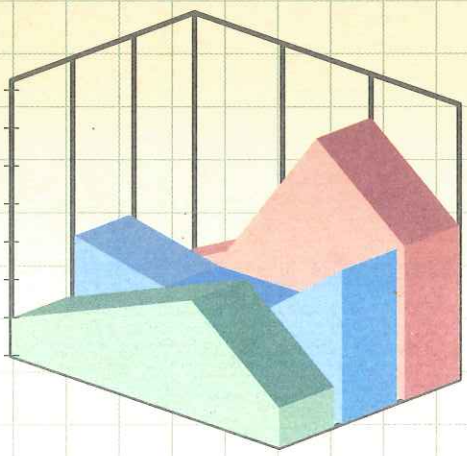
MATEMÁTICA

Ministério da Educação
Departamento do **Ensino Secundário**

Estatística

10º ano de escolaridade

Maria Eugénia Graça Martins
Cecília Monteiro
José Paulo Viana
Maria Antónia Amaral Turkman



ÍNDICE

PREFÁCIO

Este guia tem por objectivo apoiar o professor de Matemática na leccionação da componente Estatística do programa do 10º ano. Foi considerado importante que esse apoio se orientasse em duas dimensões: uma científica proporcionando informação actualizada relativamente a conceitos fundamentais indicados no pro-grama e uma dimensão didáctica onde são sugeridas actividades que possam facilitar a aprendizagem dos alunos.

Na componente científica houve a preocupação de aprofundar um pouco mais os assuntos do que o programa sugere, de modo a que, com mais facilidade e flexibi-lidade, o professor possa planificar e desenvolver as actividades de aprendizagem.

Na componente a que chamamos "Sugestões didácticas e comentários" são apresentadas, a título de exemplo, algumas actividades que podem enriquecer a aprendizagem dos alunos, na medida em que alertam para possíveis erros que normalmente são cometidos por estes, ou ainda actividades que alargam a dimensão estritamente técnica dos cálculos. Sugerimos ainda a utilização de uma calculadora de modo a que, ao libertar o aluno dos cálculos, ele mais fácil e rapidamente compreenda os conceitos. Em alguns exemplos evidenciamos o modo como uma calculadora gráfica pode ser um instrumento útil e necessário para uma melhor compreensão das diversas situações em estudo (qualquer outra calculadora gráfica pode ser utilizada, com as necessárias adaptações).

Cada vez mais é reconhecida a importância da Estatística no currículo dos alunos. Ela tem sido inserida nos programas de Matemática e é encarada como uma área favorável ao desenvolvimento de certas capacidades expressas nos currículos, tais como interpretar e intervir no real; formular e resolver problemas; comunicar; manifestar rigor e

espírito crítico; e ainda a aquisição de uma atitude positiva face à Ciência. Deste modo, ensinar Estatística não pode limitar-se ao ensino de técnicas e fórmulas e aprender Estatística não pode ser aprender a aplicar rotineiramente procedimentos desinseridos de contextos, sem ter de interpretar, de analisar e de criticar.

Uma das finalidades da escola é preparar os alunos para as necessidades e problemas do mundo real onde vivemos, necessidades e problemas esses que todos os dias aparecem nos meios de comunicação social, televisão, rádio e jornais. Alfabetizar estatisticamente os alunos de modo a perceberem as notícias que ouvem e lêem, é desenvolver-lhes o sentido crítico, a capacidade de argumentar sobre elas e inclusivamente serem capazes de intervir e tomar decisões.

Outro aspecto importante no ensino da Estatística é a compreensão da importância da ciência e da investigação como um meio de resolver problemas do homem e obter benefícios para a sociedade. A Estatística é relevante para áreas como a Economia, a Medicina, a Política, a Geografia, a Psicologia e muitas outras. A procura do conhecimento tem sido uma das motivações das pessoas que se dedicam a investigar e a Estatística tem vindo a desempenhar um papel cada vez mais importante na seriedade dos processos utilizados nessa procura da "verdade". Por exemplo, as questões relativas aos processos de amostragem devem ser discutidas e bastante trabalhadas com os alunos, visto depender da amostra e do processo da sua selecção a validade das conclusões que se podem tirar de um estudo.

Ao nível do 10º ano de escolaridade a Estatística assume um carácter puramente descritivo, onde o ênfase é dado à organização e interpretação de dados qualitativos e quantitativos. No entanto é uma parte do currículo de Matemática que mais permite o desenvolvimento das capacidades nele enunciadas, que proporciona o desenvolvimento de projectos significativos, que permite a ligação da Matemática à realidade e portanto a outras áreas do saber.

Na Educação Estatística deverão seguir-se os seguinte princípios metodológicos:

1. Os conceitos estatísticos deverão ser sempre abordados em contextos significativos de modo a que a sua análise e interpretação possa ser feita de modo

inserido. Não tem interesse que o aluno se limite apenas a saber calcular um desvio padrão, por exemplo, mas sim que entenda o significado do valor encontrado na situação proposta.

2. A comunicação dos resultados de actividades práticas e de problemas deverá ser acompanhada de relatórios escritos e de discussão na turma, onde os alunos expliquem as conclusões por palavras suas. Cada vez mais é reconhecida na Educação Matemática a importância da comunicação escrita e oral por parte do aluno e da discussão entre pares na construção e compreensão dos conceitos e dos procedimentos.

3. O desenvolvimento de projectos de carácter investigativo pelos alunos deve ser levado a cabo através de trabalho de grupo, porque é também através do trabalho colaborativo que surge a discussão e portanto, muitas vezes a clarificação dos conceitos.

Não consideramos que esta obra seja definitiva. Contamos, assim, com a vossa colaboração no sentido de nos enviarem críticas e sugestões, que possam contribuir para o seu melhoramento.

Sabendo que a componente de Estatística do programa de Matemática é, de um modo geral, uma das preferidas pelos alunos, esperamos que este guia contribua para o professor desenvolver na sala de aula actividades e projectos significativos para eles, e portanto motivantes, contribuindo assim para o sucesso em Mate-mática.

Os autores

Capítulo 1

RECENSEAMENTO E SONDAGEM POPULAÇÃO E AMOSTRA

1.1 - Recenseamento e sondagem

Estes dois termos, que com certeza fazem já parte do vocabulário do estudante, são suficientemente interessantes para iniciar o aluno no estudo da Estatística, e suficientemente motivadores para o Professor introduzir os conceitos mais gerais de população e amostra, fundamentais a qualquer análise estatística.

O termo recenseamento está, em regra geral, associado à contagem oficial e periódica dos indivíduos de um País, ou parte de um País. Ele abrange, no entanto, um leque mais vasto de situações. Assim pode definir-se recenseamento do seguinte modo:

> ***Recenseamento*** *- Estudo científico de um universo de pessoas, instituições ou objectos físicos com o propósito de adquirir conhecimentos, observando todos os seus elementos, e fazer juízos quantitativos acerca de características importantes desse universo.*

Para a maioria das pessoas a palavra *recenseamento* ou *censo* encontra-se associada à enumeração dos elementos da população de um País. O recenseamento geral de uma população é uma prática que remonta à antiga Roma e Egipto, onde já há conhecimento de recenseamentos da população, feitos a intervalos regulares, com o objectivo principal de obter informação para a colecta de impostos, chamada para o serviço militar e outros assuntos governamentais. Apesar disso, a sua prática corrente, com carácter periódico, só teve lugar, na maioria dos Países, a partir do sec XIX. Esses censos periódicos são feitos em geral de 10 em 10 anos e, em princípio, todos os Países são encorajados a cumprir certas normas internacionais ao elaborar um recenseamento.

11

Em Portugal a primeira operação que se conhece deste género foi levada a cabo por D. João III em 1527 e ficou conhecida pelo "numerando dos vizinhos", tendo permitido estalelecer uma estimativa da população portuguesa. Este apuramento estatístico, constitui um motivo de orgulho para os portugueses visto que foi um dos primeiros estudos deste género conhecido na Europa.

O INE, Instituto Nacional de Estatística, tem a seu cargo fazer recenseamentos da população portuguesa, o último dos quais, o XIII Recenseamento Geral da População, foi reali-zado em 1991. Neste recenseamento ficaram a conhecer-se variadas características do nosso povo como por exemplo: a situação civil, a habitacional, a população emigrante, etc. Os dados relativos aos censos são extremamente importantes pois têm influência directa na decisão em assuntos de interesse nacional e local, tal como seja na educação, emprego, saúde, transportes, recursos naturais, etc, etc. Comparando resultados de recenseamentos sucessivos pode-se extrapolar e predizer padrões futuros da população. Podemos obter informação sobre, por exemplo, a estrutura da idade da população e crescimento populacional, fundamental para o planeamento na construção de novas escolas, alojamento para idosos, etc.

A realização de um recenseamento geral da população, além de implicar gastos muito elevados, é extremamente difícil de conduzir. Há problemas associados com a recolha adequada da informação, seu armanezamento, tratamento, posterior divulgação, etc. É de referir que esta prática se pode estender a outras situações, tais como, às habitações (recenseamento da habitação), às indústrias (recenseamento industrial), à Agricultura (recenseamento agrícola), etc. É importante que fique claro que a palavra recenseamento está associada à análise de todos os elementos da população em causa e que tem por objectivo não só a enumeração dos seus elementos, como também o estudo de características importantes . Não é contudo viável nem desejável, principalmente quando o número dos elementos da população é muito elevado, inquirir todos os elementos da população sempre que se quer estudar uma ou mais características particulares dessa população. Assim surge o conceito de *sondagem*, que se pode tentar definir como:

> *Sondagem - Estudo científico de uma parte de uma população com o objectivo de estudar atitudes, hábitos e preferências da população relativamente a acontecimentos, circunstâncias e assuntos de interesse comum.*

12

A realização de sondagens é uma actividade da segunda metade do séc XX. Embora antes de 1930 já se tenham realizado sondagens, estas eram feitas de um modo muito pouco científico. Foi necessário um desenvolvimento adequado de métodos e técnicas estatísticas para que as sondagens pudessem ser realizadas e os resultados analisados cientificamente.

Só em 1973 é que, pela 1ª vez , apareceu publicado nos orgãos de comunicação social o resultado de uma sondagem realizada em Portugal, nomeadamente, "63% dos Portugueses nunca votaram" (Paula Vicente *et al*, 1996). Embora as sondagens se tenham popularizado devido a questões políticas, elas não são apenas um importante instrumento político; acima de tudo constituem um instrumento de importância vital em estudos de natureza, quer económica, quer social . Assim, se nos meios políticos as sondagens são usadas para obter informação acerca das atitudes dos eleitores, de modo a planear campanhas, etc, elas são importantes também em estudos de mercado, para testar as preferências dos consumidores, descobrir o que mais os atrai nos produtos existentes ou a comercializar, tendo como objectivo o de satisfazer os clientes e aumentar as vendas. Também na área das ciências sociais as sondagens são importantes para, por exemplo, estudar as condições de vida de certas camadas da população.

É fundamental referir que, contrariamente ao recenseamento, as sondagens inquirem ou analisam apenas uma parte da população em estudo, isto é, restringem-se a uma amostra dessa população, mas com o objectivo de extrapolar para todos os elementos da população os resultados observados na amostra.

Uma sondagem realiza-se em várias fases: escolha da amostra, obtenção da informação, análise dos dados e relatório final. Para que os resultados de uma sondagem sejam válidos há necessidade de essa amostra ser representativa da população. O processo de recolha da amostra, a amostragem, tem de ser efectuada com os cuidados adequados. Quando são usadas técnicas apropriadas e a amostra é suficientemente grande, os resultados obtidos encontram-se em geral perto dos resultados que se obteriam, se fosse estudada toda a população.

Há certos livros de texto do ensino secundário que identificam amostragem com sondagem. Isto não é correcto. Com efeito, a amostragem diz respeito ao procedimento de recolha de amostras qualquer que seja a natureza do estudo estatístico que se

pretenda fazer. A sondagem, por sua vez, pressupõe a existência de uma amostragem, isto é, a amostragem é uma das várias fases do processo de sondagem. As sondagens dizem respeito a um estudo estatístico específico. É importante referir que a sondagem visa estudar características da população tal como ela se apresenta. Por exemplo, se quisermos comparar diversas escolas relativamente ao sucesso escolar na disciplina de Matemática, realizamos uma sondagem. Se quisermos averiguar se o método de ensino A é melhor que o método de ensino B na aprendizagem da Matemática, sendo cada um dos métodos atribuído a grupos diferentes de alunos, e averiguando depois o sucesso em cada grupo, já não temos uma sondagem, pois houve intervenção no estudo da característica.

Embora o termo sondagem esteja essencialmente ligado a inquéritos à opinião pública, não há nada que impeça que a mesma técnica seja útil e aplicada para obter informação de qualquer outro tipo de populações. Assim podemos definir mais geralmente sondagem como:

> ***Sondagem*** *- Estudo estatístico de uma população, feito através de uma amostra, destinado a estudar uma ou mais das suas características tal como elas se apresentam nessa população*.

Sugestões didácticas e comentários

Discuta com os alunos as vantagens dos governos dos países efectuarem periodicamente recenseamentos das suas populações.
Discuta também o tipo de carcterísticas que convém conhecer e com que objectivos. Será que os objectivos de hoje são os mesmos de antigamente?

14

1.2 - População e Amostra

Quer se trate de uma sondagem ou não, a maior parte das situações em que é necessário utilizar técnicas estatísticas envolve a necessidade de tirar conclusões gerais acerca de um grande conjunto de indivíduos, baseando-nos num número restrito desses indíviduos. Surge assim a necessidade de definir os conceitos de *População* e *Amostra*, conceitos estes já utilizados anteriormente.

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

Exemplo 1 - Relativamente à população constituída pelos alunos da Escola Secundária Prof. Herculano de Carvalho, em Lisboa, poderíamos estar interessados em estudar as seguintes características populacionais:

- altura (em cm) dos alunos;
- notas obtidas na disciplina de Português, no 1º período;
- número de irmãos de cada aluno;
- tempo que cada aluno demora a chegar à escola;
- idade dos alunos;
- cor dos olhos.

Exemplo 2 - Uma população que pode ter interesse estudar é a constituída pelas temperaturas (em ºC) , todos os dias às 9 horas, na praia da Costa de Caparica.

Ao estudar uma população, normalmente o que se pretende é estudar algumas características numéricas a que chamamos *parâmetros*.

Exemplo 3 - Ao estudar a população constituída por todos os potenciais eleitores para as legislativas, dois parâmetros que podem ter interesse são:

- **idade média** dos potenciais eleitores que estão decididos a votar;
- **percentagem** de eleitores que estão decididos a votar.

Para conhecer aqueles parâmetros, teria de se perguntar a cada eleitor a sua idade, assim como a sua intenção no que diz respeito a votar ou não. Esta tarefa seria impraticável, nomeadamente por questões de tempo e de dinheiro.

15

Outras razões, além das apontadas anteriormente, que podem levar a que não se possa observar exaustivamente todos os elementos de uma população, prendem-se com o facto de algumas populações terem dimensão infinita - população constituída pelas temperaturas em todos os pontos de uma cidade, ou a própria observação levar à destruição da população! Por exemplo, o departamento de controlo de qualidade de uma fábrica de baterias de carros, em que o teste para verificar se a bateria está em perfeitas condições obriga ao desmantelamento da bateria, não pode verificar todas as baterias, pois destruiria toda a população!

As considerações anteriores levam-nos a concluir que, de um modo geral, não podemos determinar exactamente os parâmetros desconhecidos da população a estudar. Podemos sim estimá-los utilizando *estatísticas*, que são quantidades calculadas a partir da observação de uma amostra recolhida da população.

> ***Amostra*** - *subconjunto da população, que se observa com o objectivo de tirar conclusões para a população de onde foi recolhida.*

Tendo em consideração o objectivo com que se recolhe a amostra, o de retirar conclusões para a população, esta fase do processo estatístico, a da recolha da amostra, é muito importante, pois a amostra deve ser tão representativa quanto possível da população.

Resumindo, é importante chamar a atenção que, em toda a situação estatística envolvendo população e amostra, a característica numérica que se está a estudar aparece sob duas formas: como característica populacional ou parâmetro e como característica amostral ou estatística. No caso do exemplo 3, à característica populacional "percentagem de eleitores que estão decididos a votar" corresponde a característica amostral " percentagem dos 1000 eleitores (entretanto recolheu-se uma amostra de dimensão 1000), que interrogados disseram estar decididos a votar". Estas quantidades são conceptualmente distintas, pois enquanto a característica populacional pode ser considerada um valor exacto, embora desconhecido, a característica amostral é conhecida, embora contendo um certo erro, mas que todavia pode ser considerada uma estimativa útil da característica populacional respectiva, se efectivamente a amostra utilizada for representativa da população subjacente.

16

Quando uma amostra não é representativa da população, diz-se que é *enviesada*. A sua utilização para estimar características da população pode ter consequências graves, na medida em que a amostra tem propriedades que não reflectem as propriedades da população.

Exemplos de más amostras ou amostras enviesadas e resultado da sua utilização:

> Amostra 1 - Opiniões de alguns leitores de determinada revista técnica, para representar as opiniões dos portugueses em geral.
> Resultado - Diferentes tipos de pessoas lêem diferentes tipos de revistas, pelo que a amostra não é representativa da população. Basta pensar que, de um modo geral, a população feminina ainda não adere às revistas técnicas como a população masculina. A amostra daria unicamente indicações sobre a população constituída pelos leitores da tal revista.
> Amostra 2 - Utilizar alguns alunos de uma turma, para tirar conclusões sobre o aproveitamento de todos os alunos da escola.
> Resultado - Poderíamos concluir que o aproveitamento dos alunos é pior ou melhor do que na realidade é. As turmas de uma escola não são todas homogéneas, pelo que a amostra não é representativa dos alunos da escola. Poderia servir para tirar conclusões sobre a população constituída pelos alunos da turma.
> Amostra 3 - Utilizar os jogadores de uma equipa de basquete de uma determinada escola para estudar as alturas dos alunos dessa escola.
> Resultado - O estudo concluiria que os estudantes são mais altos do que na realidade são.

Como seleccionar uma "boa" amostra?

A selecção de uma amostra representativa da população a estudar é um problema que nem sempre é simples de abordar, mas existe um princípio que deve estar presente que é o da *aleatoriedade*. Dada uma população, uma *amostra aleatória* é uma amostra tal que, qualquer outra amostra possível, da mesma dimensão, tem igual possibilidade de ser seleccionada.

17

Este princípio pode ser exemplificado com uma população de dimensão pequena, como no exemplo seguinte.

Exemplo 4 - Consideremos a população constituída pelos 18 alunos de uma turma do 10º ano de uma determinada Escola Secundária, em que a característica de interesse a estudar é a altura média desses alunos. Uma maneira possível de recolher desta população uma amostra aleatória, seria escrever cada um dos indicadores dos elementos da população num quadrado de papel, inserir todos esses bocados de papel numa caixa e depois seleccionar tantos quantos a dimensão da amostra desejada.
Este exemplo pode ser aproveitado pelo Professor, que pedirá a cada aluno que retire da caixa 4 papéis, registe os números dos alunos seleccionados e os coloque de novo na caixa, antes do próximo aluno fazer a recolha da sua amostra. Chamar-se-á aqui a atenção que a recolha está a ser feita *sem reposição*, pois quando se retira um papel (elemento da população), ele não é reposto enquanto a amostra não estiver completa (com a dimensão desejada). Qualquer conjunto de números recolhidos desta forma dará origem a uma amostra aleatória, constituída pelas alturas dos alunos seleccionados. Cada aluno disporá assim de uma amostra de dimensão 4, que lhe vai permitir calcular uma média, que será uma estimativa do parâmetro a estudar - valor médio da altura dos alunos da turma. Obter-se-ão tantas estimativas, quantas as amostras retiradas.

Chamar-se-á então a atenção para o facto de nesta altura não se poder dizer qual das estimativas é "melhor", isto é, qual delas é uma melhor aproximação do parâmetro a estimar, já que esse parâmetro é desconhecido (obviamente que nesta população tão pequena seria possível estudar exaustivamente todos os seus elementos, não sendo necessário recolher nenhuma amostra - este exemplo só serve para exemplificar uma situação)!

O processo que acabamos de descrever é um processo que nos permite obter amostras aleatórias simples.

Nesta altura poder-se-á explorar a utilização da calculadora, para obter uma amostra aleatória.

18

Actividade - PROCESSOS PARA OBTER AMOSTRAS ALEATÓRIAS SIMPLES

Uma escola tem 123 alunos do 10º ano. Pretende-se fazer um estudo sobre os seus projectos quanto ao prosseguimento de estudos superiores. Para isso resolveu fazer-se um inquérito que abranja uma amostra de 25 alunos. Como obter essa amostra?

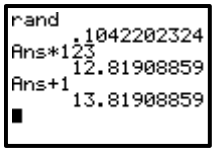
Um método elementar consiste em arranjar 123 papéis ou cartões iguais, escrever em cada um o nome de um aluno, meter tudo num saco, misturar bem e extrair 25 papeis, como já foi explicado anteriormente. Este método é pouco prático (dá bastante trabalho escrever os 123 nomes) mas funciona bem desde que se tenha o cuidado de misturar cuidadosamente os cartões.

Como quase todas as calculadoras, tanto as científicas simples como as gráficas, possuem uma função geradora de números aleatórios[1], podemos aproveitar esse facto para um novo método.

Começamos por numerar os alunos, de 1 a 123.

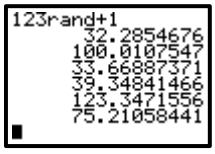
A função **rand** (ou RND em certas máquinas) gera um número aleatório pertencente ao intervalo [0 ; 1[, intervalo que tem amplitude 1. Podíamos dividir este intervalo em 123 partes iguais, fazendo corresponder a cada aluno uma das partes. Depois ver-se-ia em qual das partes calhava cada número aleatório que aparecesse. Mas isso não era nada cómodo. Então, o que vamos fazer é arranjar maneira de sortear um número aleatório num intervalo de amplitude 123.

Para isso, poderíamos começar por pedir com **rand** um número aleatório entre 0 e 1. **Multiplicando-o por 123**, passamos a ter um número aleatório pertencente ao intervalo [0 ; 123[. **Somando uma unidade**, o resultado passa a pertencer ao intervalo [1 ; 124[. Se considerarmos só a parte inteira do número obtido, ele vai corresponder exactamente ao número de um dos alunos. No exemplo da figura, seria o aluno nº 13.



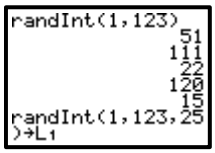
[1]Na realidade são números pseudo-aleatórios, pois são gerados a partir de um mecanismo determinista, que necessita de uma "semente" para desencadear o processo. Se se considerar a mesma semente obtém-se sempre a mesma sequência de números. O que se verifica é que normalmente estes mecanismos estão de tal modo afinados, que os números que geram se comportam como se fossem aleatórios.

19

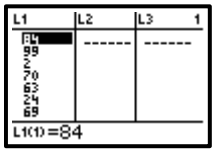


No entanto, podemos fazer isto de forma mais prática escrevendo logo a instrução completa **123** $\times$ **rand + 1** , passando a obter um número aleatório pertencente ao intervalo [1 ; 124[cada vez que carregarmos em **ENTER**.

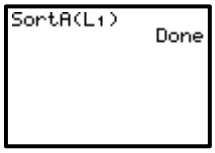
Neste exemplo, os primeiros alunos escolhidos para a amostra são os números 32, 100, 33, 39, 123 e 75. Bastava continuar até obter os 25 elementos, tendo o cuidado de verificar se não surgiam números repetidos.



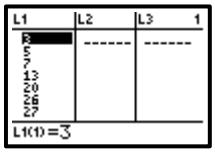
Em certas máquinas, o processo ainda pode ser melhorado do ponto de vista prático com a função **randInt(1,123)** que gera imediatamente um número inteiro aleatório entre 1 e 123 (inclusive).



Como queremos 25 números aleatórios, isso pode ser obtido de uma só vez fazendo simplesmente **randInt(1,123,25)** e guardando os números numa lista.



Depois, podemos até ordenar a lista para ser mais fácil ver quais foram os alunos seleccionados.



Contudo, novamente temos de ter o cuidado de verificar se não há números repetidos (e o mais provável é que haja). Se isso acontecer, vai ser preciso sortear mais alguns números.

Nota: Um outro processo relacionado com a recolha de uma amostra, é abordado através do exemplo seguinte. Embora seja abordada uma noção que não faz parte do programa, pensamos que é importante, porque relata uma situação que surge com frequência nas aplicações.

20

Exemplo 5 - Suponhamos que numa escola secundária se pretende averiguar, após o 1º período, a percentagem de alunos do 10º ano, com nota negativa a Matemática. Sabe-se que as turmas não são todas uniformes no aproveitamento, pois que a sua constituição obedeceu à partida a procedimentos não aleatórios. Assim, para seleccionar uma amostra representativa da população a estudar, deve-se ter o seguinte cuidado: começa-se por verificar quantas turmas e quantos alunos de cada turma constituem a população.

Para fixar ideias, admitamos que a população a estudar é constituída por 3 turmas A, B e C, com 25, 30 e 18 alunos respectivamente e que se pretende recolher uma amostra de dimensão 15. Calculando-se a percentagem de alunos de cada turma que compõem a população, entra-se com esses valores para calcular quantos alunos se deve recolher em cada turma para constituirem a amostra:

Turma	Nº elementos	% Pop.	Nº el. da amostra
A	25	25/73 = .34	.34 x 15 ≈ 5
B	30	30/73 = .41	.41 x 15 ≈ 6
C	18	18/73 = .25	.25 x 15 ≈ 4
Total	73		15

No exemplo anterior obtivemos uma *amostra estratificada*, em que os estratos são as turmas. Não teria sido correcto recolher a informação sobre os alunos de uma única das turmas, pois não havendo garantia de homogeneidade entre as turmas, a amostra recolhida seria enviesada.

Este exemplo simples pode servir ao Professor para chamar a atenção para outros casos menos simples, mas cuja técnica é análoga. Por exemplo, ao procurar estudar os rendimentos anuais da população constituída pelas famílias portuguesas, deve ser feito um planeamento prévio sobre a estrutura da população, identificando alguns estratos, como sejam o meio rural e urbano e eventualmente dentro destes estratos alguns subestratos. Por exemplo na zona de Lisboa e arredores, são facilmente identificadas algumas zonas socialmente mais favorecidas do que outras, constituindo diferentes estratos.
Outro caso, é o que se passa quando se pretende recolher informação sobre a percentagem de potenciais eleitores que votam em determinado partido. Pode-se chamar a atenção para o facto de, frequentemente, empresas diferentes apresentarem resultados bastante diferentes sobre as percentagens de cada partido, em vésperas de eleições.

21

Esta discrepância entre os resultados apresentados prende-se, normalmente, com a falta de cuidado na selecção da amostra, que não é representiva da população.

Outra técnica de amostragem que por vezes se utiliza, é a da *amostragem sistemática*, que pressupõe que a população se apresenta numerada de 1 a N, por alguma ordem. Para a recolha de uma amostra de dimensão n, tomamos um elemento da população de entre os k primeiros e depois selecciona-se a partir daí todos os que se distanciam dele k unidades. No caso do exemplo 4, considerando k = 5, se começassemos por escolher o elemento 3, os outros elementos escolhidos seriam 8, 13 e 18.

Qual a dimensão que se deve considerar para a amostra?

Outro problema que se levanta com a recolha da amostra é o de saber qual a *dimensão* desejada para a amostra a recolher.

Este é um problema para o qual nesta fase, não é possível avançar nenhuma teoria, mas sobre o qual o Professor deve tecer algumas considerações gerais. Pode começar por dizer que, para se obter uma amostra que permita calcular estimativas suficientemente precisas dos parâmetros a estudar, a sua dimensão depende muito da variabilidade da população subjacente. Por exemplo, se relativamente à população constituída pelos alunos do 10º ano de uma escola secundária, estivermos interessados em estudar a sua idade média, a dimensão da amostra a recolher não necessita de ser muito grande já que a variável idade apresenta valores muito semelhantes, numa classe etária muito restrita. No entanto se a característica a estudar for o tempo médio que os alunos levam a chegar de casa à escola, já a amostra terá de ter uma dimensão maior, uma vez que a variabilidade da população é muito maior. Cada aluno pode apresentar um valor diferente para esse tempo.

Chama-se a atenção para a existência de técnicas que permitem obter valores mínimos para as dimensões das amostras a recolher e que garantem estimativas com uma determinada precisão exigida à partida. Uma vez garantida essa precisão, a opção por escolher uma amostra de maior dimensão, é uma questão a ponderar entre os custos envolvidos e o ganho com o acréscimo de precisão. Vem a propósito a seguinte frase (*Statistics: a Tool for the Social Sciences*, Mendenhall et al., pag. 226):

22

"Se a dimensão da amostra é demasiado grande, desperdiça-se tempo e talento; se a dimensão da amostra é demasiado pequena, desperdiça-se tempo e talento".

Convém ainda observar que a dimensão da amostra a recolher não é directamente proporcional à dimensão da população a estudar, isto é, se por exemplo para uma população de dimensão 1000 uma amostra de dimensão 100 for suficiente para o estudo de determinada característica, não se exige necessariamente uma amostra de dimensão 200 para estudar a mesma característica de uma população análoga, mas de dimensão 2000. Finalmente chama-se a atenção para o facto de que se o processo de amostragem originar uma amostra enviesada, aumentar a dimensão não resolve nada, antes pelo contrário!

Sugestões didácticas e comentários

a) Sugerir aos alunos comentários sobre a identificação da amostra e sua representatividade, relativamente à respectiva população, em algumas situações, tais como:

1. Para investigar as preferências musicais dos alunos do ensino secundário entregou-se um questionário aos alunos desse nível de ensino que frequentavam o Conservatório.
2. Uma empresa de publicidade pretendia perceber quais os anúncios da televisão que mais facilmente eram recordados pelas pessoas, tendo inquirido uma amostra de pessoas à saída de um supermercado num determinado dia.
3. O conselho directivo de uma escola secundária do Porto pretendia saber se os alunos estavam satisfeitos com a alimentação fornecida pela cantina da escola. Inquiriu todos os alunos com número ímpar.

Os exemplos apresentados devem ser simples e bastante claros. Pretende-se apenas que o aluno perceba que, para poder tirar conclusões válidas para uma determinada população, a amostra deve ser cuidadosamente seleccionada de modo a evitar possíveis enviesamentos. Por exemplo, no 1º caso apresentado a amostra seria válida apenas para tirar conclusões sobre as preferências musicais dos alunos do secundário que também frequentam o conservatório. É natural que um aluno que frequenta um Conservatório tenha uma apetência musical diferente doutro que não o frequente e portanto conclusões que se tirem de tal amostra não podem ser válidas para a população dos alunos do Ensino Secundário. No 2º exemplo a amostra não é representativa da

23

população pois é possível que as pessoas à saída do supermercado se lembrem melhor dos produtos que, ou acabaram de comprar ou que aí encontraram, sendo assim as suas respostas enviesadas. No 3º exemplo a amostra já é representativa da população. É um exemplo de amostragem sistemática. É também importante que o aluno reconheça que uma amostra pode ser representativa de uma população quando se pretende estudar uma sua característica e o deixe de ser ao estudar outra característica. Por exemplo, se se pretende estudar a característica "cor dos olhos" de uma população, pode-se recolher uma amostra constituída apenas por médicos. Esta amostra não servirá, no entanto, para estudar a característica "conhecimentos de biologia", dessa mesma população já que os médicos têm conhecimentos de Biologia diferentes dos da generalidade da população. Os conceitos população, amostra e característica (características) a estudar não se podem assim dissociar.

b) Pedir aos alunos que recolham informação nos jornais sobre notícias que envolvam recenseamentos e sondagens, aproveitando para as comentar. Por exemplo:

Sondagem

10% não sabem quem é o Presidente da República

DEZ por cento dos portugueses não sabem quem é o Presidente da República e 9 por cento desconhecem a identidade do primeiro-ministro. Uma sondagem de 2000 inquiridos EXPRESSO/Euroexpansão revela ainda índices mais desoladores para o presidente da Assembleia da República (só identificado por 39 por cento dos inquiridos), para os líderes partidários (desconhecidos de mais de metade do universo) e para os chefes dos grupos parlamentares (ignorados pela quase totalidade da amostra). Os dados da sondagem mostram ainda que os portugueses não distinguem entre António Guterres/ primeiro-ministro e António Guterres/secretário-geral do PS: 91 por cento sabem que ele é o chefe de Governo, mas 52 por cento ignoram que é ele o líder dos socialistas (**ver pág. 7**).

Ficha Técnica

Sondagem efectuada entre os dias 6 e 31 de Janeiro. O universo é constituído pela população de Portugal Continental, com idades entre os 18 e os 74 anos. A amostra é de 1964 indivíduos, entrevistados directamente, nas suas residências, seleccionados através do método de quotas resultantes da intersecção das variáveis sexo, idade e grau de instrução, e distribuídos do seguinte modo: Litoral Norte (474), Grande Porto (212) , Interior Norte (272), Litoral Centro (298), Grande Lisboa (449) e Interior Sul (259). Os resultados foram ponderados com base nas variáveis região/sexo/idade. A sondagem é da responsabilidade da Euroexpansão e a análise de resultados feita pelo EXPRESSO.

(in Expresso 15/03/97)

24

A ficha técnica, que deve vir sempre associada ao relatório dos resultados de uma sondagem, é absolutamente necessária para a identificação da população, amostra e processo de amostragem e pode ajudar o Professor a, mais uma vez, lembrar a importância da representatividade das amostras. O Professor pode aproveitar para comentar com os alunos o fenómeno, tantas vezes observado, de resultados de sondagens contraditórios, principalmente quando estão envolvidas questões políticas.

PAREDE

Recenseamento

A Junta de Freguesia da Parede está a realizar o recenseamento da população desta freguesia, afim de actualizar o número real das pessoas ali residentes. Estes dados precisos, quantitativos de população, só são actualizados de dez em dez anos, com o recenseamento geral da população.

Para o efeito, a Junta elaborou um formulário onde constam o nome e a morada, a naturalidade dos residentes, a filiação e outros dados pessoais, o ano em que se fixou na freguesia, a profissão e as habilitações literárias.

Todo este processo está a ser realizado por partes, uma vez que a Parede é constituída por vários aglomerados, abrangendo uma área considerável. Assim, foram entregues em casa de cada paredense, o número de formulários correspondente aos elementos do agregado familiar. De seguida, com um prazo máximo de oito dias, é feita a recolha dos formulários, sendo a responsabilidade da própria freguesia.

O acesso aos resultados será possível daqui a alguns meses, quando todo este processo tiver terminado, visto que, a seguir à recolha dos dados proceder-se-á ao seu tratamento.

Bárbara Bárcia

(in Jornal da Região, 12/03/97)

Que benefícios para a população podem advir dos resultados de tal recenseamento? Em que é que esses resultados podem ajudar a Junta de Freguesia da Parede na tomada de decisões? Estas são questões que o Professor pode discutir com os alunos em face de notícias desta natureza.

Seria interessante também se o Professor pudesse levar consigo um exemplar do formulário relativo ao recenseamento geral da população entregue pelo INE, de modo a poder discutir com os alunos possíveis implicações sociais e económicas que os resultados do inquérito possam trazer. Exemplos de alguns resultados extraídos do recenseamento de 1991:

- Existiam 1 235 948 famílias (de vários tipos) com pelo menos uma criança.

25

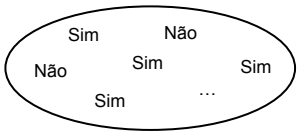
- Existiam 100 977 famílias monoparentais, com pelo menos uma criança com menos de 15 anos, em que esta ou estas viviam com o pai ou com a mãe - maioritariamente com a mãe: 89% dos casos) e em cerca de metade dos casos sem outros adultos.
- Existiam 18 034 famílias com crianças com menos de 15 anos, vivendo apenas com um ou os dois avós.
- 8 616 crianças viviam em alojamentos descritos como "barracas", especialmente junto às grandes cidades.

1.3 - Estatística Descritiva e Estatística Indutiva (Inferência Estatística)

Uma vez recolhida a amostra procede-se ao seu estudo. Este consiste em resumir a informação contida na amostra construindo tabelas, gráficos e calculando algumas características amostrais (estatísticas). Este estudo descritivo dos dados é o objectivo da *Estatística Descritiva*. No entanto, ao estudar a amostra tem-se, normalmente, como objectivo final inferir para a população as propriedades estudadas na amostra. Assim o objectivo do estudo estatístico pode ser o de estimar uma quantidade ou testar uma hipótese, utilizando-se técnicas estatísticas convenientes, as quais realçam toda a potencialidade da Estatística, na medida em que vão permitir tirar conclusões acerca de uma população, baseando-se numa pequena amostra, dando-nos ainda uma medida do erro cometido. Esta quantificação do erro cometido, ao transportar para a população as propriedades verificadas na amostra, é feita utilizando a Probabilidade. Efectivamente, é nesta fase do processo estatístico que temos necessidade de entrar com este conceito, para quantificar a incerteza associada aos procedimentos aqui considerados.

Exemplo 6 - O Senhor X, candidato à Câmara da cidade do Porto, pretende saber, qual a percentagem de eleitores que pensam votar nele nas próximas eleições. Havendo algumas limitações de tempo e dinheiro, a empresa encarregada de fazer o estudo pretendido decidiu recolher uma amostra de dimensão 1000, perguntando a cada eleitor se sim ou não pensava votar no Senhor X. Como resultado da amostragem obteve-se um conjunto de sim's e não's, cujo aspecto não é muito agradável, pois à primeira vista não conseguimos concluir nada:

26



Procede-se à redução dos dados, resumindo a informação sobre quantos sim's se obtiveram, chegando-se à conclusão que nas 1000 respostas, 635 foram afirmativas. Então dizemos que a percentagem de eleitores que pensam votar no candidato, de entre os inquiridos, é de 63.5%. A função da Estatística Descritiva acabou aqui! (Se toda a População tivesse sido inquirida, este estudo descritivo dar-nos-ia a informação necessária para o fim em vista).

Poderemos agora inferir que 63.5% dos eleitores da cidade do Porto pensam votar no Senhor X? A resposta a esta pergunta nem é sim, nem não, mas talvez. É agora que temos necessidade de utilizar o conceito de Probabilidade, para quantificar a incerteza associada à inferência. Assim, existem processos de inferência estatística que, do resultado obtido a partir da amostra, nos permitirão concluir que o intervalo [60.5%, 66.5%] contém o valor exacto para a percentagem de eleitores da cidade que pensam votar no Senhor X, com uma confiança de 95%.

Nota - A confiança de 95% deve ser entendida no seguinte sentido: se se recolherem 100 amostras, cada uma de dimensão 1000, então poderemos construir 100 intervalos; destes 100 intervalos esperamos que 95 contenham o verdadeiro valor da percentagem (desconhecida) de eleitores da cidade do Porto, que pensam votar no candidato.

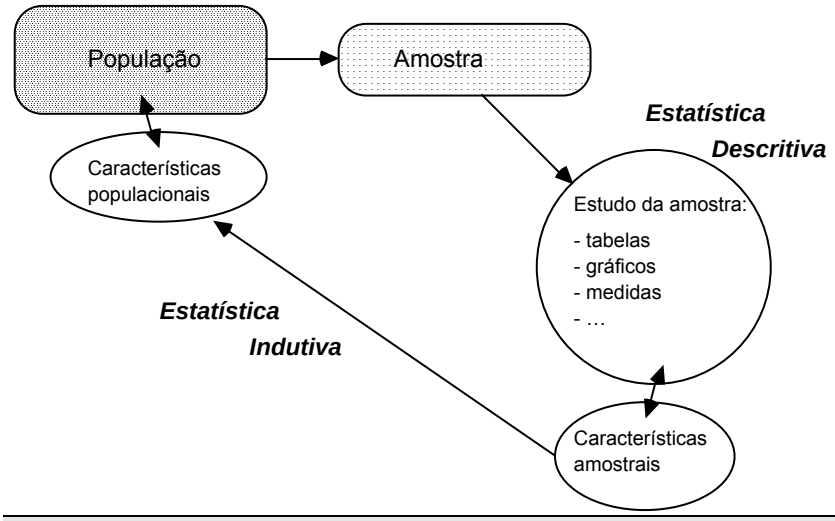
Nesta altura o Professor poderá recordar aos alunos a forma como as previsões são dadas, em noite de eleições, sob a forma de intervalos. Poderá referir que por vezes a guerra de audiências faz com que estas previsões tenham pouco sentido, por apresentarem intervalos com uma tão grande amplitude que a sua precisão, como estimativas das percentagens pretendidas, é muito pequena. Esta situação prende-se com o facto de as amostras utilizadas para a construção dos intervalos terem uma dimensão muito reduzida, havendo assim muito pouca informação disponível. No entanto, à medida que a noite vai avançando, os intervalos vão diminuindo de amplitude, estando esta diminuição da amplitude relacionada com a dimensão da amostra que

27

entretanto vai aumentando, até finalmente estarem todos os votos contados. Nesta altura, os intervalos reduzem-se a pontos, que são as percentagens pretendidas.

Poder-se-á também chamar a atenção para que a compreensão do processo estatístico nos permitirá compreender melhor notícias que, com muita frequência, se lêem nos jornais ou ouvem na televisão. Por vezes alguns estudos sobre os mesmos assuntos, apresentam resultados que chegam a ser contraditórios! Isto acontece nomeadamente no estudo de certos aspectos do comportamento humano, utilizando testes psicológicos, ou no estudo de certas doenças utilizando cobaias. Muitas das inferências feitas são imperfeitas, a maior parte das vezes por terem como base dados imperfeitos.

O seguinte esquema pretende resumir as diferentes etapas que normalmente são seguidas num procedimento estatístico:



Sugestões didácticas e comentários

Das situações a seguir indicadas refira quais constituem exemplos de Estatística Descritiva e de Inferência Estatística:

1. Um lote de 100 aparelhos de televisão considera-se em bom estado para venda se ao serem testados 10 eles não apresentarem deficiências.

Temos aqui um exemplo de Inferência Estatística. De uma amostra de 10 televisores infere-se para a população do lote de 100. Acredita-se, com base na teoria da Inferência Estatística, que se 10 televisores aleatoriamente seleccionados (seleccionados ao acaso) estiverem todos bons, então o mesmo deve acontecer aos restantes.

28

2. Um teste à opinião pública revelou que 65% da população portuguesa apoiava um determinado candidato para Presidente da República. Se esse candidato se apresentar às eleições, é de esperar que ele ganhe.

Temos novamente aqui um exemplo de Inferência Estatística. Sendo a amostra representativa da população de todos os eleitores Portugueses, então é de esperar que o que se passa na amostra também se passe na população e portanto que mais do que 50% dos Portugueses votem nesse candidato.

3. Os 120 empregados de um fabrica ganha em média 100 mil escudos por mês.

Aqui temos apenas um problema de Estatística Descritiva visto que a informação foi feita com base nos dados relativos ao salário de todos os empregados da empresa.

4. Baseados numa amostra de 500 trabalhadores de uma empresa de construção civil, acredita-se que a média dos salários dos trabalhadores de esse ramo é de 110 000$00.

Como apenas se estudou o salário de uma amostra de trabalhadores da empresa, estamos perante um problema de Inferência Estatística.

Nota: Ao discutir cada exemplo, o Professor deve lembrar que há sempre um erro, medido em termos de probabilidade, associado a qualquer Inferência Estatística que se faça. Esse erro depende, além de outros factores, da dimensão da amostra. Assim, no 1º exemplo a inferência que fizermos é tanto mais segura quanto mais televisores forem inspeccionados, sendo certa apenas se inspeccionarmos todos os televisores. Repare-se que também, no exemplo 2, a inferência será tanto mais segura quanto mais eleitores se inquirirem. No entanto, nunca podemos ter uma garantia de 100% que o Candidato ganhe as eleições pois pode haver sempre alteração de opinião.

1.4 - Exemplos de aplicação da Estatística

Estudos de mercado - O gerente de uma fábrica de detergentes pretende lançar um novo produto para lavar a loiça, pelo que encarrega uma empresa especialista em estudos de mercado, de "estimar" a percentagem de potenciais compradores desse produto.

População - conjunto de todos os agregados familiares do País.

Amostra - conjunto de alguns agregados familiares, inquiridos pela empresa.

29

Problema - pretende-se a partir da percentagem de respostas afirmativas, de entre os inquiridos, sobre a compra do novo produto, obter uma *estimativa* do número de compradores na população.

Medicina - Pretende-se estudar o efeito de um novo medicamento para curar determinada doença. É seleccionado um grupo de 20 doentes, administrando-se o medicamento a 10 desses doentes escolhidos ao acaso, e o medicamento habitual aos restantes.

População - conjunto de todos os doentes com a doença que o medicamento a estudar pretende tratar.

Amostra - conjunto dos 20 doentes seleccionados.

Problema - pretende-se, a partir dos resultados obtidos, realizar um *teste de hipóteses* para tomar uma decisão sobre qual dos medicamentos é melhor.

Controlo de qualidade- O administrador de uma fábrica de parafusos pretende assegurar-se de que a percentagem de peças defeituosas, não excede um determinado valor, a partir do qual determinada encomenda poderia ser rejeitada.

População - conjunto de todos os parafusos fabricados ou a fabricar pela fábrica.

Amostra - conjunto de alguns parafusos, escolhidos ao acaso, de entre o lote de produzidos.

Problema - pretende-se, a partir da percentagem de parafusos defeituosos presentes na amostra, estimar a percentagem de defeituosos em toda a produção.

Pedagogia - Um conjunto de pedagogos desenvolveu uma técnica nova para a aprendizagem da leitura na escola primária, a qual, segundo dizem, encurta o tempo de aprendizagem relativamente ao método habitual.

População - conjunto dos alunos que entram para a escola primária sem saber ler.

Amostra - conjunto de alunos de algumas escolas, seleccionadas para o estudo. Os alunos foram separados em dois grupos para se aplicarem as duas técnicas em confronto.

Problema - a partir dos tempos de aprendizagem obtidos verificar se existe evidência significativa para afirmar que os tempos com a nova técnica são menores.

30

Capítulo 2

ANÁLISE, REPRESENTAÇÃO E REDUÇÃO DE DADOS TABELAS E GRÁFICOS

2.1 - Introdução

A forma como se organiza e reduz a informação obtida a partir da observação da amostra utilizando *tabelas*, *gráficos* e *medidas*, depende em grande parte do tipo de dados a estudar. Estes processos de análise procuram responder a algumas questões, tais como:

- Serão os dados quase todos iguais?
- Serão muito diferentes uns dos outros?
- De que modo é que são diferentes?
- Existe alguma estrutura subjacente ou alguma tendência?
- Existem alguns agrupamentos especiais?
- Existem alguns dados muito diferentes da maior parte?

Estas questões não podem ser respondidas rapidamente, olhando unicamente para um conjunto de dados! No entanto, se estiverem organizados sob a forma de tabelas ou gráficos, já a resposta às questões anteriores se torna mais simples.

Seguidamente começaremos por dar uma possível classificação para os dados e os processos adequados para a sua representação. Estes processos de redução dos dados permitem realçar as características principais e a estrutura subjacente, à custa de alguma informação que se perde, mas que não é relevante para o estudo em vista.

2.2 - Tipos de dados. Frequência absoluta e relativa

Como se sabe o objectivo da Estatística é o estudo de Populações com características comuns. A uma característica comum que possa assumir valores ou modalidades diferentes, de indivíduo para indivíduo, chamamos *variável*. As variáveis podem ser de dois tipos: *qualitativas* e *quantitativas*. Para os *dados estatísticos* - resultado da observação de uma variável, também se usa a mesma terminologia, conforme resultem da observação de variáveis qualitativas ou quantitativas.

2.2.1 - Dados qualitativos

> ***Dados qualitativos*** *- Representam a informação que identifica alguma qualidade, categoria ou característica, não susceptível de medida, mas de classificação, assumindo várias modalidades.*

Por exemplo, o estado civil de um indivíduo é um dado qualitativo, assumindo as categorias : solteiro, casado, divorciado e viúvo.

Ao conjunto de dados, resultantes da observação de alguns elementos da Popu-lação dá-se o nome de amostra observada ou simplesmente amostra. Assim, no que se segue utilizaremos o termo amostra com o significado de conjunto de dados.

Dado um conjunto de dados, estes são organizados na forma de uma *tabela de frequências*, que apresenta o número de elementos - *frequência absoluta* (ou só frequência) de cada uma das modalidades ou *classes*, que os dados assumem.

Numa tabela de frequências, além das frequências absolutas, também se apresentam as *frequências relativas,* onde

frequência relativa = **Erro!**

entendendo-se por *dimensão* da amostra o número de elementos da amostra.

Exemplo 1: Perguntou-se a cada um dos 100 habitantes de uma determinada aldeia, qual a telenovela preferida, do seguinte conjunto:

CI - Cinzas **PP** - Pedra sobre Pedra **CA**- Corpo e Alma

MP - Mico Preto **BA** - Barriga de Aluguer **PL** - Plumas e Lantejoulas

Obtiveram-se os seguintes resultados (Obviamente que ninguém respondeu " Não gosto de nenhuma" …):

Classes	Freq. abs.	Freq. rel.
CI	11	0.11
PP	31	0.31
BA	8	0.08
CA	21	0.21
PL	13	0.13
MP	16	0.16
Total	100	1.00

A redução dos dados anteriores segundo uma tabela de frequências permite concluir imediatamente que:

A novela preferida por mais pessoas é a Pedra sobre Pedra

A novela preferida por menos pessoas é a Barriga de Aluguer

Estas conclusões não seriam tão evidentes a partir dos dados inicialmente recolhidos. Ao fazer a redução, sob a forma de uma tabela de frequências, a única informação que se perdeu foi a ordenação inicial dos dados.

Quando se constrói uma tabela de frequências, a partir de uma amostra, um processo de fácil verificação de que as frequências estão bem calculadas consiste em somá-las para todas as classes consideradas, pois:

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

Exemplo 2: A seguinte tabela apresenta a distribuição de pessoal docente (freq. absolutas), segundo os ramos de ensino, em Portugal Continental, durante os anos de 1985-1986 e 1986-1987 (Fonte: Anuário Estatístico de Portugal - 1992)

	Básico		**Secundário**				
	Primário	Preparat.	Sec. Unific	Sec. comp.	12ºano	Liceal	Técnico
1985-1986	41534	29189	28675	14187	3584	3069	2216
1986-1987	41553	31742	28751	15171	4136	3454	2656

(cont)

	Cursos Profission	**Artístico**	**Médio** Mag.Infantil	Mag.Primário	Total
1985-1986	1281	629	535	571	125470
1986-1987	969	602	414	485	129933

Observação: Não foram considerados os ensinos pré-escolar e superior por não haver informação disponível completa.

A utilização das frequências relativas é preferível, relativamente às frequências absolutas, pois assim é possível fazer a comparação de conjuntos de dados de dimensões diferentes. É o que se passa no caso do exemplo presente, em que as dimensões dos conjuntos relativamente a 1985-1986 e 1986-1987 são respectivamente 125470 e 129933.

	Básico Primário	Preparat.	**Secundário** Sec. Unific	Sec. comp.	12ºano	Liceal	Técnico
1985-1986	0.331	0.233	0.229	0.113	0.029	0.024	0.018
1986-1987	0.320	0.244	0.221	0.117	0.032	0.027	0.020

(cont)

	Cursos Profission.	**Artístico**	**Médio** Mag.Infantil	Mag.Primário	Total
1985-1986	0.010	0.005	0.004	0.005	1
1986-1987	0.007	0.005	0.003	0.004	1

Da tabela das frequências relativas, podemos concluir qual a evolução, em termos percentuais dos docentes dos diferentes tipos, de um ano para o outro. Repare-se que embora o nº de docentes do ensino Secundário Unificado tenha aumentado, em termos percentuais houve uma diminuição.

2.2.2 - Dados quantitativos

> ***Dados quantitativos*** *- Representam a informação resultante de características susceptíveis de serem medidas, apresentando-se com diferentes intensidades, que podem ser de natureza* ***discreta*** *- dados discretos, ou* ***contínua*** *- dados contínuos.*

Uma variável é **discreta** se só pode tomar um nº finito (ou infinito numerável) de valores distintos. É o caso, por exemplo, do nº de acidentes, por dia, num determi-nado cruzamento.

No caso de uma variável **contínua**, esta pode tomar todos os valores numéricos, compreendidos no seu intervalo de variação. É o caso, por exemplo, do peso, da altura, etc.

Nota: Chama-se a atenção para que a classificação de uma variável em discreta ou contínua, é por vezes susceptível de algumas dúvidas. Por exemplo a variável idade, ao contrário do que possa parecer à primeira vista, já que só utilizamos nú-meros inteiros para a representar, é uma variável contínua, pois a diferença de ida-de entre dois indivíduos pode ser tão pequena quanto se queira - um ano, um mês, uma hora, um minuto, …. Podemos dizer que a variável é contínua quando, para se passar de um valor a outro, se tem de passar por todos os pontos intermédios.

Como organizar os dados?

Os dados são organizados na forma de uma *tabela de frequências*, do mesmo modo que os dados qualitativos. No entanto convém fazer distinção entre os dados discretos e contínuos, já que a construção da tabela de frequências se processa, de um modo geral, de forma diferente.

Assim, no caso de dados discretos, a construção da tabela é análoga à que foi feita para os dados qualitativos, mas em vez das categorias consideram-se os valores distintos que surgem na amostra, os quais vão constituir as *classes*.

Exemplo 3: Numa turma do 10º ano da Escola Secundária Professor Herculano de Carvalho, em Lisboa, os alunos registaram o nº de irmãos, tendo-se obtido o seguinte conjunto de dados:

1 2 2 1 3 0 0 1 1 2

1 1 1 0 0 3 4 3 1 2

Tabela de frequências

Classes	Freq. abs.	Freq. rel.	Freq.rel.acum
0	4	0.20	0.20
1	8	0.40	0.60
2	4	0.20	0.80
3	3	0.15	0.95
4	1	0.05	1.00
Total	20	1.00	-

Introduzimos na tabela de frequências mais uma coluna, com as frequências relativas acumuladas. Pode servir, por exemplo, para calcular a mediana e os quartis, como veremos um pouco mais tarde.

Podemos no entanto dispor de uma amostra de dados discretos, mas estes assumi-rem muitos valores distintos, que torne pouco prático a construção de uma tabela de frequências, onde se consideram todos esses valores como classes. Neste caso procede-se a um agrupamento conveniente para os dados, como se exemplifica a seguir.

Exemplo 4: No Distrito Sanitário de Chicago, a escolha dos técnicos é feita mediante um exame. Em 1966, havia 223 candidatos para 15 lugares. 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 (Freedman *et al*., 1991 *Statistics*, pag.51):

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 poderia processar-se do mesmo modo que no exemplo anterior; resultaria, no entanto, 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.

Tabela de frequências

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	0.999

Definição das classes

Enquanto que no caso dos dados discretos a construção da tabela de frequências é, de um modo geral, muito simples, no caso de variáveis contínuas o processo é um pouco mais elaborado, já que a definição das classes não é tão imediata. Efectivamente não tem sentido considerar, para classes, os diferentes valores que surgem na amostra, pois eventualmente eles são todos diferentes.

De um modo geral, as classes vão ser intervalos fechados à esquerda e abertos à direita, todos eles com a mesma amplitude. As classes não se devem sobrepor nem deixar intervalos entre elas. O valor mínimo da amostra deve pertencer à primeira classe e o máximo deve pertencer à última.

O número total de classes e a amplitude da cada classe estão relacionados entre si: se a amplitude aumentar, o número de classes diminui, e vice-versa.

Normalmente, é conveniente que os extremos de cada classe sejam números de fácil leitura de modo a que, quando se observa uma tabela ou um gráfico, se tenha imediatamente ideia do significado de cada classe.

Em certos casos, não é conveniente que as classes tenham todas a mesma amplitude. Nessa altura é preciso não esquecer que as classes são disjuntas duas a duas e que a sua união contém todos os elementos da amostra.

Quantas classes se devem considerar no estudo de uma amostra?

Não há uma regra definitiva, sendo esta precisamente uma das etapas que pode causar mais dificuldades na organização dos dados na forma de uma tabela de frequências. Um número exagerado de classes não permite sobressair a forma da distribuição subjacente aos dados, isto é não permite ter uma ideia global da situação; por outro lado um número muito pequeno de classes, despreza muita informação e pode esconder algumas características interessantes que não são realçadas.

Existe uma regra empírica que nos dá um valor **aproximado** para o número de classes:

– Para uma amostra de dimensão n, o número de classes k é o menor inteiro tal que $2^k \geq n$.

Esta regra deve ser encarada como uma ajuda para iniciar o estudo de um conjunto de dados, quando não há qualquer outra indicação à partida que nos ajude a decidir em quantas classes vamos organizar os dados.

Exemplo 5: Os dados seguintes (que se encontram ordenados) referem-se ao tempo de vida (em anos) de 50 doentes que nasceram com uma certa doença rara :

0.8	1.7	2.5	4.8	9.7	16.2	23.5	28.1	33.2	45.0
0.9	1.9	2.6	6.3	13.5	18.2	23.6	29.7	36.6	45.1
1.0	2.0	2.6	6.9	13.5	18.2	23.7	30.9	36.7	61.7
1.1	2.0	3.2	7.6	14.4	20.7	27.1	31.2	38.0	66.4
1.1	2.4	3.5	9.0	15.5	21.8	27.6	31.7	40.2	67.4

Dimensão da amostra: 50

De acordo com a regra empírica apresentada anteriormente teríamos:

Número de classes: k = 6, pois $2^6>50$, mas $2^5<50$

Amplitude de classe ≈ **Erro!**≈ **Erro!**≈ 11.1

Podemos escolher para amplitude de classe h=10 (é mais sugestivo considerar intervalos com amplitude de 10 anos do que um valor próximo do sugerido).

Por outro lado vamos começar por construir as classes, considerando para limite inferior da 1ª classe o valor 0, já que o mínimo da amostra está próximo desse valor. Com esta escolha obtemos 7 classes, em vez do valor 6 sugerido pela regra:

Tabela de frequências

Classes	Freq. abs.	Freq. rel.
[0, 10[	21	0.42
[10, 20[	7	0.14
[20, 30[	9	0.18
[30, 40[	7	0.14
[40, 50[	3	0.06
[50, 60[	0	0.00
[60, 70[	3	0.06
Total	50	1.00

Nota 1: Um erro que se comete com muita frequência é considerar a última classe fechada à direita. Este procedimento não é correcto. Todas as classes devem ser construídas segundo a mesma metodologia, isto é, fechadas à esquerda e abertas à direita.

Nota 2: Para definir um conjunto de classes associado a um conjunto de dados, de-ve-se ter em conta que, de um modo geral, quanto mais elementos tiver a amostra, maior será o número de classes que se deve considerar (o que está de acordo com a regra indicada). No entanto, mesmo que a dimensão da amostra seja suficiente-mente grande, não é aconselhável considerar um número de classes superior a 15.

Exemplo 6 - Foram inquiridos 75 agregados familiares de uma determinado zona residencial, com o objectivo de tomar decisões a muito curto prazo sobre as necessidades da rede escolar. Cada agregado familiar deu indicações sobre as idades dos filhos entre os 3 e os 18 anos. Obteve-se uma amostra de dimensão 133, a qual se organizou na seguinte tabela de frequências:

Tabela de frequências

Classes	Freq.abs.	Freq.rel.
[3, 6[	44	0.33
[6, 10[	36	0.27
[10, 12[	28	0.21
[12, 15[	15	0.11
[15, 19[	10	0.08
Total	133	1.00

Qual o critério utilizado na definição das classes? O que ressalta da tabela quanto à classe etária da população da dita zona residencial e quanto às necessidades, no que diz respeito à rede escolar?

Comentário: Na definição das classes anteriores teve-se em conta o objectivo do estudo sobre as necessidades da rede escolar. Assim, consideraram-se como classes as classes etárias que correspondem, de uma maneira geral, aos diferentes graus de ensino. Da análise da tabela conclui-se que na dita zona residencial a população é relativamente jovem, havendo predominância de crianças em idade pré-escolar, pelo que se deve começar a pensar em criar meios, para daqui a alguns anos, essas crianças terem acesso à escolaridade obrigatória e eventualmente ao secundário.

Sugestões didácticas e comentários

1. É importante que os alunos interpretem a situação dada, a fim de criticarem o critério que foi usado para a definição das classes e que também sejam solicitados a decidir qual o número de classes e amplitude de classe mais adequada para um determinado conjunto de dados. Por exemplo, sugira que se discuta o critério usado em cada uma das seguintes situações:

a) Tempo de reacção muscular a um impulso medido em milésimas de segundo (classes de amplitude .005):

0.206 0.209 0.218 0.226 0.239 0.224 0.207 0.215 0.219 0.222
0.225 0.219 0.218 0.245 0.220 0.237 0.207 0.245 0.207 0.222

b) Pontuações de um teste de Matemática numa escala de 0 a 100, onde houve classificações entre 24 e 65 (amplitude 3).

c) Idades dos professores de uma escola portuguesa do 1º ciclo, com idades compreendidas entre 24 e 65 anos (amplitude 3).

Nota: As situações das alíneas b) e c) são propositadamente ambíguas, pois não se sabe qual a dimensão da amostra, nem o que se pretende com os dados a analisar. Por exemplo, no caso do exemplo b) não é indiferente se as pontuações se referem a uma turma ou à escola toda. No primeiro caso pode não ter qualquer interesse considerar classes com aquela amplitude, pois correr-se-ia o risco da maior parte das classes consideradas ter frequência nula. Por outro lado pareceria muito mais interessante considerar classes de amplitude 5, já que nos transmite informação de uma forma mais

sugestiva. No caso da alínea c) será que tem interesse saber quantos professores estão perto da reforma, para fazer uma programação atempada das necessidades? Se sim, talvez se justifique considerar classes com aquela amplitude. São estas condicionantes que devem ser objecto de discussão.

2. A discussão à volta dos possíveis critérios utilizados nestes exemplos para a amplitude das classes, permite que os alunos se apercebam que a melhor escolha depende por vezes dos objectivos do estudo.

Dada um determinado conjunto de dados solicite aos alunos, em trabalho de gru-po, que escolham as classes que lhes parecem mais apropriadas a uma deter-minada situação. Peça para irem apresentar a sua solução, justificando a escolha que fizeram. Confronte as diferentes soluções e promova a discussão na aula.

2.3 - Representação gráfica de dados

2.3.1 - Variáveis discretas. Diagrama de barras.

Vimos que, no caso de dados discretos, a construção da tabela de frequências se resume, de um modo geral, a considerar como classes os diferentes valores que surgem na amostra. Uma representação gráfica adequada para estes dados, é o diagrama de barras.

> ***Diagrama de barras*** *- Representação gráfica, que consiste em marcar num sistema de eixos coordenados, no eixo dos xx, o valor das classes e nesses pontos barras verticais de altura igual à frequência absoluta ou à frequência relativa.*

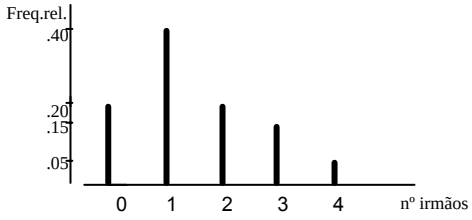
Algumas considerações sobre os passos a seguir na construção do diagrama de barras:

1 - Ordenar a amostra e considerar para classes os diferentes valores aí conside-rados. Marcar essas classes no eixo dos xx, num sistema de eixos coordenados.

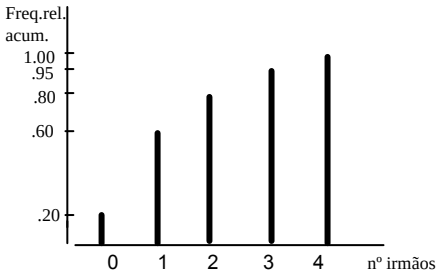
2 - Nos pontos onde se consideraram as classes, marcar barras de altura igual à fre-quência absoluta ou relativa, da respectiva classe. De preferência utilizar as fre-

quências relativas, pois para comparar diagramas de barras de amostras diferentes, temos a garantia de que a soma das barras é igual a 1.

Exemplo 3 (cont): O diagrama de barras que representa a distribuição das frequên-cias do nº de irmãos dos alunos da turma considerada, tem o seguinte aspecto:

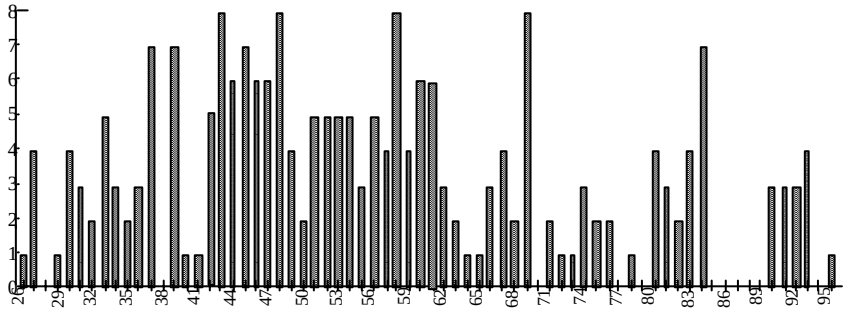


Para representar graficamente as frequências relativas (absolutas) acumuladas, considera-se um diagrama de barras em que as barras têm comprimento igual às frequências acumuladas.



Quer as tabelas, quer os gráficos das frequências acumuladas são úteis na determinação de certas medidas de localização a que chamamos mediana e quartis.

Exemplo 4 (cont) - A partir da tabela de frequências, considerando todos os valores distintos que compõem o conjunto de dados, construiu-se o seguinte diagrama de barras:



Da análise do gráfico anterior verifica-se a existência de uma lacuna, não havendo classificações iguais a 85, 86, 87, 88 e 89 e o nº de classificações iguais ou superiores a 90 ser de 15, precisamente igual ao nº de lugares vagos, para os 223 candidatos. Não terá havido batota da parte dos examinadores?

Nota: Não se aconselha pedir aos alunos a construção de gráficos que envolvam tantas classes como o exemplo anterior, se eles não dispuserem de meios computacionais.

2.3.2 - Variáveis contínuas. Histograma. Função cumulativa.

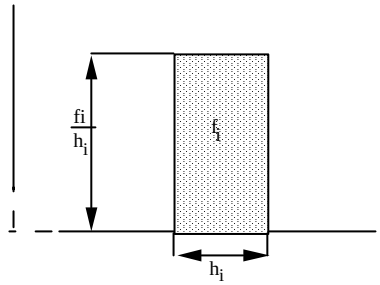
2.3.2.1 - Histograma

Já vimos anteriormente a forma de obter a tabela de frequências de uma amostra de dados contínuos. Ao contrário do caso anterior, agora as classes já não são pon-tos isolados, mas intervalos. Assim, a representação gráfica já não pode ser o dia-grama de barras, pois não existem pontos isolados, onde colocar as barras! Vejamos como construir a representação gráfica adequada, que se chama *histograma*.

> ***Histograma*** *- Para a representação gráfica de dados contínuos, usa-se um diagrama de áreas ou histograma, formado por uma sucessão de rectângulos adjacentes, tendo cada um por base um intervalo de classe e por área a frequência relativa (ou a frequência absoluta). Deste modo, a área total coberta pelo histograma é igual a 1 (respectivamente igual a n, a dimensão da amostra).*

Para construir o histograma, quais as alturas que se devem considerar para os rectângulos?

Se se pretende que a área do rectângulo, correspondente à classe de ordem i, seja a frequência relativa f_i (ou absoluta n_i), então a altura desse rectângulo deverá ser **Erro!**, onde h_i representa a amplitude da classe i.



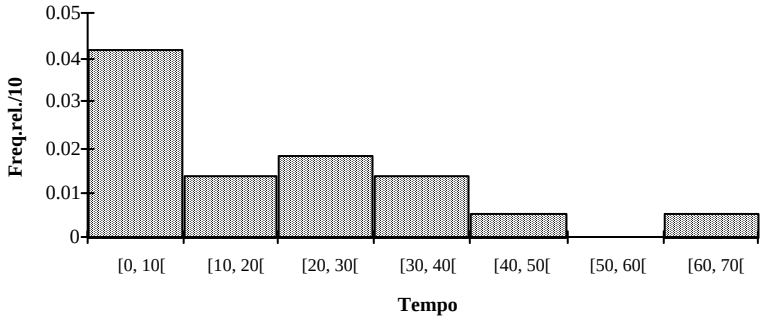
Nota 1: Se todas as classes tiverem a mesma amplitude, então $h_i = h$. Neste caso, por vezes constroem-se os rectângulos com alturas iguais às frequências relativas (absolutas) das respectivas classes, vindo as áreas dos rectângulos proporcionais e não iguais às frequências. A constante de proporcionalidade é a amplitude de classe. No entanto, se se pretender comparar várias amostras através de histogramas, deve-se ter o cuidado de os construir de forma indicada inicialmente, de modo que a área total ocupada por cada um dos histogramas seja 1.

Nota 2: Um erro que se costuma cometer com muita frequência é construir o histograma com os rectângulos separados! Este procedimento não é correcto, pois os rectângulos são adjacentes, dando no seu conjunto a ideia de uma área.

Exemplo 5 (cont) - Para tornar mais simples a construção do histograma, incluímos na tabela de frequências uma nova coluna em que para cada classe se considerou a frequência relativa a dividir pela amplitude de classe:

Tabela de frequências

Classes	Freq. abs.	Freq. rel.	Freq.rel.acum	Freq.rel./h
[0, 10[	21	0.42	0.42	0.042
[10, 20[	7	0.14	0.56	0.014
[20, 30[	9	0.18	0.74	0.018
[30, 40[	7	0.14	0.88	0.014
[40, 50[	3	0.06	0.94	0.006
[50, 60[	0	0.00	0.94	0.000
[60, 70[	3	0.06	1.00	0.006
Total	50	1.00	-	-



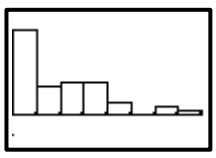
A área total ocupada pelo histograma é igual a 1.

Actividade - Construção do **HISTOGRAMA** utilizando a máquina de calcular.

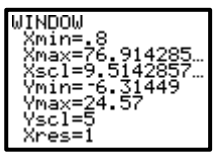
Podemos obter um histograma com a calculadora gráfica. Para isso, começamos por inserir os dados numa lista, normalmente em L1.

Depois vamos a **STAT PLOT**, escolhemos **1:Plot 1** e seleccionamos as opções indicadas na figura.

Se em **ZOOM** escolhermos **9:ZoomStat**, a máquina traça um histograma com um certo número de classes.

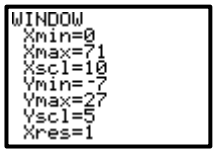


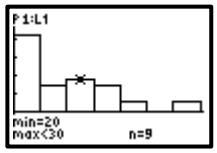
Carregando em **WINDOW** vemos que a primeira classe começa em 0.8 e a última termina em 76.914, sendo a amplitude das classes, indicada em **Xscl**, de aproxima-damente 9.514.



Se quisermos escolher a amplitude das classes e o início da primeira classe, basta alterar em **WINDOW** os respectivos valores.

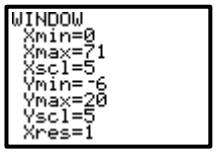
Por exemplo, começando em 0 com amplitude de 10, obtemos este histograma.

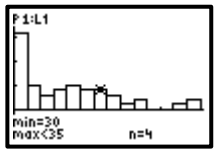




Se teclarmos **TRACE** e deslocarmos o cursor, podemos ver quantos elementos exis-tem em cada classe. No caso indicado na figura, a classe [20;30[tem 9 elementos.

Se quisermos classes de ampli-tude 5, basta fazer **Xscl=5** e ada-ptar os valores no eixo dos YY de modo a obter um histograma com aspecto aceitável.





Fazendo TRACE vemos que, por exemplo, a classe [30 ; 35[tem 4 elementos.

Muitas vezes fazemos um estudo de uma certa amostra na calculadora gráfica e depois não nos convém apagar os dados introduzidos porque iremos precisar deles mais tarde. Temos por isso de guardá-los numa lista própria.

Para isso, teclamos **2nd L1** e **STO–>** e depois criamos uma lista com a seguinte sequência:

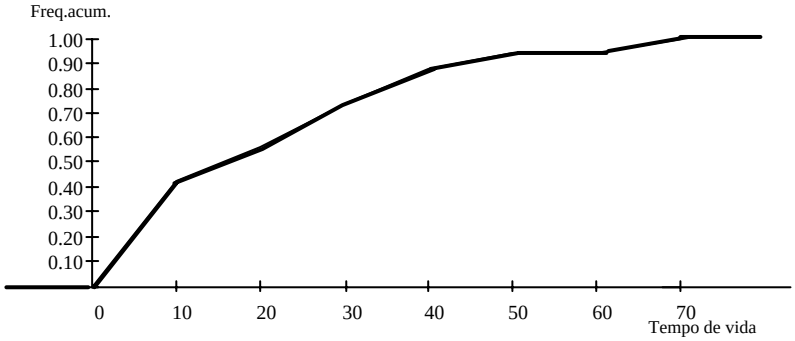
2nd LIST OPS B: L

e escrevemos a seguir o nome que queremos dar a esta lista, com um máximo de 5 caracteres (escolhemos DOENT).

Teclando **ENTER** os dados que estavam em L1 ficam guardados na lista LDOENT. Quando quisermos voltar a usar estes dados basta ir buscar esta lista a **LIST**.

2.3.2.2 - Função cumulativa

Para representar graficamente as frequências acumuladas considera-se a função cumulativa cuja construção se exemplifica a seguir:

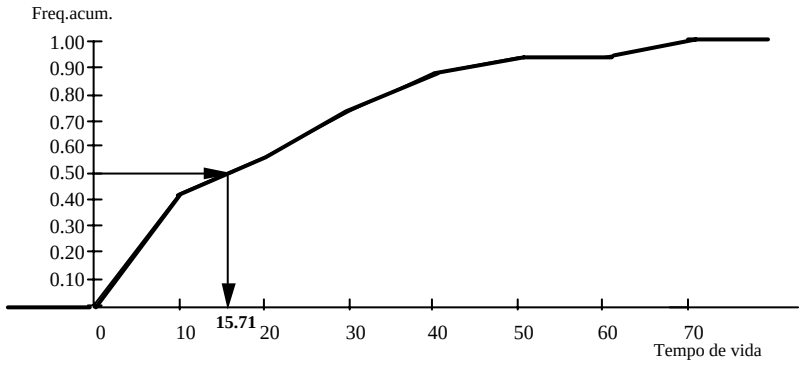


- Antes do limite inferior da 1ª classe, isto é o ponto 0, a frequência acumulada é nula, pelo que se traça um segmento sobre o eixo dos xx, até esse ponto.
- No limite inferior da 2ª classe, isto é o ponto 10, a frequência acumulada é a frequência da classe anterior, ou seja 0.42. Agora, admitindo que a frequência se distribui uniformemente sobre o intervalo da classe, unimos o ponto (0, 0) com o ponto (10, 0.42).
- No limite inferior da 3ª classe, a frequência acumulada é a soma das frequências das duas classes anteriores, sendo portanto 0.56. Então, unimos o ponto (10, 0.42) com o ponto (20, 0.56).
- Quando chegarmos à última classe, temos a garantia que a frequência acumulada correspondente ao seu limite superior é igual a 1, pelo que nesse ponto marcamos 1 e continuamos com um segmento de recta paralelo ao eixo dos xx.

Pode-se chamar a atenção para algumas propriedades da função cumulativa, tal como foi construída:

- Está definida para todo o x real;
- É sempre não decrescente;
- Só assume valores no intervalo [0, 1].

A partir da representação gráfica anterior é possível, por exemplo, saber qual o valor aproximado da variável tempo de vida a que corresponde uma frequência relativa acumulada igual a 50%.



Uma vez que se admite que a frequência se distribui uniformemente sobre a amplitude de classe, isto é, a frequência 0.14 (=0.56-0.42) distribui-se uniforme-mente sobre o intervalo de amplitude 10, através da resolução de uma equação de proporcionalidade, obtém-se o ponto que andávamos à procura:

$$\textbf{Erro!} = \textbf{Erro!}x = \textbf{Erro!} = 5.71$$

Então o valor procurado é 10 + 5.71 = 15.71.

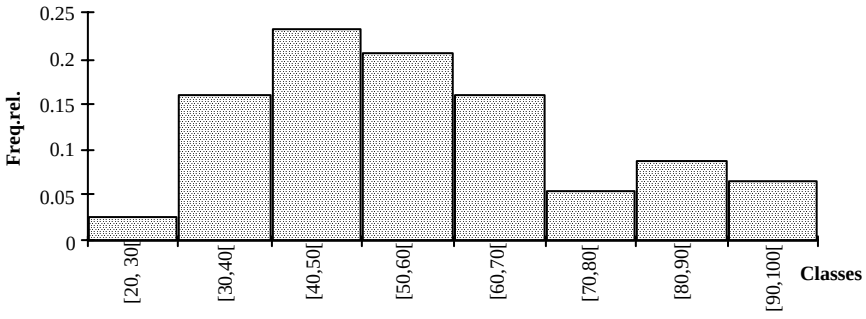
Ao valor obtido anteriormente, a que corresponde uma frequência acumulada de 50%, chamamos *mediana*. A mediana divide a distribuição das frequências em duas partes iguais, já que 50% dos dados são menores ou iguais a ela e os restantes 50% são maiores ou iguais a ela. Recordamos que a técnica utilizada permitiu-nos obter um valor aproximado para a mediana, e não o valor exacto da mediana do conjunto de dados originais, antes de proceder ao agrupamento. Mais à frente, quando falarmos de medidas de localização, veremos como determinar a mediana a partir dos dados, sem estarem agrupados.

Nota: Visto que a partir dos dados agrupados só se pode obter um valor aproximado para a mediana e não o valor exacto, aconselhamos que se peça aos alunos para obter esse valor unicamente através da representação gráfica da função cumulativa, sem ser necessário estar a proceder à interpolação.

Em vez de pretendermos determinar o valor a que corresponde a percentagem de 50%, poderíamos procurar os valores a que correspondem as percentagens de 25% ou 75%, a que chamamos quartis, respectivamente 1º quartil e 3º quartil. A técnica é análoga à seguida para a determinação da mediana.

Nota: Embora o histograma seja uma representação gráfica essencialmente para dados contínuos, também se pode utilizar para representar dados discretos, quando estes assumem muitos valores distintos, como fizemos no Exemplo 4.

Exemplo 4 (cont) - Dado o agrupamento proposto, tem-se o seguinte histograma:



Na construção dos rectângulos que formam o histograma, utilizámos para altura de cada rectângulo a frequência relativa, em vez do quociente entre a frequência relativa e a amplitude de classe, já que as classes tinham todas a mesma amplitude. Chama-se no entanto a atenção para que a área total ocupada pelo histograma já não é 1, mas sim 10 (amplitude de classe).

A representação deste conjunto de dados sob a forma do histograma, embora faça perder alguma informação, por outro lado faz sobressair a estrutura subjacente, no que diz respeito à forma da distribuição das frequências. Verifica-se que essa distri-buição apresenta uma classe, [40, 50[com maior frequência, havendo um decréscimo nas classes anteriores e posteriores, para tornar novamente a ter um "pico" na penúltima classe. Isto é sintoma de que se deve investigar um pouco mais atentamente esta cauda, já que com uma distribuição de classificações é natural esperar algumas classes centrais com maior frequência, a qual irá dimi-nuindo à medida que as classes se afastam dessas classes centrais. Algum do deta-lhe perdido diz respeito ao conjunto das 15 classificações isoladas das restantes, que se observava no diagrama de barras. Por outro lado o demasiado detalhe apresentado de um modo geral no diagrama de barras, não permite sobressair tão bem como no histograma, a estrutura subjacente à distribuição das classificações.

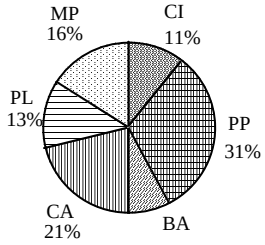
2.3.3 - Outras representações gráficas

Além das representações gráficas anteriormente consideradas, isto é, o diagrama de barras e o histograma, especialmente adequadas, respectivamente para dados discretos ou contínuos (embora o histograma também se possa utilizar para dados discretos), há outras representações, que passamos a descrever.

2.3.3.1 - Diagrama circular

Como o nome sugere esta representação é constituída por um círculo, em que se a-presentam vários sectores circulares, tantos quantas as classes consideradas na ta-bela de frequências da amostra em estudo. Os ângulos dos sectores são proporcio-nais às frequências das classes. Por exemplo uma classe com uma frequência re-lativa igual a 0.20, terá no diagrama circular um sector com um ângulo igual a 360x0.20=72 graus. É uma representação utilizada essencialmente para dados qualitativos.

Exemplo 1 (cont): O diagrama circular para este caso tem o seguinte aspecto:



Nesta representação, juntamente com a identificação da categoria, indica-se a frequência relativa da respectiva classe.

2.3.3.2 - Caule-e-folhas

É um tipo de representação que se pode considerar entre a tabela e o gráfico, uma vez que são apresentados os verdadeiros valores da amostra, mas numa apresentação sugestiva, que faz lembrar um histograma. Consiste em escrever do lado esquerdo de uma linha vertical, o dígito (ou dígitos) da classe de maior grandeza, seguidos dos restantes. Exemplificamos seguidamente a construção de uma representação em caule-e-folhas.

Exemplo 6 - Num determinado teste realizado a 48 estudantes, obtiveram-se as seguintes pontuações:

75	98	42	75	84	87	65	59	63	86	78	37
99	66	90	79	80	89	68	57	95	55	79	88
76	60	77	49	92	83	71	78	53	81	77	58

93 85 70 62 80 74 69 90 62 84 64 73

Para fazer a representação caule-e-folhas, começamos por traçar uma linha verti-cal e do lado esquerdo os dígitos dominantes, que no nosso caso é o das dezenas:

1º passo

3
4
5
6
7
8
9

2º passo

3	
4	
5	
6	
7	5
8	
9	

3º passo

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

No 1º passo limitamo-nos a colocar os dígitos dominantes, que são os caules. Agora teremos de pendurar em cada caule as folhas respectivas. O 1º número do conjunto de dados é o 75, pelo que vamos pendurar o 5 no caule 7 (2º passo). O processo repete-se até termos esgotados todos os elementos da amostra (passo 3). Finalmente é usual apresentar as folhas de cada caule ordenadas:

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

Esta representação é muito útil para ordenar amostras, pois basta agora percorrer a representação de cima para baixo, para recuperar a amostra ordenada.

Exemplo 7: No seguinte quadro, apresenta-se o número de concelhos de cada um dos distritos de Portugal Continental e das Regiões Autónomas de Açores e Madeira (Anuário Estatístico de Portugal, 1992):

Região	Nº concelhos	Região	Nº concelhos
Aveiro	19	Lisboa	15
Beja	14	Portalegre	15
Bragança	13	Porto	17
Braga	12	Santarém	21
Cast.Branco	11	Setúbal	13
Coimbra	17	Viana Cast.	10
Évora	14	Vila Real	14
Faro	16	Viseu	24

Guarda	14	Açores	19
Leiria	16	Madeira	11

Uma representação de caule-e-folhas, possível para o conjunto de dados considerado é a seguinte:

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

Nesta representação utilizamos 5 caules para o número 1, pendurando o 0 e o 1 no primeiro caule, o 2 e o 3 no segundo caule, etc. Procedeu-se de modo análogo com o 2.

Chama-se a atenção para que, embora o 2º caule correspondente ao 2 não tenha folhas penduradas, ele deve estar lá, precisamente para dar a ideia da existência de lacunas naqueles valores. Por exemplo, na representação anterior, sobressai um distrito com um número de concelhos "substancialmente" superior aos restantes, que é o distrito de Viseu com 24 concelhos.

Se não pretendêssemos tantos caules, uma alternativa seria considerar 2 caules para cada dígito dominante, pendurando no primeiro caule as folhas 0, 1, 2, 3 e 4 e no 2º caule as folhas 5, 6, 7, 8 e 9:

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

Repare-se que, em qualquer das modalidades apresentadas, cada caule tem sempre a possibilidade de ter penduradas o mesmo número de folhas diferentes: na primeira representação 2 folhas e na última representação 5 folhas.

Nota: A representação em caule-e-folhas é muito sugestiva para a representação de dois conjuntos de dados referentes à mesma característica, mas de populações diferentes, como se exemplifica a seguir.

Exemplo 8: Utilizaram-se 45 ratos de ambos os sexos, no estado adulto, e mediu-se o tempo (em segundos) de reacção a determinada droga, sendo os resultados sumariados no quadro seguinte:

Sexo	Tempo	Sexo	Tempo	Sexo	Tempo	Sexo	Tempo	Sexo	Tempo
M	142	M	142	M	151	M	121	M	152
M	126	M	128	M	141	M	115	M	127
M	134	M	132	M	120	M	99	M	138
M	112	M	107	M	55	M	120	M	130
M	199	M	118	M	123	M	101	M	95
M	97	M	108	F	33	F	37	F	30
F	90	F	58	F	41	F	65	F	102
F	52	F	55	F	68	F	61	F	66
F	53	F	50	F	64	F	71	F	74

F		M
7 3 0	3	
1	4	
8 5 3 2 0	5	5
8 6 5 4 1	6	
4 1	7	
	8	
0	9	5 7 9
2	10	1 5 7
	11	2 5 8
	12	0 0 1 3 6 7 8
	13	0 2 4 8
	14	1 2 2
	15	1 2
	16	
	17	
	18	
	19	9

Para comparar o tempo de reacção dos ratos de ambos os sexos, construimos o diagrama de caule-e-folhas, considerando os mesmos caules e dispondo as folhas para um e outro lado, conforme o sexo. Da representação anterior ressalta imediatamente o maior tempo de reacção observado, de um modo geral, nos ratos do sexo masculino, quando comparado com o do sexo feminino.

Sugestões didácticas e comentários

Não se pretende que os alunos se limitem à representação gráfica de dados, mas principalmente que os interpretem no contexto onde estão inseridos. Assim, é importante acrescentar ao pedido de elaboração do gráfico, questões que possam ajudar os alunos a fazer essa interpretação, como no exemplos seguintes:

1. Um professor de Estatística procura o método mais eficiente para ensinar Estatís-tica aos seus alunos. Assim, resolveu pôr em prática dois métodos diferentes, um em cada uma das duas turmas que leccionava. Na turma A usava o método expositivo tradicional, enquanto que na turma B promovia a discussão dos assuntos na aula e resolução de alguns problemas em grupo. Os resultados foram:

Turma A:	73	84	76	70	69	69	46	81	92	66
	87	81	78	45	67	73	88	79	95	86
Turma B:	79	75	98	81	82	70	60	82	77	81
	81	87	88	94	79	92	77	70	74	71

Através da representação em diagramas de caule-e-folhas procura-se avaliar a situação das duas turmas, relativamente aos resultados obtidos (Runyon *et al.*, 1996).

2. Para estudar o comportamento da sua turma relativamente aos dois últimos testes de Matemática, o professor pode pedir aos alunos a representação gráfica em diagramas de caule-e-folhas, dos resultados dos dois testes. Para um dos lados consideram-se os resultados do 1º teste, enquanto que para o outro lado se consideram os resultados do 2º teste. Será que a turma melhorou, piorou, ou, de um modo geral, não houve alterações significativas?

3. Para comparar as idades dos pais e das mães, pode-se pedir aos alunos da turma que indiquem as idades do pai e da mãe. Depois representam-se os dois conjuntos de dados, constituídos pelas idades das mães e pelas idades dos pais, num sistema de caule-e-

folhas, como no exemplo 8. Estes dados poderão ser mais tarde utilizados para verificar a existência de correlação.

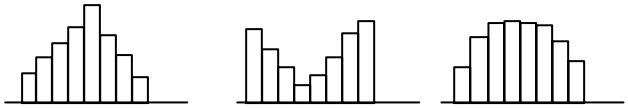
Que característica é que se pretende realçar, quando se representa um conjunto de dados, sob a forma de um histograma ou de uma representação de caule-e-folhas?

Dada uma amostra, o aspecto do histograma reflecte a forma da distribuição da População subjacente aos dados observados. Este é um dos aspectos da redução dos dados, em que se perde alguma informação contida nesses dados, mas em contrapartida obtemos a estrutura da População que eles pretendem representar.

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

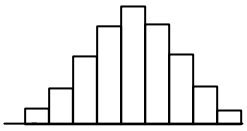
Distribuições simétricas

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



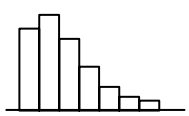
Caso especial de uma distribuição simétrica

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

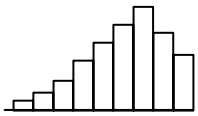


Distribuições enviesadas

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



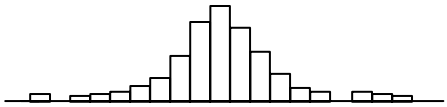
Cauda direita mais longa



Cauda esquerda mais longa

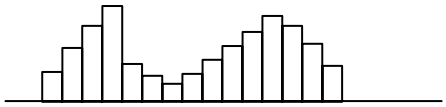
Distribuições com caudas longas

A distribuição das frequências faz-se de tal forma que existem algumas classes nos extremos, cujas frequências são muito pequenas, relativamente às classes centrais, apresentando algumas classes intermédias com frequência nula:



Distribuições com vários "picos" ou modas

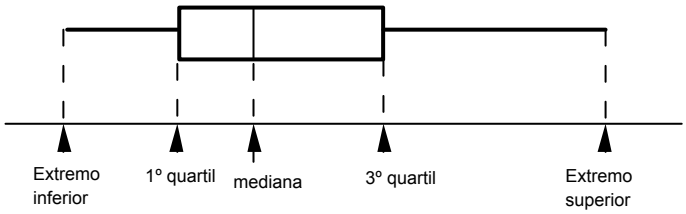
A distribuição das frequências apresenta 2 ou mais "picos" a que chamamos modas, sugerindo que os dados são constituídos por vários grupos distintos:



2.3.3.3 - Diagrama de extremos e quartis

É um tipo de representação gráfica, em que se realçam algumas características da amostra. O conjunto dos valores da amostra compreendidos entre o 1º e o 3º QUARTIS, que vamos representar por Q_1 e Q_3 é representado por um rectângulo (caixa) com a MEDIANA indicada por uma barra. A largura do rectângulo não dá qualquer informação, pelo que pode ser qualquer. Consideram-se seguidamente duas linhas que unem os meios dos lados do rectângulo com os extremos da amostra. Para obter esta representação, começa por se recolher da amostra, informação sobre 5 números, que

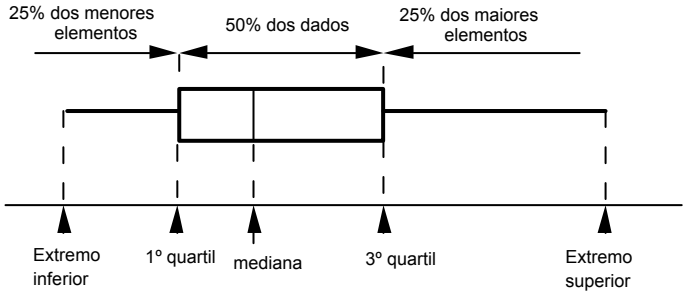
são: os 2 extremos (mínimo e máximo), a mediana e o 1º e 3º quartis. A representação do diagrama de extremos e quartis tem o seguinte aspecto:



O extremo inferior é o mínimo da amostra, enquanto que o extremo superior é o máximo da amostra.

Qual a importância da representação do diagrama de extremos e quartis?

Realça informação importante sobre os dados, nomeadamente sobre o centro da amostra (mediana), variabilidade e simetria. Repare-se que da forma como o diagrama se constrói, se pode retirar imediatamente a seguinte informação:

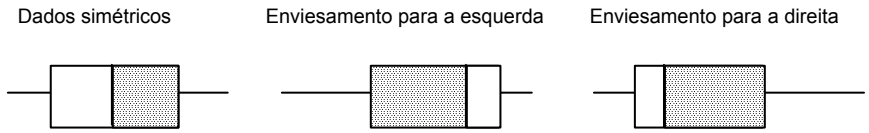


Como é que se pode reconhecer a simetria ou o enviesamento dos dados, a partir desta representação?

Existem fundamentalmente três características da representação extremos e quartis, que nos dão ideia da *simetria ou enviesamento* dos dados e da sua *maior ou menor concentração*:

- distância entre a linha indicadora da mediana e os lados do rectângulo;
- comprimento da caixa;
- comprimento das linhas que saem dos lados dos rectângulos.

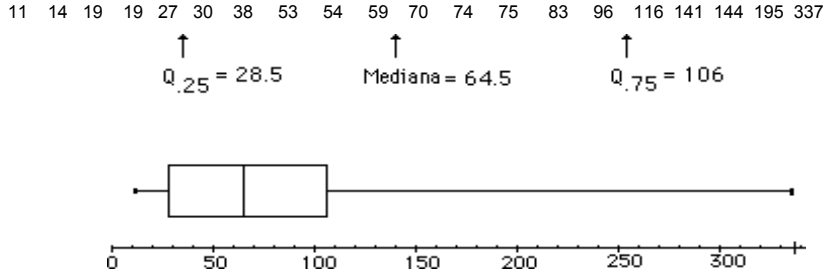
Apresentamos seguidamente 3 exemplos de diagramas de extremos e quartis, correspondentes a tipos diferentes de distribuição dos dados.



Exemplo 9 - Num inquérito à comunidade científica sobre a utilização de meios informáticos, realizado pela Fundação para o Desenvolvimento dos Meios Nacionais de Cálculo Científico, obtiveram-se os seguintes resultados quanto ao tipo de problemas tratados:

Ajustamento de dados	337	Eq. Diferenc. Ordinárias	54
Análise de Fourier	195	Gráfica Computacional	53
Anál. Estatíst. de Dados	144	Integração Numérica	38
Desenv. de Software	116	Inteligência Artificial	30
Diferenças Finitas	96	Interpolação	27
Diferenciação Numérica	83	Método Monte Carlo	19
Elementos de Fronteira	75	Métodos Numéricos	19
Elementos Finitos	74	Simulação	14
Eq. Algébricas Lineares	70	Valores e Vect. Próprios	11
Eq. Algéb. não Lineares	59	Outros	141

Uma representação de extremos e quartis para estes dados, tem o seguinte aspecto[1]:



Da análise da representação anterior, verifica-se que os 50% dos dados centrais são um pouco enviesados para a direita, havendo um grande enviesamento nos 25% dos dados superiores, provocado pelo valor 337.

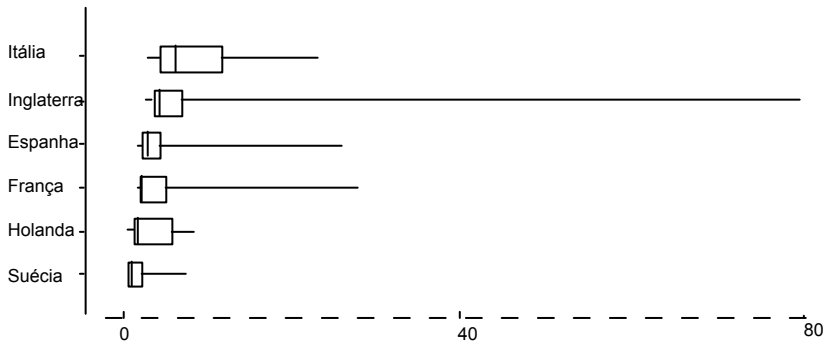
Nota: A representação de extremos e quartis é muito útil para a comparação de vários conjuntos de dados, como se exemplifica a seguir.

[1] Na secção destinada às características amostrais indicaremos a maneira de calcular a mediana e os quartis.

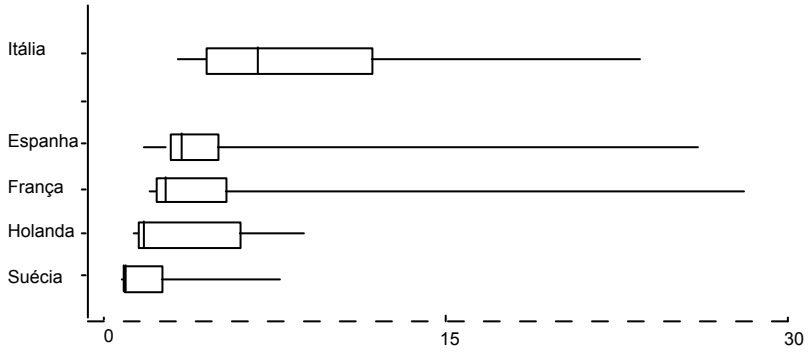
Exemplo 10: As tabelas seguintes referem-se à população (em centenas de milhar) de 10 grandes cidades de 6 países europeus, reportada no World Almanac de 1967 e usando o último censo acessível.

1) Suécia		(2) Holanda		(3) França	
Estocolmo	7.87	Amesterdão	8.68	Paris	28.11
Gotemburgo	4.22	Roterdão	7.31	Marselha	7.83
Malmo	2.49	Haia	6.02	Lyon	5.35
Norrkoping	0.94	Utrecht	2.64	Toulouse	3.30
Vasteras	0.89	Eindhoven	1.75	Nice	2.94
Uppsala	0.87	Haarlem	1.72	Bordéus	2.54
Orebro	0.81	Groningen	1.51	Nantes	2.46
Halsingborg	0.78	Tilburg	1.42	Estrasburgo	2.33
Linkoping	0.71	Enschede	1.31	St. Etienne	2.03
Boras	0.69	Arnhem	1.29	Lille	1.99
(4) Espanha		(5) Inglaterra		(6) Itália	
Madrid	25.99	Londres	79.86	Roma	23.59
Barcelona	16.96	Birmingham	11.02	Milão	15.80
Valencia	5.01	Liverpool	7.22	Nápoles	11.82
Sevilha	4.74	Manchester	6.38	Turim	11.14
Saragoça	3.57	Leeds	5.09	Génova	7.84
Bilbao	3.34	Sheffield	4.88	Palermo	5.90
Málaga	3.12	Bristol	4.30	Florença	4.54
Murcia	2.64	Coventry	3.30	Bolonha	4.44
Córdova	2.14	Nottingham	3.10	Catânia	3.61
Palma	1.69	Kingston	2.99	Veneza	3.36

Para comparar os conjuntos de dados anteriores, utilizamos diagramas de extremos e quartis paralelos



Na representação anterior, as caixas aparecem com um comprimento muito pequeno, devido ao valor exagerado correspondente à cidade de Londres. Quando retiramos a Inglaterra, já se torna mais simples a comparação dos restantes países, sendo de realçar ainda as cidades de Paris, Madrid e Roma substancialmente mais populosas dos que as restantes. De notar também o enviesamento, com cauda mais longa para a direita, apresentado por todos os países:



Nas representações anteriores apresentam-se os diagramas de extremos e quartis dos diferentes conjuntos de dados, por ordem crescente da respectiva mediana. Imediatamente se conclui que existe um enviesamento para a direita, isto é, há menor dispersão no grupo das 50% cidades menos populosas, quando comparadas com as 50% cidades mais populosas. Também se verifica que (de entre os países considerados) a Itália é o país que tem, de um modo geral, as cidades mais populosas .

Sugestões didácticas e comentários

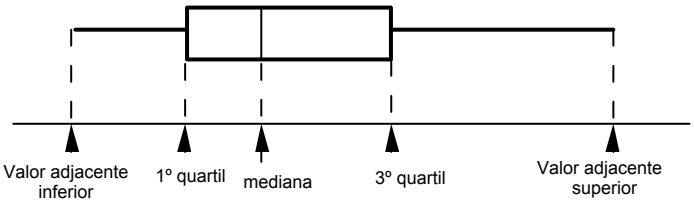
1. Sugerir a um grupo de alunos que investigue quais as 20 serras mais altas de Portugal continental e que façam a respectiva representação num diagrama de extremos e quartis.

2. Pedir a um grupo de alunos que durante 2 semanas tome nota das temperaturas máximas e mínimas registadas diariamente, em várias cidades de Portugal continental, assim como no Funchal e Ponta Delgada. Depois dos dados recolhidos utilizar diagramas de extremos e quartis paralelos para comparar a distribuição das temperaturas máximas das diferentes cidades. Repetir para as temperaturas mínimas. Para cada cidade, comparar a distribuição das temperaturas máximas com a das mínimas.

Nota: Caixa-dos-bigodes (Box-plot) - Uma outra representação análoga à anteriormente considerada, mas um pouco mais elaborada é a caixa dos bigodes, que se apresenta a seguir.

Tal como no diagrama de extremos e quartis o conjunto dos valores da amostra compreendidos entre o 1º e o 3º QUARTIS, é representado por um rectângulo (caixa)

com a MEDIANA indicada por uma barra. Consideram-se seguidamente duas linhas que unem os meios dos lados dos rectângulos com os chamados *valores adjacentes*, que definiremos a seguir.



Define-se valor *adjacente inferior AI*, como sendo o *menor* valor da amostra (eventualmente o mínimo), que é maior ou igual que

$$Q_1 - 1.5 \times (Q_3 - Q_1)$$

Define-se valor *adjacente superior AS*, como sendo o *maior* valor da amostra (eventualmente o máximo), que é menor ou igual que

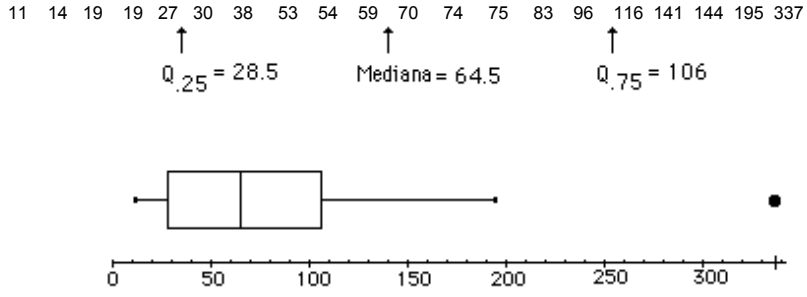
$$Q_3 + 1.5 \times (Q_3 - Q_1)$$

Por vezes surgem na amostra valores, que se distinguem dos restantes por serem muitos grandes ou muito pequenos. A esses valores chamamos *outliers*. Dizemos que um valor é outlier, quando não está compreendido no intervalo [AI, AS]. Os outliers representam-se na caixa-dos-bigodes por uma notação que pode ser um traço, um asterisco ou um ponto.

Tal como a representação extremos e quartis, a caixa-dos-bigodes realça informação importante sobre os dados, nomeadamente sobre o centro da amostra (mediana), variabilidade, simetria, dando-nos ainda informação sobre a existência de outliers (valores que se distinguem dos restantes, dando a ideia de não pertencerem ao mesmo conjunto de dados).

Repare-se que esta representação coincide com o diagrama de extremos e quartis, quando não existem outliers.

Exemplo 9 (cont) - Uma representação caixa-dos-bigodes para estes dados, tem o seguinte aspecto:



Da análise da representação anterior, verifica-se que os dados são um pouco enviesados para a direita e existe um outlier correspondente ao valor 337, que diz respeito à utilização dos meios informáticos para o ajustamento de dados.

Sugestões didácticas e comentários

1 - Considere as seguintes tabelas que apresentam as "Despesas dos agregados familiares por categoria sócio-económica: principais rubricas", relativamente aos anos de 1981 e 1990. (Fonte: A situação social em Portugal, 1960 - 1995, Orga-nização de António Barreto, Instituto de Ciências Sociais, Universidade de Lisboa)

Ano 1981

	Produtores agrícolas	Assalaria-dos agrícolas	Pessoal operário	Empresári-os não agrícolas	Pessoal admi-nistrativo	Quadros técnicos, científicos e de direcção	Profission. liberais	Não activos
Desp. média anual total	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
Aliment.	52.7	51.9	42.9	36.6	34.2	22.6	27.9	47.3
Vestuário	10.4	10.5	10.9	10.4	10.6	9.4	7.9	8.9
Habitação	16.5	17.5	18.0	15.1	18.7	19.0	15.1	18.6
Saúde	2.7	2.6	1.9	2.4	2.5	1.9	1.7	4.2
Transporte	9.2	8.2	12.0	19.0	14.7	22.5	28.2	9.4
Educação e cultura	1.4	1.9	3.3	3.9	4.9	6.9	5.5	2.9
Outros	7.2	7.4	11.0	12.6	14.5	17.7	13.8	8.8

Ano 1990

	Produtores agrícolas	Assalaria-dos agrícolas	Pessoal operário	Empresári-os não agrícolas	Pessoal admi-nistrativo	Quadros técnicos, científicos e de direcção	Profission. liberais	Não activos
Desp. média anual total	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
Aliment.	44.3	44.4	35.7	29.8	26.9	19.9	19.0	40.7
Vestuário	9.9	10.8	9.7	9.9	10.2	9.4	9.7	8.1
Habitação	17.3	17.2	19.0	19.5	18.5	19.7	21.8	20.3
Saúde	2.1	2.1	2.4	2.4	2.5	2.6	2.5	4.8
Transporte	13.1	10.1	14.1	16.2	19.0	22.4	19.5	11.9
Educação e cultura	2.0	3.1	3.6	3.6	4.5	6.8	3.6	2.5
Outros	11.3	12.3	15.5	18.6	18.4	19.3	24.0	11.8

a) Fixando-se num dos anos, considere dois grupos sócio-económicos à sua escolha. Faça representações gráficas adequadas para os dados relativos aos grupos que considerou e compare-os no que diz respeito às despesas nas diferentes rubricas.

b) Considerando o mesmo grupo para os dois anos, estude a evolução das despesas nas diferentes rubricas.

2 - Em 1960 e novamente em 1980 foi feito um inquérito às mulheres americanas sobre o nº de filhos. Os resultados obtidos foram os seguintes (Freedman *et al*., 1991, *Statistics*):

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

Construa uma representação gráfica adequada para os dados anteriores e tire conclusões, no que diz respeito à evolução da natalidade.

3 - 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 de 1880 e o de 1970 (Freedman *et al*., 1991, *Statistics*)

Dígito	*1880*	*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

a) Da consulta da tabela verifica a existência de algumas anomalias?

b) Construa diagramas de barras relativamente aos dois censos.

c) Em 1880 havia uma nítida preferência pelos dígitos 0 e 5. Tem alguma explicação para este facto?

d) Em 1970 essa preferência é quase despercebida. Como explica esse facto?

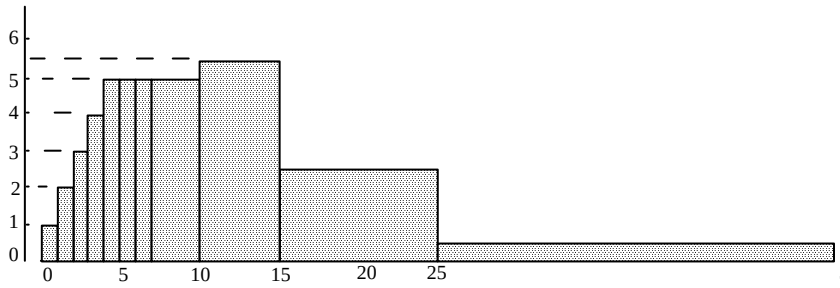
4 - Considere a seguinte tabela de frequências correspondente aos resultados de uma prova específica de Literatura Portuguesa, no ano de 1995.

a) Da consulta da tabela verifica a existência de algumas anomalias?

b) Faça um agrupamento conveniente para os dados, assim como uma represen-tação gráfica.

Nota	Freq.abs	Nota	Freq.abs	Nota	Freq.abs	Nota	Freq.abs	Nota	Freq.abs
0	17	20	115	40	149	60	40	80	11
1	2	21	51	41	82	61	38	81	13
2	12	22	56	42	98	62	30	82	7
3	6	23	73	43	81	63	52	83	7
4	10	24	61	44	64	64	38	84	6
5	22	25	115	45	104	65	34	85	4
6	27	26	64	46	54	66	26	86	8
7	26	27	76	47	69	67	22	87	1
8	42	28	69	48	64	68	37	88	6
9	25	29	59	49	38	69	19	89	1
10	59	30	114	50	186	70	27	90	6
11	25	31	57	51	74	71	19	91	2
12	37	32	83	52	101	72	14	92	2
13	33	33	80	53	61	73	34	93	1
14	50	34	62	54	63	74	15	94	1
15	73	35	118	55	80	75	18	95	4
16	43	36	62	56	52	76	14	96	1
17	62	37	96	57	48	77	1	97	0
18	65	38	94	58	37	78	22	98	1
19	56	39	74	59	39	79	13	99	0

5 - O histograma seguinte representa o rendimento familiar, em milhares de dólares de famílias americanas (Freedman *et al*., 1991, *Statistics*)



Cerca de 1% das famílias têm rendimentos entre 0 e 1000 dólares. Estime a percentagem de famílias com rendimentos:

I) a) Entre 1000 dólares e 2000 dólares b) Entre 2000 dólares e 3000 dólares

c) Entre 3000 dólares e 4000 dólares d) Entre 4000 dólares e 5000 dólares

e) Entre 4000 dólares e 7000 dólares f) Entre 7000 dólares e 10000 dólares

II) a) Haverá mais famílias com rendimentos entre 6000 dólares e 7000 dólares ou entre 7000 dólares e 8000 dólares ? Ou será aproximadamente o mesmo?

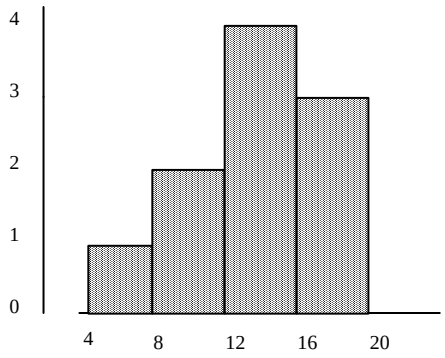
b) Haverá mais famílias com rendimentos entre 10000 dólares e 11000 dólares ou entre 15000 dólares e 16000 dólares ? Ou será aproximadamente o mesmo?

R: I) a) 2%b) 3% c) 4% d) 5% e) 15% f) 15%

II) a) O mesmo b) Mais entre 10000 dólares e 11000 dólares

Comentário: Chama-se a atenção para que neste histograma, a escala do eixo das ordenadas tem unicamente como função permitir o cálculo das áreas dos rectângulos que formam o histograma. Assim, a informação relevante é dada pela percentagem de 1% de famílias com rendimentos entre 0 e 1000 dólares, o que significa que a uma área igual a 1 corresponde uma frequência relativa de 1%. Por exemplo a percentagem de famílias com rendimentos entre 15 e 25 será de 25% (a área correspondente a esta classe é 10x2.5=25).

6 - O histograma seguinte mostra a distribuição das notas finais de Matemática de uma determinada turma:



a) Algum aluno teve nota inferior a 4?

b) 10% dos alunos da turma tiveram nota entre 4 e 8. Qual a % de alunos com nota entre 8 e 12? (Ver comentário do exercício anterior)

c) Qual a percentagem de alunos com nota superior a 12?

R: a) Não b) 20% c) 70%

7 - Seguidamente apresentam-se 4 "manchas" de histogramas, que apresentam os resultados do estudo, numa pequena cidade, das 4 características seguintes (Free-dman *et al*., 1991, *Statistics*):

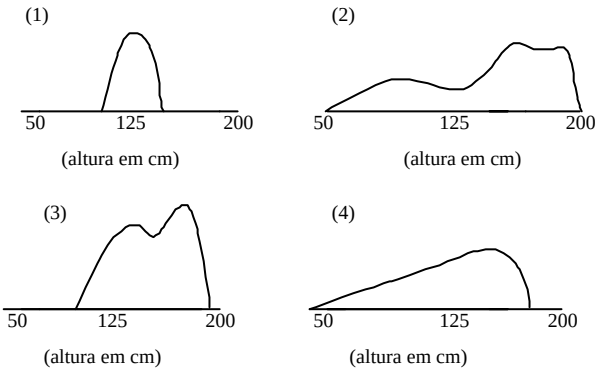
a) Alturas de todos os elementos das famílias em que os pais tinham idade inferior a 24 anos.

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? Explique porquê.

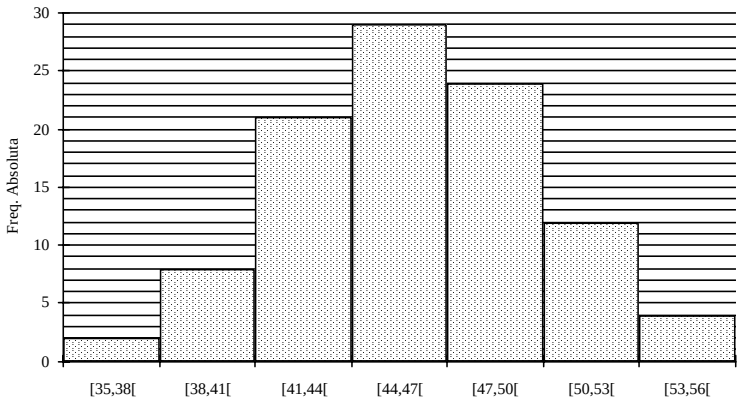


R:a) - (2) b) - (3) c) - (4) d) - (1)

8 - Num viveiro dos Serviços Florestais, está-se a estudar o crescimento, no nosso clima, de um novo tipo de pinheiro (PN). Passados dois meses sobre o lançamento à terra das sementes, mediu-se a altura atingida pelos pinheiros, tendo-se recolhido uma amostra de dimensão 100, a partir da qual se construiu o seguinte histograma (a unidade de medida é o mm):

a) Qual a percentagem de pinheiros com tamanho inferior a 44 mm?

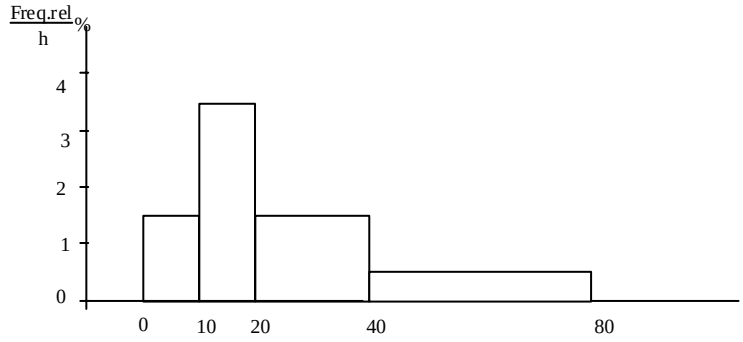
b) Pensa-se que o pinheiro habitual (PH) tem um crescimento muito mais lento que esta nova espécie ensaiada, admitindo-se até que a velocidade do crescimento do PH seja metade da do PN. Por outro lado, pensa-se que se se utilizar um fertilizante adequado, o PN cresce mais 10 mm do que se não se utilizar o fertilizante. Tendo em consideração o histograma apresentado pela amostra de PN, esboce histogra-mas que representem uma amostra de PH e outra amostra de PN com fertilizante. Justifique os esboços apresentados.



Comentário: No histograma anterior utilizaram-se como alturas dos rectângulos que formam o

histograma, as frequências absolutas. Deve-se chamar a atenção para que a área total ocupada pelo histograma é igual a 300.

9 - Um serviço de saúde registou o nº médio de cigarros fumados por dia por cada doente (homem) assistido nesse serviço. Os dados recolhidos permitiram construir o seguinte histograma:



a) A percentagem de fumadores que fuma menos de 10 cigarros por dia é aproximadamente: 1.5%; 15%; 30%; 50%?

b) A percentagem de fumadores que fuma um maço ou mais por dia, mas menos de 2 maços é aproximadamente: 1.5%; 15%; 30%; 50%?

c) A percentagem de fumadores que fuma um maço ou mais por dia, é aproximadamente: 1.5%; 15%; 30%; 50%?

d) A percentagem de fumadores que fuma três maços ou mais por dia, é aproximadamente: 0.25%; 0.5%; 10%?

e) A percentagem de fumadores que fuma 15 cigarros por dia, é aproximadamente:

0.3%; 0.5%; 1.5%; 3.5%; 10%?

R: a) 15% b) 30% c) 50% d) 10% e) 3.5%

10 - A seguinte tabela apresenta os índices gerais de produção industrial, nos dife-rentes países da comunidade e noutros países (Fonte : Anuário Estatístico de Portugal - 1992):

Eur12	**1984**	**Out. países**	**1984**	**Eur12**	**1990**	**Out. países**	**1990**
Alemanha	95.3	Áustria	95.4	Alemanha	117.9	Áustria	121.2
Bélgica	97.6	Canadá	95.0	Bélgica	118.4	Canadá	107.0
Dinamarca	95.9	EUA	98.3	Dinamarca	107.8	EUA	115.7
Espanha	98.0	Finlândia	96.6	Espanha	116.1	Finlândia	114.0
França	99.8	Japão	96.5	França	113.6	Japão	125.4
Grécia	96.7	Noruega	98.0	Grécia	103.3	Noruega	141.1
Holanda	96.1	Suécia	97.3	Holanda	109.1	Suécia	105.2
Irlanda	96.7	Suiça	94.2	Irlanda	143.8	Suiça	118.0
Itália	98.6	Turquia	99.0	Itália	117.8	Turquia	138.8
Luxemb.	93.6	URSS	95.8	Luxemb.	118.0	URSS	x

Portugal	90.2	Portugal	135.2
Reino Uni.	94.8	Reino Uni.	109.3

Obs: Considerou-se como índice 100 o ano de 1985.
x - Informação não disponível

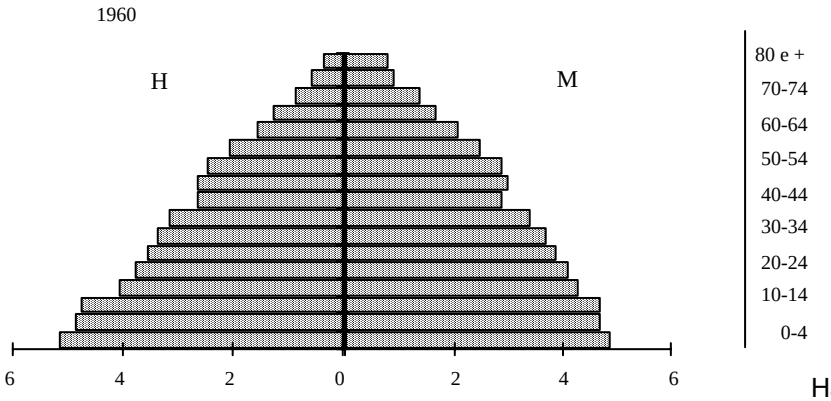
Faça uma representação gráfica adequada para os dados e tire conclusões.

11 - Na tabela seguinte apresenta-se a estrutura etária da população portuguesa em 1960, 1970, 1981 e 1991 (em percentagem) (Fonte: A situação social em Portugal, 1960-1995, Organização de António Barreto, Instituto de Ciências Sociais, Universi-dade de Lisboa):

Construa pirâmides de idade para Portugal em 1960, 1970, 1981 e 1991 e tire conclusões quanto à evolução da população. Será que a população portuguesa está a envelhecer? Discuta algumas implicações sociais.

Grupos etários	1960			1970			1981			1991		
	H	M	HM	H	M	HM	H	M	HM	H	M	HM
0-4	5.2	4.9	10.1	4.7	4.5	9.2	4.1	3.9	8.1	2.8	2.7	5.5
5-9	4.9	4.7	9.6	5.0	4.8	9.9	4.5	4.3	8.8	3.4	3.2	6.5
10-14	4.8	4.7	9.4	4.8	4.7	9.4	4.4	4.3	8.7	4.0	3.9	7.9
15-19	4.1	4.3	8.4	4.1	4.4	8.5	4.4	4.3	8.7	4.3	4.2	8.6
20-24	3.8	4.1	7.9	3.5	3.8	7.3	3.9	3.9	7.8	3.9	3.8	7.8
25-29	3.6	3.9	7.6	2.8	3.2	6.0	3.4	3.5	6.9	3.6	3.7	7.4
30-34	3.4	3.7	7.2	2.9	3.3	6.2	3.1	3.3	6.4	3.5	3.6	7.0
35-39	3.2	3.4	6.7	3.1	3.4	6.5	2.7	3.0	5.8	3.3	3.4	6.7
40-44	2.7	2.9	5.6	3.0	3.4	6.4	2.8	3.1	5.8	3.1	3.3	6.4
44-49	2.7	3.0	5.7	2.8	3.1	6.0	2.8	3.1	6.0	2.8	3.0	5.8
50-54	2.5	2.9	5.4	2.4	2.7	5.2	2.7	3.1	5.8	2.7	3.0	5.7
55-59	2.1	2.5	4.6	2.4	2.7	5.1	2.5	2.9	5.4	2.7	3.0	5.7
60-64	1.6	2.1	3.8	2.1	2.6	4.8	2.0	2.4	4.4	2.5	2.9	5.4
65-69	1.3	1.7	3.0	1.6	2.2	3.8	1.9	2.3	4.2	2.1	2.6	4.8
70-74	0.9	1.4	2.3	1.1	1.6	2.7	1.4	2.0	3.4	1.5	2.0	3.5
75-79	0.6	0.9	1.5	0.6	1.0	1.6	0.8	1.4	2.2	1.1	1.6	2.7
80 e +	0.4	0.8	1.2	0.6	1.0	1.6	0.5	1.2	1.7	0.9	1.7	2.6
Total	47.9	52.1	100.0	47.5	52.5	100.0	48.2	51.8	100.0	48.2	51.8	100.0

Sugestão - para construir uma pirâmide de idades considere um eixo vertical em que marca as classes etárias e construa para um e outro lado desse eixo os histogramas correspondentes aos homens e às mulheres. A título de exemplo considera-se a pirâmide para 1960:



Haverá mais nascimentos do sexo feminino ou masculino?

Será razoável afirmar que existem mais viúvas do que viúvos?

Capítulo 3

CARACTERÍSTICAS AMOSTRAIS
MEDIDAS DE LOCALIZAÇÃO E DISPERSÃO

3.1 - Introdução

Vimos anteriormente alguns processos de resumir a informação contida nos dados, utilizando tabelas e gráficos. Veremos agora um outro processo de resumir essa informação, utilizando determinadas *medidas*, calculadas a partir dos dados, que se chamam *estatísticas*.

Das medidas ou estatísticas que iremos definir para caracterizar os dados, destacam-se as *medidas de localização*, nomeadamente as que localizam o centro da amostra, e as *medidas de dispersão*, que medem a variabilidade dos dados.

Observemos que, ao resumir na forma de alguns números a informação contida nos dados, estamos a proceder a uma redução "drástica" desses dados. Assim, estas medidas devem ser convenientemente escolhidas, de modo a representarem o melhor possível o conjunto de dados que pretendem sumariar. Como veremos, definiremos várias medidas possíveis, mas não poderemos dizer, de uma forma geral, que uma é melhor do que outra, já que a sua utilização depende do contexto e da situação em que necessitam de ser calculadas e de como vão ser utilizadas.

Será mesmo necessário utilizar os dois tipos de medidas, isto é, de localização e de dispersão, para caracterizar um conjunto de dados? O exemplo seguinte procura responder a esta questão.

Exemplo 1 - Dois alunos do 12º ano obtiveram as seguintes notas:

Pedro	14	13	13	13	13	13	14	13	13
João	15	10	8	13	14	13	16	14	16

O Pedro e o João tiveram a mesma média de 13.2, mas o João não teve aproveitamento a todas as disciplinas. Quer dizer que utilizámos uma medida de redução dos dados, a média, que não é suficiente para caracterizar e diferenciar os dois conjuntos de dados.

Efectivamente, se representarmos num diagrama de caule-e-folhas os dois conjuntos, obtemos duas representações com aspecto diferente, já que na segunda representação se verifica uma maior variabilidade, isto é, os dados estão mais dispersos.

1	3 3 3 3 3 3 3
1	4 4

0	8
1	0
1	
1	
1	3 3
1	4 4
1	5
1	6 6

Antes de começar a definir as medidas que vão ser utilizadas para resumir a informação contida nos dados (e lembramos mais uma vez que estamos na fase da análise estatística conhecida por ESTATÍSTICA DESCRITIVA), vamos introduzir uma notação conveniente para representar a amostra. Assim, o conjunto de dados ou observações que constituem a amostra será representado por

$$\mathbf{x_1, x_2, x_3, \ldots, x_n}$$

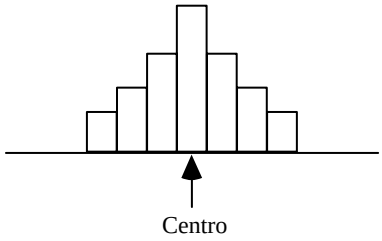
onde x_1, x_2,...., x_n, representam, respectivamente, os resultados da 1ª observação,d a 2ª observação, da n-ésima observação, a serem recolhidas para constituir uma amostra de dimensão n. Esta notação não pressupõe uma ordenação.

3.2 - Medidas de localização

De entre as medidas de localização, merecem destaque especial as que localizam o *centro de uma amostra*.

Vimos anteriomente que uma representação gráfica adequada para um conjunto de dados contínuos era, por exemplo, o histograma. Vimos também que um histograma pode ter vários aspectos, nomeadamente pode apresentar uma forma simétrica ou

enviesada. No caso particular do histograma ser perfeitamente simétrico, não haveria dúvida em dizer qual o centro dessa distribuição:



No entanto, a situação anterior é muito rara, pois devido à aleatoriedade presente nos dados, os histogramas não apresentam aquele aspecto. Por outro lado, quando o histograma é enviesado, a situação ainda se torna mais complicada, pois é difícil de dizer o que é o centro. Existem então vários processos para definir o centro, cujas medidas não dão normalmente o mesmo resultado. Destas medidas destacamos a média e a mediana, a definir seguidamente.

3.2.1 - Média

A média amostral ou simplesmente média, é a medida de localização do centro da amostra, mais vulgarmente utilizada. Representa-se por $\bar{x}$ e calcula-se utilizando o seguinte processo:

- Somam-se todos os elementos da amostra
- Divide-se o resultado da soma pelo número de elementos da amostra

Utilizando a notação introduzida anteriormente para representar a amostra, de dimensão n, a média obtém-se a partir da expressão:

$$\bar{x} = \textbf{Erro!}$$

E se os dados se encontram agrupados?

Neste caso podem-se verificar duas situações:

- Os dados são discretos e as diferentes classes são os diferentes valores que surgem na amostra. Então ainda se pode calcular a média a partir da seguinte expressão

$$\bar{x} = \textbf{Erro!}$$

onde: k é o número de classes do agrupamento

n_i é a frequência absoluta da classe i, n = **Erro!**

y_i é o valor correspondente à classe i

- Os dados são discretos ou contínuos e as classes são intervalos. Então já não temos um valor exacto para a média, mas sim um valor aproximado, o qual é dado pela expressão

$$\bar{x} \approx \textbf{Erro!}$$

onde: k é o número de classes do agrupamento

n_i é a frequência absoluta da classe i

y_i é o ponto médio da classe i, o qual é considerado como elemento representativo da classe.

Observação importante: Ao calcular a média a partir de dados agrupados, em que as classes são intervalos, não se obtém o verdadeiro valor da média, mas sim um valor aproximado. Para se obter o valor exacto da média terá de se considerar os dados originais, caso estejam disponíveis.

Ao contrário do que o novo programa de Matemática (Matemática - Programas, 10º, 11º e 12º anos - Ministério da Educação, Departamento do Ensino Secundário, Janeiro 1997) faz crer, para calcular a média de dados contínuos, não tem que se proceder a qualquer agrupamento. Pode acontecer que os dados nos sejam fornecidos já agrupados e nesse caso não temos outra alternativa senão calcular um valor aproximado para a média.

A média será sempre uma medida representativa dos dados?

Ao determinar a média dos seguintes dados

12.4 13.5 13.6 11.2 15.1 10.6 12.4 14.3 113.5

obteve-se o valor $\bar{x} = 24.1$.

Embora todos os dados, menos um, estejam no intervalo [10.6, 15.1], o valor obtido para a média está "bem afastado" daquele intervalo! Uma medida que se pretendia representativa dos dados, não está a conseguir esses objectivos, pois se nos disserem

que um conjunto de dados tem média 24.1, imediatamente pensamos em valores que não se afastem muito daquele valor.

O que acontece é que *a média é muito sensível a valores muito grandes ou muito pequenos.*

No caso do exemplo foi o valor 113.5 que inflacionou a média. Além disso temos alguma razão para pensar que pode ter havido um erro ao digitar o valor 113.5, digitando um 1 a mais!

E se em vez de 113.5 o valor correcto fosse 13.5, qual o valor da média? Neste caso para a média dos seguintes dados

12.4 13.5 13.6 11.2 15.1 10.6 12.4 14.3 13.5

obteve-se o valor $\bar{x}$ = 13.0, significativamente diferente do obtido no caso anterior!

Sendo a média uma medida tão sensível aos dados, é preciso ter cuidado com a sua utilização, pois pode dar uma imagem distorcida dos dados que pretende representar!

Para além do facto de ser uma medida muito simples de calcular, existirá alguma outra razão que a torne uma medida tão "popular"?

Pode-se mostrar (e essa demonstração faz parte da Inferência Estatística) que quando a distribuição dos dados é "normal" (o histograma correspondente tem a forma aproximada de um sino), então a melhor medida de localização do centro é a média. Ora sendo a Distribuição Normal uma das distribuições mais importantes e que surge com mais frequência nas aplicações, esse facto justifica a grande utilização da média.

A média tem uma outra característica, que torna a sua utilização vantajosa em certas aplicações:

Quando o que se pretende representar é a *quantidade total expressa pelos dados*, utiliza-se a média. Na realidade, ao multiplicar a média pelo nº total de elementos, obtemos a quantidade pretendida.

Obervação: Chama-se a atenção para que só tem sentido calcular a média para dados de tipo quantitativo.

Sugestões didácticas e comentários

Actividade 1 - Média (Statistical Tools and Statistical Literacy: the Case of the Average - Teaching Statistics, vol 17, n. 3, 1995)

Pretende-se que os Professores insistam não só no conhecimento do conceito de média, mas também na sua interpretação. O estudo da média deve envolver muito mais do que aprender as propriedades matemáticas da média (por exemplo a soma dos desvios é igual a zero), ou fazer com que os estudantes calculem a média de qualquer conjunto de dados que lhes apareça, independentemente se isso tem ou não sentido.

Como futuros consumidores da informação estatística, os estudantes devem ter bem presente as várias interpretações da palavra "média". Uma actividade interessante pode ser a de pedir aos estudantes que procurem nos dicionários o significado para esta palavra. Uma quantidade de interpretações legítimas, a acrescentar à " aquilo que se obtém somando os dados todos e dividindo pelo número deles", surpreender-nos-á!

Seguidamente pedir-se-á aos estudantes que interpretem e comentem algumas frases onde entra a palavra média, como por exemplo:

1 - Um adulto médio come 5 kg de gelado por ano.
2 - Em média, os adultos comem 5 kg de gelado por ano.
3 - Um adulto come uma média de 5 kg de gelado por ano.

Comentário: Em 1, a palavra médio não está empregue com o significado estatístico de média como característica amostral. Pretende significar um adulto "normal", que se utiliza como referência. Em 2, pretende-se dizer que ao recolher a informação sobre a quantidade de gelado comida por ano, por uma amostra de vários adultos, se concluiu que a média dos valores observados é de 5 kg. Obteve-se a média dividindo a soma das quantidades observadas pelo número de adultos inquiridos. Finalmente em 3, o que se observou foi a quantidade de gelado comida, por ano, por um adulto, escolhido ao acaso, e durante vários anos. Obteve-se a média dividindo a soma das quantidades obtidas pelo número de anos observados.

Os estudantes ao elaborarem um inquérito podem discutir como é que os inquiridos interpretarão frases alternativas para uma questão, tais como:

4 - Em média, quanto gelado come por semana?
5 - Qual a quantidade média de gelado que come por semana?
6 - Numa semana média, quanto gelado come?

Esta actividade pode também ajudar os estudantes a aperceberem-se que, para compreender o significado do termo "média" quando usado num sentido estatístico, é necessário saber muito mais do que somar e dividir! É necessário obter informação, por exemplo, acerca do contexto e objectivo do estudo.

Actividade 2 - Um pai tinha 5 depósitos a prazo (de montantes diferentes) que pensou sortear pelos seus 5 filhos. Depois pensou melhor e decidiu que eles tinham que receber todos a mesma quantia. Então como é que ele deve proceder? Se ele tivesse começado por utilizar o primeiro processo, alguns dos irmãos teriam de devolver dinheiro, enquanto que os outros teriam de receber mais. Será que as quantias devolvidas chegam para pagar aos que ainda têm de receber?

Concretize a situação anterior admitindo que as quantias (em milhares de contos) em jogo eram 10 , 11, 14, 15 e 16.

1º caso - O pai dá uma das quantias a cada filho, tendo o resultado do sorteio sido o seguinte:

José - 10 mil contos ; Joana - 11 mil contos; Maria - 14 mil contos;
João - 15 mil contos; Luís - 16 mil contos

2º caso - O pai dá uma quantia igual a cada um dos filhos

Então terá que dar a cada um a média das quantias, pelo que cada filho recebe

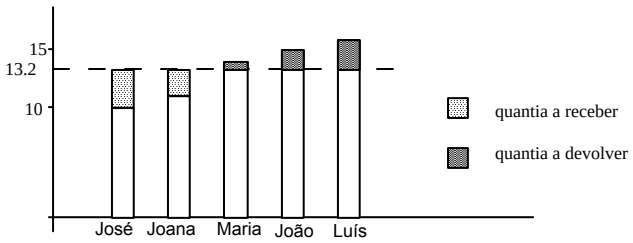
Erro!= 13.2.

Assim

José	-	tem a receber 3 mil e 200 contos
Joana	-	tem a receber 2 mil e 200 contos
Maria	-	tem a devolver 800 contos
João	-	tem a devolver mil e 800 contos
Luís	-	tem a devolver 2 mil e 800 contos

A soma das quantias a receber é (3.2 + 2.2) = 5.4, enquanto que a soma das quantias a devolver é (0.8 + 1.8 + 2.8) = 5.4, pelo que efectivamente as quantias devolvidas chegam para pagar as quantias a receber.

Graficamente temos



Propriedade: Dado um conjunto de dados a soma dos desvios de cada um, relativamente à média, é igual a zero.

Problema: "Todos os jovens levaram bolos para uma festa. Durante a festa todos os jovens comeram a mesma quantidade de bolos, por isso houve alguns que tiveram de dar bolos e outros que receberam bolos. O número total de bolos dados foi igual ao número total de bolos recebidos" Isto é verdade? SIM ou NÃO?" (Leon e Zawojewsk, *in* Hawkins, 1992)

Actividade 3 - Sendo uma medida importante, a média muitas vezes permite que se façam afirmações menos correctas. Comente com os alunos casos onde a média é "mal" utilizada, como por exemplo:

Um jornalista publicou no seu jornal a seguinte notícia relativamente aos atrasos das camionetas que partiam de Sintra para Lisboa: "As camionetas da empresa VIAJANTE com destino a Lisboa e partindo de Sintra, têm em média meia hora de atraso". O

jornalista baseou-se na seguinte informação: As camionetas com partida às 10h30m verificaram os seguintes atrasos (em minutos), durante a semana de 24 a 30 de Março:

2ª feira	3ª feira	4ª feira	5ª feira	6ª feira	Sábado	Domingo
5	11	170	8	sem atraso	6	10

Nota: Na 4ª feira houve uma ameaça de bomba no terminal de camionagem.

Vamos ver de seguida uma outra medida de localização do centro da amostra, alternativa à média e que é a mediana.

3.2.2 - Mediana

A mediana é uma medida de localização do centro da distribuição dos dados, definida do seguinte modo: ordenados os elementos da amostra, a mediana é o valor (pertencente ou não à amostra) que a divide 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.

Para a determinação da mediana, utiliza-se a seguinte regra, depois de *ordenada* a amostra de *n* elementos:

- Se **n** é **ímpar**, a mediana é o elemento médio.
- Se **n** é **par**, a mediana é a semi-soma dos dois elementos médios.

Uma forma simples de aplicar a regra anterior é considerar o quociente **Erro!**:

- Se este quociente for um nº inteiro, considera-se para mediana o elemento nessa posição;
- Se este quociente terminar em 0.5, considera-se a sua parte inteira e faz-se a semi-soma do elemento a que corresponde essa ordem, com o seguinte.

Exemplo 2: Considere o seguinte conjunto de notas de um aluno do 10º ano

10 10 10 11 11 11 11 12

A média e a mediana deste conjunto de dados são, respectivamente,

$\bar{x} = 10.75$ e $m = 11$

Admitamos que uma das notas de 10 foi substituída por uma de 18. Então neste caso a mediana continuaria a ser 11, enquanto que a média subiria para 11.75!

Como medida de localização, a mediana é mais resistente do que a média, pois não é tão sensível aos dados!

Então qual destas medidas é preferível? Média ou mediana?

- Quando a distribuição é simétrica, a média e a mediana coincidem.
- A mediana não é tão sensível, como a média, às observações que são muito maiores ou muito menores do que as restantes (*outliers*). Por outro lado, a média reflecte o valor de todas as observações.

Assim, não se pode dizer em termos absolutos, qual destas medidas é preferível, dependendo do contexto em que estão a ser utilizadas.

Quando a distribuição dos dados é simétrica ou aproximadamente simétrica, as medidas de localização do centro da amostra, média e mediana, coincidem ou são muito semelhantes. O mesmo não se passa quando a distribuição dos dados é assimétrica, facto que se prende com a pouca resistência da média, como já se referiu anteriormente. 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, chamados *outliers*, são os responsáveis pela má utilização da média em muitas situações em que teria mais significado utilizar a mediana.

Exemplo 3: Os salários dos 160 empregados de uma determinada empresa, distribuem-se de acordo com a seguinte tabela de frequências:

Salário (milhares escudos)	45	60	70	80	120	380
Nº empregados	23	58	50	20	7	2

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

Cálculo da média:

$\bar{x} = (23 \times 45 + 58 \times 60 + \ldots + 7 \times 120 + 2 \times 380)/ 160$

$= 71.5$

Cálculo da mediana:

Como n é par, a mediana é a semi-soma dos elementos médios

m = semi-soma dos elementos de ordem 80 e 81

= 60

A média é muito superior à mediana, pois 2 dos valores do conjunto de dados são muito grandes, quando comparados com os restantes, tendo assim inflacionado a média. Efectivamente, dos 160 empregados, só 29 é que têm salário superior à média.

A mediana dá-nos uma ideia mais correcta do nível dos salários, que são de um modo geral muito baixos. Assim, dá-nos a indicação de que 50% dos salários são menores ou iguais a 20 mil escudos, enquanto que os restantes são maiores ou iguais àquele valor.

Sugestões didácticas e comentários

Suponhamos que ao pretender digitar, num computador, o seguinte conjunto de dados

5, 2, 10, 6, 9

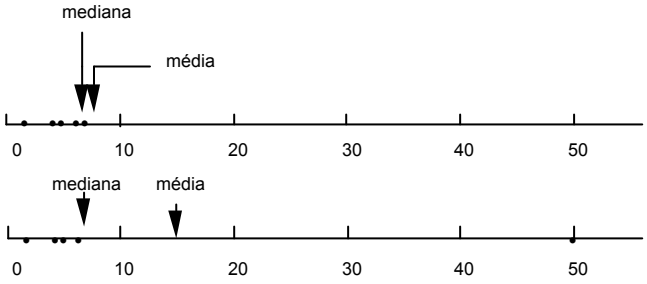
se digitou

5, 2, 50, 6, 9

Estude o comportamento das duas medidas de localização do centro da amostra, relativamente ao outlier introduzido (o valor 50):

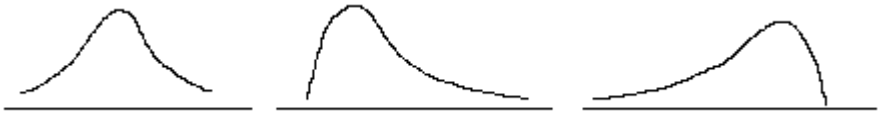
Resolução:

	Dados originais (5,2,10,6,9)	Dados copiados (5,2,50,6,9)
Média	6.4	14.4
Mediana	6	6



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:



média ≈ mediana média > mediana média < mediana

Deve ser então chamada a atenção que o simples cálculo da média e da mediana nos pode dar informação sobre a forma da distribuição dos dados.

Como calcular a mediana a partir de dados agrupados?

Por vezes os dados apresentam-se agrupados, sendo necessário calcular a mediana a partir das tabelas ou das representações gráficas correspondentes.

Consideremos de novo o exemplo 5 do capítulo 2.

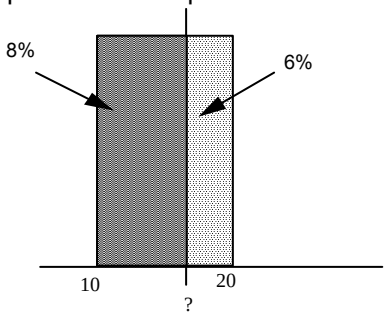
Exemplo 5 (cont) - A partir da tabela de frequências pretende-se calcular a mediana:

Tabela de frequências

Classes	Rep. classe	Freq. abs.	Freq. rel.	Freq.rel.acum
[0, 10[	5	21	0.42	0.42
[10, 20[	15	7	0.14	**0.56**
[20, 30[	25	9	0.18	0.74
[30, 40[	35	7	0.14	0.88
[40, 50[	45	3	0.06	0.94
[50, 60[	55	0	0.00	0.94
[60, 70[	65	3	0.06	1.00
Total		50	1.00	-

Considerando a coluna correspondente às frequências relativas acumuladas, verificamos que a frequência de 50% corresponde à classe [10, 20[, sendo então esta a classe que contém a mediana: *classe mediana*. Para obter um valor aproximado para a mediana, partimos do princípio que a frequência de 14% correspondente a esta classe se distribui uniformemente sobre o intervalo de amplitude 10. Assim, fazendo uma regra

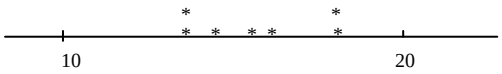
de três simples, como já exemplificámos com a função cumulativa, vamos a esta classe procurar o valor a que corresponda uma frequência de 8%:



O valor aproximado para a mediana, obtido por este processo, é 15.71.

A partir dos dados originais, o valor obtido para a mediana é a semi-soma entre os elementos das posições 25ª e 26ª, ou seja, **Erro!**= 15.85

Observação: Como se verifica, existe uma diferença entre o valor aproximado da mediana, obtido a partir dos dados agrupados e o valor exacto da mediana obtido a partir dos dados originais. Efectivamente, a hipótese de que, dentro de cada classe, os dados se distribuem uniformemente é, muitas vezes, pouco realista. Repare-se como se distribuem os 7 elementos da classe [10, 20[:



Ainda para este exemplo, vamos calcular o valor aproximado para a média a partir dos dados agrupados. Substituimos os elementos de cada classe pelo ponto médio da classe, que elegemos como ponto representativo :

$$\bar{x} \approx 5 \times 0.42 + 15 \times 0.14 + 25 \times 0.18 + 35 \times 0.14 + 45 \times 0.06 + 65 \times 0.06$$

$$\bar{x} \approx 20.02$$

Por outro lado o valor exacto para a média será:

$$\bar{x} = \textbf{Erro!}$$

$$= 19.46$$

Comparando os valores da mediana e da média, verifica-se que a média é superior à mediana. Isto é sintoma de que os dados não se distribuem de forma simétrica, mas sim de forma enviesada para a direita, havendo alguns valores grandes que estão a

inflacionar a média. Efectivamente esta característica já havia sido realçada pela forma do histograma.

E se a tabela de frequências tivesse o seguinte aspecto

Tabela de frequências

Classes	Freq. abs.	Freq. rel.	Freq.rel.acum
[0, 10[	21	0.42	0.42
[10, 20[	4	0.08	**0.50**
[20, 30[	12	0.24	0.74
[30, 40[	7	0.14	0.88
[40, 50[	3	0.06	0.94
[50, 60[	0	0.00	0.94
[60, 70[	3	0.06	1.00
Total	50	1.00	-

como calcular um valor aproximado para a mediana? Neste caso considerávamos o valor 20, pois é o menor valor a que corresponde uma frequência acumulada de 50%.

Ainda para exemplificar o cálculo da mediana para dados agrupados vejamos o seguinte exemplo correspondente a dados discretos (as classes são pontos):

Tabela de frequências

Classes	Freq. abs.	Freq. rel.	Freq.rel.acum
0	4	0.20	0.20
1	6	0.30	**0.50**
2	5	0.25	0.75
3	3	0.15	0.90
4	2	0.10	1.00
Total	20	1	-

O valor 1 satisfaz a condição para ser mediana, mas qualquer valor entre 1 e 2 também satisfaz essas condições! É ou não verdade que se escolhessemos para mediana 1.2, 50% dos elementos da amostra são menores ou iguais a 1.2 e os restantes são maiores ou iguais a 1.2? No entanto, para fixar ideias costuma-se escolher para mediana o ponto médio entre 1 e 2, de forma que a mediana seria neste caso 1.5, o que está de acordo com a metodologia indicada para o cálculo da mediana a partir dos dados antes de agrupados.

Nota: Deve-se chamar a atenção para que, com dados de tipo qualitativo, as únicas características amostrais que se podem calcular são a *moda*, categoria com maior

frequência, e por vezes a *mediana*, quando for possível estabelecer uma hierarquia entre as diferentes categorias ou modalidades que a variável em estudo possa assumir. Por exemplo, numa grande empresa em que os trabalhadores podem assumir um de 5 postos possíveis, representados pelas letras A, B, C, D e E, em que o posto A é o mais importante e E o menos importante, recolheu-se uma amostra de 15 empregados, registando-se as respectivas categorias:

A, E, E, E, B, C, E, D, D, E, B, D, D, E, E

Ordenando a amostra anterior, por ordem crescente de importância do posto de trabalho, obtém-se:

E, E, E, E, E, E, E, D, D, D, D, C, B, B, A

mediana

Se a amostra anterior não tivesse o elemento A, então a mediana seria o posto de trabalho E, pois 50% dos elementos da amostra têm categoria inferior ou igual a E.

3. 2.3 - Quartis

A noção de quartil já foi abordada, quando falamos no diagrama de extremos e quartis. Assim o quartil de ordem 1 ou 1º quartil (respectivamente ordem 3 ou 3º quartil), Q_1 (Q_3), será o valor tal que 25% (75%) dos elementos da amostra são menores ou iguais a ele e os restantes são maiores ou iguais.

Há vários processos para a determinação dos quartis, que nem sempre conduzem aos mesmos resultados. Um dos processos pode ser o de utilizar a mesma metodologia aplicada para a obtenção da mediana, isto é, consideram-se os *quartis* como as medianas das duas partes em que ficou dividida a amostra inicial pela mediana. A parte inferior é dividida pelo 1º quartil, enquanto que a parte superior é dividida pelo 3º quartil.

Exemplo 4: Dada a seguinte amostra

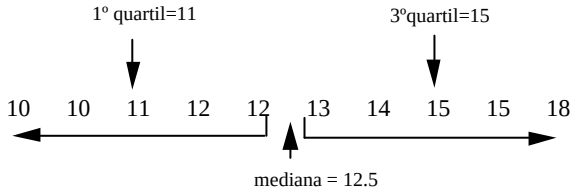
12 10 11 17 18 14 13 10 15 12

pretende-se calcular o 1º quartil e o 3º quartil.

1º - A primeira operação consiste em ordenar a amostra:

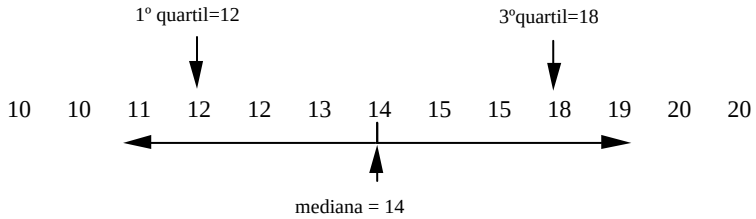
10 10 11 12 12 13 14 15 17 18

2º - Depois, uma vez que o número de elementos é 10 (**par**), a mediana será a semi-soma dos elementos de ordem 5 e 6:



3º - Finalmente o 1º quartil (3º quartil) será a mediana da parte inferior (parte superior) em que ficou dividida a amostra pela mediana.

Suponhamos que a amostra tinha mais 3 elementos (nº **ímpar** de elementos):



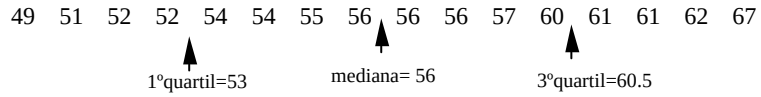
Comentário: Mesmo na utilização deste processo podem-se levantar algumas dúvidas, quando o número de elementos da amostra é ímpar. Efectivamente pode-se optar por considerar o elemento da amostra, seleccionado para mediana, como não pertencente a nenhuma das partes, ao contrário do que foi feito no exemplo, em que a mediana conta para as duas partes.

Exemplo 5: Tendo-se decidido registar os pesos dos alunos de uma determinada turma de Matemática do 12º ano, obtiveram-se os seguintes valores (em kg):

52 56 62 54 52 51 60 61 56 55 56 54 57 67 61 49

Um aluno com o peso de 62kg, pode ser considerado "normal" , isto é nem demasiado magro, nem demasiado gordo?

Ordenando a amostra anterior, cuja dimensão é 16, temos



Um aluno com o peso de 62 Kg é um bocado "forte", pois só 25% dos alunos é que têm um peso superior ou igual a 60.5 Kg.

3.2.4 -Moda

Para um conjunto de dados, define-se **moda** como sendo o valor que surge com mais frequência, se os dados são discretos, ou o intervalo de classe com maior frequência, se os dados são contínuos e estão agrupados.

Esta medida merece referência por ser especialmente útil para reduzir a informação de conjuntos de dados qualitativos, portanto apresentados sob a forma de nomes ou categorias, para os quais não se pode calcular a média e por vezes nem a mediana (se não forem susceptíveis de ordenação).

Sugestões didácticas e comentários

a) Com as sugestões apresentadas nos pontos 1 a 5, pretende-se obter uma maior familiaridade com a noção dos quartis.

1. Considere os dados do exemplo 4. Determine os quartis, utilizando tabelas de frequência em que as classes são os diferentes valores que surgem na amostra. Verifique que os resultados obtidos são idênticos aos obtidos no exemplo, tanto para a amostra de dimensão 10 como 13.

2. Considere uma amostra de dados discretos de dimensão 15. Verifique que a determinação dos quartis, pelos dois processos sugeridos (dados originais e dados agrupados), não conduz aos mesmos resultados. Como curiosidade, adianta-se que os dois processos só não conduzem aos mesmos resultados quando a dimensão da amostra é um múltiplo de 4 menos 1.

Comentário: Se os dados forem contínuos, ou no caso de serem discretos o agrupamento em classes foi feito utilizando intervalos, não se espera que os dois processos conduzam aos mesmos resultados.

3. Pode-se referir aos alunos que os quartis, assim como outras medidas deste género a que chamamos percentis (os quartis são os percentis 25 e 75) são largamente utilizadas pelos pediatras. Quando uma mãe leva o bebé ao pediatra, ele pesa e mede a criança. Depois pergunta à mãe quantos meses tem o filho, consulta umas tabelas e diz em que percentil é que o filho está, relativamente ao peso e à altura, tecendo alguns comentários sobre a condição física da criança. Assim, por exemplo, se o peso estiver no percentil 60, significa que o bebé está muito "bonzinho"! Se estiver perto do percentil 75, combina com a mãe uma dieta adequada, pois o bebé está a ficar um pouco gordo!

4. Falar nas tabelas de pesos, que sobretudo as raparigas gostam de consultar para saber se estão "na linha"!

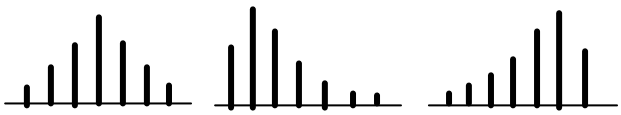
5. Falar no processo utilizado para a definição da nota mínima do exame nacional, para os alunos candidatos à Universidade. No ano de 96/97, pela 1ª vez funcionaram os exames nacionais como provas específicas. Como nota mínima, exigiu-se para cada prova a nota correspondente ao percentil 25. Isto significava que os 75% melhores alunos dessa prova se poderiam candidatar.

b) **1** - Considere os seguintes conjuntos de números:

1 2 3 4 5 2 3 4 5 6 3 5 7 9 11

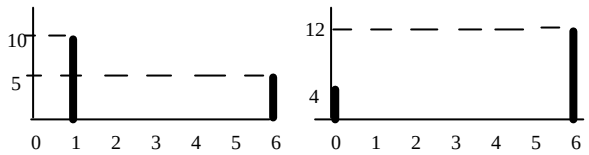
Para cada um destes conjuntos calcule a média. Identifique qual a relação existente entre os conjuntos e diga como poderia obter a média do último conjunto, a partir da média dos dois primeiros conjuntos.

2 - Considere os seguintes diagramas de barras:



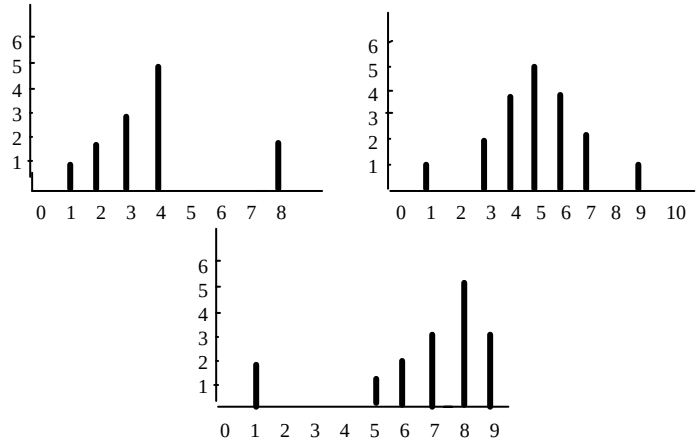
Para cada um deles assinale a posição aproximada da média.

3 - Faça o mesmo que no exercício anterior para os seguintes diagramas de barras:



Suponha que as barras representam miúdos, em que as frequências absolutas são os respectivos pesos, e o eixo horizontal a tábua de um balancé. O que representa o ponto onde marcou a média, relativamente ao balancé, se este estiver em equilíbrio?

4 - Considere os seguintes diagramas de barras. Relativamente a cada uma das representações:



a) Diga quais os dados observados e a frequência com que foram observados.

b) Assinale a posição da média e da mediana. O que conclui?

5 - Numa sala de aulas de 21 alunos, 20 desses alunos têm em média a altura de 145 cm. a) Se o outro aluno, que no dia em que se fez as medições das alturas tinha faltado, tiver de altura 150, qual é a altura média da turma? **b)** Qual deve ser a altura do outro aluno, que no dia em que se fez as medições das alturas tinha faltado, para que a altura média da turma aumente de 1 cm?

6 - Num ponto de Matemática com 5 questões, cada uma valendo 4 valores, verificaram-se os seguintes resultados:

5% dos alunos tiveram	0	40% dos alunos tiveram	12
10% " " "	4	15% " " "	16
25% " " "	8	5% " " "	20

a) Se o teste foi realizado por 10 alunos, qual a pontuação média obtida?

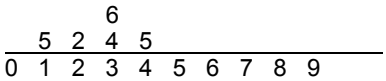
b) Se o teste foi realizado por 20 alunos, qual a pontuação média obtida?

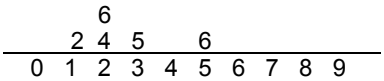
c) Será que pode calcular a média sem saber o número de alunos? Deduza uma expressão para o cálculo da média, quando os dados estão agrupados em classes e para cada classe é dada a respectiva frequência relativa.

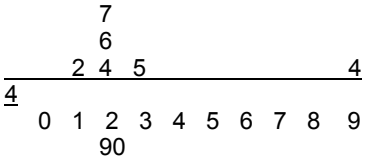
d) Qual o valor da mediana?

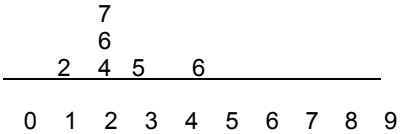
7 -Considere os seguintes diagramas caule-e-folhas:

7 7









Para cada um dos conjuntos de números representados anteriormente, calcule a média e a mediana.

Notas: 1) Nas representações anteriores desenharam-se os traços que separam os caules das folhas horizontalmente, o que torna a representação em caule-e-folha semelhante ao histograma. 2) Na última representação de caule-e-folha, utilizou-se uma notação diferente da habitual, pois um dos valores do correspondente conjunto de dados é muito maior do que os outros, optando-se por interromper o traço que separa os caules das folhas.

8 - Pretende-se iniciar uma nova cultura numa certa região agrícola. Sendo a pluviosidade um dos factores determinantes, recorreu-se aos valores da precipitação diária nos últimos 3 anos e elaborou-se a seguinte tabela:

Pluv. (mm)	Nº dias
[0,5[	105
[5,10[	148
[10,15[	220
[15,20[	193
[20,25[	184
[25,30[	123
[30,35[	95
[35,40	27

Suponha que só se deve introduzir a cultura no caso de, em pelo menos 50% dos dias a pluviosidade ultrapassar os 18 mm. Será ou não razoável, cultivar nesta região o produto em causa?

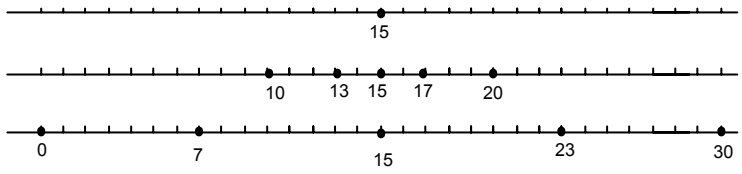
3.3 - Medidas de dispersão

Um aspecto importante no estudo descritivo de um conjunto de dados é o da determinação da variabilidade ou dispersão desses dados relativamente à medida de localização do centro da amostra. Efectivamente as medidas de localização que estudamos não são suficientes para caracterizar completamente um conjunto de dados.

Considerem-se 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

Embora tenham a mesma média e mediana, têm um aspecto bem diferente no que diz respeito à variabilidade.



Como a medida de localização mais utilizada é a média, será relativamente a ela que se define a principal medida de dispersão - o desvio padrão, apresentado a seguir. Começamos, no entanto, por definir variância, que serve de base à definição de desvio padrão.

3.3.1 - Variância

Define-se a variância, e representa-se por $\mathbf{s^2}$, como sendo a medida que se obtém somando os quadrados dos desvios das observações, relativamente à média, e dividindo pelo número de observações:

$$s^2 = \textbf{Erro!}$$

Estamos a utilizar a notação já introduzida anteriormente, para representarmos a amostra.

Quais as razões que nos levam a considerar aquela definição para a variância?

- Se afinal pretendemos medir a dispersão relativamente à média, porque é que não somamos simplesmente os desvios, em vez de os quadrar?
O que acontece é que a soma dos desvios é igual a zero, como já vimos no estudo da média

$$(x_1 - x,^{-}) + (x_2 - x,^{-}) + (x_3 - x,^{-}) + ... + (x_n - x,^{-}) = 0$$

Poderíamos ter utilizado módulos, para evitar que a soma dos desvios positivos cancelasse com a dos desvios negativos, mas pode-se mostrar que, sob o ponto de vista estatístico, é preferível trabalhar com os quadrados do que com os módulos!

Nota: Por vezes utiliza-se uma outra fórmula, muito semelhante à anterior, mas em que a soma dos quadrados dos desvios aparece a dividir por (n-1):

$$s^{*2} = \textbf{Erro!}$$

Na realidade, só aparentemente é que temos n desvios independentes, isto é, se calcular (n-1) desvios, o restante fica automaticamente calculado, uma vez que a sua soma é igual a zero! Costuma-se referir este facto, dizendo que se perdeu um grau de liberdade. Esta definição, embora preferível por razões que se prendem com a Inferência Estatística, é contudo menos intuitiva, e não é objectivo desta análise proceder a qualquer tipo de Inferência Estatística. Assim, a opção entre as duas expressões pode ser deixada ao critério do Professor, que poderá por exemplo escolher a que for utilizada

no manual indicado para os alunos. Não poderá é deixar de referir a existência das duas expressões, tanto mais que elas coexistem na máquina de calcular. Também referirá que a diferença entre as duas expressões é muito pequena, sobretudo se a dimensão da amostra for suficientemente grande.

Uma vez que a variância envolve a soma de quadrados, a unidade em que se exprime não é a mesma que a dos dados. Por exemplo, ao recolhermos informação sobre a característica altura, em cm, a variância virá em cm^2, que é uma medida de área, portanto dificilmente interpretável como medida de variabiliadde. Assim, para obter uma medida da variabilidade ou dispersão com as mesmas unidades que os dados, e portanto de mais fácil interpretação, tomamos a raiz quadrada da variância e obtemos o desvio padrão.

3.3.2 - Desvio padrão

Pelas razões apontadas anteriormente, a medida de dispersão que se costuma utilizar é o desvio padrão, que se representa por **s** e é a raiz quadrada da variância:

$$s = \textbf{Erro!}$$

ou

s* = **Erro!**

O desvio padrão é uma medida que *só pode assumir valores não negativos* e quanto *maior for*, *maior será a dispersão* dos dados.

Relativamente aos três conjuntos de dados apresentados no início do estudo das medidas de dispersão, verificamos que:

- O conjunto 1 apresenta um desvio padrão igual a zero, como seria de esperar, pois se os valores são todos iguais, a dispersão é nula.
- Os conjuntos 2 e 3 apresentam um desvio padrão **s** igual, respectivamente a 3.4 e 10.8.

Sugestões didácticas e comentários

O desvio padrão (The standard deviation: some drawbacks of an intuitive approach - *Teaching Statistics*, vol 7, n.3, 1985)

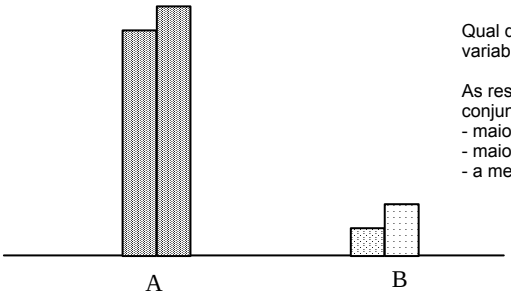
O que mede o desvio padrão? Que tipo de variabilidade?

A variabilidade apresentada por um conjunto de observações pode-se interpretar como:

- uma medida da diferença entre as obervações, umas relativamente às outras;
- uma medida da diferença entre as observações relativamente a uma medida padrão.

A seguinte experiência dá conta de que nem sempre o desvio padrão é entendido pelos alunos como uma medida da variabilidade relativamente à média.

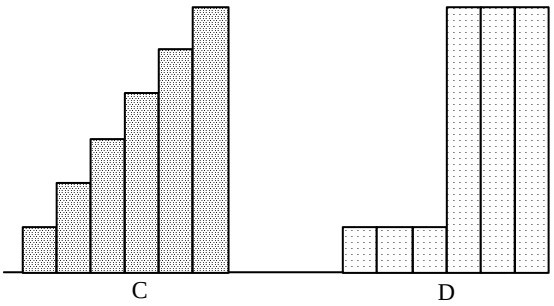
Consideremos dois conjuntos formados cada um por dois blocos: no 1º conjunto os blocos têm altura 45 e 50 cm. No 2º conjunto as alturas dos blocos são 5 e 10 cm:



Qual dos conjuntos apresenta maior variabilidade?

As respostas apresentadas por um conjunto de 154 estudantes, foram:
- maior variab. o A - 11%
- maior variab. o B - 69%
- a mesma variab. - 20%

Apresentou-se seguidamente aos mesmos alunos outros dois conjuntos C e D. No conjunto C os blocos têm alturas 10, 20, 30, 40, 50 e 60 cm; no conjunto D há 3 blocos de altura 10 cm e outros 3 blocos de altura 60 cm:

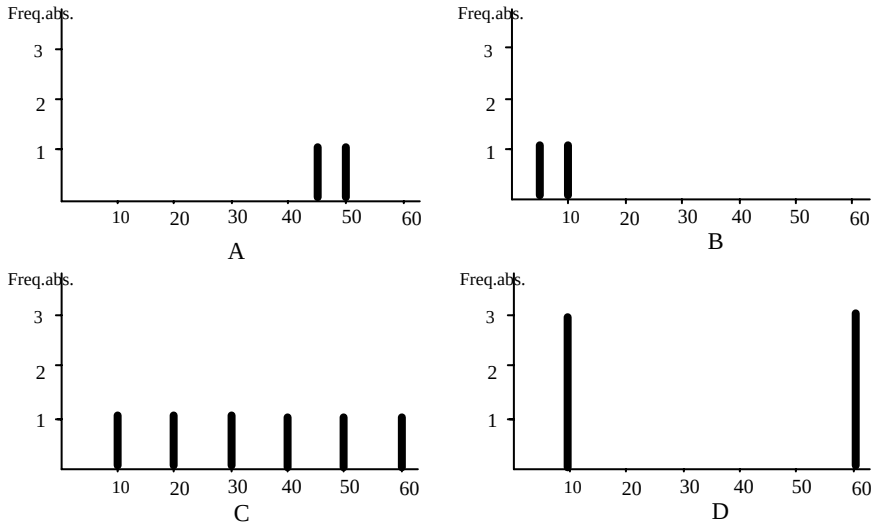


Qual dos conjuntos apresenta maior variabilidade?

As respostas apresentadas pelo conjunto de 154 estudantes, foram:
- maior variab. o C - 50%
- maior variab. o D - 36%
- a mesma variab. - 14%

Comentário: o resultado da experiência mostra que intuitivamente os estudantes enten-dem, de um modo garal, a variabiliadde em termos de "mais ou menos iguais uns rela-tivamente aos outros", independentemente de considerarem um ponto padrão como referência, nomeadamente a média.

Assim para visualizar convenientemente o conceito de variabilidade medida pelo desvio padrão, apresentam-se diagramas de barras. A partir destes gráficos os estudantes podem ver que a variabilidade das alturas pode ser expressa em termos dos desvios relativamente à média:



Pedindo para calcular o desvio padrão das alturas de cada um dos conjuntos os estudantes facilmente verificam que:

desvio padrão de A = desvio padrão de B

desvio padrão de C < desvio padrão de D

Confrontados com os resultados intuitivos, os estudantes concluem que o desvio padrão é uma medida muito específica da variabilidade.

O desvio padrão, da mesma forma que a média, é *muito sensível à presença de outliers*, sendo portanto uma medida de dispersão pouco resistente. Assim, um valor elevado para o desvio padrão pode ser devido ou a uma grande variabilidade nos dados, ou então a uma pequena variabilidade com a existência de um ou mais outliers.

3.3.3 - Amplitude inter-quartil

A medida mais simples para medir a variabilidade é a amplitude, que se representa por um R (range) e se define como a diferença entre o máximo e o mínimo da amostra:

R = máximo - mínimo

A medida anterior tem a grande desvantagem de ser muito sensível à existência, na amostra, de uma observação muito grande ou muito pequena. Assim, define-se uma outra medida, a amplitude inter-quartil, que é, em certa medida, uma solução de compromisso, pois não é afectada, de um modo geral, pela existência de um número pequeno de observações demasiado grandes ou demasiado pequenas. Esta medida é definida como sendo a diferença entre os 1º e 3º quartis:

amplitude inter-quartil = 3º quartil - 1º quartil

ou, utilizando a notação que introduzimos quando falamos nos quartis,

$$\text{amplitude inter-quartil} = Q_3 - Q_1$$

Do modo como se define a amplitude inter-quartil, concluimos que 50% dos elementos do meio da amostra estão contidos num intervalo com aquela amplitude. Esta medida já foi, aliás, utilizada na construção da box-plot.

Esta medida é *não negativa* e será *tanto maior* quanto *maior for a variabilidade* nos dados. Mas, ao contrário do que acontece com o desvio padrão, uma amplitude inter-quartil nula, não significa necessariamente, que os dados não apresentem variabilidade.

Por exemplo, o seguinte conjunto de dados

10 20 30 30 30 30 30 30 40 50

tem desvio padrão igual a 10.5 e amplitude inter-quartil igual a zero.

Qual das medidas de dispersão utilizar? Desvio padrão ou amplitude inter-quartil?

Do mesmo modo que a questão foi posta relativamente às duas medidas de localização mais utilizadas - média e mediana, também aqui se pode por o problema de comparar aquelas duas medidas de dispersão.

- A amplitude inter-quartil é mais resistente, relativamente à presença de outliers, do que o desvio padrão, que é mais sensível aos dados. Por outro lado, a amplitude inter-quartil não reflecte o conjunto de todos os dados, como o desvio padrão.
- Se a distribuição é enviesada pode acontecer que o desvio padrão seja muito superior à amplitude inter-quartil, sobretudo se se verificar a existência de "outliers".

Sugestões didácticas e comentários

1. INFLUÊNCIA DA ALTERAÇÃO DOS VALORES DA VARIÁVEL NA MÉDIA E NO DESVIO PADRÃO

Os 30 alunos de uma turma tiveram de fazer um trabalho de História. O professor resolveu ver quantas páginas tinha cada trabalho e obteve a seguinte lista:

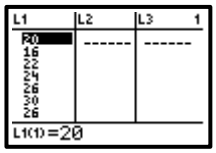
20	16	22	24	26	30	26	18	23	35
22	42	23	8	28	20	40	29	26	15
33	27	26	25	14	16	28	19	19	14

Podemos fazer um estudo estatístico sobre esta situação.

A população é constituída pelos 30 trabalhos da turma.

A variável em estudo é o número de páginas.

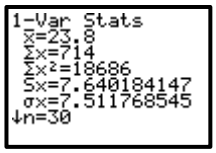
Introduzimos os dados numa calculadora gráfica e rapidamente obtemos as medidas estatísticas desta distribuição.

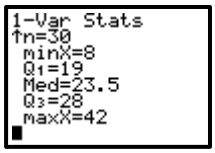


A média é 23.8.

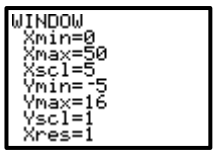
O desvio padrão é ≈ 7.512.

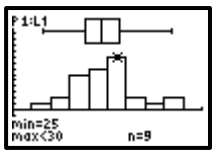
A mediana é 23.5.





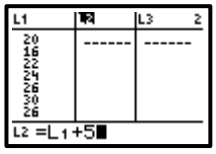
A visualização da distribuição pode ser feita num histograma e num diagrama de extremos e quartis.

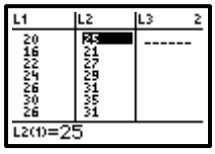




O professor achou que os trabalhos precisavam de uns anexos que ocupariam 5 páginas. Que influência terá este aumento de 5 páginas na média e no desvio padrão?

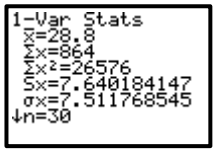
Para evitar o trabalho de escrever na calculadora todos os novos valores, podemos criar, a partir da lista L1, uma nova lista L2 em que cada elemento tem mais 5 unidades.

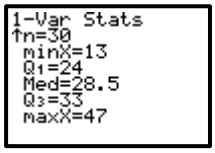




A média passou para 23.8 + 5 = 28.8.

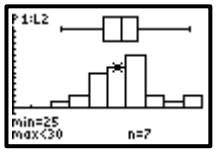
O desvio padrão manteve-se ≈ 7.512.

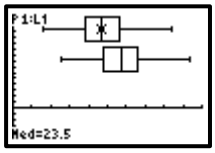




A mediana passou para 28.5.

A visualizaçãoi da distribuição pode ser feita num histograma e num diagrama de extremos e quartis.





A sobreposição no mesmo ecrã dos diagramas de extremos e quartis das duas listas mostra claramente que os dois diagramas são iguais, tendo havido apenas um deslocamento de 5 unidades.

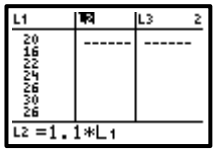
Vemos então que um aumento de 5 em todos os valores fez com que a média e a mediana também aumentassem de 5, enquanto que o desvio padrão se não alterou.

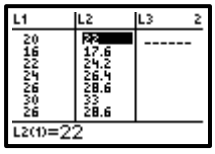
No caso geral, se todos os valores de uma população aumentarem de uma quantidade **b**, a média também aumenta **b**, mas o desvio padrão não se altera.

Imaginemos que o professor, em vez de pedir o aumento de 5 páginas, quisesse que os alunos desenvolvessem mais o trabalho, de modo que cada um deles ficasse 10% maior. Qual seria agora a influência sobre a média e o desvio padrão?

Podemos aproveitar a lista L1 que tem os valores iniciais, apagar a lista L2 e colocar aí os novos valores. Como se sabe, um aumento de 10% de uma certa quantidade corresponde a multiplicar essa quantidade por 1.1.

Basta então pôr em L2 uma lista obtida a partir de L1, multiplicando-a por 1.1.

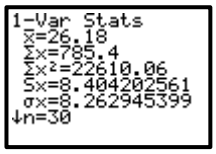


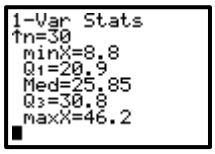


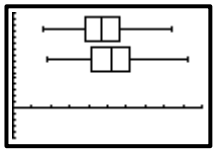
A média agora é 23.8 x 1.1 = 26.18.

O desvio padrão é 7.512 x 1.1 ≈ 8.263.

A mediana passou para 25.85.







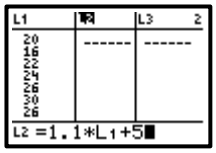
A sobreposição no mesmo ecrã dos diagramas de extremos e quartis para os dois casos mostra que o segundo diagrama sofreu um alongamento. Cada novo valor é 1.1 vezes maior que o valor correspondente da primeira distribuição. Como o diagrama é mais alongado, o desvio padrão é maior.

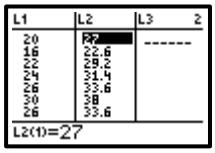
Assim, neste caso, a média aparece multiplicada por 1.1 e o desvio padrão também.

No caso geral, se todos os valores de uma população forem multiplicados por uma constante **a**, também a média e o desvio padrão aparecem multiplicados por **a**.

Que aconteceria se o professor exigisse não só que os trabalhos aumentassem 10% como também que se lhes acrescentasse o anexo de 5 páginas?

Vamos criar, a partir da pimeira lista L1, a lista L2 em que cada valor se obtém multiplicando o valor correspondente por 1.1 e somando 5.

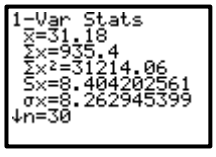


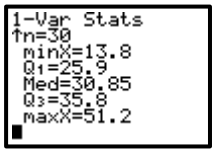


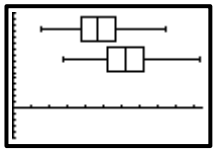
A média é 23.8 x 1.1 + 5 = 31.18.

O desvio padrão é 7.512 x 1.1 ≈ 8.263.

A mediana passou para 25.85.







A sobreposição no mesmo ecrã dos diagramas de extremos e quartis para os dois casos mostra que o segundo diagrama sofreu um alongamento e um deslocamento.

No caso geral, se todos os valores de um conjunto de dados sofrerem uma transformação do tipo ***ax+ b***, a média sofre uma transformação idêntica enquanto que o desvio padrão aparece multiplicado por **a**.

2 - Suponha que adicionou 100, a cada um dos valores de uma amostra. O que acontece ao:

a) Desvio padrão

b) Amplitude inter-quartil

c) Amplitude

d) Média

e) Mediana

3 - Suponha que obteve o valor -40.5 para a variância. O que conclui?

4 - Suponha que a amplitude de uma amostra é 105.4 e que ao calcular o desvio padrão obteve o valor 160.6. O que conclui?

5 - Suponha que tem os números 0, 1, 2, 3, …, 8, 9, 10. Pretende-se que escolha 4 destes números, sendo permitidas repetições, tal que (Moore, 1995):

a) i) Os 4 números escolhidos tenham o menor desvio padrão possível.

ii) Os 4 números escolhidos tenham o maior desvio padrão possível.

b) Haverá mais do que uma escolha possível em i) e ii)?

6 - O Sr. Malaquias, cujas habilitações literárias não vão além do 4º ano de escolaridade, respondeu a 2 anúncios de ofertas de emprego. As empresas trabalhavam no mesmo ramo, pelo que o serviço que o Sr. Malaquias iria fazer seria semelhante em qualquer das empresas. Resolveu perguntar alguma coisa sobre os ordenados processados nos dois sítios, tendo obtido a seguinte informação:

	Empresa A	Empresa B
Média	89 000$00	95 000$00
Mediana	80 000$00	70 000$00
Desvio padrão	3 200$00	3 800$00

Qual das empresas aconselharia o Sr. Malaquias a escolher, e porquê?

7 - Algumas pessoas preocupam-se com quantas calorias consomem. A revista Consumer Reports, num estudo sobre cachorros quentes, mediu as calorias em 20 tipos de salsichas de carne de vaca, 17 tipos de salsicha de carne de porco e 17 tipos de salsichas de carne de aves. Apresentam-se os "output" das estatísticas descritivas correspondentes a cada uma das variedades estudadas (Moore, 1995):

Carne de vaca:

Mean = 156.8	Standard deviation = 22.64	Min = 111	Max = 190
N = 20	Median = 152.5	Quartiles = 140, 178.5	

Carne de porco:

Mean = 158.7	Standard deviation = 25.24	Min = 107	Max = 195
N = 17	Median = 153	Quartiles = 139, 179	

Carne de aves:

Mean = 122.5	Standard deviation = 25.48	Min = 87	Max = 170
N = 17	Median = 129	Quartiles = 102, 143	

Construa diagramas de extremos-e-quartis paralelos, e faça uma comparação dos três tipos de cachorros, quanto às calorias.

8 - Suponha que pretende ter uma ideia da velocidade dos veículos numa autoestrada, por onde está a seguir. Ajusta a sua velocidade até que o nº de veículos que o ultrapassam consiga igualar o nº de veículos que ultrapassou. Com este procedimento obtém um valor aproximado para a velocidade média ou velocidade mediana (Moore, 1995)?

Capítulo 4

DADOS BIVARIADOS CORRELAÇÃO E REGRESSÃO

4.1 - Introdução

Por vezes o que se pretende estudar da População não é uma característica isolada, mas duas ou mais características que se supõe relacionadas entre si. No caso de se pretender estudar duas características conjuntamente, os valores observados aparecem sob a forma de pares de valores, isto é, cada indivíduo ou resultado experimental contribui com um conjunto de dois valores. É o que acontece, por exemplo, quando se considera para cada aluno candidato ao Ensino Superior, a classificação interna final e a nota do exame de uma disciplina. Outros exemplos são a altura e peso de alunos de uma escola primária; as notas de Física e Matemática dos alunos do 10º de uma dada escola; as alturas de pais e filhos; o consumo de gasolina e a cilindrada de um carro, etc. Então, para estudar duas características conjuntas, recolhe-se uma amostra de dados bivariados, a qual po-de ser representada da seguinte forma:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i,y_i), \ldots, (x_n,y_n)$$

Para representar e organizar este tipo de informação considera-se uma representação gráfica a que se dá o nome de nuvem de pontos ou diagrama de dispersão.

> ***Diagrama de dispersão*** *- É uma representação gráfica para os dados bivariados, em que cada par de dados* (x_i,y_j) *é representado por um ponto de coordenadas* (x_i,y_j)*, 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 x e y.

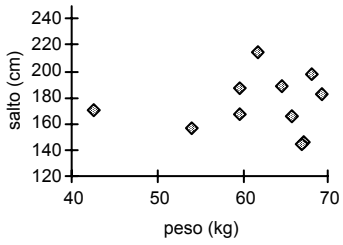
Consideremos alguns exemplos detalhadamente:

Exemplo 1: Com o objectivo de averiguar se a distância atingida no salto em comprimento está relacionada com o peso dos estudantes, um Professor de Educação Física seleccionou aleatoriamente 11 estudantes do sexo masculino para uma prova, tendo obtido os seguintes resultados:

Salto (cm)	187.5	182.5	214.0	147.0	167.0	157.5	170.0	198.5	145.0	166.5	189.0
Peso (Kg)	59.6	69.2	61.8	67.0	59.6	54.0	42.7	68.0	66.9	65.8	64.5

Que pode ele concluir? Note-se que aqui não estamos interessados no estudo estatístico de uma característica da população isoladamente, mas sim no modo como uma característica da população (a distância do salto em comprimento) está relacionada com outra característica da mesma população (o peso).

Para melhor compreendermos estes dados podemos fazer a representação gráfica adequada, obtendo uma *nuvem de pontos*, em que representamos nas ordenadas a variável de interesse (distância atingida no salto em comprimento) e em abcissa a variável explicativa (peso do estudante).



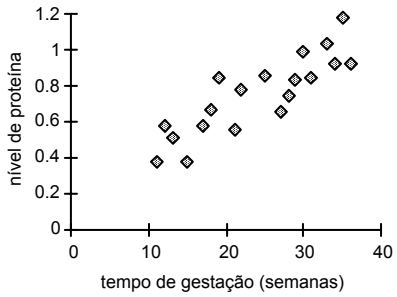
Observamos que não há uma relação clara entre estas duas características. A nuvem de pontos encontra-se bastante dispersa. Diz-se que então as duas características estão fracamente correlacionadas. Não é de esperar que o facto de sabermos o peso do aluno nos indique de algum modo a distância que ele vai saltar. Pode ser pesado e saltar bastante, como pode saltar pouco.

Exemplo 2: Um grupo de investigadores está interessado em saber se nas futuras mães o nível de uma proteína se altera (e no caso afirmativo, de que modo) ao longo da

gravidez. Seleccionou-se para o estudo 19 mulheres saudáveis, todas em estado diferente de gravidez (tempo de gestação), e mediu-se o nível de proteína em cada uma delas, tendo-se obtido os seguintes resultados (Bowman *et al.* 1987):

nível de proteína	Gestação (semanas)	nível de proteína	Gestação (semanas)	nível de proteína	Gestação (semanas)	nível de proteína	Gestação (semanas)
0.38	11	0.67	18	0.65	27	1.04	33
0.58	12	0.84	19	0.74	28	0.92	34
0.51	13	0.56	21	0.83	29	1.18	35
0.38	15	0.78	22	0.99	30	0.92	36
0.58	17	0.86	25	0.84	31		

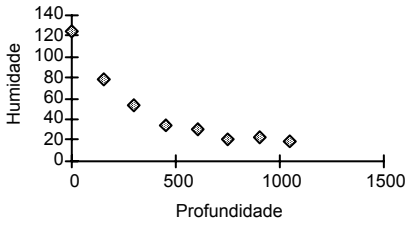
O objectivo desta experiência é averiguar como é que uma variável (nível de proteína) é afectada por uma outra variável (tempo de gestação). Se representarmos estes dados graficamente através da nuvem de pontos vemos claramente que o nível da proteína aumenta com o tempo de gestação. Podemos traçar uma recta no gráfico de modo que os pontos se encontrem próximos da recta e bem distribuídos para um lado e outro dela. Diz-se então que as variáveis estão positivamente correlacionadas. É pois de esperar que se consiga saber, através do tempo de gestação, qual o nível provável de proteína no sangue.



Exemplo 3: Recolheram-se amostras de solo do estuário do rio Tejo a 8 profundidades distintas e mediram-se os respectivos graus de humidade (gramas de água/ 100g solo) obtendo-se os seguintes resultados (Davis, 1973):

Profundidade (cm)	0	150	300	450	600	750	900	1050
Humidade (gr. água/ 100g solo)	124	78	54	35	30	21	22	18

Representando os dados graficamente obtém-se:



Observamos que quando a profundidade aumenta, a humidade diminui. Diz-se, neste caso, que as duas variáveis, estão negativamente correlacionadas, pois variam em sentidos opostos.

4.2 - Coeficiente de correlação linear

O grau de associação linear entre duas variáveis é traduzido matemáticamente por uma estatística a que se dá o nome de *correlação linear*, ou *coeficiente de correlação linear*, a qual se representa geralmente por r. Se representarmos por x_i os valores das observações correspondentes a uma das variáveis e por y_i os valores das observações correspondentes à outra variável, então o coeficiente r obtém-se através da expressão,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde $\bar{x}$ é a média das observações x_i e $\bar{y}$ é a média das observações y_i.

Prova-se que o valor desta estatística está entre -1 e 1.

Note-se que quando as variáveis variam no mesmo sentido, se $x_i > \bar{x}$, então também se espera ter, em geral, $y_i > \bar{y}$, e que quando $x_i < \bar{x}$, também $y_i < \bar{y}$, o que faz com que o produto no numerador seja, em geral, positivo. O caso r>0 corresponde assim à situação em que as variáveis variam no mesmo sentido, isto é, estão positivamente correlacionadas.

Quando as variáveis variam em sentido contrário, então valores positivos da diferença entre x_i e $\bar{x}$, aparecem associados, em geral, a valores negativos da diferença entre y_i e $\bar{y}$ e vice-versa, o que faz com que o produto no numerador venha negativo. Assim, o

caso r <0 corresponde à situação em que a variação é em sentidos opostos, ou seja as variáveis estão negativamente correlacionadas.

O caso r = 0 corresponde à situação em que aquele produto tende a ter valores quer positivos, quer negativos. Isto acontece quando um valor positivo ou negativo da diferença entre x_i e $\overline{x}$, aparece associado com valores quer positivos quer negativos da diferença entre y_i e $\overline{y}$. Diz-se então que as variáveis não estão correlacionadas.

Os valores extremos da correlação, r =1 ou -1, correspondem à situação em que os valores das variáveis se encontram sobre uma recta com declive positivo ou negativo.

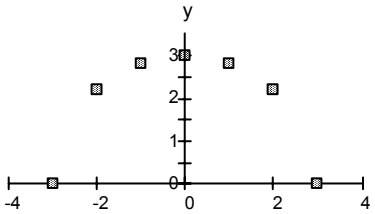
Nos exemplos apresentados os valores da estatística r são:

- r = 0.077 para o 1º exemplo traduzindo uma muito fraca associação entre o peso dos estudantes e a distância conseguida no salto em comprimento;
- r = 0.86 para o 2º exemplo, traduzindo uma forte associação positiva entre o tempo de gestação e o nível da proteína no sangue;
- r = - 0.891 para o 3º exemplo traduzindo uma forte associação negativa entre a humidade e a profundidade.

Observação: A expressão do coeficiente de correlação é aqui apresentada como mera informação para os Professores. Os alunos devem obter os valores dos coeficientes de correlação para várias situações através da máquina de calcular. O que se pretende é que eles apenas relacionem o valor de r com o grau e o tipo de associação linear existente entre as variáveis em estudo.

É também importante frisar que o coeficiente de correlação traduz apenas o grau de relação linear existente entre duas variáveis. O facto de o coeficiente de correlação ser zero, não implica que as variáveis não estejam relacionadas. Com efeito, no exemplo que se segue, r = 0 e, no entanto, as variáveis x e y estão relacionadas pela relação determinística não linear $x^2 + y^2 = 9$

x	-3	-2	-1	0	1	2	3
y	0	$\sqrt{5}$	$2\sqrt{2}$	3	$2\sqrt{2}$	$\sqrt{5}$	0



4.3 - Recta de regressão

Quando a correlação entre duas variáveis é elevada (quer seja positiva, quer seja negativa), isso significa que se conhecermos o valor de uma das variáveis então é possível ter uma ideia do valor que a outra variável irá tomar. Em linguagem estatística, diz-se que podemos inferir o valor da outra variável.

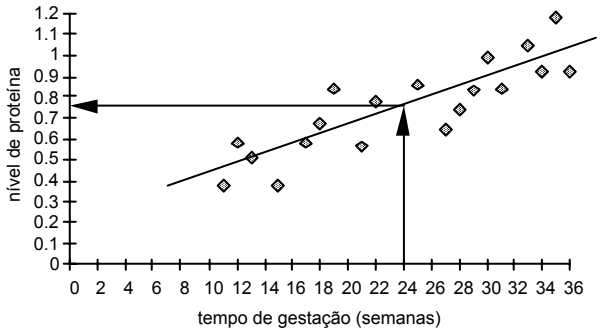
Assim, voltando ao exemplo da proteína, consideremos uma senhora grávida com 24 semanas de gestação. Qual será o valor que o nível de proteína deve apresentar?

Para respondermos a esta questão podemos construir uma recta que "melhor" aproxime os pontos que constituem a nuvem de pontos. Claro que há muitas rectas possíveis. Um dos critérios mais usados para definir esta recta, é o de tornar mínima a soma dos quadrados dos desvios dos pontos em relação à recta[1]. Essa recta é a chamada *recta de regressão* (dos mínimos quadrados). Matematicamente pode-se encontrar essa recta. Prova-se que ela passa pelo centro de gravidade da distribuição, isto é, pelo ponto $(\overline{x},\overline{y})$ e que o declive está relacionado com o coeficiente de correlação e tem o mesmo sinal.

[1] Designamos por desvio no ponto de abcissa x_i à diferença entre o valor observado y_i e o valor correspondente sobre a recta.

Para o exemplo da proteína, a recta de regressão é: $y = 0.023\,x$ +0.0202.

Construída a recta, podemos responder à pergunta formulada. O valor que inferimos para o nível da proteína correspondente a 24 semanas de gravidez é o valor sobre a recta correspondente a x_i =24, isto é 0.754:



Sugestões didácticas e comentários

Haverá alguma relação entre o número de médicos e a taxa de mortalidade infantil? À primeira vista, parece que sim. É provável que quanto mais médicos houver, menos crianças morram no primeiro ano após o nascimento. Para investigar se esta hipótese está correcta, recolheram-se os dados referentes a alguns países (Anuário Estatístico, Planeta De Agostini, 1994).

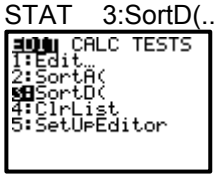
País	Médicos por 10000 habitantes	Taxa de mortalidade infantil (por 1000 nados vivos)
Bélgica	15.38	8
Honduras	2.65	49
Irão	3.06	68
México	6.75	36
Nicarágua	4.65	56
Peru	5.21	53
Polónia	14.29	15
Portugal	9.01	11
Quénia	1.25	67
Roménia	11.90	27
Uruguai	10.99	21
Venezuela	8.93	34

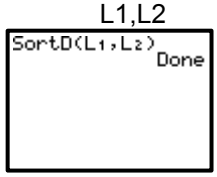
A representação destes dados num gráfico de correlação vai permitir-nos visualizar a situação. Usando uma calculadora gráfica, isso pode ser feito rapidamente.

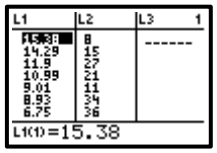
Introduzimos os dados referentes ao número de médicos por 10000 habitantes numa lista (L1) e a taxa de mortalidade infantil noutra (L2).

Se quisermos, embora não seja necessário, podemos ordenar os dados, relativamente ao número de médicos, fazendo:

STAT 3:SortD(... L1,L2 STAT 1:Edit...



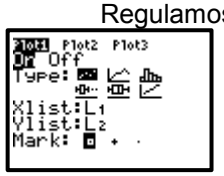




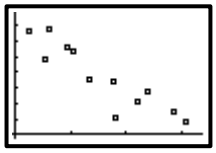
Os dados ficam por ordem decrescente da lista 1. Repare-se que os valores correspondentes das duas listas continuam associados porque, ao dar a instrução SortD(L1,L2), a máquina ordenou a lista 1, alterando ao mesmo tempo a lista 2.

Para obter a nuvem de pontos fazemos:

STAT PLOT 1:Plot1 Regulamos o gráfico ZOOM 9:ZoomStat

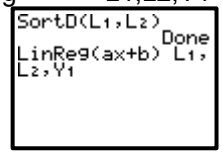


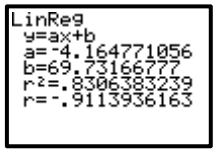
Vê-se claramente que existe uma correlação negativa. Os pontos dispõem-se de tal modo que, genericamente, ao aumento de uma variável corresponde a diminuição da outra.



Vamos agora procurar a recta de regressão e o coeficiente de correlação **r**.

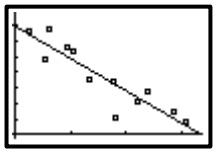
STAT CALC 4:LinReg L1,L2,Y1 ENTER





–

GRAPH



Ao pedirmos LinReg(ax+b) L1,L2,Y1 a máquina não só determina a equação da recta de regressão como também a coloca imediatamente no editor de funções. Assim, se agora pedirmos o gráfico, vamos ter a nuvem de pontos e a recta de regressão.

A equação da recta é, usando valores aproximados, $y = -4.165x + 69.73$.

A correlação é relativamente forte. O seu coeficiente é $r \approx -0.911$.

Se nos disserem que num país há 13 médicos por 10000 habitantes, qual será a sua taxa de mortalidade infantil?

Para encontrar a correlação entre as duas variáveis, só usámos os valores referentes a 12 países e não sabemos se eles são uma amostra representativa da população. Se tivermos a certeza que sim, então podemos usar a recta de regressão para encontrar um valor aproximado da taxa de mortalidade infantil:

$y = -4.165 \times \mathbf{13} + 69.73 \approx 15.6$

É de prever que a taxa de mortalidade infantil seja próxima de 16.

4.4 - Análise preliminar dos dados, antes de construir a recta de regressão

Para avaliar da necessidade de uma análise cuidada dos dados antes da obtenção da recta de regressão consideremos o seguinte exemplo (Sen *et al*., 1990, pg 24).

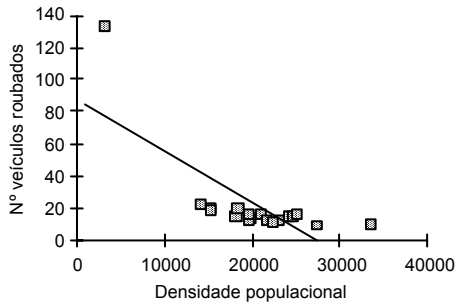
Exemplo: Fez-se um estudo para averiguar a relação existente entre o número de veículos roubados por cada mil habitantes e a densidade populacional na cidade de Chicago. Seleccionaram-se, para o efeito, 18 distritos dessa cidade. Registou-se, para cada distrito, a sua densidade populacional (DP) e o número de veículos aí roubados (NVR) por cada mil habitantes, tendo-se obtido os seguintes resultados:

DP	NVR	DP	NVR	DP	NVR
3235	132.8	19581	16.5	21675	12.5
24182	14.9	14077	22.2	22315	11.8
20993	16.7	18137	15.8	18402	19.6
15401	20.0	22919	13.3	33445	10.5
19749	14.2	24534	15.1	27345	10.1
19487	13.5	24987	16.2	15358	19.0

O coeficiente de correlação é -0.74 e a recta de regressão de mínimos quadrados é

$$NVR = 88.195 - 0.00326\ DP$$

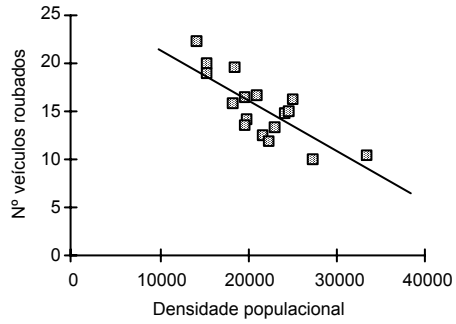
Se fizermos a representação gráfica destes dados vemos que há um distrito que tem um comportamento totalmente diferente dos outros.



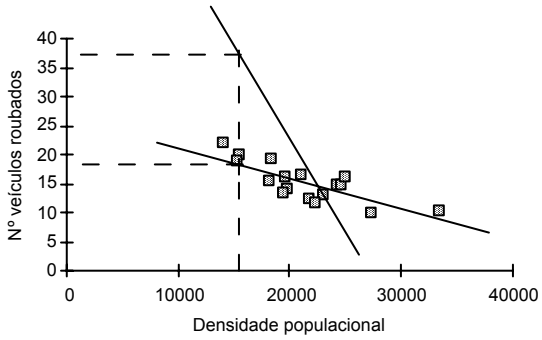
O 1º distrito que aparece na tabela tem uma densidade populacional muito baixa, mas um elevado número de veículos roubados. Uma averiguação mais cuidada levou à conclusão que aquele distrito correspondia ao Centro de Chicago, uma área essencialmente de comércio e de escritórios, e consequentemente uma área em que a densidade de veículos não tem a ver com a densidade populacional. Este distrito não deveria ter sido incluído na amostra. Assim, retirando este distrito, podemos construir uma nova recta de regressão. Obtém-se agora a recta

$$NVR = 27.36 - 0.00056\ DP$$

sendo o coeficiente de correlação -0.79.



Repare-se na alteração verificada. As conclusões extraídas de uma recta e de outra podem ser bem diferentes. Por exemplo, se considerarmos o valor de 15401 para a densidade populacional, que é um dos valores tabelados, o valor previsto para o número de carros roubados, utilizando a primeira recta de regressão é 38.0, enquanto que o previsto pela segunda recta de regressão é 18.7, bem mais próximo de 20 (valor observado).



As considerações anteriores levam-nos a concluir que a recta de regressão não é resistente, pois é muito influenciada por valores estranhos- outliers, da amostra (o facto da determinação da recta de regressão estar ligada ao ponto $(\overline{x},\overline{y})$, conduz-nos imediatamente à conclusão anterior, pois como sabemos a média não é uma medida resistente). Daí a necessidade de analisar cuidadosamente os dados, antes de se proceder a uma análise de regressão.

Sugestões didácticas e comentários

A tabela seguinte apresenta 3 conjuntos de dados A, B e C, preparados pelo estatístico Frank Anscombe, para ilustrar os perigos de calcular medidas sem primeiro representar os dados. Os conjuntos de dados A, B e C têm a mesma correlação e a mesma recta de regressão (Moore, 1995):

_

A											
x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.6
B											
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
C											
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

a) Calcule o coeficiente de correlação e a recta de regressão para cada um dos conjuntos de dados e verifique que são iguais.

b) Para cada um dos conjuntos de dados faça o diagrama de pontos e represente a recta de regressão.

c) Em qual das situações acha que pode utilizar a recta de regressão para predizer y para x=13.5? Justifique a resposta.

Capítulo 5

NOTAS FINAIS

5.1 - Introdução

Sendo objectivo da Estatística o de retirar informação a partir de **dados**, gostaríamos, como nota final, de chamar a atenção para o que diz David Moore, em The Basic Practice of Statistics, " *…Data are numbers, but they are not "just numbers". **Data are numbers with a context**. The number 10.5, for example, carries no information by itself. But if we hear that a friend's new baby weighed 10.5 pounds at birth, we congratulate her on the healthy size of the child. The context engages our background knowlwdge and allows us to make judgments. We know that a baby weighing 10.5 pounds is quite large, and that it isn't possible for a human baby to weigh 10.5 ounces or 10.5 kilograms. The context makes the number informative".*

Assim, mais uma vez observamos que deve ser incentivado nos alunos o gosto pela análise e interpretação, mais do que a simples utilização dos dados para a manipulação de gráficos e fórmulas. Aliás, aproveitamos para observar, mais uma vez, que é precisamente neste tema da Estatística que os alunos devem ser aconselhados a utilizar a **calculadora** para não serem sobrecarregados com cálculos pesados e desnecessários.

Também, tendo em consideração o que dissemos no primeiro parágrafo, a **avaliação** deste tema merece uma observação especial. Sempre que possível, essa avaliação dever-se-á centrar na realização de pequenos projectos, que se desenvolverão ao longo das aulas, à medida que os conceitos forem introduzidos, evitando, unicamente, os testes clássicos de uma disciplina de Matemática. Assim, e meramente a título de exemplo, damos algumas sugestões de pequenos trabalhos, que podem ser objecto de trabalhos de grupo.

5.2 - Sugestões para projectos a desenvolver pelos alunos

1. Pedir aos alunos da turma que recolham a informação referente à altura de cada um deles e dos respectivos pais. Utilizar esses dados para estudar, por exemplo, as alturas referentes aos homens e às mulheres, uma eventual relação de dependêncial linear entre as alturas dos maridos e das mulheres, ou entre os pais e os filhos, etc.

2. Recolher informação, junto de alguns alunos da escola, seleccionados ao acaso, sobre o nº de faltas e o dia da semana em que se deu a falta. Será que os alunos faltam uniformemente nos diferentes dias da semana, ou haverá dias com maior incidência de faltas?

3. Recolher informação sobre as notas da disciplina de Matemática de duas turmas de alunos do mesmo ano e do mesmo professor. Haverá evidência de que as turmas não tenham o mesmo aproveitamento?

4. Recolher informação sobre as notas (do 1º período) de alguns alunos do 10º ano, nas disciplinas de Matemática e Português. Haverá indícios de que os aproveitamentos sejam diferentes? Haverá tendência para que os alunos que têm boa nota a Português também tenham boa nota a Matemática?

5. Haverá relação entre o número de negativas a Português e a Matemática nas várias turmas? Recolher os dados relativos ao número de negativas nestas duas disciplinas em todas as turmas da escola e referentes ao período anterior. Estudar a possível correlação entre as duas variáveis. Elaborar um pequeno relatório.

6. Recolher informação sobre as notas da disciplina de Matemática de alguns alunos do 12º ano do ano lectivo anterior e as respectivas notas no exame nacional de Matemática. Haverá indícios que levem a afirmar que os exames nacionais foram demasiado simples ou demasiado complicados, ou pelo contrário, ajustavam-se aos alunos a que se destinavam?

7. Comparar dois autores no que diz respeito à frequência de utilização de determinadas palavras ou comprimentos das frases dos seus textos.

5.3 - Sugestões para actividades na sala de aula

1. COMPRIMENTO (1)

O professor escolhe um comprimento (por exemplo: a largura do quadro da sala, a altura da sala, o comprimento da janela). Cada aluno escreve a sua estimativa desse comprimento, com aproximação ao centímetro.

- Faz-se a recolha e a organização dos dados.
- Calculam-se as principais medidas de localização e dispersão.
- Fazem-se as representações gráficas adequadas.
- Mede-se o verdadeiro valor do comprimento e situamo-lo em relação à média e à mediana.
- Vê-se quem foi o aluno que fez a melhor estimativa.

2. COMPRIMENTO (2)

Faz-se um estudo semelhante ao anterior para as estimativas de um novo comprimento indicado. Comparar as medidas de dispersão com as do caso anterior. Desta vez a dispersão deve ser bastante menor visto os alunos terem a informação do comprimento do primeiro estudo.

3. BOLA AO CESTO

Na aula de Educação Física, escolhe-se uma certa distância à tabela de basquetebol. Cada aluno faz 20 lançamentos e regista o número de vezes que conseguiu introduzir a bola no cesto. Os alunos podem estar organizados aos pares: enquanto um lança, o outro faz os registos.

- Organizar os dados em tabelas e gráficos.
- Determinar as principais medidas de localização e dispersão.
- Fazer um relatório sobre a capacidade de encestar dos alunos da turma.

4. TEMPO

O professor tem um cronómetro, mas os alunos não podem olhar para os seus relógios. Num determinado momento o professor diz “Começou” e passado algum tempo (entre 20 e 60 segundos) diz “Fim”.

- Cada aluno regista a estimativa que faz do tempo decorrido.
- Os dados são recolhidos e tratados estatisticamente.
- No fim, o verdadeiro valor é comparado com os tempos estimados pelos alunos. A melhor estimativa pode receber um prémio.

5. M & M’s

As embalagens de M&M's trarão todas o mesmo número de pastilhas?

Cada aluno traz de casa uma embalagem pequena de M&M's fechada.

As embalagens são abertas na aula e cada aluno conta quantas pastilhas de chocolate tem a sua embalagem.

• Recolhem-se os dados referentes a todas as embalagens.

• Faz-se o estudo estatístico do número de pastilhas por embalagem.

6. SOBREVIVÊNCIA DOS M & M's

Material por cada grupo de 2 alunos:

1 copo de plástico

2 pratos de plástico

40 pastilhas de chocolate M & M's

• Colocam-se as 40 pastilhas no copo e lançam-se para um dos pratos. As pastilhas que não ficarem com a pequena inscrição "M&M" virada para cima são eliminadas e colocadas no 2º prato. As que ficaram com a inscrição virada para cima são as "sobreviventes" e voltam a ser colocadas no copo.

• Repete-se o processo com as sobreviventes.

• Ao fim de 4 lançamentos do copo, a experiência termina e regista-se o número de pastilhas que não foram eliminadas.

• Cada grupo de 2 alunos faz esta experiência 10 vezes.

• Faz-se a recolha dos resultados de todas as experiências da turma.

• Estuda-se estatisticamente o número de sobreviventes (medidas de localização e de dispersão, gráficos, etc.).

• No fim, cada um come os seus "dados estatísticos"...

7. MOEDAS

• Cada aluno regista o número de moedas que tem e a respectiva quantia em escudos.

• Recolher os dados referentes a toda a turma.

• Fazer o estudo estatístico referente à variável "número de moedas".

• Fazer o estudo estatístico referente à variável "quantia".

• Estudar a correlação entre as variáveis "número de moedas" e "quantia".

(Retirado de Bastos *et al*., 1997)

BIBLIOGRAFIA

BARRETO, A. (1996) - *A Situação Social em Portugal, 1960-1995*, Instituto de Ciências Sociais, Universidade de Lisboa.

BASTOS, R.; BERNARDES, A.; LOPES, A. V.; LOUREIRO, C.; VARANDAS, J. M.; VIANA, J. P. (1997) - *Matemática 10,* Edições Contraponto, Porto.

BOWMAN, A. W.; ROBINSON, D. R. (1987) - *Introduction to Statistics*, Adam Hilgor, Bristol.

BOWMAN, A. W.; ROBINSON, D. R. (1987) - *Regression and Analysis of Variance*, Adam Hilgor, Bristol.

CLEGG, F. (1995) - *Estatística para Todos*, Gradiva, Lisboa.

DAVIS, J. C. (1973) - *Statistics and Data Analysis in Geology*, Wiley.

FREEDMAN, D.; PISANI, R.; PURVES, R.; ADHIKARI, A. (1991) - *Statistics*, Second Edition, W.W. Norton & Company, New York.

GAL, I. (1995) - Statistical Tools and Statistical Literacy: The Case of The Average, *Teaching Statistics*, Vol. 17, Number 3.

GRAÇA MARTINS, M. E. (1995) - *Introdução às Probabilidades e à Estatística* - Edição da Sociedade Portuguesa de Estatística, Lisboa.

Grupo Azarquiel, (1993) - *Estatística no 3º Ciclo do Ensino Básico*, Associação de Professores de Matemática, Lisboa.

HAWKINS, A.; JOLLIFFE, GLICKMAN, L. (1992) - *Teaching Statistical Concepts*, Longman, London.

HOLMES, P. (1994) - Classroom Practicals, Centre for Statistical Education, University of Sheffield.

HOLMES, P. (1994) - Stem and Leaf, Centre for Statistical Education, University of Sheffield.

HOLMES, P.; WORSNOP, R. (1993) - Bottles and Things, Centre for Statistical Education, University of Sheffield.

HOLMES, P.; WORSNOP, R. (1992) - Canteen Choice, Centre for Statistical Education, University of Sheffield.

HOLMES, P.; WORSNOP, R. (1993) - Growing Up, Centre for Statistical Education, University of Sheffield.

Instituto Nacional de Estatística (1991) - *Anuário Estatístico de Portugal*, INE, Lisboa.

LOOSEN, F.; LION, M.; LACANTE, M. (1985) - The Standard Deviation: Some Drawbacks of an Intuitive Approach, *Teaching Statistics*, Vol. 7, Number 3.

MENDENHALL, W.; OTT, L.; LARSON, R. (1974) - *A Tool for the Social Sciences*, Duxbury Press, Belmont, California.

MOORE, D. (1995) - *The Basic Practice of Statistics,* W. H. Freeman adn Company, New York.

ROUNCEFIELD, M. (1994) - Box Plots, Centre for Statistical Education, University of Sheffield.

RUNYON, R. P.; HABER, A.; PITTENGER, D.; COLEMAN, K. A. (1996) - *Fundamen-tals of Behavioral Statistics*, MacGraw-Hill Companies, U.S.A..

SEN, A.; SRIVASTAVA, M. (1990) - *Regression Analysis*, Springer-Verlag, New York.

VICENTE, P.; REIS, E; FERRÃO, F. (1996) - *A amostragem como factor decisivo de qualidade*, Edições Sílabo, Lda, Lisboa.

WEISS, N. A. (1989) - *Elementary Statistics*, Addison-Wesley Publishing Company U.S.A..