariamente que não exista variabilidade. Por exemplo, o conjunto de dados

  1 2 5 5 5 5 5 5 5 14 21

  tem amplitude interquartis nula, apesar de apresentar variabilidade.

Na secção 3.2.3 calculámos os quartis da amostra constituída pelo peso dos 15 alunos de uma turma do 2.º ano. Vimos que o 1.º quartil $Q_1=27$ e o 3.º quartil $Q_3=31$, donde a amplitude interquartis = 4

Recorde-se que a representação de um conjunto de dados num diagrama de extremos e quartis, dá uma informação imediata sobre a variabilidade existente nos 50% dos elementos centrais, através do comprimento da caixa, que é igual à amplitude interquartis.

3.3.3 Desvio-padrão

Tal como a mediana, que é calculada unicamente a partir de um ou dois valores da amostra, também a amplitude interquartis é calculada unicamente a partir dos quartis, ignorando assim muita informação sobre a forma como os dados se distribuem. Quando a distribuição dos dados é aproximadamente simétrica, situação em que tem sentido falar da média como medida de localização do centro de distribuição dos dados, utiliza-se como medida de variabilidade ou dispersão dos dados, o desvio-padrão, que no seu cálculo tem em conta os desvios de todos os dados relativamente à média.

Consideremos então a amostra $(x_1,x_2,...,x_n)$ com média $\overline{x}$. Para medir a variabilidade dos dados relativamente à média, começa-se por calcular, para cada dado, a diferença entre ele e a média, a que chamamos desvio:

$$x_1 - \overline{x},\ x_2 - \overline{x},\ x_3 - \overline{x},\ \ldots,\ x_n - \overline{x}$$

Para obter a variabilidade de todos os dados, seria natural somar todos os desvios. Acontece que a soma destes desvios é sempre igual a zero, pois os desvios positivos anulam com os negativos, pelo que esta solução não serve. Então, vamos considerar não os próprios desvios, mas os seus quadrados:

$$(x_1 - \overline{x})^2,\ (x_2 - \overline{x})^2,\ (x_3 - \overline{x})^2,\ \ldots,\ (x_n - \overline{x})^2$$

Define-se variância e representa-se por s^2, a medida que se obtém somando os quadrados dos desvios e dividindo pelo número de observações menos uma:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}$$

O motivo que nos leva a considerar os quadrados dos desvios já tem uma explicação. Mas então porque é que não consideramos a média desses desvios ao quadrado, dividindo a sua soma por n em vez de (n-1), como está proposto? A este nível, a resposta que pode ser dada prende-se de certo modo com o motivo que nos levou a considerar os quadrados, em vez dos próprios desvios: como a soma dos n desvios é igual zero, basta conhecer (n-1) desses desvios, para que o n-ésimo fique automaticamente determinado. Assim, como só temos (n-1) desvios independentes, dividimos por (n-1) em vez de n.

A variância, como medida de variabilidade tem um problema que é o facto de não vir nas mesmas unidades que os dados originais. Resolve-se este problema considerando a raiz quadrada, a que se dá o nome de desvio-padrão:

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}}$$

Da forma como o desvio padrão é obtido, imediatamente se conclui que:

- O desvio padrão é sempre maior ou igual a zero e será tanto maior quanto maior for a variabilidade presente nos dados. Se não houver variabilidade, isto é, se os dados forem todos iguais, então o desvio-padrão é nulo;
- por outro lado, se o desvio padrão de um conjunto de dados é nulo, então não existe variabilidade.

Exemplo:

Tempo de realização da ficha – Na turma, o professor estava interessado em saber qual o tempo médio de realização de uma determinada ficha e desejava também saber se os tempos que seus alunos demoravam a fazer a ficha, variavam muito. Registou esses tempos

13 15 14 18 25 14 15 14 16 17 20 17

e de seguida calculou a média e o desvio padrão:

Tempo (em minutos) x_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$
13	-3,5	12,25
15	-1,5	2,25
14	-2,5	6,25
18	1,5	2,25
25	8,5	72,25
14	-2,5	6,25
15	-1,5	2,25
14	-2,5	6,25
16	-0,5	0,25
17	0,5	0,25
20	3,5	12,25
17	0,5	0,25
Soma = 198 $\bar{x} = \frac{198}{12} = 16,5$	Soma =0	Soma =123 $s^2 = \frac{123}{11} \approx 11,18$

Calculando a raiz quadrada de 11,18, vem para o desvio-padrão s = 3,34

O professor concluiu, assim, que o tempo médio de resolução da ficha foi de 16 minutos e meio, com um desvio padrão de 3,34 minutos.

Quais as medidas que se devem utilizar para resumir a informação contida numa amostra?

As medidas de localização, juntamente com as medidas de variabilidade, descrevem o comportamento dos dados. Uma questão que se pode colocar é a de saber quais as medidas de localização e de variabilidade que se devem utilizar. Recordamos que, do mesmo modo que a média só deve ser utilizada para distribuições aproximadamente simétricas, também o desvio padrão só deve ser utilizado nestas condições. Assim, quando pretendemos descrever um conjunto de dados de tipo quantitativo, pode-se utilizar a seguinte metodologia:

1. Fazer uma representação gráfica dos dados;

2. Uma vez a representação gráfica obtida:

   - Se a distribuição dos dados se apresentar aproximadamente simétrica, então utilizar a média e o desvio padrão para descrever os dados;
   - Se a distribuição apresentar enviesamento, então utilizar a mediana e a amplitude interquartis. Pode-se ainda calcular a média e verificar que esta se afasta da mediana: ou é maior ou menor que a mediana, conforme o enviesamento for para a direita (positivo) ou para a esquerda (negativo).
   - Se se verificar a existência de algum(s) *outlier*(s) e se estiver a utilizar a média e o desvio padrão, recalcular estas medidas sem o(s) *outlier*(s) e fazer um pequeno relatório sobre o assunto.

Exemplo

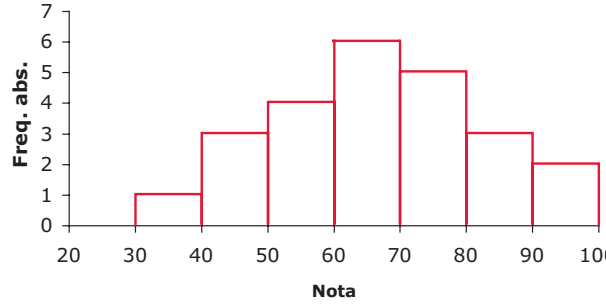
Nota mal digitada - Um professor ao digitar, numa folha de cálculo, as notas (numa escala de 0 a 20) que os seus 38 alunos tiveram no teste de Matemática, enganou-se e ao escrever 15, escreveu 155. Como é que este erro pode afectar o valor das medidas de localização, média e mediana e das medidas de dispersão, desvio padrão e amplitude interquartis?

Nitidamente o valor 155 é um *outlier*, que provocará um aumento (substancial) da média, relativamente ao valor que seria o correcto com a nota de 15. A mediana possivelmente não virá alterada e se houver alteração, não será significativa. No que diz respeito às medidas de dispersão, o desvio-padrão também virá inflacionado, enquanto que a amplitude interquartis não deve ser afectada.

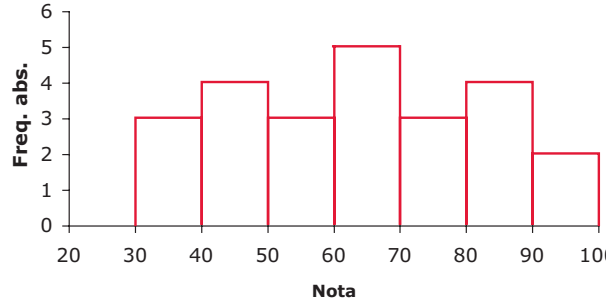
Exemplo:

Notas de três turmas - Três turmas do 10.º ano fizeram o mesmo teste de Matemática, tendo-se construído os seguintes histogramas para as classificações obtidas:

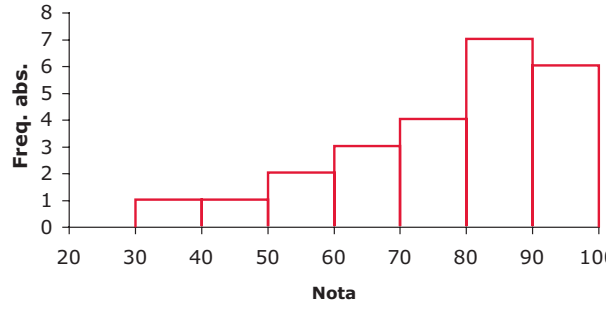
Turma 1



Turma 2



Turma 3

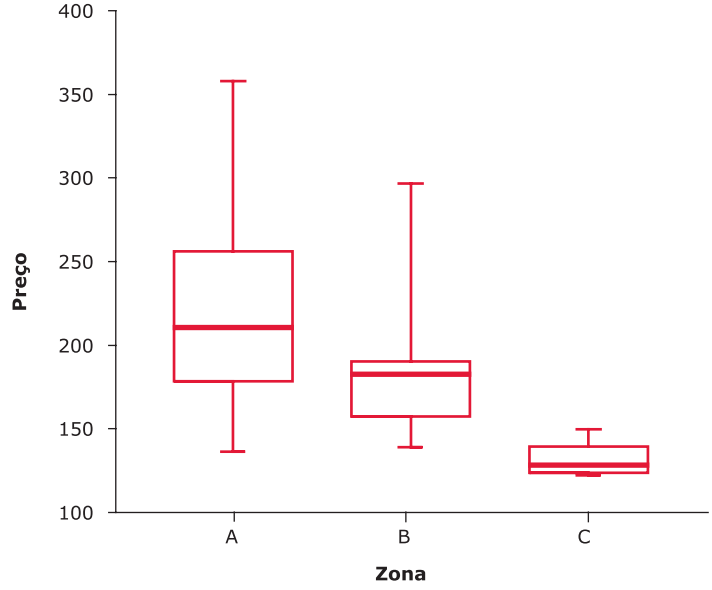


A partir das representações gráficas anteriores será possível dizer em qual das turmas se espera maior média para o teste? E maior mediana? E em qual das turmas se espera que a mediana esteja mais afastada da média?

A turma 3, teve, de um modo geral notas mais altas. Aliás, houve mais alunos a ter notas altas do que baixas, como se depreende pelo enviesamento. Assim, é de esperar que seja a turma 3 a ter maior média. Também para esta turma se espera maior mediana que para as outras turmas e além disso espera-se que a mediana seja maior que a média, pelo tipo de enviesamento apresentado.

Exemplo:

Preço das casas - Recordemos, de novo, o exemplo dos Dados sobre casas. Na secção 2.5.1 apresentámos um gráfico com 3 diagramas de extremos e quartis, referentes ao preço das casas, nas 3 zonas A, B e C:



Gráficos de extremos e quartis paralelos, para os preços das casas nas zonas A, B e C

A partir das representações anteriores verificamos que as casas da zona C são as que têm o preço mais baixo. As casas da zona A são, de um modo geral, mais caras e os preços apresentam uma grande variabilidade. A distribuição dos preços da zona B é, de certo modo, atípica, porque na zona central dos dados apresenta um enviesamento para a esquerda, uma vez que a mediana está mais perto do 3.º quartil que do 1.º quartil, enquanto que os dados mais afastados do centro apresentam um enviesamento para a direita. Esta situação não ocorre com muita frequência, sendo mais vulgares os casos apresentados pelas distribuições das zonas A e C. Para esta situação, apontada para a zona B, já não podemos dizer que a média é maior ou menor que a mediana, pois os dois tipos de enviesamento provocam efeitos contrários, enquanto que para a zona A e C esperamos que a média seja superior à mediana. Calculando estas medidas para as 3 zonas, obteve-se o seguinte quadro:

Zona	Média	Mediana
A	219,14	208,88
B	181,82	181,06
C	131,72	126,80

Analisando a tabela, verifica-se que, como se esperava, os preços das casas das zonas A e C, têm médias superiores às medianas. Para a zona B obteve-se um valor para a média muito próximo da mediana.

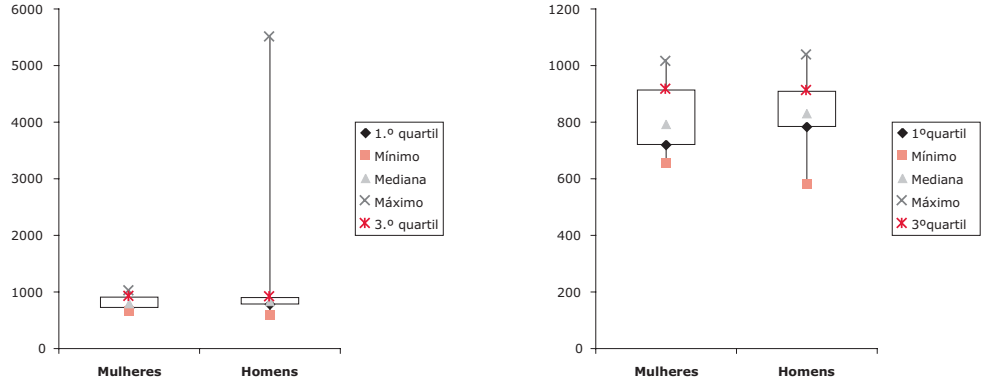
Como se comportarão as medidas de variabilidade? Sugere-se a construção de uma tabela análoga à anterior, com as medidas do desvio padrão e da amplitude interquartis, para analisar as diferenças obtidas.

***Exemplo**:*

Trabalhadores da Empresa Fio de Ouro - Um grupo de trabalhadores, constituído por mulheres, foi-se queixar ao sindicato da indústria têxtil, alegando que, na empresa Fio de Ouro, pertencente ao Sr. Silva, o salário médio dos homens era superior ao das mulheres. Será que tinham razão para se queixarem? A trabalho igual, o Sr. Silva estava a pagar de forma diferente aos homens e às mulheres? Com base na tabela fornecida pela contabilidade, vamos averiguar o que se passa com a questão anterior:

Nome	**Cargo**	**Salário**	**Nome**	**Cargo**	**Salário**	**Nome**	**Cargo**	**Salário**
António	Técnico	809	Emília	Administ.	687	Miguel	Técnico	840
Filipe	Técnico	864	Pedro	Técnico	836	Pedro	Técnico	837
Pedro	Técnico	959	João	Técnico	807	Telmo	Administ.	662
Paulo	Administ.	621	Luísa	Administ.	665	João	Técnico	884
José	Técnico	949	Cristiano	Administ.	582	Luís	Técnico	861
Ana	Técnico	770	Ronaldo	Administ.	712	Fernanda	Técnico	963
Maria	Administ.	655	Cristina	Técnico	915	Eugénia	Administ.	756
Rosa	Técnico	762	Valdemar	Técnico	927	Rita	Técnico	797
João	Técnico	783	Vasco	Administ.	702	Ana	Administ.	712
Filipa	Técnico	853	Vanessa	Técnico	909	Filipa	Técnico	967
Luís	Técnico	807	Cátia	Administ.	746	Raquel	Técnico	1013
Joaquim	Técnico	974	Bruno	Técnico	853	Rute	Técnico	816
Veríssimo	Técnico	821	Raquel	Técnico	853	Pedro	Administ.	731
Eduardo	Técnico	1037	Miguel	Técnico	1028	Ivete	Administ.	670
Fernando Silva	Sócio-gerente	5500	Ricardo	Técnico	847	João	Administ.	742
Eugénio	Técnico	1006	Túlio	Técnico	926	Miguel	Administ.	628
Álvaro	Técnico	893	Tiago	Administ.	747	Eduardo	Técnico	799
Alberto	Técnico	1031	Isabel	Administ.	719	Tiago	Técnico	803
Beto	Técnico	787	Dinis	Técnico	911	Armando	Técnico	802
Anacleto	Técnico	801	Daniela	Técnico	945	Valente	Técnico	831
António	Administ.	695	Antónia	Técnico	970	Susana	Técnico	788

Pretende-se comparar os salários dos homens e das mulheres, utilizando medidas de localização e de dispersão adequadas.
Construíram-se os diagramas de extremos e quartis paralelos e obteve-se a representação do lado esquerdo da figura seguinte:



Estamos numa situação em que existe um *outlier*, o salário de 5500 euros auferido pelo sócio-gerente. Retirou-se este valor dos salários dos homens e construiu-se de novo os diagramas de extremos e quartis paralelos, que se apresentam no lado direito da figura anterior. As representações obtidas não apresentam praticamente enviesamento, pelo que vamos utilizar a média como medida de localização do centro dos dados.

Para explorar um pouco mais os dados (sem o *outlier*), calcularam-se as médias para os empregados do sexo feminino e masculino, separando ainda os técnicos dos administrativos, tendo-se obtido a tabela seguinte:

	Cargo		
Sexo	Administrativo	Técnico	
Feminino	701	880	815
Masculino	682	877	828
	691	878	823

Analisando os resultados apresentados na tabela, conclui-se que:

- O salário médio dos empregados do sexo feminino (=815 euros) é um pouco inferior ao salário médio dos empregados do sexo masculino (=828);
- No entanto, analisando pelo tipo de cargo, verifica-se que, tanto para os administrativos como para os técnicos, o salário médio do sexo feminino é superior ao do sexo masculino, pois

Salário médio administ. feminino (=701)> Salário médio administ. masculino (=682)
Salário médio técnicos feminino (=880)> Salário médio técnicos masculino (=877)

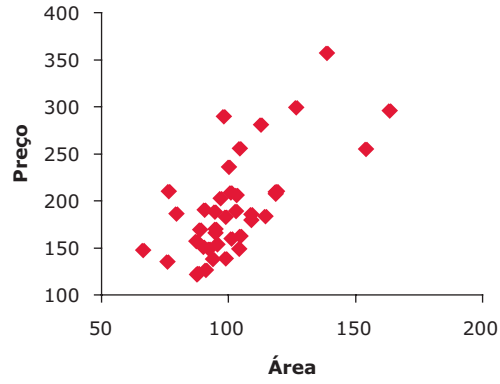
Afinal as mulheres não tinham razão de queixa, pois dentro de cada categoria, o salário médio que auferiam é até um pouco superior ao dos homens!

Esta situação paradoxal que acabámos de descrever é conhecida como o paradoxo de Simpson e pode acontecer quando se analisam os dados segundo um determinado critério e depois se entra em linha de conta com um novo critério para discriminar os dados.

Coeficiente de correlação

Vimos na secção 2.7, que quando temos dados bivariados, uma forma de os representar graficamente é através do diagrama de dispersão, em que cada par de dados (x,y) é representado, num sistema de eixos ortogonais, por um ponto de coordenadas (x,y). Obtém-se assim uma nuvem de pontos que nos permite avaliar de imediato se há ou não uma forte associação entre as duas variáveis. A associação mais simples que os pontos podem apresentar é a *associação linear e* o maior ou menor grau de proximidade dos pontos a uma linha recta pode ser traduzido numericamente por um coeficiente a que se dá o nome de **coeficiente de correlação linear**.

No diagrama de dispersão seguinte, estão representados os pares (*Área*, *Preço*) das 40 casas que constituem a amostra dos Dados sobre casas. A nuvem de pontos apresenta-se um pouco dispersa, mas não deixa por isso de ser bem patente a sua forma alongada que se desenvolve em torno de uma recta com um declive positivo:



Como se vê, verifica-se uma **tendência** para que casas de maior área tenham preços mais elevados.

Tipo de associação linear entre duas variáveis

- ***Associação positiva*** – duas variáveis dizem-se associadas positivamente se aos maiores valores de uma correspondem, em média, os maiores valores da outra.
- ***Associação negativa*** – duas variáveis dizem-se associadas negativamente se aos maiores valores de uma correspondem, em média, os menores valores da outra e vice-versa.

O *coeficiente de correlação* mede a maior ou menor força com que as variáveis se associam, quer positiva, quer negativamente.

Cálculo do coeficiente de correlação:

O *coeficiente de correlação*, representa-se por **r** e calcula-se para os pares de valores (x_1,y_1), (x_2,y_2), …, (x_n,y_n), através da seguinte fórmula:

$$r = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y})}{\sqrt{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}\sqrt{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \ldots + (y_n - \overline{y})^2}}$$

que vai ser utilizada, ainda, para justificar graficamente o maior ou menor valor obtido para o coeficiente de correlação, conforme o aspecto da nuvem de pontos.

Propriedades do coeficiente de correlação:

- O valor do coeficiente de correlação **r** varia entre -1 e 1.
- Quanto **maior** for o valor absoluto de **r**, **mais forte** será a **relação linear** existente entre os x´s e os y´s.
- O facto de **r** ser **positivo**, significa que a relação entre os x's e os y's é do **mesmo sentido**, isto é, a valores grandes de x, correspondem, em média, valores grandes de y e vice-versa - associação positiva. Quando **r** é **negativo**, a relação entre os x's e os y's é de **sentido contrário**, o que significa que a valores grandes de x, correspondem, em média, valores pequenos de y e vice-versa - associação negativa.
- A correlação não é afectada por uma mudança de unidades das variáveis.
- Uma vez que no cálculo da correlação se utilizam medidas não resistentes, como é o caso da média e do desvio padrão, então a correlação também pode ser afectada por *outliers*. Assim, deve-se começar por fazer a representação gráfica do diagrama de dispersão e verificar se não existem pontos discrepantes, que possam influenciar a correlação.

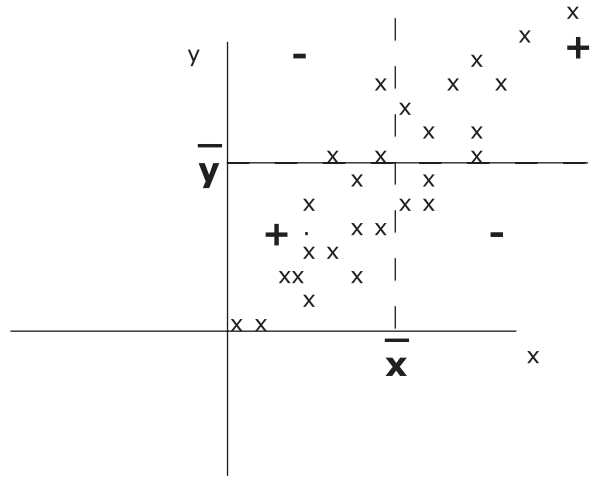
Interpretação geométrica:

- Se aos maiores valores de x, estão associados, de um modo geral, os maiores valores de y, então r>0.

Efectivamente, quando pensamos num valor grande de x, será um valor acima da média. Por outro lado, um valor pequeno de x é um valor abaixo da média. Então, se existe tendência a que, aos valores grandes de x, estejam associados os valores grandes de y, e aos valores pequenos de x estejam associados os valores pequenos de y, os produtos

$$(x_i - \overline{x})\ (y_i - \overline{y})$$

são de um modo geral positivos, já que ambos os factores são positivos ou negativos. Então o facto de somarmos grande número de parcelas positivas, faz com que o valor do coeficiente de correlação seja positivo e tanto maior quantas mais parcelas positivas houver.

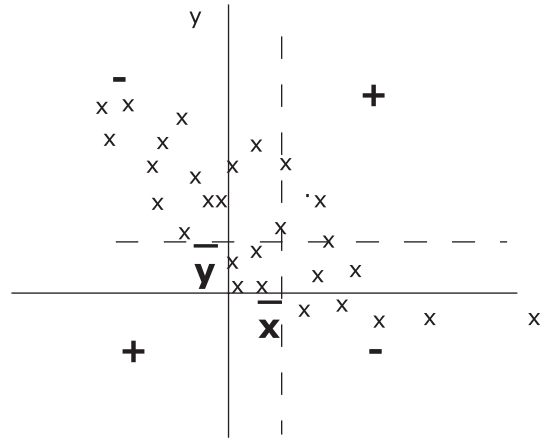


- Se aos maiores valores de x, estão associados, de um modo geral, os menores valores de y, então r<0.

Fazendo o raciocínio como no ponto anterior, verificamos que agora as parcelas são maioritariamente negativas, já que quando x é grande (superior à média dos x's), então existe tendência para que o y seja pequeno (inferior à média dos y's). Assim, os produtos

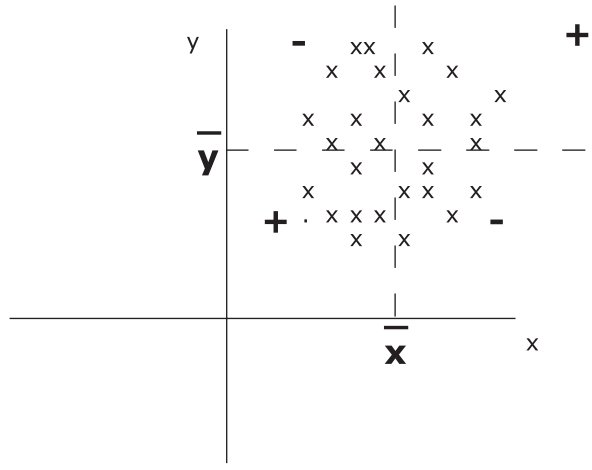
$$(x_i - \bar{x})\ (y_i - \bar{y})$$

são, de um modo geral, negativos.

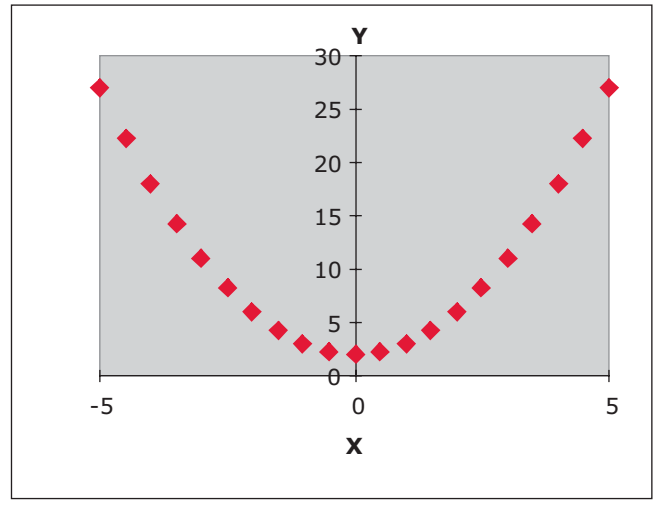


- Se não existe associação linear entre os x's e os y's, então r=0.

Neste caso tanto podem surgir produtos negativos, como positivos, distribuindo-se de forma mais ou menos equitativa. Então o valor de r vem próximo de zero.



Mais uma vez chamamos a atenção que *o coeficiente de correlação mede unicamente a relação linear existente entre as variáveis e não outro tipo de ligação*. Por exemplo, o seguinte diagrama de pontos indica uma forte associação entre as variáveis x e y:



As variáveis estão relacionadas pela equação $y = 2 + x^2$, e no entanto o coeficiente de correlação r = 0.

Na interpretação do coeficiente de correlação deve-se chamar a atenção para o facto de que a existência de correlação elevada entre duas variáveis não significa necessariamente uma relação de causa-efeito. Pode verificar-se a existência de uma ou mais variáveis relacionadas com as variáveis em estudo, a provocar aquelas correlações referidas como correlações falsas.

Para a nuvem de pontos referente aos pares (Área, Preço) obteve-se como coeficiente de correlação linear o valor r=0,68. Este valor evidencia uma correlação positiva não muito forte, confirmando a observação feita anteriormente de que a nuvem se apresenta bastante dispersa e com uma inclinação positiva – há tendência para que casas de maior área tenham preços mais elevados, mas a área, por si só, não consegue explicar na sua totalidade o preço da casa.

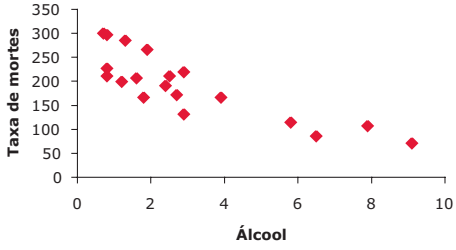
***Exemplo**:*

Será que o vinho é bom para o coração? (Moore, 1997) – Há a convicção de que o consumo moderado de vinho ajuda a prevenir ataques cardíacos. Na tabela seguinte apresentamos, para 19 países desenvolvidos, alguns valores respeitantes ao consumo anual de vinho (litros de álcool obtidos a partir do consumo de vinho, por pessoa) e a taxa de mortes anuais por doenças cardíacas (mortes por 100 000 pessoas):

País	Álcool	Taxa de mortes	País	Álcool	Taxa de mortes
Austrália	2,5	211	Holanda	1,8	167
Áustria	3,9	167	N.Zelândia	1,9	266
Bélgica	2,9	131	Noruega	0,8	227
Canadá	2,4	191	Espanha	6,5	86
Dinamarca	2,9	220	Suécia	1,6	207
Finlândia	0,8	297	Suíça	5,8	115
França	9,1	71	R. Unido	1,3	285
Islândia	0,8	211	EUA	1,2	199
Irlanda	0,7	300	Alemanha	2,7	172
Itália	7,9	107			

Faça um estudo sobre o assunto, a partir dos dados anteriores.

Começamos por construir um diagrama de dispersão que nos dá uma ideia de uma *associação linear negativa* entre o consumo de vinho e a taxa de mortes por ataques cardíacos, pois aos maiores valores da variável consumo de vinho, aqui representada por "Álcool", correspondem, de um modo geral, os menores valores da variável "Taxa de mortes".



Para medir a força desta associação calculámos o coeficiente de correlação, tendo obtido r = -0,84, o que traduz inequivocamente uma forte associação negativa entre as duas variáveis.

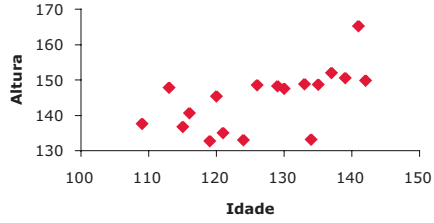
Então podemos concluir que quanto mais álcool consumirmos, menor é a probabilidade de morrer de um ataque cardíaco? Não! A associação não deve ser interpretada como *causa – efeito*. Pode, eventualmente, haver outras variáveis, com que não estamos a entrar em linha de conta, que contribuam para a associação linear verificada entre as variáveis cuja associação se está a estudar. Se formos, por exemplo, estudar para uma determinada época do ano, a associação entre o consumo diário de gelados e o número diário de incêndios, podemos obter uma forte associação positiva entre aquelas duas variáveis. Poderemos assim concluir que comer gelados provoca incêndios? Obviamente que não. O que acontece é que com o aumento de calor verifica-se o aumento do consumo de gelados, o mesmo acontecendo com o número de incêndios.

Exemplo:

Idade e altura das crianças (Graça Martins, 1999) - A tabela seguinte apresenta os valores das idades em meses e das alturas de algumas crianças de uma escola:

Criança	Idade (meses)	Altura (cm)
1	109	137,6
2	113	147,8
3	115	136,8
4	116	140,7
5	119	132,7
6	120	145,4
7	121	135,0
8	124	133,0
9	126	148,5
10	129	148,3
11	130	147,5
12	133	148,8
13	134	133,2
14	135	148,7
15	137	152,0
16	139	150,6
17	141	165,3
18	142	149,9

Representando os dados num diagrama de pontos obtém-se:



Este gráfico mostra a existência de uma certa associação linear, não muito forte, dando a indicação de que existe tendência para que quanto maior for a idade, maior seja a altura.
O valor do coeficiente de correlação é 0,60, o que vai de encontro ao que se disse anteriormente.

PROBABILIDADE

Neste capítulo faz-se uma breve introdução à interpretação frequencista de Probabilidade, de uma forma que se pretende elementar e intuitiva. Dão-se algumas indicações sobre o cálculo de probabilidades de alguns acontecimentos, em situações especiais de simetria.

Introdução

A abordagem do conceito de Probabilidade só se justifica, a este nível, de forma muito elementar e intuitiva. Todos nós, no dia a dia, fazemos conjecturas sobre a realização de um acontecimento futuro. É comum ouvir-se dizer:

- é pouco provável que amanhã chova...;
- a probabilidade de haver uma pessoa com 3 metros de altura é zero;
- a probabilidade do próximo bebé, de uma determinada família, ser do sexo masculino é aproximadamente 50%;
- a probabilidade de lançar uma moeda de 1 euro ao ar e sair a face com o 1, é 50%;
- a probabilidade de amanhã o sol nascer é um; a probabilidade de ganhar no Euromilhões é quase nula; etc.

Ao exprimirmo-nos da forma anterior, não estamos mais do que a anunciar o nosso *grau de convicção* na realização de algum acontecimento. Para exprimir esta convicção estamos a recorrer, embora intuitivamente, à frequência relativa com que o acontecimento se pode repetir.

Consideremos de novo o exemplo dos Dados sobre casas e suponhamos que na região onde se recolheu a informação da tabela, se recolhia informação sobre mais uma casa, escolhida ao acaso. Algumas questões que se podem colocar sobre essa outra casa são as seguintes:

- Será mais provável que essa casa seja nova ou usada?
- Qual será um valor aproximado para a probabilidade de a casa ser usada?

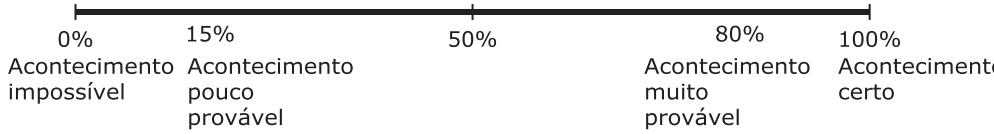
Na página 24, em que organizámos a informação constante da tabela com os dados sobre casas, verificamos que das 40 casas, 31 são usadas e 9 são novas. Então, é natural esperar que seja mais provável que esta outra casa seja usada. Por outro lado, esperamos que a probabilidade de, na dita região, encontrar à venda uma casa usada, esteja próxima de 80%, já que a frequência relativa obtida para o acontecimento "Casa usada" foi 77,5%.

> A probabilidade de um determinado acontecimento aleatório dá-nos a percentagem de vezes que se espera que ele aconteça, se se repetir a experiência, um grande número de vezes, nas mesmas condições.

No exemplo das casas, a experiência consiste em seleccionar uma casa ao acaso e em verificar se a casa é usada ou nova. Existem dois acontecimentos possíveis para o estado da casa e é por essa razão que o resultado da experiência é aleatório: antes de verificar a casa, não temos informação suficiente para saber qual dos acontecimentos se vai verificar, se é usada ou nova.

Quando lançamos um dado ao ar, também não sabemos de antemão, qual a face que vai sair. Sabemos à partida, antes de realizar a experiência de lançar o dado ao ar, que pode sair qualquer uma das faces numeradas de 1 a 6, mas não temos informação suficiente para dizer qual das faces é que vai sair, na próxima realização da experiência. É por essa razão que se diz que a saída da face com 5 pintas, por exemplo, é um acontecimento aleatório.

As probabilidades assumem valores numa escala de 0% a 100%. Se um acontecimento é impossível, atribui-se-lhe uma probabilidade de 0% (ou 0). Se temos a certeza que um acontecimento se vai verificar, então atribui-se-lhe a probabilidade de 100% (ou 1).



A probabilidade de um acontecimento não se verificar é igual a 100% menos a probabilidade de se verificar.

Assim, como atribuímos anteriormente um valor aproximado de 80% ao acontecimento "A casa é usada", podemos dizer que um valor aproximado para a probabilidade do acontecimento "A casa é nova" é 20%.

Exemplo:

Qual a probabilidade? (Freedman *et al.,* 1991) – Um computador está programado para calcular várias probabilidades. Associe as respostas numéricas com as descrições verbais seguintes:

(a)	-50%	(i) É tão provável acontecer, como não acontecer
(b)	0%	(ii) É muito provável que aconteça, mas não é certo
(c)	10%	(iii) Isto não pode acontecer
(d)	50%	(iv) Pode acontecer, mas é pouco provável
(e)	90%	(v) Isso acontecerá, de certeza
(f)	100%	(vi) Há um erro no programa
(g)	200%	

Nos valores numéricos, existem 2 que não podem ser probabilidades. Assim, só podem ser atribuídos a um erro no programa, donde (vi) corresponde a (a) e (g). Se um acontecimento é tão provável de acontecer, como de não acontecer, então temos que (i) corresponde a (d). As outras associações são (ii) a (e); (iii) a (b); (iv) a (c) e (v) a (f).

4.2 Cálculo de probabilidades numa situação especial

O argumento utilizado na secção anterior para exprimir um valor para a probabilidade de um acontecimento se verificar, exige que a experiência se possa repetir um grande número de vezes, nas mesmas condições.

Por exemplo, suponha que tem uma caixa com 10 rifas, numeradas de 1 a 10, em que 2 das rifas, por exemplo as rifas 9 e 10, dão prémio. Qual a probabilidade de ao retirar uma rifa, ao acaso, ela ter prémio? Admitindo que as rifas são iguais e se baralharam antes de retirar uma, qualquer uma delas tem igual possibilidade de ser retirada. Imagine que retira uma rifa, verifica se tem prémio e repõe a rifa novamente na caixa, repetindo este processo muitas vezes. Ao fim de muitas extracções, cada uma das rifas é extraída cerca de 10% das vezes, pelo que as rifas premiadas serão extraídas cerca de 20% das vezes.

Suponha agora que a caixa tem 100 rifas, numeradas de 1 a 100, e as 20 rifas numeradas de 81 a 100 dão prémio. Qual a probabilidade de retirar uma rifa premiada? Repetindo o processo como anteriormente, cada rifa sai cerca de 1 vez em 100, pelo que as premiadas sairão aproximadamente 20 vezes em 100, ou seja cerca de 20% das vezes.

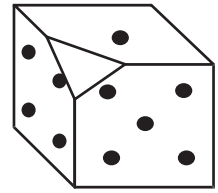
Em ambas as situações anteriores a probabilidade de tirar rifa com prémio, é idêntica, já que a proporção de rifas premiadas era a mesma nas duas caixas.

Suponha agora que tem um dado de 6 faces que, em vez de terem pintas estão pintadas: 3 faces estão pintadas de cor cinza e as outras 3 de vermelho. Suponha que lança o dado 600 vezes. Quantas vezes se espera que saia a face cor cinza? E a face vermelha? Se o dado estiver bem construído, cada face tem igual possibilidade de sair, pelo que como existem 3 faces de cor cinza, esperamos que elas saiam, aproximadamente, metade das vezes, ou seja cerca de 300 vezes. Assim, existe igual probabilidade de sair a face de cor cinza ou a face vermelha.

E se o dado tiver 2 faces de cor cinza e 4 vermelhas? Quantas vezes se espera que saia da cor cinza, nos 600 lançamentos? Como agora só temos duas faces de cor cinza, esperamos que um terço das vezes saia a cor cinza, ou seja, aproximadamente 200 vezes. Então agora a probabilidade de sair a cor cinza é de 1 em 3, ou seja 1/3.

Nos exemplos anteriores, no raciocínio utilizado para calcular as probabilidades dos acontecimentos desejados, colocámo-nos sempre numa situação especial – *situação de simetria*, em que todos os resultados possíveis das experiências estavam em igualdade de circunstâncias e não tínhamos razão para privilegiar algum(s) resultado(s) relativamente aos outros. Quando falámos em retirar uma rifa, estávamos a dar igual possibilidade a cada uma das rifas, da caixa, de ser seleccionada.

O mesmo acontece no lançamento do dado (equilibrado), em que damos igual possibilidade de sair cada uma das 6 faces, em cada lançamento. No entanto, se tivéssemos cortado um vértice ao dado

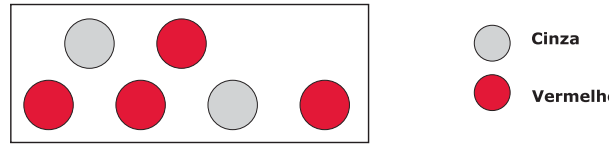


já as faces não estariam em igualdade de circunstâncias, pelo que já não poderíamos esperar que, em muitos lançamentos, se verificasse a mesma proporção de vezes para cada face. Então neste caso, como é que podemos estimar a probabilidade de sair cada face, no próximo lançamento do dado? A única solução é fazer muitos lançamentos, calcular a proporção de vezes que se verificou a saída de cada uma das faces e utilizar esse valor, para estimar a probabilidade desejada.

Existem situações em que gostaríamos de conhecer a probabilidade de se verificar determinado acontecimento, mas em que não estamos numa situação de simetria, nem é possível repetir a experiência um grande número de vezes, nas mesmas circunstâncias, de forma a utilizar a frequência relativa com que o acontecimento acontece, para estimar a probabilidade de ele se verificar. Nestas situações teremos de utilizar alguma informação que nos possa ajudar a exprimir o nosso grau de convicção na realização desses acontecimentos. Por exemplo, não é igualmente provável que o próximo Presidente da República seja homem ou mulher. Com a informação que temos do passado, é natural que se atribua ao acontecimento "o próximo presidente é homem" uma probabilidade de 100%.

Exemplo:

O jogo com berlindes – Numa caixa estão 6 berlindes, 2 de cor cinza e 4 vermelhos. Quando retira o berlinde anota a cor e repõe outra vez na caixa.



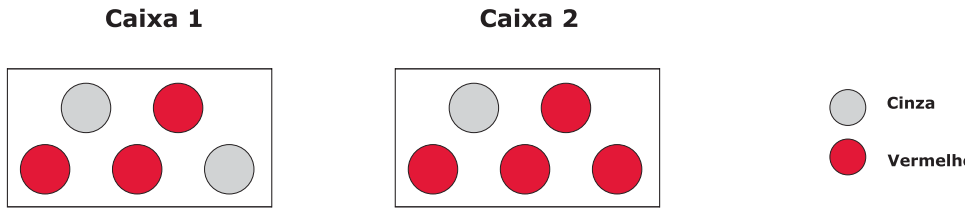
Ao fim de 300 extracções, quantos euros espera ganhar:

- a) Se por cada berlinde cinza que sair, ganhar 1 euro?
- b) Se por cada berlinde cinza ganhar 2 euros e por cada berlinde vermelho perder um euro?

Como nas 300 extracções (com reposição) se espera que saia cerca de 100 vezes berlinde cinza, e nas outras vezes berlinde vermelho, na primeira situação espera-se ganhar 100 euros, enquanto que na segunda situação se espera ganhar 200 euros e perder outros 200 euros, pelo que nesta segunda hipótese não é de esperar ganhar nem perder.

Exemplo:

As duas caixas de berlindes – Suponha que tem as seguintes caixas, cada uma com 5 berlindes cinza e vermelhos. Quando se retira um berlinde, se ele for cinza ganham-se 2 euros, se for vermelho ganha-se 1 euro:



Dão-lhe a possibilidade de escolher uma das 2 caixas para fazer 100 extracções, com reposição. Qual das caixas prefere?

Em cada extracção existem 2 possibilidades em 5 de sair um berlinde de cor cinza, se se fizer a extracção da caixa 1, enquanto que se for da caixa 2, essas possibilidades diminuem para metade. Assim, nas 100 extracções, espera-se que a cor cinza saia cerca de 40 vezes ou 20 vezes se fizermos as extracções da caixa 1 ou da caixa 2, respectivamente. É então preferível a Caixa 1, já que com esta esperamos ganhar 140 euros (40x2+60x1), enquanto que com a outra só esperamos ganhar 120 euros (20x2 + 80x1).

Tarefa

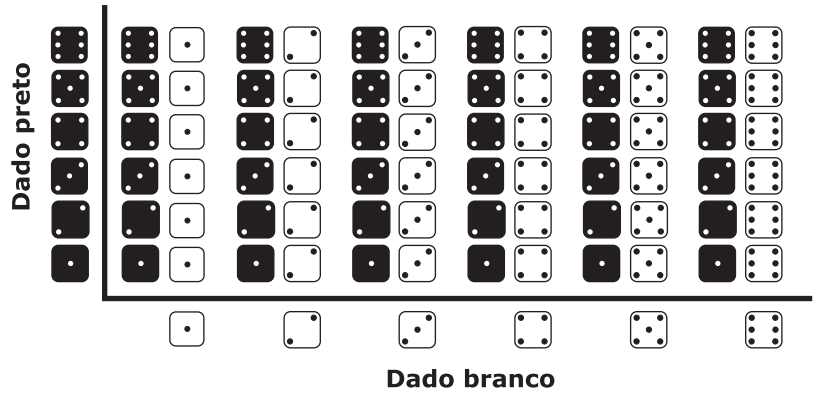
Vamos lançar dois dados

Na escola o professor propôs o seguinte jogo para ser jogado entre a Rita, o João e o Miguel: lançam-se 2 dados de 6 faces e verifica-se a soma das pintas dos dados, que pode ir de 2 a 12. Se a soma for 2, 3, 4 ou 5 o João ganha um ponto; se for 6, 7 ou 8 ganha a Rita um ponto; finalmente, se for 9, 10, 11 ou 12, ganha o Miguel. A Rita ficou muito zangada com o professor, dizendo que este a estava a desfavorecer, uma vez que aos outros colegas dava 4 possibilidades, enquanto que a ela só dava 3. Será que ela tinha razão?

Para ser mais fácil de descrever a actividade, vamos considerar dois dados em que um é preto e o outro é branco. Vamos esquematizar todas as situações possíveis de se verificarem, quando se lançam os dois dados:

Suponhamos que no dado preto saiu 1 pinta. Então no dado branco pode ter saído qualquer valor de 1 a 6:

Repetindo o processo, mas agora com 2, 3, ..., 6 pintas no dado preto, obtemos a figura seguinte, onde temos esquematizado todos as situações possíveis, em número de 36:



Vamos considerar uma tabela com os números das pintas e a soma respectiva:

6+1=**7**	6+2=**8**	6+3=**9**	6+4=**10**	6+5=**11**	6+6=**12**
5+1=**6**	5+2=**7**	5+3=**8**	5+4=**9**	5+5=**10**	5+6=**11**
4+1=**5**	4+2=**6**	4+3=**7**	4+4=**8**	4+5=**9**	4+6=**10**
3+1=**4**	3+2=**5**	3+3=**6**	3+4=**7**	3+5=**8**	3+6=**9**
2+1=**3**	2+2=**4**	2+3=**5**	2+4=**6**	2+5=**7**	2+6=**8**
1+1=**2**	1+2=**3**	1+3=**4**	1+4=**5**	1+5=**6**	1+6=**7**

Analisando com cuidado a tabela anterior, verificamos que existem algumas somas que surgem com mais frequência do que outras. Por exemplo a soma 12 só aparece quando sair 6 pintas nos dois dados

enquanto que a soma 5 aparece nas seguintes situações

Então concluímos que enquanto a probabilidade de o 12 sair é de 1 em 36, o 5 tem uma probabilidade maior, de 5 em 36. A partir da tabela anterior podemos construir uma outra tabela, com o número de vezes que pode sair cada resultado para a soma das pintas, quando se lançam 2 dados:

Resultado da soma das pintas	Número de vezes que se pode verificar	Quem ganha
2	1	João
3	2	João
4	3	João
5	4	João
6	5	Rita
7	6	Rita
8	5	Rita
9	4	Miguel
10	3	Miguel
11	2	Miguel
12	1	Miguel

Então quando se lançam os dois dados, de acordo com as regras estipuladas para o jogo:

- o João tem 10 (1+2+3+4) possibilidades de ganhar;
- a Rita tem 16 (5+6+5) possibilidades de ganhar;
- o Miguel tem 10 (4+3+2+1) possibilidades de ganhar.

Afinal a Rita não tinha razão, pois estava a ser privilegiada neste jogo, que não era um jogo justo.

O professor então propôs que redistribuissem os resultados possíveis pelos 3 colegas, de forma a transformarem um jogo que não era justo, num jogo justo. Depois de alguma discussão, propuseram a seguinte regra: se a soma for 2, 7 ou 8 o João ganha um ponto; se for 4, 5 ou 6 ganha a Rita um ponto; finalmente, se for 3, 9, 10, 11 ou 12, ganha o Miguel. Será que chegaram a uma boa solução?

Na Sala de Aula

Ao nível do 1.º ciclo do ensino básico, a forma como se trabalha a noção de probabilidade deve ser alicerçada em exemplos simples e intuitivos. Podem começar por se apresentar exemplos idênticos ao considerado na Introdução deste capítulo, nomeadamente quando se refere uma casa escolhida ao acaso, na região onde se recolheu a informação que consta do ficheiro Dados sobre casas.

Sugerimos ainda questões como a que apresentamos na Tarefa – Quais os nossos animais domésticos, do Capítulo 2. Ou ainda questões como a que apresentamos de seguida:

Na Sala de Aula

Tarefa

O que é mais provável?

Numa turma com 28 alunos, 20 são raparigas e 8 são rapazes. Dos 28 alunos, 14 têm olhos castanhos e os outros 14 têm olhos de outra cor. Também se sabe que 10 dos alunos (rapazes ou raparigas) são louros. O professor que usava fichas, cada uma com o nome de um dos alunos, um dia chegou à turma, baralhou as fichas como quem baralha um baralho de cartas e seleccionou uma ao acaso, para que o aluno cujo nome constava da ficha seleccionada, fosse ao quadro fazer um problema.

- a) É mais provável que tenha sido seleccionado um rapaz ou uma rapariga?
- b) O que é que é mais provável: que o aluno tenha olhos castanhos ou de outra cor?
- c) O que é que é mais provável: que o aluno seja louro ou não seja louro?

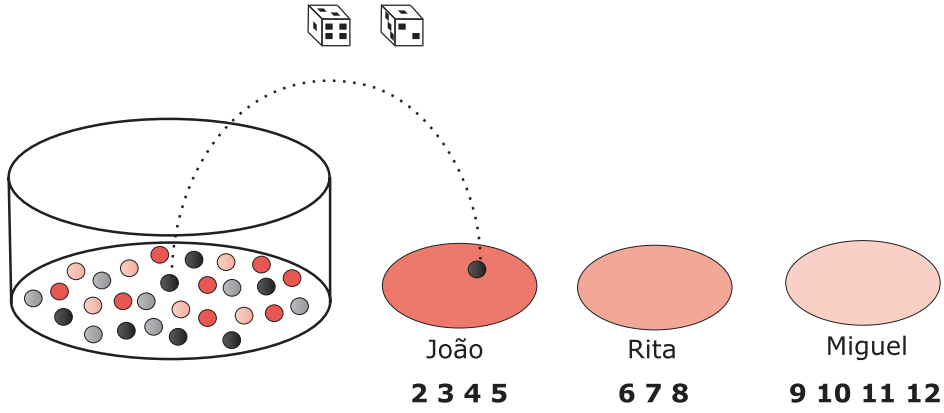
Para responder a estas questões, os alunos devem ter sensibilidade para verificar que quantos mais alunos houver pertencentes a determinada categoria, mais provável é ser seleccionado um aluno pertencente a essa categoria. Assim, será mais provável ser seleccionada uma rapariga, existe igual probabilidade de ser seleccionado um aluno de olhos castanhos e um que não tenha olhos castanhos, etc.

Na Sala de Aula

Tarefa

Vamos lançar dois dados (cont.)

Uma versão desta tarefa pode ser realizada na sala de aula da seguinte forma: o professor coloca numa taça de plástico transparente alguns smarties (em número superior ao número de alunos da turma). O professor lança 2 dados e conforme o número que se verificar para a soma das pintas das faces que ficam voltadas para cima, retira um smartie da taça e coloca no prato do João, da Rita ou do Miguel (na figura, exemplificamos uma situação em que a soma é igual a 3, pelo que o smartie foi colocado no prato do João). Quando se esgotarem os smarties da taça, ganha aquele que tiver maior número de smarties no seu prato. Quem é que se espera que ganhe?



No fim do jogo todos os alunos têm direito a um smartie, ficando o aluno ganhador com os que sobram.

A seguir apresentamos outras actividades, que o professor decidirá da oportunidade de as realizar ou não, na sala de aula.

Na Sala de Aula

Tarefa

Será que a moeda é equilibrada?

Na turma, constituída por 2 rapazes – o Tiago e o Ricardo, e 16 raparigas, era necessário escolher um aluno rapaz, para pertencer a uma comissão que tinha de integrar os dois sexos. Como só havia dois rapazes decidiram atirar uma moeda de 1 euro ao ar. Se saísse a face Euro (E) seria escolhido o Ricardo, caso contrário, se saísse a face Nacional (N) seria o Tiago. Antes de lançarem a moeda, o Tiago questionou o professor sobre se esse processo de selecção seria justo. Quem é que lhe garantia que houvesse 50% de possibilidade de ser ele o escolhido? Ou por outras palavras, o que ele desejava saber era se a moeda era equilibrada.

Decidiram fazer uma experiência de lançar a moeda algumas vezes e registar os resultados obtidos. Ao fim de 10 lançamentos, os resultados obtidos foram os seguintes:

N E N N N E E E E E

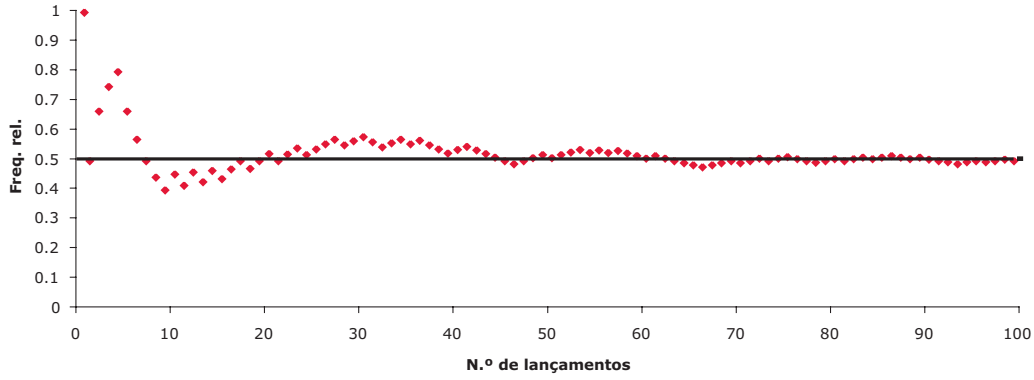
Estes resultados não sossegaram o Tiago, pois ele começou a pensar que só teria 40% de possibilidades de ser seleccionado, uma vez que em 10 vezes a moeda só lhe foi favorável 4 vezes!

N.º do lanç.	Result.	N.º de faces N	Freq. Rel. da face N	N.º do lanç.	Result.	N.º de faces N	Freq. Rel. da face N
1	N	1	1,000	26	N	14	0,538
2	E	1	0,500	27	N	15	0,556
3	N	2	0,667	28	N	16	0,571
4	N	3	0,750	29	E	16	0,552
5	N	4	0,800	30	N	17	0,567
6	E	4	0,667	31	N	18	0,581
7	E	4	0,571	32	E	18	0,563
8	E	4	0,500	33	E	18	0,545
9	E	4	0,444	34	N	19	0,559
10	E	4	0,400	35	N	20	0,571
11	N	5	0,455	36	E	20	0,556
12	E	5	0,417	37	N	21	0,568
13	N	6	0,462	38	E	21	0,553
14	E	6	0,429	39	E	21	0,538
15	N	7	0,467	40	E	21	0,525
16	E	7	0,438	41	N	22	0,537
17	N	8	0,471	42	N	23	0,548
18	N	9	0,500	43	E	23	0,535
19	E	9	0,474	44	E	23	0,523
20	N	10	0,500	45	E	23	0,511
21	N	11	0,524	46	E	23	0,500
22	E	11	0,500	47	E	23	0,489
23	N	12	0,522	48	N	24	0,500
24	N	13	0,542	49	N	25	0,510
25	E	13	0,520	50	N	26	0,520

O professor chamou então a atenção para o facto de se ter de realizar a experiência um grande número de vezes, pois com 10 lançamentos não podemos tirar qualquer conclusão. Fizeram então mais 90 lançamentos, tendo obtido os seguintes resultados:

N.º do lanç.	Result.	N.º de faces N	Freq. Rel. da face N	N.º do lanç.	Result.	N.º de faces N	Freq. Rel. da face N
51	E	26	0,510	76	N	39	0,513
52	N	27	0,519	77	E	39	0,506
53	N	28	0,528	78	E	39	0,500
54	N	29	0,537	79	E	39	0,494
55	E	29	0,527	80	N	40	0,500
56	N	30	0,536	81	N	41	0,506
57	E	30	0,526	82	E	41	0,500
58	N	31	0,534	83	N	42	0,506
59	E	31	0,525	84	N	43	0,512
60	E	31	0,517	85	E	43	0,506
61	E	31	0,508	86	N	44	0,512
62	N	32	0,516	87	N	45	0,517
63	E	32	0,508	88	E	45	0,511
64	E	32	0,500	89	E	45	0,506
65	E	32	0,492	90	N	46	0,511
66	E	32	0,485	91	E	46	0,505
67	E	32	0,478	92	E	46	0,500
68	N	33	0,485	93	E	46	0,495
69	N	34	0,493	94	E	46	0,489
70	N	35	0,500	95	N	47	0,495
71	E	35	0,493	96	N	48	0,500
72	N	36	0,500	97	E	48	0,495
73	N	37	0,507	98	N	49	0,500
74	E	37	0,500	99	N	50	0,505
75	N	38	0,507	100	E	50	0,500

O gráfico seguinte mostra a evolução da frequência relativa da saída da face N, à medida que se fazem os sucessivos lançamentos da moeda:



Tendo em conta os resultados anteriores, verifica-se que a frequência relativa da saída da face Nacional, tende a estabilizar à volta dos 50%. Assim, não temos razão para rejeitar a hipótese de a moeda ser equilibrada, dando 50% de probabilidade a cada face.

Na Sala de Aula

Tarefa

Quem é que ganha o jogo?

Na escola o professor propôs o seguinte jogo a ser jogado por dois alunos, o David e o António: lançam-se duas moedas e em cada lançamento, se sairem faces diferentes, o David ganha um ponto; caso contrário ganha o António o ponto. Ganha o jogo, aquele que, ao fim de 50 jogadas tiver ganho mais pontos. Quem é que ganhará o jogo?

Vamos agora simular o lançamento de 2 moedas equilibradas, generalizando o exemplo anterior, em que se lançou uma moeda.

Jogada	1.ª moeda	2.ª moeda	António ganha	David ganha	Pontos António	Pontos David
1	N	N	1	0	1	0
2	E	E	1	0	2	0
3	E	N	0	1	2	1
4	N	N	1	0	3	1
5	E	N	0	1	3	2
6	N	N	1	0	4	2
7	E	E	1	0	5	2
8	E	N	0	1	5	3
9	E	E	1	0	6	3
10	E	E	1	0	7	3
11	E	N	0	1	7	4
12	N	E	0	1	7	5
13	N	N	1	0	8	5
14	E	N	0	1	8	6
15	N	N	1	0	9	6
16	E	N	0	1	9	7
17	E	N	0	1	9	8
18	E	E	1	0	10	8
19	N	E	0	1	10	9
20	N	N	1	0	11	9
21	N	E	0	1	11	10
22	N	E	0	1	11	11
23	N	E	0	1	11	12
24	N	N	1	0	12	12
25	N	N	1	0	13	12
26	E	E	1	0	14	12
27	E	E	1	0	15	12
28	N	N	1	0	16	12
29	N	E	0	1	16	13
30	E	N	0	1	16	14
31	N	E	0	1	16	15
32	E	E	1	0	17	15
33	E	N	0	1	17	16

Jogada	1.ª moeda	2.ª moeda	António ganha	David ganha	Pontos António	Pontos David
34	N	E	0	1	17	17
35	E	E	1	0	18	17
36	E	N	0	1	18	18
37	N	E	0	1	18	19
38	E	E	1	0	19	19
39	E	N	0	1	19	20
40	E	N	0	1	19	21
41	E	E	1	0	20	21
42	N	E	0	1	20	22
43	E	E	1	0	21	22
44	E	E	1	0	22	22
45	N	E	0	1	22	23
46	N	N	1	0	23	23
47	E	E	1	0	24	23
48	E	E	1	0	25	23
49	E	E	1	0	26	23
50	E	N	0	1	26	24

Neste jogo ganhou o António, pois ao fim de 50 jogadas tinha alcançado 26 pontos, enquanto que o David tinha 24 pontos. Resolveram jogar novamente o mesmo jogo, tendo obtido os resultados seguintes:

Jogada	1.ª moeda	2.ª moeda	António ganha	David ganha	Pontos António	Pontos David
1	E	E	1	0	1	0
2	E	E	1	0	2	0
3	N	N	1	0	3	0
4	N	N	1	0	4	0
5	E	N	0	1	4	1
6	N	E	0	1	4	2
7	E	N	0	1	4	3
8	E	N	0	1	4	4
9	N	E	0	1	4	5
10	N	E	0	1	4	6
11	N	N	1	0	5	6
12	N	E	0	1	5	7
13	N	E	0	1	5	8
14	E	E	1	0	6	8
15	E	N	0	1	6	9
16	N	N	1	0	7	9
17	N	N	1	0	8	9
18	E	N	0	1	8	10
19	N	E	0	1	8	11
20	N	E	0	1	8	12
21	N	N	1	0	9	12
22	N	N	1	0	10	12
23	N	E	0	1	10	13
24	E	E	1	0	11	13
25	E	E	1	0	12	13
26	N	N	1	0	13	13
27	N	E	0	1	13	14
28	N	N	1	0	14	14
29	N	E	0	1	14	15
30	E	E	1	0	15	15

Jogada	1.ª moeda	2.ª moeda	António ganha	David ganha	Pontos António	Pontos David
31	E	E	1	0	16	15
32	N	N	1	0	17	15
33	E	E	1	0	18	15
34	N	N	1	0	19	15
35	N	E	0	1	19	16
36	E	E	1	0	20	16
37	N	E	0	1	20	17
38	E	E	1	0	21	17
39	N	E	0	1	21	18
40	E	N	0	1	21	19
41	E	N	0	1	21	20
42	E	N	0	1	21	21
43	E	N	0	1	21	22
44	N	E	0	1	21	23
45	E	E	1	0	22	23
46	E	N	0	1	22	24
47	N	E	0	1	22	25
48	N	E	0	1	22	26
49	N	N	1	0	23	26
50	E	E	1	0	24	26

Desta vez ganhou o David! Resolveram fazer ainda um 3.º jogo para a desforra e obtiveram os seguintes resultados:

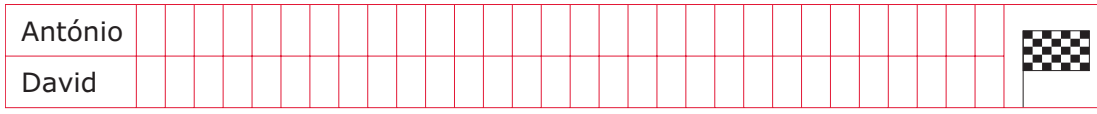
Jogada	1.ª moeda	2.ª moeda	António ganha	David ganha	Pontos António	Pontos David
1	N	N	1	0	1	0
2	E	N	0	1	1	1
3	E	N	0	1	1	2
4	E	N	0	1	1	3
5	E	N	0	1	1	4
6	N	E	0	1	1	5
7	E	E	1	0	2	5
8	E	E	1	0	3	5
9	N	E	0	1	3	6
10	N	N	1	0	4	6
11	E	N	0	1	4	7
12	N	N	1	0	5	7
13	N	E	0	1	5	8
14	N	N	1	0	6	8
15	E	N	0	1	6	9
16	E	N	0	1	6	10
17	N	N	1	0	7	10
18	E	N	0	1	7	11
19	N	N	1	0	8	11
20	N	N	1	0	9	11
21	N	E	0	1	9	12
22	N	E	0	1	9	13
23	N	N	1	0	10	13
24	N	N	1	0	11	13
25	E	N	0	1	11	14
26	E	E	1	0	12	14
27	E	E	1	0	13	14
28	N	E	0	1	13	15

Jogada	1.ª moeda	2.ª moeda	António ganha	David ganha	Pontos António	Pontos David
29	N	N	1	0	14	15
30	E	E	1	0	15	15
31	E	N	0	1	15	16
32	E	E	1	0	16	16
33	N	N	1	0	17	16
34	N	N	1	0	18	16
35	N	E	0	1	18	17
36	N	E	0	1	18	18
37	E	E	1	0	19	18
38	E	N	0	1	19	19
39	E	N	0	1	19	20
40	E	E	1	0	20	20
41	E	N	0	1	20	21
42	N	N	1	0	21	21
43	E	N	0	1	21	22
44	N	N	1	0	22	22
45	E	E	1	0	23	22
46	E	N	0	1	23	23
47	E	N	0	1	23	24
48	N	E	0	1	23	25
49	N	N	1	0	24	25
50	N	N	1	0	25	25

Agora empataram!

Afinal não se pode dizer à partida quem é que sairá o vencedor, pois qualquer um dos dois alunos tem igual "chance" de ganhar o jogo.

Uma versão deste exemplo pode ser realizado na sala de aula, em que o professor desenha no quadro um trajecto que será percorrido pelos alunos que estão a jogar, da seguinte maneira: sempre que um dos alunos ganha, avança uma quadrícula. Ganhará o que chegar mais rápido à meta.



Tarefa proposta

Moedas não equilibradas

(Adaptado de Rossman, A. *et al.*, 2001)

Na turma onde se realizou a actividade anterior, os alunos ficaram muito entusiasmados com a experiência feita, de tal modo que o professor resolveu propor ainda uma outra actividade relacionada com moedas. Tinha consigo 6 moedas, 5 das quais não tinham passado nos testes de controlo de qualidade e tinham sido rejeitadas por alegadamente não serem dadas como equilibradas. Para cada uma destas 6 moedas, a probabilidade de sair a face Nacional era:

Moeda A: 1 em 4 ou 1/4 **Moeda B**: 1 em 3 ou 1/3 **Moeda C**: 1 em 2 ou 1/2

Moeda D: 3 em 4 ou 3/4 **Moeda E**: 4 em 5 ou 4/5 **Moeda F**: 99 em 100 ou 99/100

Com o objectivo de identificar qual das moedas seria a A, B, ..., F, lançou-se cada moeda 5 vezes, tendo-se obtido os seguintes resultados:

N.º do lançamento	1.ª moeda	2.ª moeda	3.ª moeda	4.ª moeda	5.ª moeda	6.ª moeda
1	N	N	E	N	N	E
2	N	N	E	N	N	N
3	E	N	N	N	N	E
4	N	N	E	E	N	E
5	N	N	E	N	N	E
Freq. relativa Qual é a moeda?						

- a) Preencher a linha das frequências relativas com a proporção de faces nacionais obtidas nestes 5 lançamentos, de cada uma das moedas. Preencher a seguir a última linha com a letra da moeda que suspeita ter sido a 1.ª, 2.ª, …, ou 6.ª.
- b) Tem confiança que as suas suspeitas estejam correctas? Explique porquê.
- c) Suponha que se fizeram mais 5 lançamentos para cada uma das moedas, sendo agora as frequências relativas as apresentadas na tabela seguinte. Com esta informação adicional, tente novamente associar as moedas com as probabilidades respectivas.

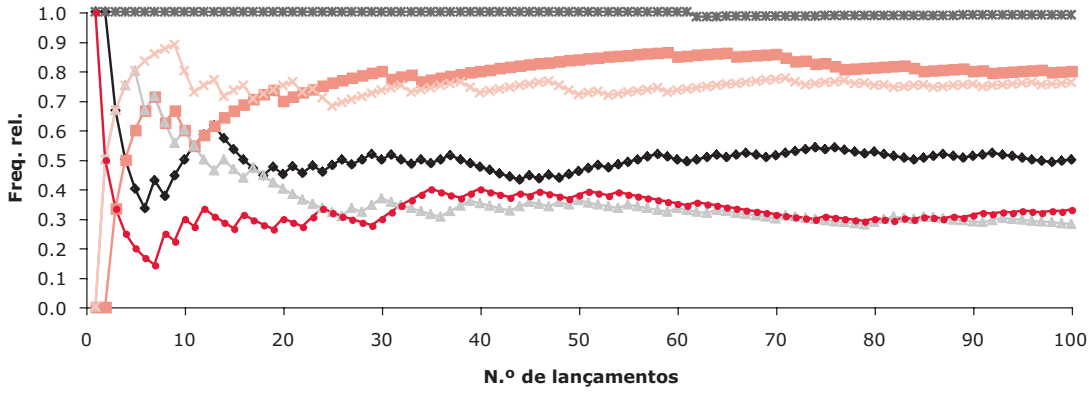
10 lançamentos	1.ª moeda	2.ª moeda	3.ª moeda	4.ª moeda	5.ª moeda	6.ª moeda
Freq. relativa Qual é a moeda?	0,70	0,90	0,20	0,80	1,00	0,20

- d) Suponha agora que lança as moedas mais 15 vezes e posteriormente mais 25 vezes, obtendo as frequências relativas apresentadas nas tabelas seguintes. Mais uma vez se pede que preencha a última linha das tabelas:

25 lançamentos	1.ª moeda	2.ª moeda	3.ª moeda	4.ª moeda	5.ª moeda	6.ª moeda
Freq. relativa	0,56	0,88	0,28	0,88	1,00	0,20
Qual é a moeda?						

50 lançamentos	1.ª moeda	2.ª moeda	3.ª moeda	4.ª moeda	5.ª moeda	6.ª moeda
Freq. relativa	0,58	0,92	0,26	0,78	1,00	0,32
Qual é a moeda?						

- e) Depois dos 50 lançamentos, estar-se-á razoavelmente seguro que as moedas estão correctamente identificadas? Explique porquê.
- f) O seguinte gráfico mostra a evolução da frequência relativa para as 6 moedas, à medida que o número de lançamentos aumenta:



Comente o que é que este gráfico revela sobre a probabilidade, como um conceito sobre o comportamento de um processo aleatório a longo-termo e não a curto-termo.

Referências BIBLIOGRÁFICAS

Na preparação destas folhas, seguiu-se essencialmente a seguinte bibliografia:

Bereska, C. *et al.* (1999) – *Exploring Statistics* in the Elementary Grades, Dale Seymour Publications

De Veaux, R. D. *et al.* (2004) – *Intro Stats*, Pearson – Addison Wesley.

Freedman, D. *et al.* (1991) – *Statistics*, W.W. Norton & Company, Inc.

Graça Martins, M.E. (2005) – *Introdução à Probabilidade e à Estatística* – Com complementos de Excel, Sociedade Portuguesa de Estatística.

Graça Martins, M. E. *et al.* (1999) – Introdução às Probabilidades e à Estatística, Universidade Aberta.

Graça Martins, M. E. *et al.* (1999) – Probabilidades e Combinatória, Ministério da Educação, Departamento do Ensino Secundário.

Graça Martins, M. E. *et al.* (2005) – *Estatística Computacional* – Anexo para apoio à interpretação do program, Módulo B2 para os Cursos Profissionais. Departamento de Estatística e Investigação Operacional, FCUL.

Rossman, A. *et al.* (2001) - *Workshop Statistics – Discovery with Data*, Key College Publishing.

Tanenbaum, P. *et al.* (1998) – *Excursions in Modern Mathematics*, Prentice-Hall, Inc.

Artigos da revista Teaching Statistics

Neville, H. (2003) – Handling Continuous Data in Excel, Vol 25, 2, pag. 42-45.

Neville, H. (2004) – Charts in Excel, Vol 26, 2, pag. 49-53.

Neville, H. (2006) – Boxplot in Excel, www.mis.coventry.ac.uk/~nhunt/boxplot.htm

Recursos na Internet

Projecto ALEA: www.alea.pt

Alguma bibliografia relacionada com o ensino da estatística, não exclusivamente no 1.º ciclo do Ensino Básico

Abrantes, P.; Serrazina, L. e Oliveira, I. (1999). *A Matemática na Educação Básica.* Lisboa: Ministério da Educação.

Azarquiel (1993). *Estatística no 3.º ciclo do Ensino Básico*. Lisboa: APM.

DEB (2001). *Currículo Nacional do Ensino Básico – Competências Essenciais.* Ministério da Educação. Departamento da Educação Básica.

Ministério da Educação (1990). Programa do 1.º ciclo do Ensino Básico. Lisboa: Ministério da Educação.

NCTM (1991). *Normas para o currículo e a avaliação em Matemática escolar*. Lisboa: APM.

NCTM (1993). *Normas para o currículo e a avaliação em Matemática Escolar - Colecção de adendas (do 1.º ao 6.º ano de escolaridade).* Lisboa: APM.

NCTM (2001). *Normas para o currículo e a avaliação em Matemática Escolar. Lidar com dados e probabilidades (anos de escolaridade 5-8).* Lisboa: APM.

NCTM (1994). *Normas Profissionais para o Ensino da Matemática*. Lisboa: APM.

NCTM (1999). *Normas para a Avaliação em Matemática Escolar*. Lisboa: APM.

NCTM (2000). *Principles and Standards for School Mathematics*. Reston: NCTM.

Palhares, P. (coord.). (2004). *Elementos de Matemática para professores do Ensino Básico*. Lisboa: Lidel.

Ponte, J.P. e Serrazina, M.L. (2000*). Didáctica da Matemática do 1.º Ciclo*. Lisboa: Universidade Aberta.

Revista *Educação e Matemática*, da APM: Associação de Professores de Matemática.

Ministério da Educação

dgidc
Direcção-Geral de Inovação
e de Desenvolvimento Curricular