escola superior de tecnologia e gestão
instituto politécnico de leiria

escola superior de educação
e ciências sociais
instituto politécnico de leiria

Projeto

Mestrado em Educação e Tecnologia em Matemática

ESTATÍSTICA
NO ENSINO BÁSICO E SECUNDÁRIO

Maria Alice da Silva Martins

Leiria, maio de 2012

escola superior de tecnologia e gestão
instituto politécnico de leiria

escola superior de educação
e ciências sociais
instituto politécnico de leiria

Projeto

Mestrado em Educação e Tecnologia em Matemática

ESTATÍSTICA
NO ENSINO BÁSICO E SECUNDÁRIO

Maria Alice da Silva Martins

Dissertação de Mestrado realizada sob a orientação da Doutora Helena Ribeiro, Professora da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, e co-orientação do Doutor Rui Santos, Professor da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria.

Leiria, maio de 2012

Agradecimentos

Ao meu marido e às minhas filhas, pela compreensão que tiveram comigo e pela força que me transmitiram ao longo desta caminhada.

Aos meus pais, pelo apoio dado, estando presentes sempre que foi necessário.

Aos meus orientadores, Doutora Helena Ribeiro e Doutor Rui Santos, pelo tempo dispensado, pelas sugestões feitas e todo o apoio prestado, que possibilitaram a realização deste trabalho.

Aos meus colegas de mestrado, pela ajuda e força dadas, mesmo quando a vontade de continuar escasseava.

À minha amiga Teresa, pelo estímulo dado ao longo deste percurso.

Aos meus colegas de trabalho, que direta ou indiretamente contribuíram com as suas palavras de encorajamento.

A todos os meus sinceros agradecimentos.

RESUMO

A sociedade da informação exige que todos os cidadãos tenham conhecimentos de Estatística para poderem intervir de forma crítica e fundamentada. Esta situação conduziu a Estatística a um lugar de relevo no currículo dos alunos, que exige um novo olhar sobre o seu ensino, como preconizam os atuais programas.

É neste contexto que surge o presente trabalho que, numa primeira parte, apresenta uma revisão dos conceitos estatísticos lecionados no ensino básico e secundário, uma ferramenta importante para o trabalho dos professores, permitindo-lhes uma clarificação desses conceitos, num texto que se pretende cientificamente rigoroso.

De forma a alertar para incorreções, gralhas e/ou erros comuns, segue-se uma análise crítica a alguns materiais disponíveis, nomeadamente manuais escolares atuais, onde o estudo da regressão linear assume uma análise mais detalhada.

Com o intuito de enriquecer os materiais existentes, numa perspetiva inovadora, capaz de promover aprendizagens significativas, apresenta-se um conjunto de propostas de trabalho para a sala de aula onde a tecnologia, nomeadamente o *GeoGebra*, adquire um papel de relevo na compreensão dos conceitos. De forma a facilitar a utilização deste *software* surge, no início da terceira etapa deste trabalho, uma explicação detalhada sobre o uso do *GeoGebra* na estatística descritiva.

Em suma, este trabalho pretende contribuir para a melhoria do ensino da Estatística, quer no que se refere à preparação do corpo docente, quer através da inclusão de propostas de trabalho para utilização em sala de aula.

Palavras chave: ensino de estatística, estatística descritiva, *GeoGebra*, regressão linear.

ABSTRACT

The media industry requires that all citizens have a knowledge of Statistics in order to play a critical and fundamented role in society. This situation has led Statistics to a prominent place in the curriculum of students and demands a new look to education, as recommended by current programs.

It is in this context that the present work has been elaborated. The first part presents a review of statistical concepts taught at an elementary and secondary level, an important tool for teachers' work, and allows the clarification of these concepts in a text intended to be scientifically rigorous.

In order to draw one's attention to mistakes, typos and/or common errors, the first part is followed by a critical analysis of some available materials, which includes current textbooks and where the study of linear regression assumes a more detailed analysis.

Having as main target the improval of the existing materials, with an innovative approach, which can foster meaningful learning, a set of suggested tasks are presented to the classroom. In this space, technology, namely *GeoGebra*, assumes a relevant role in the understanding of concepts. In order to make the use of this software easier, a detailed explanation of *GeoGebra* in descriptive statistics comes in the begining of the third stage of this work.

In summary, this paper aims to contribute to the improvement of Statistic teaching, whether it concerns the preparation of mathematics teachers or by the inclusion of tasks applied in the classroom's context.

Keywords: Statistic teaching, descriptive statistics, *GeoGebra*, linear regression.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Desde as primeiras grandes civilizações que a Estatística tem sido utilizada de forma profícua. Em 3050 A.C. já os egípcios recorriam à Estatística para apurar os recursos humanos e materiais disponíveis para a construção das pirâmides. Outras civilizações, tais como os Chineses, os Gregos ou os Romanos, utilizaram a Estatística para conhecerem os bens que o estado possuía bem como a sua distribuição pela população. Desde então, muitos foram os desenvolvimentos das aplicações da Estatística, nomeadamente no que se refere à recolha, organização e análise de dados, quer esta seja restrita ao resumo da sua informação quer seja com o intuito de inferir ou efetuar previsões. Atualmente, a Estatística é não só um instrumento indispensável na política de qualquer estado (no século XVII, em Inglaterra, a Estatística era a "Aritmética do Estado") como é também um instrumento importante em muitas outras áreas, tais como a Psicologia, a Sociologia, a Medicina, a Economia, o Desporto, a Biologia, a Física, a Educação, a Meteorologia, entre muitas outras. Deste modo, presentemente a Estatística desempenha um papel fundamental na vida do cidadão, não só pela utilidade das suas múltiplas aplicações, mas igualmente por ser indispensável para a análise, interpretação e compreensão da informação que os meios de comunicação social divulgam.

Como refere Fernandes "a influência da Estatística na vida das pessoas e nas instituições tem-se tornado cada vez mais visível, o que implica que todos os cidadãos devam ter conhecimentos de Estatística para se poderem integrar na sociedade actual" (Fernandes, 2009, p. 1). Assim, tem havido um reconhecimento da importância da Estatística no currículo dos alunos, de tal modo que esta tem vindo a ocupar, cada vez mais, um lugar de destaque no ensino, desde o ensino básico ao ensino secundário, pois segundo Martins *et al.* a Estatística "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 sentido crítico e ainda a aquisição de uma certa 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" (Martins *et al.*, 1997, pp. 7–8). A coletânea de artigos "Ensino e Aprendizagem da Estatística" (Loureiro *et al.* (2000)) é um bom exemplo das reflexões sobre a importância do Ensino da Estatística no currículo do ensino secundário e ensino básico, da utilidade da inclusão da tecnologia na lecionação destes conteúdos, bem como a partilha de experiências e materiais didáticos no início do século XXI.

No atual programa do ensino básico o ensino da Estatística inicia-se no primeiro ciclo assim como os aspetos elementares da Probabilidade, constituindo, em conjunto, desde 2007, o tema "Organização e Tratamento de Dados" (OTD)[(1)], conforme as orientações internacionais presentes em NCTM[(2)] (2008). No ensino secundário o tema aparece no 10.º ano com a designação "Estatística", onde se pretende ampliar os conhecimentos adquiridos anteriormente. É de referir que não é objetivo deste projeto abordar conceitos relativos à Probabilidade, pelo que nos restringiremos à análise dos conteúdos referentes a Estatística.

Nos últimos anos tem havido um esforço por parte de alguns autores em fornecer suporte teórico e didático que possa ajudar os professores nas suas aulas e que se encontram de acordo com os programas atuais. Martins *et al.* (2007) elaboraram uma brochura direcionada aos professores do 1.º ciclo que surgiu no âmbito do Programa Nacional de Formação Contínua em Matemática. Esta brochura, além de conter os conceitos e procedimentos fundamentais para um professor deste nível de ensino, apresenta várias propostas de tarefas a implementar na sala de aula com a respetiva exploração. A brochura elaborada por Martins & Ponte (2010) desenvolve as orientações metodológicas relativas à OTD, apresentando quatro capítulos reservados à Estatística, onde também são sugeridas tarefas a propor aos alunos na sala de aula.

Um *site* de referência na área da Estatística é o ALEA — Acção Local Estatística Aplicada, `www.alea.pt` (podemos encontrar mais informações em INE (2009)). Este projeto existe desde 1999 e tem-se mantido sempre em crescimento, apresentando várias componentes, tais como entretenimento, cursos de estatística e dados estatísticos. É direcionado a

(1) Consulte-se, por exemplo, Loura (2009) para uma análise crítica ao novo programa.

(2) NCTM — National Council of Teachers of Mathematics.

alunos, professores e público em geral e pretende contribuir para melhorar a literacia estatística[3] e fomentar situações e experiências de aprendizagem recorrendo às Tecnologias da Informação e Comunicação (TIC). É de referir que aquando da comemoração do seu 10.º aniversário foi editado um livro constituído por 5 dossiers produzidos pelo ALEA e que é mais um documento útil nesta temática. Há, contudo, diversos outros materiais para o ensino e aprendizagem da Estatística disponíveis (cf. Nascimento (2009)).

Este projeto, intitulado Estatística no Ensino Básico e Secundário, tem como principais objetivos:

- apresentar os conceitos mais elementares de Estatística lecionados nestes ciclos de ensino, num texto cientificamente rigoroso, direcionado a professores, permitindo-lhes a clarificação destes conceitos;
- analisar alguns materiais disponíveis para o ensino da Estatística nestes ciclos, apresentando as incorreções, gralhas e/ou erros detetados, bem como materiais que consideremos insuficientes e/ou inadequados para os objetivos para os quais foram concebidos;
- criar novos materiais bem como sugerir metodologias que possam ser utilizados no ensino da Estatística.

Neste sentido, o segundo capítulo apresentará uma revisão dos conceitos de Estatística lecionados no ensino básico e no ensino secundário: tabelas de frequências, representações gráficas e principais medidas de estatística descritiva, conforme definido no Programa de Matemática do Ensino Básico (Ponte *et al.*, 2007) e no Programa de Matemática A do Ensino Secundário (Silva *et al.*, 2001). Este capítulo irá basear-se em livros de autores especialistas na área, tais como Murteira (1993), Reis (1998), Batanero & Godino (2003), Pestana & Velosa (2009) e Maria Eugénia Martins com os seus diversos trabalhos nesta área (como por exemplo Martins (2005), Martins & Cerveira (1999), Martins & Ponte (2010) ou Martins *et al.* (2007, 1997)). Não se pretende que este capítulo seja direcionado para os alunos, mas antes para professores que podem procurar neste trabalho a clarificação de qualquer conceito, entre os que são abordados no ensino básico e secundário. Sendo assim, a principal preocupação será a precisão e clareza dos conceitos, com um ou outro exemplo, mas sem a preocupação didática que um manual para alunos do ensino básico ou secundário deve conter. Por esta razão não

[3] Para uma clarificação deste conceito podemos consultar, por exemplo, Branco & Martins (2002).

haverá preocupação em colocar os conceitos por ordem de lecionação, mas antes pela ordem que se considere mais adequada para obtermos uma clara exposição dos mesmos.

No terceiro capítulo do projeto pretende-se fazer uma análise crítica aos materiais disponíveis para o ensino e compreensão dos conceitos mais elementares de Estatística. Serão analisados vários manuais escolares, nomeadamente os adotados na escola onde a autora leciona, e apresentadas e fundamentadas as incorreções, gralhas ou erros detetados nos temas OTD e Estatística. Analisaremos deste modo a adequação dos materiais de forma a serem identificadas lacunas nos materiais usualmente utilizados. Será igualmente importante e pertinente a comunicação dos erros encontrados, via *e-mail*, às entidades responsáveis. É de referir que este procedimento já teve início no que diz respeito ao manual adotado na escola da autora, para o 10.º ano, através da representante da editora. Salienta-se ainda que o objetivo desta análise aos manuais não se restringe à deteção de gralhas que, embora tenham de ser corrigidas, por vezes podem até não comprometer o ensino dos conteúdos previstos nos programas. Pretende-se igualmente identificar conteúdos cujas metodologias propostas nos manuais nos pareçam insuficientes para a compreensão dos conceitos ensinados, dos quais destacamos a regressão linear. É de referir um estudo realizado por Martinho & Viseu (2009) que consistiu na análise de dois manuais do 7.º ano quanto às dimensões: interpretação, crítica e produção. A primeira dimensão refere-se à capacidade de ler e compreender a informação (textos, tabelas, gráficos). A segunda dimensão abrange a capacidade de avaliar criticamente a informação estatística. Por fim, a dimensão designada de produção contempla a capacidade de argumentar, de comunicar a informação estatística e de tomar decisões. Estes autores concluíram que a dimensão mais presente nos dois manuais é a de interpretação. Concluiu-se ainda que num manual as dimensões crítica e de produção são quase inexistentes e no outro, apesar de mais expressivas, não promovem o desenvolvimento da atitude crítica no aluno. É de salientar a relevância subjacente à análise de manuais escolares uma vez que a maioria dos professores recorre, habitualmente, a eles quando prepara as suas aulas. Segundo o estudo "Matemática 2001" realizado pela Associação de Professores de Matemática (APM (1998)), no que diz respeito às práticas profissionais dos professores e no item "materiais usados na preparação das aulas", concluiu-se que 87% dos professores utiliza *sempre ou muitas vezes* o manual adotado na escola e 68% outros manuais.

Tendo em conta o trabalho desenvolvido no terceiro capítulo do projeto, nomeadamente as situações analisadas e que, de alguma forma, exigem uma correção, uma explicação, exemplos

mais enriquecedores, entre outros, pretende-se no capítulo quatro do projeto apresentar novos materiais e propor metodologias que possam ser utilizados no ensino da Estatística, de forma a facilitar a compreensão e exploração dos principais conceitos por parte dos alunos. Uma vez que uma das vertentes fundamentais para o sucesso do ensino da Estatística é termos professores preparados para o seu ensino, pretendemos que estes materiais vão também ao encontro das necessidades dos professores, como ilustrará o caso da regressão linear simples (consultar secção 3.4 na página 58), pois trata-se de um erro encontrado em vários manuais consultados. Sendo um erro tão generalizado só pode advir da falta de compreensão dos conceitos, pelo que iremos elaborar propostas para a sua clarificação. Acrescente-se ainda que dos materiais a elaborar podem constar: propostas de trabalho para realizar na sala de aula; a explicação de como construir materiais interativos de forma a possibilitar a construção de apresentações a utilizar pelos professores; metodologias direcionadas ao ensino da estatística, sendo que o *software* utilizado na exploração destes materiais será o *GeoGebra*.

De referir igualmente que a escolha deste projeto foi fortemente motivada pelo gosto da autora pela Estatística, o qual é bem patente na contínua participação em diversos trabalhos realizados dentro e fora da sala de aula com os seus alunos, nas suas orientações a alunos para a participação em concursos a nível nacional para estudantes (como ilustra o Prémio Pedro Matos organizado pelo IPL, o Prémio Estatístico Júnior organizado pela SPE e os Desafios do ALEA promovidos pelo site www.alea.pt), bem como o seu olhar, sempre atento e crítico, aos manuais adotados para a lecionação dos conteúdos programáticos da disciplina de Matemática, essencialmente do 7.º ano ao 12.º ano. Por outro lado, a autora considera um desafio explorar as potencialidades do *GeoGebra* na Estatística, uma vez que as experiências que tinha deste *software* eram no âmbito da Geometria e da Álgebra. Acrescente-se ainda que, a propósito do Plano da Matemática e das sessões em que participa regularmente, tem tido oportunidade de analisar o programa atual do ensino básico com mais pormenor, estando mais sensibilizada para as ideias aí preconizadas.

Este projeto pretende contribuir para melhorar o ensino da Estatística, uma vez que qualquer professor do ensino básico ou secundário poderá encontrar neste trabalho esclarecimentos relativos a conceitos que tem de lecionar, chamadas de atenção para erros comuns na área da Estatística e sugestões de materiais e metodologias para o ensino da Estatística. Também poderá contribuir para que os manuais venham a ser cada vez melhores, uma vez que as incorreções, gralhas ou erros detetados serão comunicados aos seus autores.

Capítulo 2

Conceitos de Estatística

A Estatística é uma ciência atual, com múltiplas funções e útil à humanidade. Desde sempre que o homem procura o conhecimento e, para tal, recolhe dados com determinadas intenções, nas mais diversas áreas do saber. Todos os dias a comunicação social faz-nos chegar notícias baseadas em estudos estatísticos, apresentando-nos as conclusões mais relevantes. Para se obter essas conclusões há um caminho a percorrer, mais ou menos longo, consoante o estudo realizado, mas bastante facilitado com o recurso à tecnologia, à Teoria das Probabilidades e nomeadamente à Inferência Estatística. Tendo em conta os programas do ensino básico e do ensino secundário (Matemática A) este projeto vai incidir sobre Estatística Descritiva cuja finalidade é descrever os dados recolhidos a partir de uma amostra ou população, resumindo a informação através de gráficos, tabelas e algumas medidas estatísticas, sem esquecer as comparações, por exemplo, entre dois conjuntos de dados. O desenvolvimento destes conteúdos acompanhado de alguns exemplos será objeto deste capítulo. Relembramos ainda que no programa do ensino básico a Estatística aparece desde o primeiro ciclo e designa-se por "Organização e Tratamento de Dados", desde 2007, data em que o Programa de Matemática do ensino básico foi homologado. No programa de Matemática A do ensino secundário o tema aparece no 10.º ano com a designação "Estatística" onde se pretende ampliar os conhecimentos adquiridos anteriormente. Neste capítulo os conceitos serão apresentados sem que se faça referência ao ano de escolaridade em que se lecionam, utilizando a ordem que nos parece mais adequada para a sua exposição.

Tendo em conta que os dados que pretendemos resumir e interpretar devem, preferencialmente, estar associados a um contexto, vamos começar por enumerar as etapas de um estudo estatístico. Deste modo, um estudo estatístico inclui as seguintes etapas:

1. definição do problema a estudar, formulando as questões às quais se pretende dar resposta;

2. planeamento da recolha de dados tendo em vista o estudo a realizar. É nesta fase que devemos decidir se recorremos à população ou à amostra e definir as variáveis com rigor;

3. organização e tratamento dos dados através de tabelas de frequência, gráficos e algumas medidas estatísticas;

4. interpretação dos resultados obtidos e estabelecimento de conclusões.

Exemplo 2.1. Exemplos de estudos estatísticos: hábitos alimentares dos alunos do 9.º ano; a crise económica na vida dos torrejanos; o peso e a altura dos alunos do 8.º ano da Escola Artur Gonçalves e a durabilidade (em quilómetros percorridos) dos pneus de uma determinada marca.

Quando se faz um estudo estatístico pode obter-se informação de todos os elementos (**indivíduos**) do universo (**população**) sobre o qual incide o estudo e, neste caso, faz-se um censo; ou recorre-se a uma parte representativa da população (**amostra**) e o estudo efetuado denomina-se sondagem.

Definição 2.1. Uma **população** é uma coleção de unidades individuais, que podem ser pessoas, animais, objetos, acontecimentos ou resultados experimentais com uma ou mais características comuns que se pretendem estudar. A cada elemento da população chama-se **indivíduo** ou **unidade estatística**. O número de elementos da população é representado por N (caso esta seja finita).

Definição 2.2. Uma **amostra** é um subconjunto representativo da população que se obtém através de métodos apropriados. A sua dimensão é representada por n.

Nas situações em que o estudo implica a destruição dos elementos a observar (por exemplo, quando pretendemos estudar a fiabilidade dos pneus ou a existência de bactérias nos iogurtes) recorre-se sempre a uma amostra. Também é aconselhável recorrer a uma amostra por razões económicas ou de tempo, pois observar todos os elementos da população pode implicar custos elevados ou a obtenção tardia dos resultados.

A determinação dos elementos que constituem a amostra, com vista à obtenção dos dados para a realização do estudo estatístico, também designado por processo de amostragem, deverá se objeto de especial cuidado. Sempre que pretendemos estender os resultados de um estudo estatístico a toda a população devemos observar o princípio da aleatoriedade. Quando, para todo o elemento da população existe uma probabilidade positiva de pertencer à amostra, dizemos que estamos perante uma amostra aleatória. Caso particular é o processo de amostragem simples onde cada grupo de dimensão n tem igual probabilidade de ser selecionado (com probabilidade igual a $\frac{1}{\binom{N}{n}}$, uma vez que existem $\binom{N}{n}$ amostras distintas com igual probabilidade de serem selecionadas). Neste caso prova-se que cada indivíduo tem a mesma probabilidade de ser selecionado, sendo esta probabilidade igual a $\frac{n}{N}$. Para mais informações consultar INE (2009, p. 43-71).

Quando existem elementos da população que podem não ser selecionados para a amostra estamos perante um processo de amostragem não aleatória. Neste caso dizemos que o processo de recolha da amostra é inviesado e poderá conduzir a interpretações erradas.

Exemplo 2.2. Exemplos de situações que originam **amostras enviesadas**: perguntar aos alunos do 9.º ano que almoçam diariamente no refeitório da escola os hábitos alimentares e generalizar a todos os alunos do 9.ºano; perguntar aos torrejanos que trabalham numa empresa os efeitos da crise económica e generalizar a todos os torrejanos; perguntar o clube preferido à porta do Estádio da Luz e generalizar a toda a população e efetuar um inquérito, num Hospital, sobre a saúde dos portugueses.

No primeiro caso ilustrado no exemplo 2.2 não obteríamos qualquer informação relativa aos alunos que almoçam no bar da escola, em casa ou nos arredores da escola. No segundo caso, pelo facto de as pessoas questionadas trabalharem numa empresa, não estavam incluídos, por exemplo, indivíduos desempregados. No terceiro caso, os indivíduos seriam todos ou quase todos do Benfica e no último caso as pessoas inquiridas estariam doentes (ou eram acompanhantes dos doentes).

As situações anteriores evidenciam fontes de enviesamento na recolha dos dados, pelo que esses estudos não conduziriam a resultados eficientes, nem permitiriam efetuar generalizações. No entanto há diversos estudos possíveis de realizar ao nível do ensino básico e secundário em que, recorrendo a amostras não enviesadas, se pode generalizar os resultados obtidos. Por exemplo, se selecionarmos aleatoriamente 5 alunos de cada turma de uma escola (supondo que cada turma tem um número de alunos aproximado) e estudarmos o número de

irmãos, o número do sapato ou a altura, faz sentido generalizar para toda a escola. Neste caso estamos a considerar que a população, constituída pelos alunos da escola, está dividida em várias subpopulações (designadas de estratos) mais ou menos homogéneas e em cada uma destas subpopulações recolhe-se uma amostra aleatória simples (amostra aleatória estratificada). É de salientar que os alunos mais novos, nomeadamente do 1.º ciclo, poderão usar a sua turma como população em estudo de modo a facilitar os seus projetos nesta área.

Em qualquer estudo estatístico, é necessário identificar e classificar as características em análise, de acordo com os objetivos traçados.

2.1 Variáveis

Tendo em conta a amostra sobre a qual recai o estudo estatístico e os objetivos fixados, definem-se as caraterísticas a analisar (**variáveis estatísticas**), as quais devem ser comuns a todos os elementos da população.

Definição 2.3. Variável estatística (ou atributo) é a propriedade ou característica comum que se observa em cada uma das unidades estatísticas. Representa-se, habitualmente, por uma das últimas letras do alfabeto, por exemplo, x ou y.

As variáveis estatísticas podem classificar-se em quantitativas ou qualitativas. A variável **quantitativa** é aquela que se refere a uma característica mensurável, isto é, que se pode contar ou medir. Por conseguinte, traduz-se por valores numéricos.

Exemplo 2.3. Exemplos de variáveis quantitativas: número de divisões de uma habitação; idade de um indivíduo; número de alunos por turma e tempo necessário para chegar de casa à escola.

Uma variável quantitativa pode ser discreta ou contínua. Quando a característica em estudo se pode apenas contar e não medir, a variável é discreta, como por exemplo o número de alunos por turma. Por outro lado, uma variável quantitativa que se pode medir é uma variável contínua, como por exemplo o tempo necessário para chegar de casa à escola.

A variável **qualitativa** é aquela que se refere a uma característica que não é susceptível de medição ou contagem e, como tal, traduz-se por diferentes modalidades (possíveis respostas que a variável pode assumir).

Exemplo 2.4. Exemplos de variáveis qualitativas: ano de escolaridade dos alunos; profissão dos pais; meio de transporte utilizado para chegar de casa à escola e cor preferida dos estudantes de uma turma.

As variáveis qualitativas podem ser classificadas em nominais ou ordinais. Um variável estatística é nominal quando não se pode estabelecer uma relação de ordem entre as modalidades e é ordinal no caso em que as modalidades apresentam uma ordem subjacente. Como exemplo de uma variável qualitativa ordinal podemos considerar as habilitações literárias de um indivíduo. Um exemplo de uma variável qualitativa nominal pode ser a cor preferida dos estudantes.

2.2 Dados estatísticos

Sempre que se observa uma variável estatística, quantitativa ou qualitativa, obtemos determinados resultados que designamos por dados estatísticos.

Definição 2.4. Dado estatístico é o resultado de cada observação da variável numa unidade estatística.

Se a variável em estudo for, por exemplo, o número de alunos por turma, os dados estatísticos podem ser:

$$22, 28, 24, 24, 28, 20, 28, \cdots,$$

mas no caso da variável ser o ano de escolaridade dos alunos, alguns dados estatísticos podem ser:

$$5.^{o} \text{ ano; } 6.^{o} \text{ ano; } 6.^{o} \text{ ano; } 8.^{o} \text{ ano; } 7.^{o} \text{ ano; } 12.^{o} \text{ ano, } \cdots$$

Os dados estatísticos são muito mais do que números ou modalidades. Eles estão sempre associados a um contexto. Na primeira situação apresentada, em que a variável é o número de alunos por turma, cada unidade estatística é uma turma. Para cada turma observou-se o número de alunos, obtendo-se, assim, os dados estatísticos. Na segunda situação, em que a variável é o ano de escolaridade, cada unidade estatística é um aluno. Para cada aluno registou-se o seu ano de escolaridade. Deste modo, obtiveram-se os dados estatísticos que neste caso são modalidades, pois a variável é qualitativa. Aproveitamos para reforçar a importância de,

num estudo estatístico, saber quais os dados que queremos obter de modo a dar resposta às questões levantadas, apresentando conclusões pertinentes.

Neste capítulo vamos recorrer, como exemplo, ao resultado de um inquérito efetuado a 25 alunos de uma escola, selecionados de entre as três turmas do 8.º ano, relativamente à sua cor preferida; seus hábitos de leitura; idade (em anos); altura (em centímetros - cm); notas do teste diagnóstico de Matemática (Mat.) e do teste diagnóstico a Ciências Físico-Químicas (C.F.Q.) numa escala de 0 a 100 valores. Os dados obtidos estão apresentados na Tabela 2.1.

Notemos que as variáveis cor preferida e hábitos de leitura são variáveis qualitativas, as variáveis nota a Matemática, nota a Ciências Físico-Químicas e idade[(1)] são variáveis quantitativas discretas e a variável altura é uma variável quantitativa contínua.

Relativamente à notação que utilizaremos ao longo deste trabalho, as observações da amostra serão representadas por

$$x_1,\ x_2,\ \ldots,\ x_n,$$

onde x_i representa a resposta do indivíduo i relativamente à variável x. Para o tratamento estatístico é usual agrupar os indivíduos cujas respostas são iguais, sendo as diferentes respostas presentes na amostra representadas por

$$x'_1,\ x'_2,\ \ldots,\ x'_p$$

onde p representa o número de respostas distintas e, naturalmente, $p \leq n$. Desta forma podemos construir as tabelas de frequências.

2.3 Tabelas de frequências

Após a recolha de dados, outra fase muito importante é a sua organização em tabelas de frequências. Com este objetivo devemos atender às seguintes definições.

Definição 2.5. A **frequência absoluta** é o número de vezes que cada valor da variável (ou cada modalidade) aparece num conjunto de dados. Representa-se por n_i que corresponde ao número de vezes que se observou x'_i.

(1) Apesar dos dados resultarem de uma medição, a forma como são apresentados (número inteiro de anos) têm a aparência de dados discretos. Contudo, por exemplo, o valor 14, refere-se a todas as idades maiores ou iguais a 14 e menores que 15. Por esta razão podemos também considerar a variável idade como uma variável contínua que foi discretizada.

Aluno	Cor preferida	Hábitos de leitura	Idade (anos)	Altura (cm)	Nota a Mat.	Nota a C.F.Q.
1	azul	leio todos os dias	15	150	42	60
2	branco	não costumo ler	14	159	37	43
3	azul	leio todas as semanas	13	146	80	81
4	branco	leio todos os dias	14	157	78	80
5	azul	leio todas as semanas	14	163	79	85
6	amarelo	não costumo ler	13	158	63	60
7	azul	leio todos os dias	15	160	63	64
8	azul	leio todas as semanas	15	165	45	53
9	amarelo	leio todas as semanas	14	154	50	55
10	verde	só leio nas férias	13	149	32	35
11	branco	só leio nas férias	14	153	60	70
12	branco	leio todas as semanas	14	166	38	39
13	cor-de-rosa	leio todas as semanas	13	153	45	47
14	cor-de-rosa	só leio nas férias	13	152	60	62
15	azul	leio todos os dias	14	159	71	70
16	amarelo	leio todos os dias	14	155	25	27
17	amarelo	só leio nas férias	13	156	60	64
18	cor-de-rosa	leio todas as semanas	14	152	64	65
19	cor-de-rosa	leio todas as semanas	14	163	65	60
20	cor-de-rosa	só leio nas férias	14	157	37	38
21	verde	leio todos os dias	15	164	64	65
22	azul	leio todas as semanas	15	169	87	90
23	cor de rosa	leio todas as semanas	14	157	87	88
24	verde	não costumo ler	16	164	48	54
25	cor-de-rosa	só leio nas férias	14	155	50	55

Tabela 2.1: Dados obtidos através de um questionário a 25 alunos do 8.º ano

A soma das frequências absolutas é igual à dimensão da amostra, isto é,

$$\sum_{i=1}^{p} n_i = n. \tag{2.1}$$

Definição 2.6. A **frequência relativa** de cada modalidade x_i' é a proporção de observações iguais a x_i', isto é, o quociente que se obtém dividindo a frequência absoluta de um valor (ou modalidade) pelo número total de dados. Representa-se por f_i sendo

$$f_i = \frac{n_i}{n}, \tag{2.2}$$

onde n é o número de elementos da amostra.

Uma propriedade importante das frequências relativas é a soma das frequências relativas ser igual a 1, isto é,

$$\sum_{i=1}^{p} f_i = 1. \tag{2.3}$$

Notemos que quando pretendemos efetuar comparações devemos usar a frequência relativa, pois muitas vezes as amostras têm dimensões diferentes e, nestes casos, não faz sentido usar a frequência absoluta. Por exemplo, quando se pretende comparar o número de aprovações de duas turmas com dimensões distintas não se deve comparar as frequências absolutas mas antes as frequências relativas.

Definição 2.7. A **frequência absoluta acumulada** de cada valor x_i' da variável é o número total de dados com valor menor ou igual a x_i'. Representa-se por N_i.

A frequência absoluta acumulada obtém-se adicionando as frequências absolutas desde o primeiro até ao último valor considerado da variável, isto é,

$$N_i = \sum_{j=1}^{i} n_j, \tag{2.4}$$

onde, naturalmente, $N_p = n$.

Definição 2.8. A **frequência relativa acumulada** de cada valor x_i' da variável é a soma das frequências relativas de todos os dados com valor menor ou igual a x_i'. Representa-se por F_i.

A frequência relativa acumulada obtém-se adicionando as frequências relativas desde o primeiro até ao último valor considerado da variável, isto é,

$$F_i = \sum_{j=1}^{i} f_j, \tag{2.5}$$

ou utilizando as frequências absolutas acumuladas

$$F_i = \frac{N_i}{n}, \tag{2.6}$$

onde $F_p = 1$.

Uma das formas de organizarmos os dados estatísticos é através da construção de tabelas de frequências, pois trazem-nos fortes vantagens na leitura dos dados. Agora que já definimos os diferentes tipos de frequência, apresentamos na Tabela 2.2 uma tabela de frequências geral. Na primeira coluna são apresentados os diferentes valores ou modalidades, da variável estatística, presentes na amostra e nas colunas seguintes as correspondentes frequências absolutas, relativas e acumuladas. Na última linha da tabela é apresentada a soma da respetiva coluna, sempre que tal tenha significado.

Variável x	Frequência absoluta n_i	Frequência absoluta acumulada N_i	Frequência relativa f_i	Frequência relativa acumulada F_i
x'_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$
x'_2	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x'_i	n_i	$N_i = n_1 + \cdots + n_i$	$F_i = f_i$	$f_1 + \cdots + f_i$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x'_p	n_p	$N_p = n_1 + \cdots + n_p = n$	f_n	$F_p = f_1 + \cdots + f_p = 1$
Total	n		1	

Tabela 2.2: Tabela de frequências

É comum designar a tabela de frequências de acordo com as frequências que a compõem. Por exemplo, uma tabela de frequências absolutas simples apresenta apenas estas frequências para cada um dos valores ou modalidades.

2.3.1 Tabela de frequências de uma variável qualitativa

Quando a variável em estudo é qualitativa os dados estatísticos apresentam-se na forma de modalidades. Como tal, deve proceder-se à contagem das diferentes modalidades e organizar os dados numa tabela de frequências absolutas e relativas simples. Vamos ilustrar esta situação com as cores preferidas dos alunos, dadas na Tabela 2.1 presente na página 13. Os dados apresentam 5 modalidades diferentes ($p = 5$) e encontram-se organizados na Tabela 2.3.

Cor preferida dos alunos	Frequência absoluta n_i	Frequência relativa f_i
branco	4	4:25 = 0,16 (16%)
amarelo	4	4:25 = 0,16 (16%)
cor-de-rosa	7	7:25 = 0,28 (28%)
azul	7	7:25 = 0,28 (28%)
verde	3	3:25 = 0,12 (12%)
Total	25	1

Tabela 2.3: Tabela de frequências da variável cor preferida dos alunos

Refira-se que, caso estivéssemos perante uma variável qualitativa ordinal, por exemplo, habilitações literárias, faria sentido o cálculo das frequências acumuladas, uma vez que neste caso a ordem das modalidades tem significado.

2.3.2 Tabela de frequências de uma variável quantitativa discreta

Tratando-se de uma variável quantitativa discreta deve proceder-se à contagem dos diferentes valores e à organização dos dados numa tabela de frequências absolutas ou relativas, simples ou acumuladas. Neste tipo de variável os dados estatísticos apresentam-se na forma de valores sendo, sempre possível, a sua ordenação.

Para ilustrar esta situação consideremos agora as idades dos 25 alunos. Verificamos que existem quatro valores diferentes da variável (13, 14, 15 e 16), logo $p = 4$. Os dados podem organizar-se numa tabela de frequências conforme a apresentada como Tabela 2.4.

Função cumulativa para dados discretos

Associada a cada uma das frequências acumuladas podemos definir a **função cumulativa**. No caso das frequências absolutas, associa a cada valor de x o número total de dados observados com valor menor ou igual a x, isto é,

$$N(x) = \sum_{x_i' \leq x} n_i. \tag{2.7}$$

Idade dos alunos	Frequência absoluta n_i	Frequência absoluta acumulada N_i	Frequência relativa f_i	Frequência relativa acumulada F_i
13	6	6	6:25 = 0,24 (24%)	0,24 (24%)
14	13	19	13:25 = 0,52 (52%)	0,76 (76%)
15	5	24	5:25 = 0,2 (20%)	0,96 (96%)
16	1	25	1:25 = 0,04 (4%)	1 (100%)
Total	25		1	

Tabela 2.4: Tabela de frequências da variável idade dos alunos

No caso das frequências relativas, faz corresponder a cada valor de x a frequência relativa do total de dados observados com valor menor ou igual a x, isto é,

$$F(x) = \sum_{x_i' \leq x} f_i. \tag{2.8}$$

Como exemplo vamos definir, analiticamente, as funções cumulativas das frequências absolutas e das frequências relativas para a variável idade dos alunos,

$$N(x) = \begin{cases} 0 & \text{se } x < 13 \\ 6 & \text{se } 13 \leq x < 14 \\ 19 & \text{se } 14 \leq x < 15 \\ 24 & \text{se } 15 \leq x < 16 \\ 25 & \text{se } x \geq 16 \end{cases}, \tag{2.9}$$

$$F(x) = \begin{cases} 0 & \text{se } x < 13 \\ 0,24 & \text{se } 13 \leq x < 14 \\ 0,76 & \text{se } 14 \leq x < 15 \\ 0,96 & \text{se } 15 \leq x < 16 \\ 1 & \text{se } x \geq 16 \end{cases}. \tag{2.10}$$

A representação gráfica da função (2.9) pode ver-se na secção 2.4.6 (página 24).

2.3.3 Tabela de frequências de uma variável quantitativa contínua

Quando a variável em estudo é quantitativa contínua[2], os dados estatísticos podem tomar qualquer valor de um certo intervalo, surgindo poucas repetições. Por este motivo, não faz sentido atribuir uma frequência a cada valor diferente da variável, pois a tabela assim obtida não permitiria obter conclusões importantes pelo facto de não conduzir a regularidades, como ilustra o exemplo apresentado na Tabela 2.5. Neste exemplo consideramos a altura dos alunos e tratamos a variável altura como quantitativa discreta.

Altura dos alunos	Frequência absoluta n_i
146	1
149	1
150	1
152	1
153	2
154	1
155	2
156	1
157	3
158	1
159	1
160	2
162	1
163	2
164	1
165	2
166	1
169	1

Tabela 2.5: Tabela de frequências da altura dos alunos (como uma variável discreta)

Uma vez que as regularidades são impercetíveis nesta tabela, vamos agrupar os dados,

[2] Este procedimento também deve ser aplicado em variáveis quantitativas discretas que assumam muitos valores distintos

relativos às alturas, em classes procedendo do seguinte modo:

- calcular a diferença entre a altura máxima e a altura mínima (amplitude da amostra);
- determinar o número de classes k a construir, utilizando a regra de Sturges, (onde k é o menor número inteiro tal que $2^k \geq n$);
- determinar a amplitude de cada classe que será aproximadamente (por excesso) o quociente que se obtém dividindo a amplitude da amostra pelo número de classes.

Vamos assim obter os intervalos, habitualmente denominados por intervalos de classe,

$$[l_0,\, l_1[,\ \ [l_1,\, l_2[,\ \ \cdots,\ \ [l_{k-2},\, l_{k-1}[,\ \ [l_{k-1},\, l_k],$$

onde $l_0 < l_1 < l_2 < \cdots < l_{k-1} < l_k$. Os intervalos são disjuntos dois a dois e a sua união contém todos os valores. Notemos que l_0 é menor ou igual que o valor mínimo observado e l_k é maior ou igual que o valor máximo observado (de forma a garantir que o intervalo $[l_0,\, l_k]$ contenha todas as observações).

Tendo em conta os passos anteriores, relativamente à altura dos 25 alunos, verifica-se que a altura máxima é 169, a altura mínima é 146 e a amplitude é $169 - 146 = 23$. Como $2^5 \geq 25$ (e $2^4 < 25$), iremos considerar $k = 5$ classes. Dado que $\frac{23}{5} = 4,6$, vamos construir 5 classes todas de amplitude 5. Assim obtém-se a tabela de frequências apresentada na Tabela 2.6.

Altura dos alunos	Frequência absoluta n_i	Frequência absoluta acumulada N_i	Frequência relativa f_i	Frequência relativa acumulada F_i
[145,150[	2	2	2:25 = 0,08 (8%)	0,08 (8%)
[150,155[	5	7	5:25 = 0,20 (20%)	0,28 (28%)
[155,160[	8	15	8:25 = 0,32 (32%)	0,60 (60%)
[160,165[	6	21	6:25 = 0,24 (24%)	0,84 (84%)
[165,170]	4	25	4:25 = 0,16 (16%)	1 (100%)

Tabela 2.6: Tabela de frequências da variável altura

Pode aplicar-se este processo (regra de Sturges) de modo idêntico sempre que a variável seja quantitativa contínua. Saliente-se, contudo, que não é método único e que nem sempre

conduz a resultados aceitáveis. A título de exemplo, usando este método, se analisarmos o rendimento de 100 indivíduos e um deles tiver um rendimento muito superior aos restantes podemos obter 7 classes de igual amplitude, onde na primeira estão 99 indivíduos e na última apenas um, como se ilustra no exemplo 2.5.

Exemplo 2.5. De um grupo de 100 trabalhadores de um empresa, 99 auferem entre 475 euros (rendimento mínimo) e 1100 euros e um indivíduo aufere 5000 euros (rendimento máximo). Tendo em conta os passos descritos previamente, a amplitude é igual a $5000-475=4525$ e o número de classes a considerar é igual a 7 ($k=7$) uma vez que $2^7 \geq 100$ (e $2^6 < 100$). Dado que $\frac{4525}{7} \approx 646,4$, segundo a regra de Sturges vamos construir 7 classes todas de amplitude 647, de onde se obtém a tabela de frequência apresentada como Tabela 2.7. Naturalmente, num caso como este é pertinente o uso de classes com diferentes amplitudes.

Rendimento em euros	Frequência absoluta n_i
[475,1122[	99
[1122,1769[	0
[1769,2416[	0
[2416,3063[	0
[3063,3710[	0
[3710,4357[	0
[4357,5004[	1
Total	100

Tabela 2.7: Rendimentos agrupados em classes de igual amplitude

2.4 Representações Gráficas

Para além de se poder organizar os dados estatísticos em tabelas de frequências, outra forma de os apresentar é recorrer a vários tipos de representações gráficas, tais como pictogramas, gráficos de pontos, diagramas de barras, gráficos circulares, diagramas de caule e folhas e histogramas.

Um gráfico é um instrumento de síntese apelativo e que nos dá uma ideia geral da questão abordada, sem no entanto deixar de destacar alguns aspetos particulares. Quando pretendemos representar conjuntos de dados graficamente deveremos selecionar o gráfico mais adequado a cada situação. Por outro lado, deve ter-se em conta que a informação que nele existe é suficiente para que qualquer pessoa o compreenda.

2.4.1 Pictograma

O pictograma é um tipo de representação gráfica muito sugestivo, pois na sua construção são utilizadas figuras representativas da informação. Apesar de se poder usar o mesmo símbolo, variando a área ou volume de forma a que sejam proporcionais à frequência absoluta, torna-se mais simples usar uma figura que se repete sempre da mesma maneira. As figuras devem estar igualmente espaçadas e devem apresentar-se em linhas ou colunas. Apesar de nos dar uma ideia geral da situação uma das desvantagens é, por vezes, não ser possível uma leitura rigorosa das frequências absolutas de cada valor ou modalidade. Um pictograma tem de incluir o significado do símbolo que pode ser a unidade ou não. Pode visualizar-se um exemplo de pictograma na Figura 2.1, respeitante aos hábitos de leitura dos alunos (cujos dados estão apresentados na Tabela 2.1, na página 13).

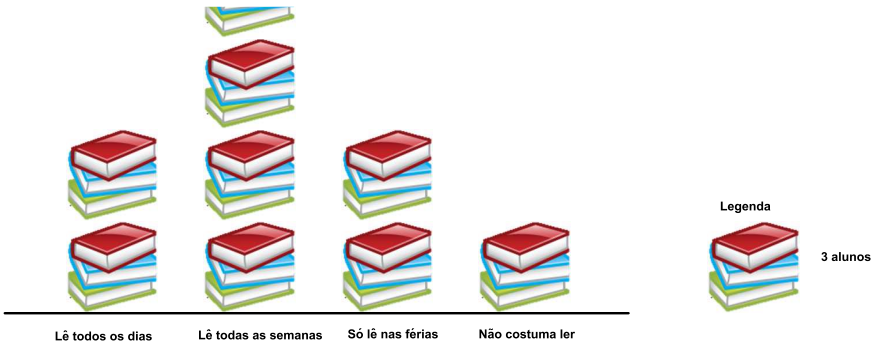


Figura 2.1: Pictograma dos hábitos de leitura dos alunos

2.4.2 Gráfico de pontos

Um gráfico de pontos é uma representação gráfica muito simples e que pode utilizar-se para variáveis qualitativas ou para variáveis quantitativas. Para a sua elaboração começa-se por desenhar um eixo horizontal onde se marcam os valores ou modalidades que a variável assume

em cada conjunto de dados. Por cima de cada valor ou modalidade marca-se um ponto sempre que um elemento da amostra for igual a esse valor ou a essa modalidade.

Na Figura 2.2 utilizamos a idade dos alunos para ilustrar um gráfico de pontos.

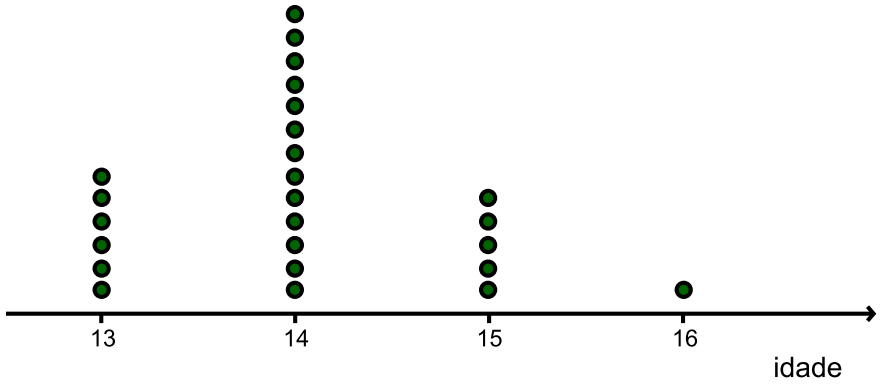


Figura 2.2: Gráfico de pontos das idades dos alunos

2.4.3 Gráfico de barras

Um gráfico de barras é um gráfico muito comum para representar informação, pelo facto de ser fácil de elaborar e interpretar. Pode ser usado quando a variável é qualitativa ou quantitativa discreta. Serve para representar um conjunto de dados ou para comparar conjuntos de dados relativamente a uma variável. Tal como no gráfico de pontos, no eixo horizontal indicam-se as modalidades ou os valores da variável. Para além deste eixo é necessário um eixo vertical onde se marcam as frequências absolutas ou relativas. Não deve haver quebra de escala no eixo vertical pois os gráficos tornam-se enganadores, mostrando aparentemente grandes variações quando na verdade não existem (ou ao contrário). As barras devem ter a mesma largura, devem estar igualmente espaçadas e a sua altura deve ser proporcional às frequências. Na maioria das vezes a altura de cada barra coincide com a frequência.

Na Figura 2.3 utilizamos a idade dos alunos para ilustrar um gráfico de barras.

2.4.4 Gráfico circular

Um gráfico circular é constituído por um círculo no qual se definem setores de área diretamente proporcional à frequência que representam. Cada um dos setores corresponde a uma

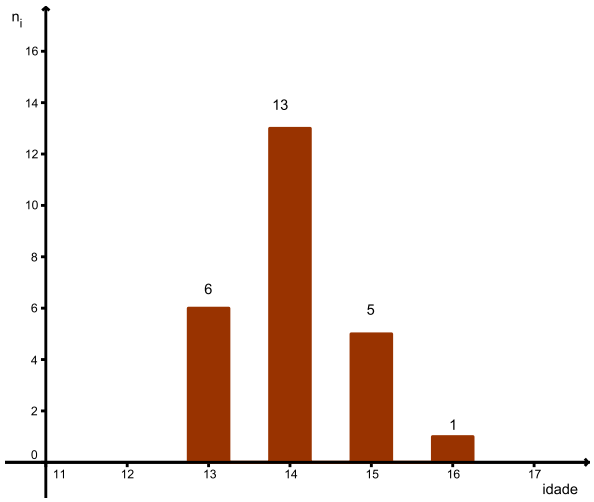


Figura 2.3: Gráfico de barras das idades dos alunos

modalidade ou a um valor da variável. Desta forma, deve ser utilizado quando a variável apresenta um número reduzido de valores ou modalidades e sempre que essas frequências não sejam próximas de 0.

Alguns programas constroem estes gráficos a partir da tabela de frequências, no entanto se utizarmos o *GeoGebra* ou se os construirmos usando papel e lápis, é necessário determinar a amplitude de cada setor. Para isso é habitual usar uma regra prática que consiste em multiplicar a frequência relativa por 360° (graus). Outra alternativa é utilizar "regras de três simples" considerando que o todo (total de elementos da amostra) corresponde a 360°. Para facilitar a leitura e interpretação dos gráficos devemos incluir as percentagens correspondentes a cada setor (como se mostra na Figura 2.4) e sempre que necessário uma legenda. Este gráfico mostra-nos as cores preferidas dos alunos, cujos dados se encontram na Tabela 2.1.

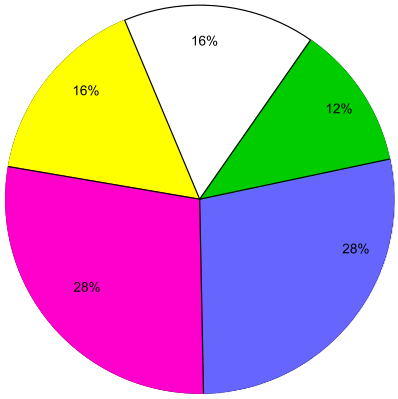


Figura 2.4: Gráfico circular das cores preferidas de 25 alunos

2.4.5 Histograma e polígono de frequências

O histograma é uma representação gráfica que se utiliza quando a variável em estudo é contínua. É composto por uma sucessão de retângulos adjacentes que têm por base um intervalo de classe e área igual à frequência[3] (absoluta ou relativa). Desta forma a área total do histograma é n (dimensão da amostra) ou 1 (soma das frequências relativas). Na primeira situação, para determinar a altura de cada retângulo, deve usar-se $\frac{n_i}{h_i}$; na segunda situação, deve usar-se $\frac{f_i}{h_i}$, onde h_i representa a amplitude da classe i, isto é, $h_i = l_i - l_{i-1}, i = 1, \cdots, k$.

É de referir que nos casos em que os intervalos de classe têm a mesma amplitude é habitual considerar as alturas dos retângulos iguais (ou proporcionais) às frequências absolutas ou relativas, como por exemplo, no gráfico presente na Figura 2.5.

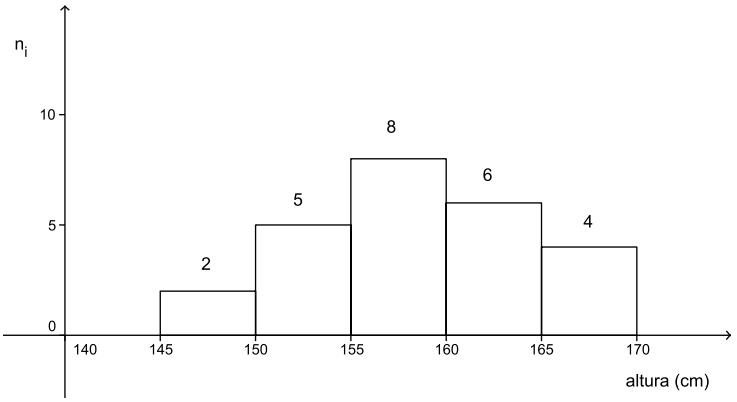


Figura 2.5: Histograma da altura (em cm) de 25 alunos

O polígono de frequências é um gráfico de linhas associado ao histograma e que se obtém unindo os pontos médios da base superior de cada retângulo. Para que o polígono comece e termine no eixo horizontal, imagina-se uma classe à esquerda da primeira (com a mesma amplitude da primeira) e outra à direita da última (com a mesma amplitude da última), ambas com frequência igual a zero, como se ilustra no gráfico presente na Figura 2.6. Notemos que, deste modo, a área do histograma será igual à área entre o polígono e o eixo Ox.

2.4.6 Representação gráfica da função cumulativa

Tratando-se de dados discretos, o gráfico da função cumulativa, quer das frequências relativas acumuladas quer das frequências absolutas acumuladas é em escada. A título de exemplo, na

[3] Não é obrigatório ser igual, pode ser proporcional. Contudo estes são os casos mais utilizados.

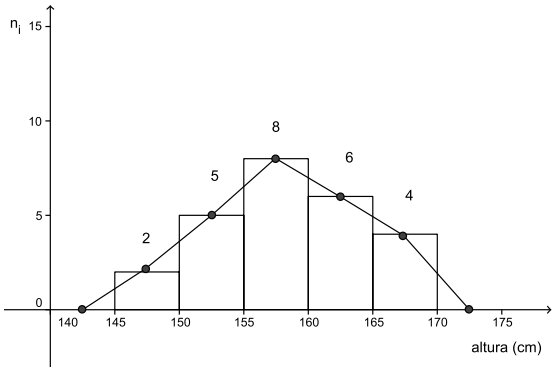


Figura 2.6: Polígono de frequências referente à altura dos alunos

Figura 2.7 apresentamos o gráfico da função cumulativa presente na página 17, referente à idade dos alunos.

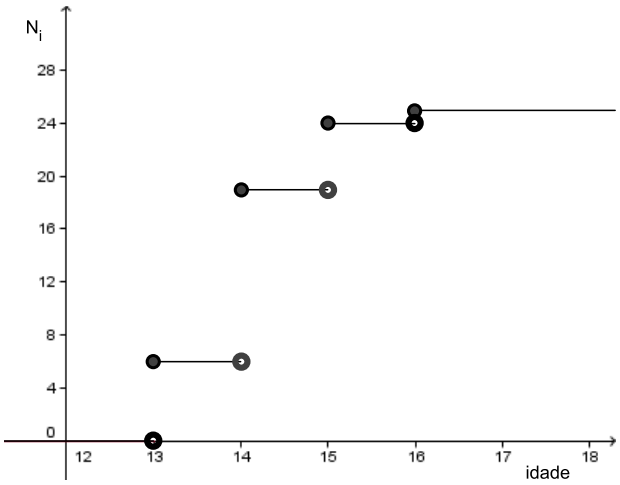


Figura 2.7: Função cumulativa relativa à idade dos alunos

No caso da variável ser contínua a representação gráfica da função cumulativa é uma linha poligonal que pode ser obtida a partir do histograma de frequências acumuladas. Considerando as alturas dos 25 alunos, vamos representar a função cumulativa a partir do histograma de frequências relativas acumuladas. Assim, como podemos ver na Figura 2.8, antes da frequência da 1.ª classe, a frequência acumulada é nula, pelo que se traça um segmento sobre o eixo Ox até ao limite inferior da 1ª classe (ficando sobreposto ao eixo Ox). A partir daqui, e admitindo que as observações se distribuem uniformemente em cada uma das classes, unimos os pontos de coordenadas $(l_i, F_i), i = 0,1,2,3,4,5$, onde l_i é o limite inferior da classe $[l_i, l_{i+1}[$ e F_i a frequência relativa acumulada da classe anterior com $F_0 = 0$. Quando chegamos à última classe temos a garantia que a frequência acumulada correspondente ao seu limite superior é

igual a 1, razão pela qual se desenha um segmento de reta paralelo ao eixo *Ox*, a partir desse ponto.

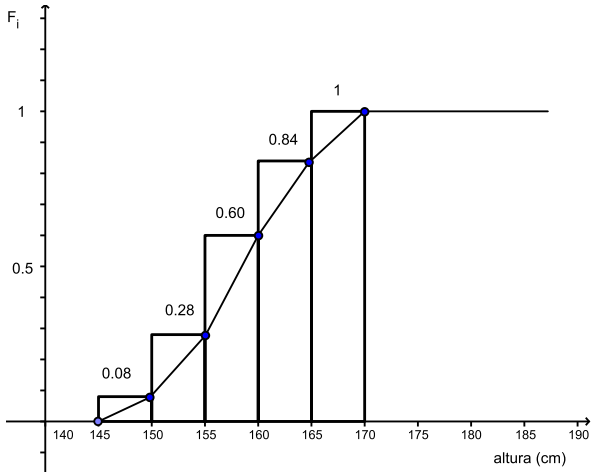


Figura 2.8: Função cumulativa relativa à altura dos alunos

2.4.7 Diagrama de caule e folhas

O diagrama de caule e folhas é uma forma de representar os dados de fácil construção. Pode situar-se entre a tabela e o gráfico, sendo uma das vantagens o facto de todos os dados observados estarem presentes e ordenados.

Para construir um diagrama de caule e folhas desenha-se uma linha vertical e coloca-se do lado esquerdo o dígito ou dígitos de maior grandeza e do lado direito os restantes dígitos.

Por exemplo, no caso da altura dos alunos, 146, 149, 150,$\cdots$, registadas na Tabela 2.1, na página 13, colocamos os dígitos das centenas e das dezenas à esquerda e os algarismos das unidades à direita, ficando o início do diagrama com este aspeto 14$|$ 6 9 (conforme Figura 2.9). Aos valores colocados à esquerda do traço vertical chamamos caule e aos valores colocados à direita denominamos por folhas.

Notemos que o diagrama de caule e folhas tem uma representação gráfica semelhante à do histograma, se fizermos uma rotação de 90°, no sentido contrário ao dos ponteiros do relógio. Deste modo, corresponderia a um histograma com três classes, respetivamente $[140,150[$, $[150,160[$ e $[160,170[$. Uma das vantagens do histograma relativamente ao diagrama de caule e folhas é o facto de haver menos restrições na construção das classes do que na construção dos caules, no entanto ao construir as classes para a elaboração do histograma perde-se informação

existente na amostra.

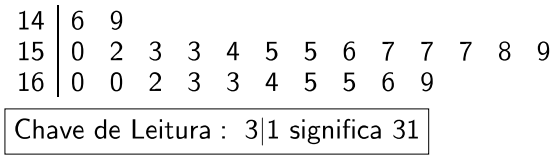


Figura 2.9: Diagrama de caule e folhas relativo à altura dos alunos

2.5 Medidas de tendência central

Nas duas secções anteriores dedicámo-nos à construção de tabelas de frequências e à representação gráfica dos dados. As tabelas de frequências permitem organizar e facilitam a leitura dos dados. Os gráficos, apesar de geralmente conterem menos informação que as tabelas de frequências, são de fácil leitura e têm um peso significativo na análise descritiva dos dados, não esquecendo que "uma imagem vale mais do que mil palavras".

A partir daqui e até ao final deste capítulo, vamos estudar um conjunto de medidas descritivas que permitem sumariar alguns dos aspetos mais importantes do conjunto de dados. Nesta secção, em particular, vamos estudar as medidas de tendência central: moda, média e mediana.

2.5.1 Moda

A moda é uma medida com especial interesse para resumir os dados no caso da variável ser qualitativa. Nesta situação nem a média nem a mediana podem ser calculadas ou são desprovidas de significado.

Definição 2.9. A **moda**, representada por Mo, é o valor ou modalidade que surge com maior frequência na amostra.

Se considerarmos o conjunto de dados constituído pelas 25 cores preferidas dos alunos verificamos que há duas modalidades com a frequência absoluta mais elevada (consultar Tabela 2.3 da página 16). Neste caso dizemos que há duas modas: cor-de-rosa e azul, e a distribuição

é bimodal. Se tivesse três modas seria trimodal. No caso de ter mais do que três modas é multimodal. Quando nenhum valor ou modalidade aparece com maior frequência do que os restantes diz-se que não há moda.

Quando a variável é quantitativa contínua e não se conhecem os dados reais podemos identificar a classe modal (classe com maior frequência por unidade de amplitude). Por exemplo, relativamente à altura dos alunos, representada na Tabela 2.6 da página 19 a classe modal é $[155,160[$, porque corresponde à classe com maior frequência absoluta por unidade de amplitude. Notemos que, quando as classes têm igual amplitude, bastará identificar a classe com maior frequência. Caso contrário devemos utilizar a classe com maior $\frac{n_i}{h_i}$, conforme ilustra o exemplo 2.6.

Exemplo 2.6. Perguntou-se a 40 indivíduos o número de dias que faltaram ao trabalho, por doença, no ano anterior. Os dados encontram-se organizados na Tabela 2.8. Como podemos verificar as classes não têm todas a mesma amplitude. Neste caso indicar a classe modal implica encontrar a classe em que o quociente entre a frequência absoluta e a respetiva amplitude, é maior. Por observação da tabela concluímos que a classe modal é a classe [4,8[(e não a classe [8,16[que corresponde à de maior frequência).

N.º de faltas por doença	Frequência absoluta	$\frac{n_i}{h_i}$
[0,4[	4	$\frac{4}{4}=1$
[4,8[	10	$\frac{10}{4}=2,5$
[8,16[	12	$\frac{12}{8}=1,5$
[16,24[	9	$\frac{9}{8}=1,125$
[24,30]	5	$\frac{5}{6}=0,8(3)$
total	40	

Tabela 2.8: Número de faltas por doença

2.5.2 Média

De entre as medidas de localização estudadas, a média[4] é a mais usada.

[4] Geralmente quando falamos em média estamos a referir-nos à média aritmética, como a definimos neste trabalho. No entanto há outros tipos de médias, como por exemplo a média geométrica, a média harmónica

Definição 2.10. A **média**, representada por $\overline{x}$, obtém-se dividindo a soma de todos os valores de uma variável pelo número total de observações,

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.11}$$

onde $x_1, x_2, \cdots, x_n$ são os n valores da variável quantitativa.

Em muitos casos os dados encontram-se organizados numa tabela como, por exemplo, na Tabela 2.4 na página 17. Nestes casos, para calcular a média utilizamos a frequência absoluta de cada valor distinto da variável, do seguinte modo:

$$\overline{x} = \frac{x'_1 n_1 + x'_2 n_2 + \cdots + x'_p n_p}{n} = \frac{1}{n}\sum_{i=1}^{p} x'_i n_i, \tag{2.12}$$

onde p representa o número de valores diferentes da variável.

Aplicando este raciocínio à idade dos alunos (Tabela 2.4 na página 17) temos:

$$\overline{x} = \frac{13\times 6 + 14\times 13 + 15\times 5 + 16\times 1}{25} = \frac{351}{25} = 14,04 \text{ anos}. \tag{2.13}$$

Por vezes, quando a variável em estudo é contínua, não conhecemos os seus valores reais. Nestes casos partimos dos dados agrupados em classes e determinamos um valor aproximado da média. Para ilustrar esta situação, vamos supor que o peso (em kg) dos 25 alunos estão organizados em 5 classes, conforme Tabela 2.9. Determina-se a marca da classe, isto é, o ponto médio de cada classe,

$$x'_i = \frac{l_{i-1} + l_i}{2}, \quad i = 1, 2, \cdots, 5 \tag{2.14}$$

e considera-se, para cada classe, que o peso de cada aluno é igual à marca da classe[5]. Depois

ou a média (aritmética) aparada. A média geométrica é muito usada, por exemplo, em economia no cálculo de taxas de variação ou de crescimento (se colocarmos no banco um montante com uma taxa de juro igual a 2% no primeiro ano e 3% no segundo ano, a taxa média de crescimento não será exactamente igual a 2,5%). A média harmónica utiliza-se quando estamos perante grandezas inversamente proporcionais, como é o caso da velocidade e do tempo (notemos que se fizermos um trajeto duas vezes, uma a 80 km/h e outra a 120 km/h, a velocidade média não será 100 km/h). Quando temos alguns valores muito distantes da média (*outliers*) é comum retirar as α% de observações menores e as α% de observações maiores, determinando-se a média aritmética das restantes $(100 - 2\alpha)$% de observações. A esta média denomina-se por média aparada a α%, sendo uma medida mais robusta que a média aritmética quando estamos perante um conjunto de dados com *outliers*.

[5] Se considerarmos que as observações estão igualmente espaçadas dentro de cada classe (uniformemente distribuídas) o resultado será exatamente o mesmo, contudo a ideia será mais difícil de transmitir aos nossos alunos.

procede-se como no caso anterior e obtemos, desta forma, um valor aproximado do peso médio dos alunos dado por:

$$\bar{x} = \frac{1452}{25} = 58,08 \text{ kg}. \tag{2.15}$$

Peso dos alunos (kg)	Frequência absoluta n_i	Marca da classe x_i'	$x_i' \, n_i$
$[40,48[$	5	44	$44 \times 5 = 220$
$[48,56[$	5	52	$52 \times 5 = 260$
$[56,64[$	9	60	$60 \times 9 = 540$
$[64,72[$	3	68	$68 \times 3 = 204$
$[72,80]$	3	76	$76 \times 3 = 228$
Total	25		1452

Tabela 2.9: Tabela de frequências do peso dos alunos

A média goza de algumas propriedades importantes. De seguida apresentaremos duas delas.

Com este propósito, consideremos que a idade dos 25 alunos foi registada no dia 1 de setembro de 2010. Por (2.13) sabemos que, nesta data, a idade média dos alunos é igual a 14,04 anos. Qual será a média das idades destes alunos no dia 1 de setembro de 2011?

Facilmente respondemos que, como passou um ano completo para cada aluno, a média aumenta uma unidade, passando de 14,04 para 15,04 anos.

Para generalizar este resultado, sejam $x_1, x_2, \cdots, x_n$ os n valores da variável quantitativa x com média $\bar{x}$.

Propiedade 2.1. Se a cada valor da variável x adicionarmos uma constante $c \neq 0$ obteremos uma nova variável, que representamos por y, cuja média é:

$$\bar{y} = \bar{x} + c. \tag{2.16}$$

De facto,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + c) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} c = \bar{x} + c.$$

Para ilustrar a segunda propriedade da média, suponhamos que todos os alunos começaram a frequentar o ginásio que se situa junto à escola no dia 1 de outubro de 2010. Suponhamos ainda que ao fim de um mês cada aluno tinha perdido 2% do seu peso. Qual é o peso médio dos alunos da turma no dia 1 de novembro do mesmo ano? A resposta, em kg, será dada por

$$58,08 - 0,02 \times 58,08 = 0,98 \times 58,08 \approx 56,92 \text{ kg}.$$

Cada peso registado no dia 1 de novembro corresponde a 98% do respetivo peso anterior. Então a média obtida será 98% da média anterior. De um modo geral, temos a seguinte propriedade.

Propiedade 2.2. Se multiplicarmos cada valor da variável x por uma constante c obteremos uma nova variável w cuja média é:

$$\overline{w} = c \times \overline{x}. \tag{2.17}$$

Efetivamente,

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n} cx_i = c\frac{1}{n}\sum_{i=1}^{n} x_i = c\overline{x}.$$

2.5.3 Mediana

Para determinar a mediana devemos previamente ordenar as observações. Neste sentido sejam $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ as observações $x_1, x_2, \cdots, x_n$ ordenadas. Desta forma $x_{(1)}$ representa a menor observação, $x_{(2)}$ a segunda menor observação, $\cdots$, $x_{(i)}$ a i-ésima menor observação, $\cdots$ e $x_{(n)}$ será a observação máxima, isto é,

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}. \tag{2.18}$$

Definição 2.11. A **mediana**, representada por Me, é um valor que divide ao meio o conjunto das observações, isto é, 50% dos valores são inferiores ou iguais à mediana e 50% dos valores são superiores ou iguais à mediana.

Existe uma regra prática para calcular a mediana. Depois de ordenar os valores por ordem crescente consideram-se os seguintes dois casos.

1. Se n é ímpar a mediana é o valor que ocupa a posição central. Então, numa amostra de dimensão n, teremos

$$Me = x_{\left(\frac{n+1}{2}\right)}. \tag{2.19}$$

2. Se n é par a mediana é a média dos dois valores que ocupam a posição central. Então, numa amostra de dimensão n, teremos

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}. \tag{2.20}$$

Notemos que, neste último caso, a mediana pode não coincidir com nenhuma das observações da amostra.

Exemplo 2.7. Consideremos os dados registados na Tabela 2.1 relativos à idade dos 25 alunos e ordenemo-los por ordem crescente,

$$13,13,13,13,13,13,14,14,14,14,14,14,14,14,14,14,14,14,14,15,15,15,15,15,16.$$

Como o número de observações é ímpar, $n = 25$, a mediana é dada por

$$Me = x_{(13)} = 14. \tag{2.21}$$

No caso de dados não agrupados, através da tabela de frequências podemos determinar facilmente o valor que ocupa a posição central (ou os dois valores que ocupam as posições centrais) recorrendo às frequências acumuladas (absolutas ou relativas). Caso haja uma modalidade x_i' onde $F_i = 0,5$ (ou $N_i = \frac{n}{2}$), então $Me = \frac{1}{2}\left(x_i' + x_{i+1}'\right)$. Caso contrário, a mediana corresponderá à primeira modalidade x_i' tal que $F_i > 0,5$ (ou $N_i > \frac{n}{2}$).

Por outro lado, quando os dados estão agrupados em classes, podemos encontrar a classe mediana de modo idêntico. Assim, a classe mediana ($[l_{i-1}, l_i[$) será aquela que corresponde à primeira classe com frequência relativa acumulada superior ou igual a 50% ($F_i \geq 0.5$) ou, o que é equivalente, à primeira classe com frequência absoluta acumulada superior ou igual a $\frac{n}{2}$ ($N_i \geq \frac{n}{2}$). Por exemplo, relativamente à altura dos 25 alunos, apresentada na Tabela 2.6, na página 19, a classe mediana é $[155, 160[$, pois esta classe acumula 60% dos valores e a anterior acumula apenas 28%. No entanto, como neste caso se conhecem todos os valores da variável, deve calcular-se o valor da mediana, em vez da classe mediana.

2.5.4 Comparação das medidas de tendência central

As medidas estatísticas média, moda e mediana são designadas por medidas de tendência central, pois são três formas distintas de representar o centro dos dados. A utilização de cada uma destas medidas apresenta vantagens e desvantagens em relação às outras medidas.

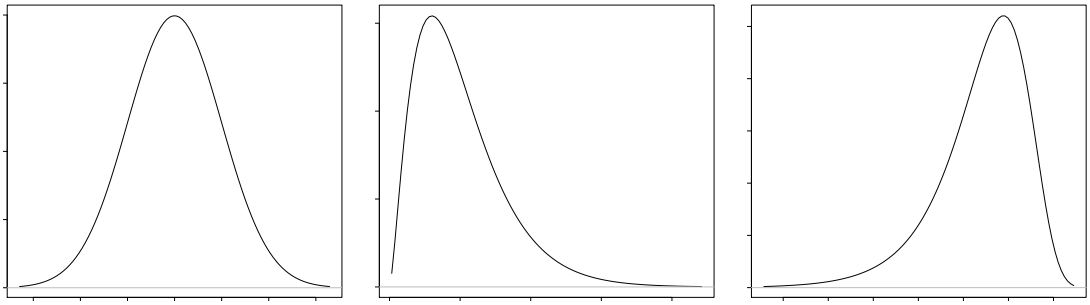
A **moda** pode ser determinada quer a variável seja qualitativa quer seja quantitativa. No entanto, um conjunto de dados pode não ter moda. Além disso, esta medida não é influenciada pelos valores extremos. Para ilustrar a relevância desta medida podemos referir, a título de exemplo, a importância que pode ter para uma empresa do setor do calçado saber qual o tamanho do sapato mais vendido, ou para uma empresa de laticínios saber o sabor dos iogurtes preferido dos clientes.

A **média** é uma medida estatística muito utilizada e no seu cálculo intervêm todos os dados. Se por um lado não há perda de informação, por outro lado qualquer alteração num dos valores produz um valor diferente no resultado da média. A média pode ser "enganadora" pois é influenciada pelos valores extremos (nomeadamente se existirem valores muito baixos ou muito altos), podendo em alguns casos "deixar de ser representativa" (isto é, exigir um cuidado particular na sua interpretação). Notemos que, por exemplo, duas turmas com a mesma média na classificação da disciplina de Matemática podem ter comportamentos muito distintos. Neste caso, a determinação das medidas de dispersão, que medem a variabilidade dos dados, podem ser fulcrais para uma melhor interpretação dos dados. Salientemos ainda que a média apenas se pode calcular quando a variável é quantitativa, podendo em alguns casos, o seu valor não coincidir com nenhum dos possíveis valores da variável (por exemplo, o número médio de elementos de um agregado familiar em determinada cidade de Portugal é 2,3, contudo não é possível haver um agregado familiar constituído por 2,3 indivíduos!).

A **mediana** divide as observações em dois grupos com igual número de indivíduos, mas o seu valor nem sempre coincide com um dos dados. Uma vez que o seu valor depende do número de observações e não de todos os valores, é uma medida estatística mais robusta do que a média no sentido em que não é influenciada pelos valores muito altos nem pelos valores muito baixos (*outliers*).

Estes valores muito elevados ou muito pequenos, comparativamente aos restantes, são comuns em algumas distribuições que designamos por assimétricas (positivas ou negativas). O tipo de assimetria decorre da comparação das medidas de tendência central. Seguidamente apresentamos três distribuições típicas, conforme se ilustra na Figura 2.10.

- distribuição simétrica se $\bar{x} = Me = Mo$.
- distribuição assimétrica positiva se $Mo \leq Me \leq \bar{x}$.
- distribuição assimétrica negativa se $\bar{x} \leq Me \leq Mo$.



Simetria	Assimetria positiva	Assimetria negativa
$\overline{x} = Me = Mo$	$Mo \leq Me \leq \overline{x}$	$\overline{x} \leq Me \leq Mo$

Figura 2.10: Tipos de assimetria

A título de exemplo vamos apresentar os diagramas de barras referentes às classificações de Matemática, obtidas no final do primeiro período, por três turmas do 7.º ano. Por observação do gráfico da Figura 2.11 podemos afirmar que a turma A apresenta uma distribuição simétrica, pois os dados estão igualmente distribuídos à direita e à esquerda do centro (valor das medidas de tendência central). De facto, a média, a moda e a mediana têm o mesmo valor, $\overline{x} = Me = Mo = 3$. Quanto às turmas B e C (ver gráficos da Figura 2.12), verificamos que na turma B, a distribuição é assimétrica positiva ou enviesada à direita, sendo a média igual a 2,76 e a moda e a mediana iguais a 2, logo $Mo \leq Me \leq \overline{x}$. Na turma C, a distribuição é assimétrica negativa ou enviesada à esquerda, sendo a média 3,92, a mediana 4 e a moda 5, logo $\overline{x} \leq Me \leq Mo$.

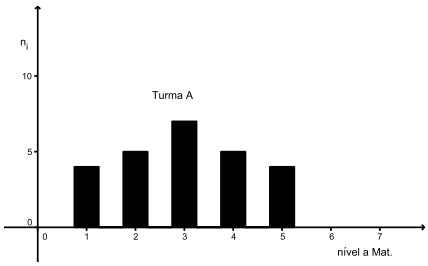
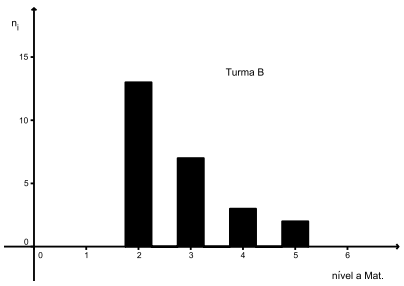


Figura 2.11: Distribuição simétrica

2.6 Medidas de tendência não central

Para além das medidas de tendência central existem outras medidas que nos informam relativamente à localização dos valores da variável. Costumam designar-se por quantis e iremos aqui abordar os quartis e os percentis, apesar dos percentis não estarem contemplados nos

Assimetria positiva



Assimetria negativa

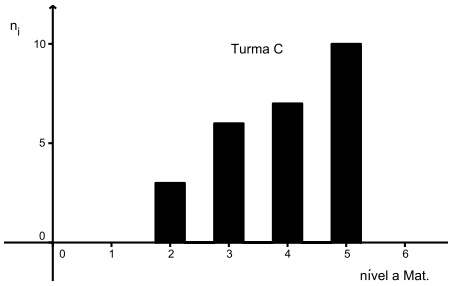


Figura 2.12: Distribuições assimétricas

programas do ensino básico e secundário (contudo são utilizados em diversas situações do nosso quotidiano).

2.6.1 Quartis

Os quartis são medidas estatísticas extremamente úteis na caracterização de uma amostra. A partir deles podemos obter uma representação gráfica designada por diagrama de extremos e quartis (subsecção 2.6.2) e calcular uma medida de dispersão (secção 2.15).

Definição 2.12. Os **quartis** são os valores que dividem o conjunto das observações, depois de ordenado, em quatro partes iguais, cada uma contendo 25% das observações. Os quartis são 3 e representam-se por Q_1, Q_2 e Q_3, sendo $Q_2 = Me$. Assim:

- Q_1 — o 1.º quartil é o valor que verifica a seguinte propriedade: 25% das observações são menores ou iguais a Q_1 e 75% são superiores ou iguais a Q_1, conforme ilustrado na Figura 2.13 (onde x_{min} e x_{max} representam o mínimo e o máximo da amostra).
- Q_2 — o 2.º quartil é igual à mediana.
- Q_3 — o 3.º quartil é o valor que verifica a seguinte propriedade: 75% das observações são menores ou iguais a Q_3 e 25% são superiores ou iguais a Q_3.

Para determinar o 1.º e o 3.º quartis de um conjunto ordenado de observações começa-se por determinar a mediana, Q_2, dividindo esse conjunto em duas partes iguais. O 1.º quartil será a mediana das observações que se encontram à esquerda de Q_2 e o 3.º quartil será a mediana das observações que se encontram à direita de Q_2.

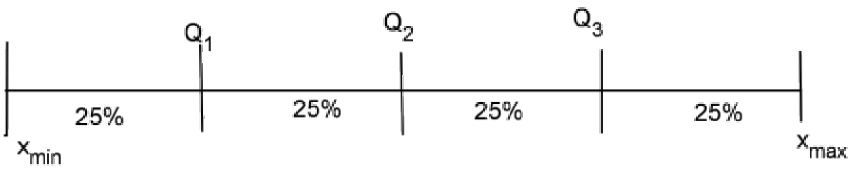


Figura 2.13: Esquema relativo aos extremos e quartis de uma distribuição

Para exemplificar, consideremos as idades dos 25 alunos, da tabela 2.1 presente na página 13. Já vimos que $Me = x_{(13)} = 14$.

13,13,13,13,13,13,14,14,14,14,14,14, **14,** 14,14,14,14,14,14,15,15,15,15,15,16

O 1.º quartil é a mediana dos 12 primeiros valores, isto é,

$$Q_1 = \frac{x_{(6)} + x_{(7)}}{2} = \frac{13+14}{2} = 13,5. \tag{2.22}$$

O 3.º quartil é a mediana dos 12 últimos valores, isto é,

$$Q_3 = \frac{x_{(19)} + x_{(20)}}{2} = \frac{14+15}{2} = 14,5. \tag{2.23}$$

Por outro lado, quando a variável é contínua podemos determinar as classes às quais pertencem o 1.º e 3.º quartis, recorrendo à frequência relativa acumulada, como procedemos com a mediana. Assim, a classe que contém o 1.º quartil será aquela que corresponde à primeira classe com frequência relativa acumulada superior ou igual a 25% ($F_i \geq 0,25$) ou, o que é equivalente, à primeira classe com frequência absoluta acumulada superior ou igual a $\frac{n}{4}$ ($N_i \geq \frac{n}{4}$). Analogamente, a classe que contém o 3.º quartil será aquela com frequência relativa acumulada superior ou igual a 75% ($F_i \geq 0,75$) ou com frequência absoluta acumulada superior ou igual a $\frac{3}{4}n$ ($N_i \geq \frac{3}{4}n$).

Voltando de novo ao exemplo relativo às alturas dos 25 alunos apresentadas na Tabela 2.6, da página 19, já vimos que a classe mediana é [155,160[, pois esta classe acumula 60% dos valores e a anterior acumula apenas 28%. Conclui-se, igualmente, que a classe à qual pertence o 1.º quartil é [150,155[, pois esta classe acumula 28% dos valores e a anterior acumula apenas 8%. Do mesmo modo, a classe à qual pertence o 3.º quartil é [160,165[pois esta classe acumula 88% dos valores e a anterior acumula apenas 60%.

2.6.2 Diagrama de extremos e quartis

Um diagrama de extremos e quartis é uma representação gráfica que podemos utilizar quando pretendemos representar esquematicamente um conjunto de dados numéricos. A sua construção depende de 5 valores: valor mínimo, valor máximo, 1.º quartil, 2.º quartil e 3.º quartil. Começa-se por traçar um eixo graduado, onde se assinalam os 5 valores. De seguida, acima desse eixo, traça-se um segmento horizontal desde o mínimo até ao 1.º quartil. Depois desenha-se um retângulo desde o 1.º quartil até ao 3.º quartil, divido pela mediana. Por fim, faz-se novamente um segmento horizontal desde 3.º quartil até ao valor máximo. No início e no fim do diagrama desenha-se, ainda, um pequeno segmento vertical. Deste modo, ficam definidas quatro zonas (contendo cada uma 25% dos dados), sendo duas delas centrais.

Este diagrama fornece informações sobre a forma como os dados estatísticos se distribuem, nomeadamente sobre a concentração/dispersão. Quanto mais estreita for uma zona, maior concentração de dados existe nessa zona. Os diagramas de extremos e quartis podem surgir na posição horizontal ou vertical. Na Figura 2.14 temos o diagrama de extremos e quartis das idades dos alunos.

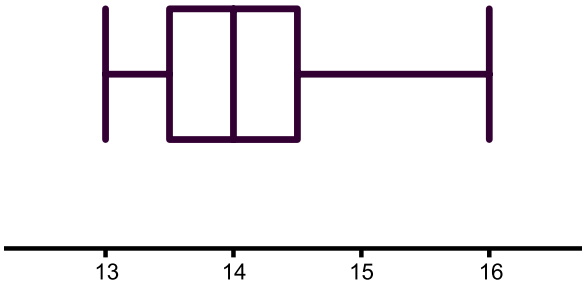


Figura 2.14: Diagrama de extremos e quartis relativo à idade dos 25 alunos

2.6.3 Percentis

Como já foi referido, os percentis estão fora do âmbito dos programas do ensino básico e secundário, no entanto optamos por fazer uma pequena abordagem pelo facto destas medidas se utilizarem na vida real, nomeadamente, para informar sobre o desenvolvimento das crianças.

Definição 2.13. Os **percentis** são os valores que dividem o conjunto das observações, depois de ordenado, em cem partes iguais, cada uma contendo 1% das observações. Os percentis são 99 e representam-se por P_1, P_2, $\ldots$, P_{99}, sendo $P_{25} = Q_1$, $P_{50} = Me$ e $P_{75} = Q_3$. Deste

modo, $P_\alpha = k$ significa que α% das observações são inferiores ou iguais a k e $(100-\alpha)$% das observações são iguais ou superiores a k.

Assim, por exemplo, se para um conjunto de dados tivermos:

- $P_{16} = 34$, significa que 16% das observações são inferiores ou iguais a 34 e 84% são iguais ou superiores a 34.
- $P_{72} = 55$, significa que 72% das observações são inferiores ou iguais a 55 e 28% são iguais ou superiores a 55.

Exemplo 2.8. Consideremos que numa consulta o pediatra, após pesar e medir a criança, afirma que esta está no percentil 80 no peso e no percentil 25 em relação à altura (algumas das medidas patentes na Caderneta de Saúde da Criança são os percentis do peso, da altura e do perímetro cefálico por idade). Qual o significado destes valores referidos pelo pediatra? Significam que, relativamente às crianças da mesma idade, existem 80% de criaças com um peso menor ou igual e apenas 20% com um peso maior ou igual. No que se refere à altura, relativamente às crianças com a mesma idade, exitem 25% de crianças mais baixas e 75% de crianças mais altas (a forma da evolução de cada um destes percentis bem como a discrepância entre eles é um dado importante na análise do desenvolvimento da criança).

2.7 Medidas de dispersão

Abordámos até agora várias medidas estatísticas que permitem caracterizar uma amostra relativamente à sua localização (seja ela central ou não central). Contudo, quando se pretende estabelecer comparações, deparamo-nos com muitas situações em que estas medidas não se revelam suficientes. A título ilustrativo consideremos as notas, numa escala de 0 a 20 valores, obtidas por dois alunos do mesmo ano de escolaridade, em dez fichas de avaliação de uma determinada disciplina.

Aluno A — 9, 9, 11, 12, 8, 7, 13, 11, 9, 11

Aluno B — 5, 15, 4, 4, 5, 17, 13, 17, 6, 14

Os dois alunos apresentam a mesma nota média, dez valores, mas da observação das suas notas poderemos dizer que os dois alunos são muito diferentes no que respeita ao aproveitamento nessa disciplina, apesar de terem tido o mesmo número de fichas com nota positiva.

Para além deste exemplo podemos considerar muitos outros, tais como duas cidades com a mesma temperatura média e a mesma temperatura mediana, apresentando amplitudes térmicas muito distintas ou duas turmas com a mesma média a Matemática, em que uma delas apresenta uma percentagem de negativas bastante superior à outra.

Nestes casos há então necessidade de calcular outras medidas estatísticas, medidas de dispersão, para conhecer de que forma os dados se encontram distribuídos.

2.7.1 Amplitude total

Uma das medidas de dispersão mais fácil de determinar é a amplitude total. O seu cálculo depende apenas dos dois valores extremos da amostra.

Definição 2.14. A **amplitude total** ou amplitude é a diferença entre o valor máximo e o valor mínimo do conjunto das observações. Representa-se por I_T e tem-se

$$I_T = x_{(n)} - x_{(1)}. \tag{2.24}$$

A amplitude dá-nos informação sobre a distância entre os valores extremos. Em duas turmas com a mesma média, a amplitude será maior na que apresenta as classificações mais dispersas. No entanto, esta situação pode resultar apenas de uma só classificação muito baixa ou muito alta. Pelo facto de a amplitude ser muito sensível aos extremos, é uma medida de dispersão pouco utilizada. Outra medida de dispersão que podemos calcular, não sensível aos valores extremos, é a amplitude interquartis.

2.7.2 Amplitude interquartis

Recorrendo à definição de quartis de uma distribuição, podemos determinar a amplitude interquartis que é uma medida de dispersão que envolve no seu cálculo o 1.º e o 3.º quartis. A amplitude interquartis não só é insensível aos valores extremos observados (máximo e mínimo), como também às 25% de observações de valores mais baixos e às 25% de valores mais elevados.

Definição 2.15. A **amplitude interquartis** é a diferença entre o 3.º e o 1.º quartis. Representa-se por I_Q e determina-se através de

$$I_Q = Q_3 - Q_1. \tag{2.25}$$

Esta medida indica-nos a amplitude do intervalo onde se situam as 50% das observações centrais, mostrando-nos a variabilidade dos dados em relação à mediana. Assim, é possível estabelecer comparações entre dois conjuntos de observações no que diz respeito à dispersão ou concentração dos valores em relação à mediana. Quanto menor for a amplitude interquartis, maior é a concentração dos valores em relação à mediana.

2.7.3 Desvio médio absoluto, variância e desvio padrão

Já conhecemos uma medida estatística que mede a variabilidade dos valores em relação à mediana, no entanto a medida de tendência central mais utilizada é a média e, por isso, faz sentido que haja uma medida de dispersão que nos dê a variabilidade dos valores em relação à média. Neste âmbito parece interessante que façamos o cálculo dos desvios de cada observação em relação à média. De seguida bastaria fazer a média dos desvios. Procedendo deste modo, verifica-se, facilmente, que a soma dos desvios é nula (conforme exemplo das idades dos alunos apresentado na Tabela 2.10), uma vez que

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})=\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}\overline{x}=\overline{x}-\frac{1}{n}\,n\,\overline{x}=0.$$

Para ultrapassar esta situação definiu-se o desvio médio absoluto, no qual se considera os valores absolutos dos desvios, impedindo que a soma dos desvios dê zero e obtendo a média das distâncias entre as observações e a média.

Definição 2.16. O **desvio médio absoluto**, representado por d_m, é a média das distâncias entre as observações e a média, isto é,

$$d_m=\frac{1}{n}\sum_{i=1}^{n}|x_i-\overline{x}|=\frac{1}{n}\sum_{i=1}^{p}n_i|x_i'-\overline{x}|. \tag{2.26}$$

Outra forma de contornar o facto de a soma dos desvios ser nula é calcular a média dos quadrados dos desvios. Surge assim outra medida de dispersão, a variância, que teve mais aceitação.

Definição 2.17. A **variância** é a média[6] dos quadrados dos desvios de cada observação da

[6] Efetivamente não é a média(na sua conceção habitual) uma vez que não dividimos por n, mas antes por $n-1$. Esta correção está ligada à inferência estatística, nomeadamente à utilização da variância da amostra como estimador da variância da população.

Idade dos alunos	Frequência absoluta n_i	Desvios $n_i(x_i'-\overline{x})$	Valores absolutos dos desvios $n_i\lvert x_i'-\overline{x}\rvert$	Quadrados dos desvios $n_i(x_i'-\overline{x})^2$
13	6	$6(13-14,04)=-6,24$	6,24	6,4896
14	13	$13(14-14,04)=-0,52$	0,52	0,0208
15	5	$5(15-14,04)=4,8$	4,8	4,608
16	1	$16-14,04=1,96$	1,96	3,8416
Total	25	0	13,52	14,96

Tabela 2.10: Cálculo do desvio médio absoluto

amostra relativamente à média. Representa-se por s^2,

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n-1}\sum_{i=1}^{p}n_i(x_i'-\overline{x})^2. \tag{2.27}$$

Pegando na definição de variância e aplicando propriedades dos somatórios podemos obter uma fórmula simplificada para o cálculo da variância, fórmula de Köning,

$$s^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right]=\frac{1}{n-1}\left[\sum_{i=1}^{p}n_ix_i'^2-n\overline{x}^2\right] \tag{2.28}$$

Notemos que apesar desta fórmula simplificar os cálculos, com o recurso a computadores ou à máquina de calcular a sua utilidade tornou-se reduzida (neste contexto).

A variância pode ser utilizada para comparar dois conjuntos de observações. Quanto maior for a variância, maior é a dispersão dos valores relativamente à média[7]; quanto menor é a variância, maior é a concentração dos valores relativamente à média. Contudo a variância apresenta uma desvantagem, uma vez que os quadrados dos desvios passam a ser de uma ordem de grandeza superior aos desvios. Por exemplo, no caso da idade, passamos de anos para anos ao quadrado; se noutra situação a variável fosse a distância em metros, passaríamos a metros quadrados. Para resolver este problema, permitindo que se volte à ordem de grandeza inicial, definiu-se outra medida de dispersão que resulta da raiz quadrada da variância e que se designa por desvio padrão.

[7] Nesta observação consideramos que os dados comparados estão expressos na mesma unidade de medida e que têm médias próximas. Em rigor quando se pretende comparar a dispersão de dois conjuntos de dados deve-se utilizar o coeficiente de dispersão de Pearson, $CD=\frac{s}{\overline{x}}$, com $\overline{x}\neq 0$.

Definição 2.18. O **desvio padrão** é a raiz quadrada da variância, sendo representado por s e determinado por

$$s = \sqrt{s^2}. \tag{2.29}$$

Por exemplo, se considerarmos a amostra dos 25 alunos e a variável idade, temos:

$$s = \sqrt{\frac{1}{24}\sum_{i=1}^{4} n_i(x_i - \overline{x})^2} = \sqrt{\frac{14,96}{24}} \approx \sqrt{0,6233} \approx 0,789. \tag{2.30}$$

O desvio padrão mede a variabilidade dos valores em relação à média e a sua interpretação é idêntica à da variância. Quanto maior for o valor do desvio padrão, maior é o afastamento dos valores em relação à média. Um dos inconvenientes do desvio padrão é ser influenciado por valores extremos, ou seja, valores muito maiores ou muito menores que os restantes.

O desvio padrão goza de algumas propriedades importantes. Consideremos, sem perda de generalidade, $x_1, x_2, \cdots, x_n$ os n valores de uma variável quantitativa com média $\overline{x}$ e desvio padrão s.

Propiedade 2.3. O desvio padrão é sempre não negativo.

Esta propriedade é consequência imediata da definição de desvio padrão.

Propiedade 2.4. Se o desvio padrão é igual a zero significa que não existe variabilidade. Consequentemente as observações são todas iguais.

De facto,

$$s = 0 \Leftrightarrow s^2 = 0 \Leftrightarrow \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0 \Leftrightarrow x_i = \overline{x},\ \forall i,$$

isto é, todos os valores observados têm de ser iguais à média.

Propiedade 2.5. Se a cada valor da variável x adicionarmos uma constante c obteremos um conjunto de dados cujo desvio padrão é $s' = s$ (os valores das observações são alterados, mas a distância entre eles não).

De facto, recorrendo à propriedade 2.1 (página 30), a variância para o novo conjunto de dados é determinada através de

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(c + x_i) - (c + \overline{x})]^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^2.$$

Propiedade 2.6. Se multiplicarmos cada valor da variável x por uma constante c obteremos um conjunto de dados cuja desvio padrão é $s' = |c|s$.

Neste caso, fazendo uso da propriedade 2.2 (página 31), obtem-se

$$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(cx_i - c\overline{x})^2 = c^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

logo $s' = |c|s$.

2.8 Distribuições bidimensionais

Em vários estudos estatísticos assume, por vezes, maior importância o estudo em simultâneo de duas variáveis da mesma amostra, pretendendo-se estudar em que medida elas se relacionam, isto é, de que forma a variação de uma influencia a variação da outra.

Exemplo 2.9. Exemplos de situações em que se estuda a relação entre duas variáveis:

- relação entre as notas de Matemática e de Ciências Físico-Químicas.
- relação entre as classificações da avaliação interna e as classificações da avaliação externa, numa determinada disciplina;
- relação entre o peso e a altura de um conjunto de adolescentes;
- relação entre a idade do bebé e o perímetro cefálico;
- relação entre a idade do pai e a idade da mãe de um conjunto de crianças.

Para estudarmos relações deste tipo há necessidade de recolher uma amostra de dados bivariados que pode ser representada na forma

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

isto é, (x_i,y_i) com $i = 1,2,\cdots,n$, onde cada indivíduo contribui com um par de valores.

Definição 2.19. Distribuição bidimensional é uma distribuição em que os dados são bivariados.

2.8.1 Diagrama de dispersão

A representação gráfica de uma distribuição bidimensional difere bastante das representações já mencionadas anteriormente. Um conjunto de dados bivariados representa-se através de uma **nuvem de pontos** ou **diagrama de dispersão**, onde são representados os pontos (x_i, y_i), num sistema de eixos coordenados.

Exemplo 2.10. Estudo da relação entre as notas obtidas nos testes diagnósticos de Ciências Físico-Químicas e de Matemática, da amostra constituída pelos 25 alunos do 8.º ano. Se representarmos graficamente os pontos de coordenadas (x_i, y_i), em que x_i é a nota de Matemática e y_i é a nota de Fisica-Química, obteremos a nuvem de pontos representada na Figura 2.15.

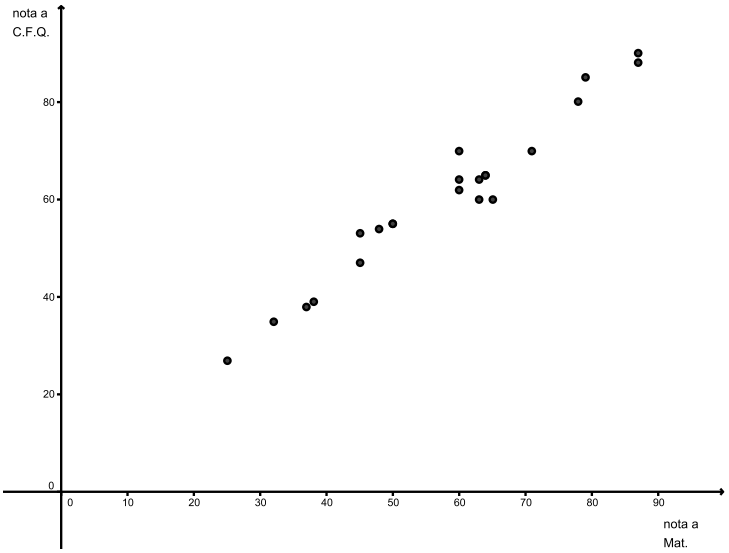


Figura 2.15: Diagrama de dispersão das notas de Matemática e de Ciências Físico-Químicas

O diagrama de dispersão é útil pois permite visualizar se existe ou não relação (ou correlação) entre as variáveis. Na situação apresentada, podemos observar que quanto melhor for o resultado no teste diagnóstico de Matemática, melhor será o resultado no teste diagnóstico de Ciências Fisico-Químicas.

Definição 2.20. O **ponto médio ou centro de gravidade** de uma nuvem de pontos é o ponto de coordenadas $(\overline{x}, \overline{y})$, em que $\overline{x}$ e $\overline{y}$ correspondem às médias aritméticas dos valores das variáveis x e y, respetivamente.

Se os pontos da nuvem se localizarem à volta de uma reta, a correlação diz-se linear. Por outro lado, a correlação será positiva quando a maioria dos pontos se situa à volta de uma reta

de declive positivo (ver Figura 2.16) e será negativa quando a maioria dos pontos se situam à volta de uma reta de declive negativo. Para medir o grau de associação linear entre duas variáveis utiliza-se o **coeficiente de correlação linear de Pearson**, usualmente designado apenas por **coeficiente de correlação**.

2.8.2 Coeficiente de correlação

Consideremos n observações bivariadas, $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ relativas ao par de variáveis quantitativas x e y.

Definição 2.21. O **coeficiente de correlação linear**, representado por r_{xy} (ou simplesmente r), é o valor que mede o grau de associação linear entre duas variáveis e calcula-se do seguinte modo:

$$r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{2.31}$$

onde s_x representa o desvio padrão da variável x e s_y representa o desvio padrão da variável y.

Notemos que no numerador da primeira fórmula que permite calcular o coeficiente de correlação, apresentada na definição anterior, encontra-se uma medida que se designa por covariância. A covariância depende das unidades de medida que estamos a considerar e por isso é muito difícil de interpretar (se uma variável for o peso e a outra a altura, a covariância terá como unidade o produto das duas).

$$s_{xy} = cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}). \tag{2.32}$$

Uma das vantagens do coeficiente de correlação relativamente à covariância é o facto de não ser influenciado pelas unidades de medida.

Propiedade 2.7. O coeficiente de correlação é invariante para alterações de unidade de medida.

Considerando que $a+bx_i$ corresponde à mudança de unidade de medida de x_i, temos então

$$\frac{\sum_{i=1}^{n}[(a+bx_i)-(a+b\bar{x})](y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}[(a+bx_i)-(a+b\bar{x})]^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{b\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{|b|\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{b}{|b|}r_{xy} = r_{xy},$$

uma vez que nas mudanças de medida, temos $b > 0$. Podemos assim concluir que não há alteração do coeficiente de correlação.

Embora se possa, por observação dos diagramas de dispersão, dizer se há correlação entre as duas variáveis e, caso exista, se é positiva ou negativa, podemos compreender melhor porquê se analisarmos o modo como se calcula o coeficiente de correlação. Para explicar esta ideia consideremos a nuvem de pontos relativa às idades dos pais e das mães de 15 bebés de uma creche. Marcou-se o ponto $(\overline{x},\overline{y})$ e um novo sistema de eixos definido pelas retas de equação $x = \overline{x}$ e $y = \overline{y}$. As coordenadas dos pontos que definem a distribuição serão $(x_i - \overline{x}, y_i - \overline{y})$.

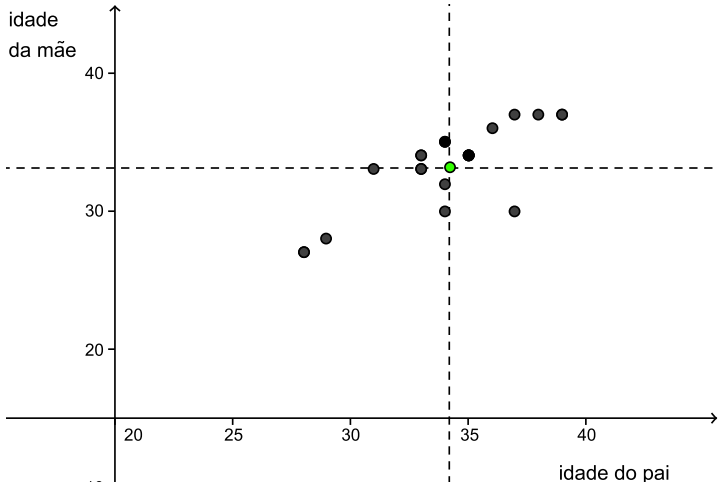


Figura 2.16: Diagrama de dispersão da idade dos pais e das mães

O sinal do coeficiente de correlação é o sinal do produto $(x_i - \overline{x})(y_i - \overline{y})$. Uma vez que nos quadrantes ímpares a abcissa e a ordenada têm o mesmo sinal, quando os pontos se concentram nestes quadrantes, o produto é maioritariamente positivo. Assim será de esperar que a soma $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ seja positiva e por consequência o valor do coeficiente de correlação será positivo. É isto que se verifica na Figura 2.16. Por outro lado, como nos quadrantes pares a abcissa e a ordenada dos pontos têm sinais contrários, quando os pontos se concentram nestes quadrantes, o produto é maioritariamente negativo. Assim será de esperar que a soma $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ seja negativa e por consequência o valor do coeficiente de correlação será negativo, como por exemplo no Gráfico V da Figura 2.17.

Em qualquer distribuição estatística o valor do coeficiente de correlação pertence ao intervalo $[-1,1]$, isto é, $-1 \leq r_{xy} \leq 1$ ou $|r_{xy}| \leq 1$. Quando a correlação é perfeita e negativa o coeficiente de correlação toma o valor -1. Quando a correlação é perfeita positiva o coeficiente de correlação toma o valor 1. A valores próximos de zero corresponde uma correlação

quase nula [8]. Quanto maior for o valor absoluto de r_{xy} mais forte será a correlação linear entre as variáveis. O sinal do coeficiente dá-nos o sentido da correlação.

Apresentamos de seguida na Figura 2.17 um conjunto de gráficos que traduzem vários tipos de relações entre duas variáveis e na Tabela 2.11 os valores dos respetivos coeficientes de correlação.

Gráfico	Coeficiente de correlação
I	$0,99$
II	$0,47$
III	$0,02$
IV	$-0,51$
V	$-0,9$
VI	$0,13$

Tabela 2.11: Exemplos de valores do coeficiente de correlação

2.9 Regressão linear simples

Após a representação da nuvem de pontos ficamos com uma ideia da correlação que existe entre as duas variáveis. Existindo correlação linear o próximo passo é traçar a reta que melhor se ajusta ao conjunto de pontos. Essa reta chama-se **reta de regressão** e passa pelo ponto médio ou centro de gravidade. A sua equação é do tipo

$$y = ax + b. \tag{2.33}$$

Para obter esta reta recorre-se à calculadora gráfica ou ao computador. Ambos usam o método dos mínimos quadrados, isto é, determinam a reta que melhor se aproxima dos valores observados, de tal modo que seja mínima a soma dos quadrados dos desvios entre os valores observados e os correspondentes na reta.

[8] Refira-se que o facto de o coeficiente de correlação ser próximo de zero, $r_{xy} \approx 0$, não significa que não se verifica outro tipo de dependência entre as variáveis (ver, por exemplo, o gráfico VI da Figura 2.17). Este valor indica-nos que não há uma dependência linear das variáveis.

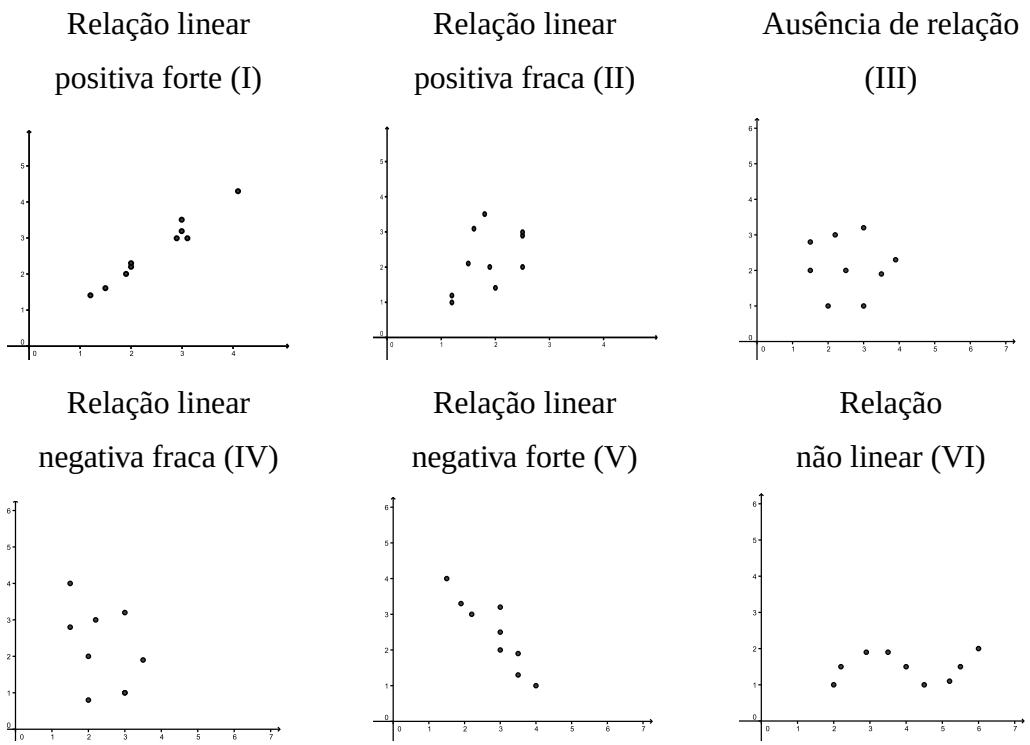


Figura 2.17: Diagramas de dispersão com diferentes relações entre as variáveis

Uma das vantagens de conhecer a equação da reta é a possibilidade de prever o comportamento da variável dependente *y*, conhecendo o valor da variável independente *x*. Como veremos no próximo capítulo, surge, frequentemente, nos manuais escolares um erro relacionado com a previsão de um valor da variável independente, conhecido um valor da variável dependente.

2.9.1 Regressão linear no Ensino Secundário

No programa de Matemática A do ensino secundário, nomeadamente no 10.º ano, a terceira unidade a ser lecionada é a Estatística. Nesta fase os alunos já possuem algumas noções que adquiriram no 3.º ciclo e já realizaram pequenos trabalhos, no entanto é a primeira vez que se fará referência a distribuições bidimensionais, sendo uma abordagem gráfica e intuitiva.

Desta forma, no 10.º ano de escolaridade é transmitida uma ideia intuitiva de reta de regressão, tentando explorar a sua interpretação e as suas limitações. Apesar de não ser objetivo deste nível de ensino explicar formalmente a reta obtida, é transmitida a ideia pela qual ela é determinada — corresponde à reta que faz com que a soma dos quadrados das distâncias de cada ponto da nuvem à reta seja mínima (método dos mínimos quadrados), sendo esta reta

unicamente determinada recorrendo a uma calculadora.

2.9.2 O método dos mínimos quadrados

Consideremos um conjunto de dados (x_i, y_i), com $i = 1, \ldots, n$ para o qual se pretende ajustar a reta

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.34}$$

onde y_i representa a variável dependente ou endógena (que o modelo pretende explicar o comportamento), x_i a variável independente ou exógena (que será uma variável explicativa para a modelação de y_i), ε_i a variável erro (que é uma variável aleatória com algumas características fundamentais para a fiabilidade da inferência estatística associada à regressão) e β_0 e β_1 os parâmetros da regressão. A reta estimada será da forma

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i \tag{2.35}$$

onde $\widehat{y_i}$, $\widehat{\beta_0}$ e $\widehat{\beta_1}$ representam respetivamente os estimadores (ou estimativas[9]) de y_i, β_0 e β_1. Desta forma, os erros ε_i correspondem à diferença entre os valores observados para y_i e os valores estimados, *i.e.*

$$\varepsilon_i = y_i - \widehat{y_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i. \tag{2.36}$$

O método habitualmente utilizado para estimar os parâmetros da reta é o método dos mínimos quadrados (consultar, por exemplo, Montgomery *et al.(2006)*que determina os estimadores $\widehat{\beta_0}$ e $\widehat{\beta_1}$ que minimizam a soma dos quadrados dos erros, isto é, que minimizam

$$f(\widehat{\beta_0}, \widehat{\beta_1}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right)^2. \tag{2.37}$$

Para determinar o mínimo teremos

$$\left\{ \begin{array}{rcl} \frac{f\left(\widehat{\beta_0}, \widehat{\beta_1}\right)}{\partial \widehat{\beta_0}} & = & 0 \\ \frac{f\left(\widehat{\beta_0}, \widehat{\beta_1}\right)}{\partial \widehat{\beta_1}} & = & 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rcl} \widehat{\beta_0} & = & \overline{y} - \widehat{\beta_1} \overline{x} \\ \widehat{\beta_1} & = & \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x^2}} = \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x} r_{xy} \end{array} \right. , \tag{2.38}$$

[9] Ao longo do presente trabalho não iremos efetuar distinção entre estimador e estimativa no que se refere à notação utilizada. Há, contudo, a referir que as estimativas agora representadas por $\widehat{\beta_0}$ e $\widehat{\beta_1}$ correspondem aos parâmetros anteriormente representados por *a* e *b* na equação (2.33).

onde s_{xy} representa a covariância entre y e x; s_x o desvio padrão de x e r_{xy} o coeficiente de correlação entre x e y. O determinante da matriz Hessiana é igual a

$$4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2 = 4n^2 s_x^2, \tag{2.39}$$

que é positivo (desde que x_i não assuma sempre o mesmo valor). Assim sendo, a matriz Hessiana é definida positiva e os valores determinados em (2.38) correspondem efectivamente ao mínimo pretendido.

Desta forma, na equação (2.38) temos as fórmulas dos estimadores dos mínimos quadrados de β_0 e β_1 com os quais podemos, recorrendo à regressão (2.35), obter estimativas para y conhecendo um valor específico de x. As propriedades da inferência estatística resultante desta aplicação dependem das características dos resíduos (variável ε). Contudo, uma vez que esta abordagem não é efetuada no ensino não superior, sublinhamos apenas que os estimadores assim obtidos gozam de excelentes propriedades (não enviesamento, eficiência e consistência) caso a variável ε satisfaça determinadas características, nomeadamente a normalidade, independência e homocedasticidade.

Algumas características da regressão são exploradas no ensino secundário, como ilustram as seguintes propriedades.

Propiedade 2.8. A soma dos erros é nula, isto é

$$\sum_{i=1}^{n} \varepsilon_i = 0. \tag{2.40}$$

De facto

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = n\left(\overline{y} - \widehat{\beta}_0 - \widehat{\beta}_1 \overline{x}\right),$$

e, utilizando $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$, obtém-se o resultado pretendido.

Propiedade 2.9. A reta estimada pelo método dos mínimos quadrados passa sempre pelo centro de gravidade dos dados $(\overline{x},\overline{y})$.

Efetivamente, utilizando (2.35) e (2.38), obtém-se

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x} = \overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 \overline{x} = \overline{y},$$

isto é, o valor de $\widehat{y}$ quando $x = \overline{x}$ é $\widehat{y} = \overline{y}$.

A grande maioria dos docentes do ensino secundário, a avaliar pelos manuais por nós consultados, desconhece que a reta obtida pelo método dos mínimos quadrados para estimar y em função de x e a reta obtida para estimar x em função de y não são, em geral, idênticas.

2.9.3 A regressão linear inversa

Podemos utilizar fórmulas análogas às (2.38) para efetuarmos a regressão de x em função de y obtendo

$$\widehat{x_i} = \widehat{\alpha}_0 + \widehat{\alpha}_1 y_i, \tag{2.41}$$

onde $\widehat{\alpha}_0 = \overline{x} - \widehat{\alpha}_1 \overline{y}$ e $\widehat{\alpha}_1 = \frac{s_x}{s_y} r_{xy}$, (cf. equação2.38) Invertendo a reta (*i.e.* resolvendo em função de y_i e trocando os papéis desempenhados pelas variáveis) obtemos

$$\widetilde{y_i} = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_i. \tag{2.42}$$

Os estimadores de β_0 e β_1 assim obtidos, supondo $r_{xy} \neq 0$ (pois caso $r_{xy} \approx 0$ a reta de regressão não terá qualquer sentido), serão dados por:

$$\widetilde{\beta}_0 = -\frac{\widehat{\alpha}_0}{\widehat{\alpha}_1} = -\frac{\overline{x} - \widehat{\alpha}_1 \overline{y}}{\widehat{\alpha}_1} = \overline{y} - \widetilde{\beta}_1 \overline{x} \quad \text{e} \quad \widetilde{\beta}_1 = \frac{1}{\widehat{\alpha}_1} = \frac{s_y}{s_x} r_{xy}^{-1}. \tag{2.43}$$

As retas (2.42) e (2.35) são coincidentes se e só se

$$\begin{cases} \widehat{\beta}_0 &= \widetilde{\beta}_0 \\ \widehat{\beta}_1 &= \widetilde{\beta}_1 \end{cases} \Leftrightarrow \widehat{\beta}_1 = \widetilde{\beta}_1 \Leftrightarrow r_{xy} = 1 \vee r_{xy} = -1. \tag{2.44}$$

Assim sendo, ambas as retas passam pelo mesmo ponto $(\overline{x}, \overline{y})$, mas só se obtém a mesma recta utilizando os dois métodos se o módulo da correlação for unitário (sendo os erros, nestes casos, todos nulos uma vez que a reta passa precisamente por todos os pontos). Deste modo, a reta de regressão de y em função de x, determinando os parâmetros da recta $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_0 x$ que minimizam $\sum_t (y_t - \widehat{y_t})^2$, será distinta (excepto se $|r_{xy}| = 1$) da reta obtida quando efectuamos uma regressão de x em função de y, determinando os parâmetros da reta $\widehat{x_t} = \widehat{\alpha}_0 + \widehat{\alpha}_1 y_t$ que minimizam $\sum_t (x_t - \widehat{x_t})^2$. Esta diferença resulta da forma como definimos os erros nas duas regressões, pois enquanto na primeira os erros são medidos paralelamente ao eixo das ordenadas (o erro é definido pela diferença entre o valor observado de y e o seu valor estimado condicionalmente a x, $\varepsilon_t = y_t - \widehat{y_t}$, cf. ilustra o primeiro gráfico da Figura 2.18), na segunda os erros são medidos paralelamente ao eixo das abcissas (o erro é definido pela diferença entre o valor observado x e o seu valor estimado pela regressão em função de y, $\varepsilon_t = x_t - \widehat{x_t}$,

cf. segundo gráfico da Figura 2.18). Desta forma, será erróneo utilizar a regressão de y em função de x para efetuar previsões para x quando conhecemos um determinado valor para y e, apesar de em algumas aplicações a diferença das duas retas poder ser diminuta, existem outras situações em que o erro pode assumir valores elevados. Há, contudo, determinadas situações específicas para as quais se justifica a necessidade de utilização de regressão inversa, como ilustram alguns modelos de calibração (*cf.* Brown (1993)).

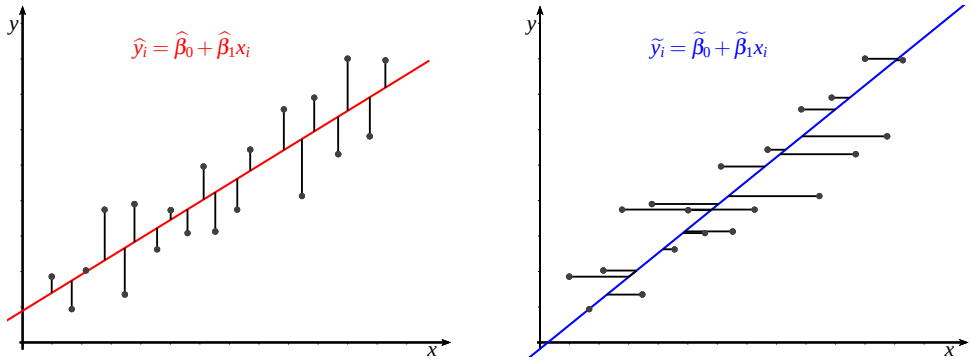


Figura 2.18: Definição dos erros na regressão linear

Os modelos de calibragem estatística (ou, por vezes, denominados por previsão inversa) são muito usados em química, engenharia, bioestatística e podem ser extremamente úteis em algumas aplicações. Contudo, estas aplicações são caracterizadas por contextos distintos dos que previamente referimos. Consideremos, então, que a variável y é medida através de processo complicado (muitas vezes fora do nosso alcance), dispendioso (em termos de tempo e/ou monetariamente) mas que os seus resultados são extremamente precisos (sem erros ou com erros negligenciáveis em relação aos erros de medição da variável x). A variável x é medida por um processo simples, rápido, barato mas com pouca precisão (obtemos valores aproximados). O objetivo é estimar novos valores de y para alguns valores de x conhecidos. Neste contexto, os erros a ter em conta no método de regressão linear deverão ser $\varepsilon_i = \widehat{x}_i - x_i$ estimando-se a regressão $\widetilde{y}_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_i$ utilizando (2.43). Com esta regressão podemos estimar novos valores para y condicionados a valores conhecidos de x. Os estimadores assim obtidos, sob determinadas condições, gozam igualmente de boas propriedades. Osborne (1991) faz uma apresentação histórica da evolução destes métodos, incluindo alguma discussão sobre a problemática inerente à sua utilização.

Contudo, estes contextos específicos nos quais a utilização da calibração é frequentemente utilizada, não são abordados no ensino secundário.

2.9.4 Estimação de y condicionada a $x = x_0$

Consideremos que se pretende estimar o valor de y quando x assume o valor x_0. Recorrendo à equação (2.35), regressão de y em função de x, obtemos

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 = \overline{y} - \frac{s_y}{s_x} r_{xy} \overline{x} + \frac{s_y}{s_x} r_{xy} x_0 = \overline{y} + \frac{s_y}{s_x} r_{xy} (x_0 - \overline{x}) \tag{2.45}$$

e recorrendo à equação (2.42), função inversa da regressão de x em função de y, obtemos

$$\widetilde{y} = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_0 = \overline{y} - \frac{s_y}{s_x} r_{xy}^{-1} \overline{x} + \frac{s_y}{s_x} r_{xy}^{-1} x_0 = \overline{y} + \frac{s_y}{s_x} r_{xy}^{-1} (x_0 - \overline{x}), \tag{2.46}$$

sendo a distância entre as duas estimativas dada por

$$|\widehat{y} - \widetilde{y}| = \frac{s_y}{s_x} |x_0 - \overline{x}| \left| r_{xy} - r_{xy}^{-1} \right| \tag{2.47}$$

que depende do quociente entre o desvio padrão de y e o de x, da distância entre x_0 e $\overline{x}$ (o que era espectável uma vez que ambas as rectas passam no ponto $(\overline{x}, \overline{y})$) e da distância entre r_{xy} e r_{xy}^{-1}.

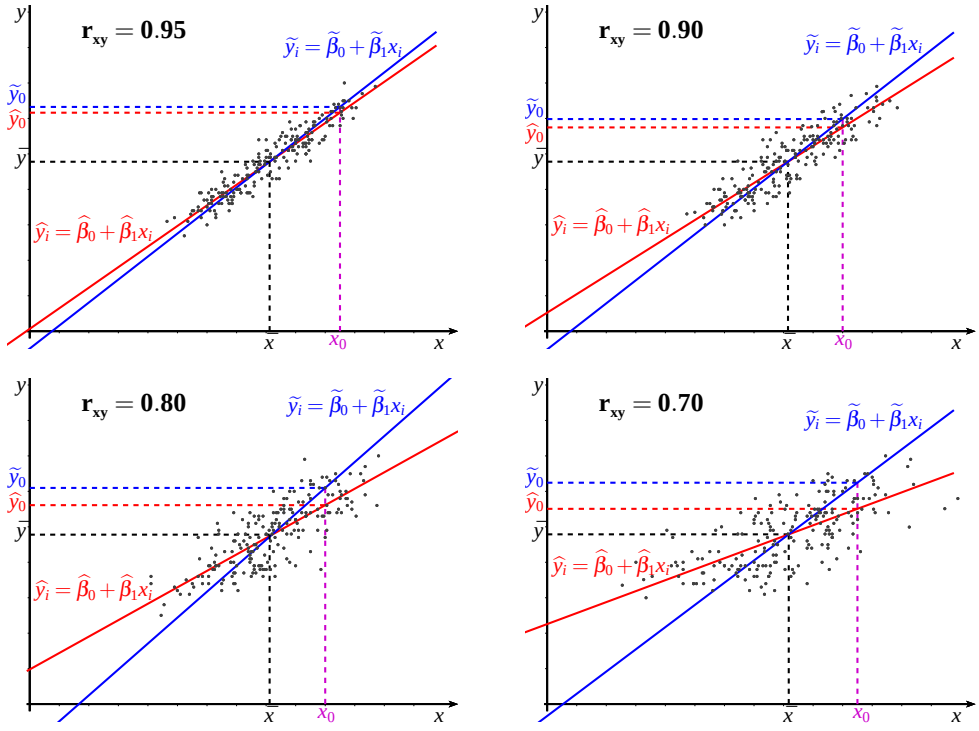


Figura 2.19: Regressão de y condicionada a x *versus* de x condicionada a y

Conforme claramente ilustram os quatro exemplos retratados na Figura 2.19, onde são apresentadas as duas retas obtidas utilizando conjuntos distintos de 200 observações (recorrendo ao *software GeoGebra*), com coeficiente de correlação igual a 0.95, 0.90, 0.80 e 0.70,

podemos constatar a distinção entre as duas retas, a diferença entre o valor estimado de y obtido pelas duas rectas quando x assume o valor x_0, bem como o aumento desta distância com a diminuição da correlação entre x e y e/ou o afastamento de x_0 relativamente a $\overline{x}$. Estes gráficos podem ser facilmente construídos em sala de aula (cf. ilustraremos na secção 4.1.5 na página 71), com a vantagem de serem dinâmicos, isto é, ao alterarmos um ponto (ou um conjunto de pontos) visualizarmos imediatamente as consequentes alterações no coeficiente de correlação e nas retas estimadas, bem como ao mudarmos a coordenada x_0 percebermos as consequentes implicações no valor de y estimado.

Capítulo 3

Análise crítica aos materiais disponíveis

Da análise efetuada a um conjunto de materiais disponíveis para o ensino da Estatística no ensino básico e secundário, serão apresentadas, neste capítulo, várias situações encontradas em manuais escolares, onde se detetaram erros e/ou gralhas relativos às representações gráficas, à notação utilizada e à exploração de alguns conceitos. Para cada situação detetada será apresentado um exemplo, mencionado o manual (ou manuais, se for o caso) e apresentada a devida justificação.

Constatamos que nos manuais do ensino básico há erros frequentes nas representações gráficas, enquanto nos manuais do ensino secundário, embora estes também apareçam em alguns casos, existe um erro comum no ensino da regressão linear e uso incorreto de notação.

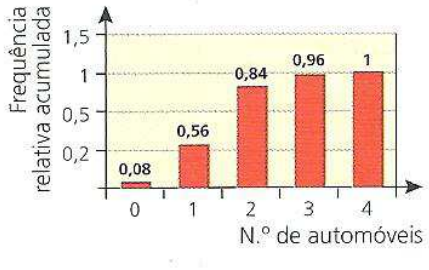
Ilustramos, de seguida, algumas das conclusões da análise efetuada.

3.1 Erros nas escalas

Um tipo de erro que surge com alguma frequência diz respeito às representações gráficas. Magro *et al.* (2010, p. 112) apresentam um gráfico de linhas em que a escala do eixo vertical não começa no zero como devia (ver Figura 3.1). Esta situação induz o leitor em erro pois parece que a temperatura aumentou muito mais do que na realidade. Uma situação muito semelhante verifica-se em Neves *et al.* (2010b, p. 30) num gráfico intitulado "Evolução da temperatura das 6 às 12 horas". Ainda neste âmbito, Negra & Martinho (2010, p. 144) apresentam um gráfico de barras de frequência relativa acumulada em que a escala do eixo vertical está errada, apesar de começar no zero (cf. Figura 3.1). Por outro lado, uma das barras

está mal construída pois a sua altura não está de acordo com o respetivo rótulo.

Negra & Martinho (2010), p. 144



Magro *et al.* (2010), p.112

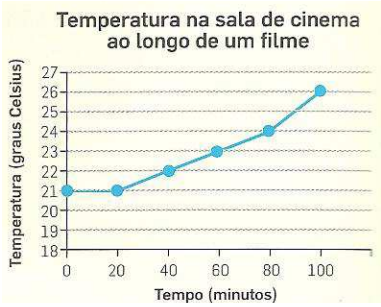


Figura 3.1: Erros de escala

Relativamente às escalas no eixo horizontal podemos ver que o gráfico de barras intitulado Carregamento de telemóveis que Duarte & Filipe (2010, p. 131) apresentam, não está correto pois a distância entre as duas últimas barras deveria ser o quádruplo daquela que aparece (cf. Figura 3.2). Da forma como está parece que o maior valor da variável é 35 euros e não 50 euros. O gráfico deveria também evidenciar que não há carregamentos mensais de 35, 40 e 45 euros, ficando no gráfico o lugar dessas barras de frequência nula. Uma situação idêntica surge no manual de 7.º ano de Costa & Rodrigues (2010b, p. 15) e ainda no manual de 8.º ano de Neves *et al.* (2011, p. 21).

Duarte & Filipe (2010), p. 131

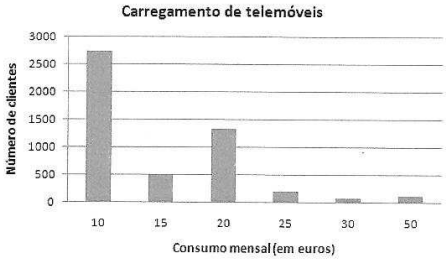


Figura 3.2: Erros de escala horizontal

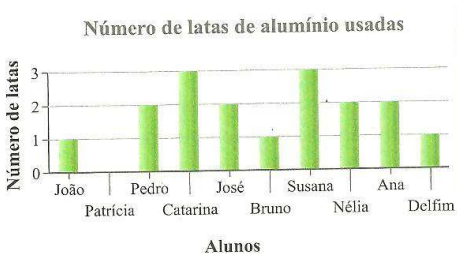
É de referir ainda o gráfico de barras no manual do 10.º ano de Costa & Rodrigues (2010a, p. 182). Este gráfico não tem qualquer escala no eixo vertical, embora as barras contenham rótulos e além disso não apresenta as barras igualmente espaçadas. Não é um bom exemplo para os alunos.

3.2 Confusão entre dados e frequência

Apresenta-se de seguida algumas situações onde há confusão entre dados e frequência e que dão origem a gráficos que, embora tenham barras, não são gráficos de barras segundo Martins & Ponte (2010), uma vez que no eixo horizontal não aparecem nem valores da variável nem modalidades.

Um exemplo surge no manual do 7.º ano de Costa & Rodrigues (2010b). A variável em estudo é o consumo de água, logo os dados são numéricos, no entanto os autores designam por gráfico de barras um gráfico que em vez de ter no eixo horizontal os valores da variável tem os meses do ano. O mesmo acontece noutro manual do 7.º ano. Segundo Faria *et al.* (2010) num estudo em que a variável é o número de pares de sapatos vendidos, o gráfico que foi considerado pelos autores como um gráfico de barras apresenta no eixo horizontal os dias da semana (cf. Figura 3.3). Trata-se de um gráfico com barras! Aparece igualmente mais um exemplo destes no manual de Passos & Correia (2010). Na página 36 os autores apresentam um gráfico em que a variável em estudo é o número de latas de alumínio usadas por 10 alunos e que no eixo horizontal tem os nomes dos alunos em vez do número de latas (cf. Figura 3.3). Mesmo assim os autores perguntam o nome do gráfico, esperando que os discentes respondam "gráfico de barras". Neste exemplo os valores da variável são 0, 1, 1, 1, 2, 2, 2, 2, 3, 3, isto é, temos 10 dados estatísticos com os quais podemos fazer uma tabela de frequências e construir o respetivo gráfico de barras. Por exemplo, a barra referente ao valor 2 terá altura 4.

Passos & Correia (2010), p. 36



Faria *et al.* (2010), p. 118

Figura 3.3: Confusão entre dados e frequência

3.3 Cálculo da média quando a variável é contínua

Quando estamos perante um conjunto de dados reais de uma variável contínua e pretendemos calcular a média devemos efetuar os cálculos usando os dados reais. Só deveremos recorrer às marcas das classes quando não conhecemos os valores que pertencem a essas classes[(1)]. No manual de Duarte & Filipe (2010, p. 141) aparece o cálculo aproximado da média sem que nada seja referido, quando os dados reais são conhecidos. Mesmo que fosse uma explicação do procedimento a fazer noutras situações, deve ser dito que a média calculada a partir das marcas das classes é um valor aproximado e que, neste caso particular, não deve ser calculada deste modo.

3.4 Um erro comum na regressão linear

Nesta secção iremos focar a atenção no erro mais comum que detetamos nos manuais por nós consultados, a de utilização da mesma reta de regressão, obtida pelo método dos mínimos quadrados, para estimar um valor de x condicionado a um dado valor de y bem como para estimar um valor de y condicionado a um valor de x quando, corretamente, dever-se-iam utilizar duas retas distintas (exceto em alguns casos muito particulares onde as duas retas são análogas). Um exemplo ilustrativo é o livro de Jorge *et al.* (2010) que na página 83, apresenta um diagrama de dispersão que relaciona o Índice de Desenvolvimento Humano (IDH) de Portugal com o número de anos decorridos após 1975 e usa a reta de regressão $y = 0,004x + 0,7926$ para estimar o IDH para o ano 2008, mas também usa a mesma reta para saber o ano em que Portugal atingiu um determinado índice. Este erro será exemplificado utilizando um *software* (que é *freeware*) frequentemente utilizado no ensino da geometria no ensino básico e secundário, o *GeoGebra*, e que é uma potencial ferramenta no ensino da regressão linear (conforme salientaremos na secção 4.1.5, na página 71). Em muitos manuais utilizados no ensino secundário é utilizada a mesma reta para estimar o valor de y quando conhecemos um valor de x (condicionada a $x = x_0$) e para estimar o valor de x em função de um valor específico da variável y ($y = y_0$), o que não deveria ocorrer pelas razões previamente

(1) Esta situação é frequente quer por não termos acesso aos dados em algumas situações, quer pelo facto de, em alguns inquéritos, as perguntas surgirem logo em classes. Um exemplo ilustrativo desta situação é o rendimento mensal (pois ninguém responde o seu rendimento exato).

apontadas. Esta confusão deriva do desconhecimento por parte de muitos docentes, incluindo autores de manuais, de que as retas obtidas são distintas. Sublinhemos que, dos manuais por nós consultados, há um que apresenta corretamente este tema e, partindo de um exemplo simples, explica aos alunos a razão pela qual não devem usar a mesma reta de regressão em ambos os casos, referindo:

> "*A recta de regressão, de equação* $y = 0,797x + 121,282$*, foi construída para, dado o peso x, em kg, de um jogador prever a altura y, em cm, do mesmo (...). Os erros cometidos, relativamente aos valores de y medidos e os valores previstos são os comprimentos dos segmentos de reta assinalados na figura (...). A equação da reta de regressão é determinada, utilizando ferramentas matemáticas, de forma a minimizar a acumulação destes comprimentos (...). Percebe-se assim, que é possível utilizar a equação desta reta para prever valores da altura dado o valor do peso, mas, no entanto não se pode utilizar esta equação para prever o valor do peso dado o valor da altura*"
>
> [Negra & Martinho, 2010, pp.181-182]

Pelo facto da maioria dos professores do ensino secundário desconhecer a existência das referidas diferenças, é importante divulgá-las de forma a serem efetuadas as correções necessárias nos manuais, bem como alertar e clarificar o corpo docente para esta situação. A exploração de exemplos dinâmicos em *GeoGebra*, *software* torna óbvia a diferença entre os dois métodos. O cálculo das duas retas é automático e alterando os valores da nossa amostra (presentes na folha de cálculo do *GeoGebra*) permite visualizar a consequente alteração (sensibilidade) das retas bem como dos valores estimados pelos dois modelos, razão pela qual nos parece uma ferramenta adequada para o ensino da regressão linear conforme os objetivos estipulados no programa de matemática do ensino secundário.

Perante um erro tão generalizado, pretendemos divulgá-lo de forma a serem efetuadas as correções necessárias nos manuais, bem como alertar e clarificar o corpo docente para esta situação, razão pela qual apresentamos um poster acerca deste assunto no XIX Congresso da Sociedade Portuguesa de Estatística (SPE) que decorreu na Nazaré em setembro de 2011 (cf. Martins *et al*.(2011)).

3.5 Definições pouco claras

Qualquer definição ao dispor do aluno deve apresentar-se da forma mais clara possível, pois servirá de base para o trabalho que o aluno irá desenvolver à volta desse tópico e o aluno terá mais segurança se souber que para relembrar, facilmente, qualquer conceito, dispõe de textos esclarecedores. Parece-nos pouco clara, quer para alunos quer para professores, a definição de mediana dada por Negra & Martinho (2010) que apresenta "A mediana é o dado que ocupa a posição central da distribuição, quando ordenada" (Negra & Martinho, p. 149). Por um lado, atendendo ao conceito de dado estatístico, a mediana pode não ser um dado. Esta situação só é garantida se o número de dados for ímpar. Por outro lado, apenas com esta definição um aluno não conseguiria calcular a mediana de um conjunto de dados par, como por exemplo: 8, 9, 9, 12, 12, 15, uma vez que não é referido na definição que tem de calcular a média dos dois valores centrais.

3.6 Erros e/ou falta de clareza na notação

Os erros relativos à notação utilizada surgem com mais frequência nos manuais do ensino secundário, pois no ensino básico a linguagem simbólica usada é muito reduzida. Para analisarmos alguns deles podemos consultar, por exemplo, Duarte & Filipe (2010) na página 152, onde na demonstração de uma propriedade da média se lê:

"Sejam $x_1, x_2, x_3, \cdots, x_n$ os N valores da distribuição X, sendo

$$\overline{x} = \frac{x_1 + x_2 + x_3 \ldots + x_n}{N}. \text{"} \tag{3.1}$$

Quando observamos o que está escrito pode parecer-nos uma gralha e que basta trocar n por N. No entanto poderá fazer alguma confusão aos alunos pois na página anterior os autores usam a letra n para designar o número de valores diferentes da variável e usam N para designar o número total de valores da distribuição.

Também podemos verificar uma situação semelhante em Neves *et al.* (2010a, p. 129), onde aparece a seguinte fórmula da variância:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 n_i}{n-1}. \tag{3.2}$$

Os autores usam a letra *n* para designar o número de elementos da amostra e simultaneamente usam-na para designar o número de valores distintos da variável, quando deveriam ter usado notações diferentes.

Também em Costa & Rodrigues (2010b) surge confusão com as notações utilizadas. Na página 182 a letra *k* é utilizada para representar o número de valores distintos da variável. Duas páginas à frente *k* já representa a posição da mediana. Por outro lado, é usada a letra *N* para designar o total de dados quando estão agrupados e usada a letra *n* quando são dados simples. Ainda no mesmo manual usam a letra *N* para designar o número de dados da amostra na página 178 e na definição de mediana referem-se a *n* dados (sem referir amostra ou população). Desta análise e tendo em conta que existem mais situações para além das que foram aqui apresentadas, conclui-se que é difícil entender o que as letras representam neste capítulo do livro e que é urgente que se realize uma correção numa próxima edição.

Deste modo, parece-nos que, porventura, seria mais fácil se os autores simplesmente utilizassem a notação internacional (deste modo a notação usada em cada manual seria consistente, além de ser consistente com os restantes manuais).

Na tabela 3.1 propomos uma notação, baseada na notação internacional e que corresponde à que utilizamos no capítulo 2.

Notação	Descrição
n	dimensão da amosta
N	dimensão da população
$x_1, \cdots, x_n$	observações da amostra
$x_{(1)}, \cdots, x_{(n)}$	observações da amostra ordenadas
$x'_1, \cdots, x'_p$	modalidades diferentes na amostra
p	número de modalidades distintas
n_i	frequência absoluta de x'_i
N_i	frequência absoluta acumulada de x'_i
f_i	frequência relativa de x'_i
F_i	frequência relativa acumulada de x'_i
$\bar{x}$	média de uma amostra
μ	média de uma população
Mo	moda
Me	mediana
Q_1	1.º quartil
Q_3	3.º quartil
P_α	Percentil α
k	número de classes
h_i	amplitude da i-ésima classe
I_T	amplitude total
I_Q	amplitude Interquartis
s^2	variância de uma amostra
σ^2	variância de uma população
s	desvio padrão de uma amostra
σ	desvio padrão de uma população
ρ	coeficiente de correlação da população
r	coeficiente de correlação da amostra

Tabela 3.1: Sugestão de notação

Capítulo 4

Materiais e sugestões metodológicas

Neste capítulo vamos apresentar algumas propostas de trabalho para a sala de aula com recurso ao *software GeoGebra* (irá ser utilizada a versão 4). Este programa é gratuito (*disponível em www.geogebra.org*) e muito utilizado em Geometria por ser dinâmico. Neste trabalho pretendemos explorar as suas potencialidades no ensino e aprendizagem da estatística e, por isso, apresentamos em primeiro lugar uma explicação mais detalhada sobre os vários comandos que serão utilizados nessas propostas de trabalho.

4.1 O *GeoGebra* no ensino da Estatística

O *GeoGebra* é uma potencial ferramenta que pode ser explorada no ensino da Estatística. Com este *software* podemos construir tabelas de frequência, vários tipos de gráficos (gráficos de barras, gráficos de pontos, histogramas, diagramas de caule e folhas, diagramas de extremos e quartis e diagramas de dispersão), calcular quase todas as medidas estatísticas que são lecionadas no ensino básico e no ensino secundário, e pelo facto de ser um *software* dinâmico, podemos alterar os dados e verificar os efeitos dessas alterações quer nos gráficos quer nas medidas estatísticas, permitindo fazer várias explorações dos conceitos. Esta possibilidade de usar a tecnologia nas aulas de Matemática permitirá melhorar as oportunidades de aprendizagem dos alunos se aproveitarmos aquilo que a tecnologia faz de forma "correcta e eficiente –construção de gráficos, visualização e cálculo." NCTM (2008, p. 27).

4.1.1 Inserir dados no *GeoGebra*

A janela principal do *GeoGebra* dispõe de uma barra de menus, uma barra de ferramentas, uma zona algébrica, uma zona gráfica e uma entrada de comandos. Também podemos visualizar uma folha de cálculo onde podemos introduzir os dados que constituem a amostra e a partir daí temos a possibilidade de construir tabelas e vários gráficos, como de seguida vamos explicar. Para visualizar a folha de cálculo deve-se selecionar na barra de ferramentas *Exibir* e de seguida escolher *Folha de Cálculo*. Surge, então, do lado direito (cf. Figura 4.1) uma folha de cálculo, contendo várias colunas $A, B, C, \cdots$. Numa delas introduzimos os dados não classificados e, utilizando o botão do lado direito do rato, selecionarmos *Criar lista*. A lista obtida aparece na folha algébrica (lado esquerdo) designada por *lista1*, havendo possibilidade de alterar o seu nome. Para isso, com o cursor colocado em *lista1*, basta utilizar o botão do lado direito do rato, clicar em *propriedades dos objetos* (também se pode obter o mesmo efeito clicando em *renomear*) e escrever o nome que se pretende na janela *nome* (consideremos, por exemplo, idade). Refira-se que os dados da Tabela 2.1 presente na página 13 podem ser introduzidos na folha de cálculo, podendo à primeira coluna chamar-se *cor*, à segunda coluna *leitura*, à terceira coluna *idade*, à quarta coluna *altura*, à quinta *Mat* e, finalmente, à sexta coluna *CFQ*.

Na folha de cálculo podemos igualmente colocar os dados classificados, usando a primeira coluna para colocar os diferentes valores da variável (lista1) e a segunda coluna para as respetivas frequências (lista2). Este procedimento pode ser útil na construção de gráficos, nomeadamente o diagrama de barras, como poderemos ver na secção 4.1.3. De seguida vamos apresentar como se constroem tabelas de frequência.

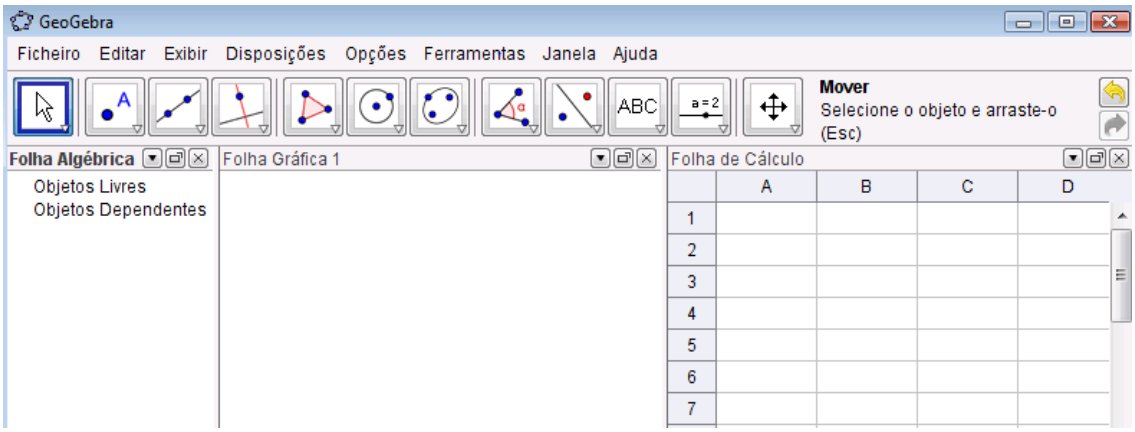


Figura 4.1: Janela principal do *GeoGebra* e folha de cálculo

4.1.2 Construção de tabelas de frequência

Após termos criado a lista de dados podemos contruir tabelas de frequências colocando na *Entrada de comandos* cada um dos comandos que a seguir apresentamos, clicando de seguida em *Enter*.

- TabelaFrequências[<Lista de Dados>] – Permite obter uma tabela de frequências absolutas a partir dos dados não classificados.
- TabelaFrequências[True, <Lista dos Dados>] – Permite obter uma tabela de frequências absolutas acumuladas a partir dos dados não classificados.
- TabelaFrequências[<Lista dos Limites das Classes>,<Lista dos Dados>] – Permite obter uma tabela de frequências absolutas para uma variável contínua.
- TabelaFrequências[True,<Lista dos Limites das Classes>,<Lista dos Dados>] – Permite obter uma tabela de frequências absolutas acumuladas para uma variável contínua.

Assim, se pretendermos uma tabela de frequências absolutas da variável *idade*, ver Tabela 2.4 com os mesmos dados (página 17), podemos usar o comando "TabelaFrequências[idade]" e obter uma tabela como a que se apresenta na Figura 4.3. Podemos proceder de modo idêntico no caso de uma variável qualitativa. No caso da variável cor preferida dos alunos, usamos a lista *cor* e, através do comando "TabelaFrequências[cor]" obtemos a tabela da Figura 4.2 cujos dados constam na Tabela 2.3 da página 16.

Valor	Contagem
amarelo	4
azul	7
branco	4
corderosa	7
verde	3

Figura 4.2: Tabela de frequências absolutas da variável cor preferida

Se pretendermos uma tabela de frequências acumuladas escrevemos na entrada "TabelaFrequências[True, idade]". Se, por outro lado, estamos interessados numa tabela com os dados organizados em classes, nomeadamente as alturas dos 25 alunos, além da lista das alturas, designada por *altura*, criamos outra lista com os limites das classes {145, 150, 155, 160,

Valor	Contagem
13	6
14	13
15	5
16	1

Figura 4.3: Tabela de frequências absolutas da variável idade

165 e 170} que se pode designar por *limites* e usa-se o comando "TabelaFrequências[limites, altura]" ou também pode ser escrita diretamente no comando do seguinte modo: "TabelaFrequências[{145, 150, 155, 160, 165 e 170}, altura]". Para obter a lista dos limites das classes podemos recorrer à Regra de Sturges. Na tabela obtida, as classes não aparecem escritas usando parênteses retos, no entanto a contagem é feita corretamente, sendo cada classe fechada à esquerda e aberta à direita. Apresentamos na Figura 4.4, a título de exemplo, a tabela de frequências absolutas para a variável *altura* construída com este *software*, e que já apresentamos de outra forma na página 19. Obteríamos de modo semelhante a tabela de frequências absolutas acumuladas, introduzindo o comando "TabelaFrequências[True, limites, alturas]"

Intervalo	Contagem
145 – 150	2
150 – 155	5
155 – 160	8
160 – 165	6
165 – 170	4

Figura 4.4: Tabela de frequências da variável altura

4.1.3 Representações gráficas

A construção de vários tipos de gráficos está muito facilitada recorrendo ao *GeoGebra*. Assim vamos apresentar de seguida os comandos utilizados para obter alguns gráficos que aparecem no capítulo 2 deste trabalho.

- GráficoPontos[<Lista de Dados Não Classificados>] – Permite obter um gráfico de pontos a partir dos dados não classificados.

- DiagramaBarras[<Lista dos Dados não classificados>, <Largura das barras>] – Permite obter um gráfico de barras de frequências absolutas a partir dos dados não classificados.

- DiagramaBarras[<Lista dos Dados classificados>, <Lista das frequências>,<Largura das barras>] – Permite obter um gráfico de barras a partir dos dados classificados.

- DiagramaCauleFolhas[<Lista>] – Permite obter um diagrama de caule e folhas a partir da lista de dados não classificados.

- DiagramaCauleFolhas[<Lista>, <Ajustamento (-1 | 0 | 1)>] – Permite obter um diagrama de caule e folhas a partir da lista de dados não classificados e fazer o ajustamento, em que o aumento de uma unidade corresponde a multiplicar por dez o valor do caule.

- DiagramaExtremosQuartis[<Ordenada>, <Semialtura>, <Lista de Dados Não Classificados>] – Permite obter um diagrama de extremos e quartis a partir da lista de dados não classificados. O valor da ordenada está relacionado com a distância ao eixo das abcissas e a semi-altura dá-nos a largura do retângulo que contém os dados à volta da mediana.

- DiagramaExtremosQuartis[<Ordenada>, <Semialtura>, <Mínimo>, <Quartil1>, <Mediana>, <Quartil3>, <Máximo>] – Permite obter um diagrama de extremos e quartis a partir dos valores dos extremos e dos quartis.

- Histograma[<Lista dos Limites das Classes>, <Lista das Frequências>] – Permite obter um histograma a partir da lista dos limites das classes e da lista das frequências.

- Histograma[<Lista dos Limites das Classes>, <Lista dos Dados>, <Densidade (true | false)>, <Escala (opcional)>] – Permite obter um histograma a partir da lista dos limites das classes e da lista dos dados não classificados. Caso se pretenda um histograma de frequências relativas , na densidade seleciona-se *true*.

- Histograma[<Acumulada (True | false)>, <Lista dos Limites das Classes>, <Lista dos Dados>, <Densidade (true | false)>, <Escala (opcional)>] – Permite obter, por exemplo, um histograma de frequências absolutas acumuladas, selecionando primeiro *true* e depois na densidade *false* ou um histograma de frequências relativas acumuladas, selecionando *true* em Acumulada e também em Densidade.

Para representar um gráfico de pontos, como por exemplo, o gráfico da Figura 2.2 presente na página 22, podemos usar o comando "GráficoPontos[idade]", onde *idade* designa a lista de todas as idades.

Para representar um gráfico de barras, como por exemplo, o gráfico da Figura 2.3 presente na página 23, podemos usar o comando "DiagramaBarras[idade, 0.5]", onde *idade* designa a lista de todas as idades. Para que as barras não sejam adjacentes, a largura das barras deve ser inferior a 1. Outro processo para obter o mesmo gráfico é usar o comando "DiagramaBarras[idade1, freq, 0.5]", onde *idade1* é a lista das idades diferentes e *freq* é a lista das frequências.

Para construir um diagrama de caule e folhas como o da Figura 2.9, que se pode visualizar na página 27, usou-se o comando "DiagramaCauleFolhas[idade]", em que *idade* designa a lista de todas as idades. Também poderíamos ter usado o comando "DiagramaCauleFolhas[idade,-1]", considerando o ajustamento igual a -1 (define o valor do caule).

Para construir um diagrama de extremos e quartis, como o da Figura 2.14 que se apresenta na página 37, usou-se o comando "DiagramaExtremosQuartis[8, 3, 13, 13.5, 14, 14.5, 16]". Também poderíamos ter usado o comando "DiagramaExtremosQuartis[8, 3, idade]".

Para representar o histograma da Figura 2.5 (página 24) uma das alternativas é usar o comando "Histograma[limites , freq1]", onde limites é a lista dos limites das classes e freq1 é a lista das frequências. Se recorrermos à lista da totalidade dos dados usamos o comando "Histograma[limites, altura, false]". Para obtermos o polígono de frequências da Figura 2.6, presente na página 25, recorremos também às potencialidades do *GeoGebra* no âmbito da Geometria, construindo segmentos de reta cujos extremos são pontos médios de outros segmentos de reta. Para obter o gráfico patente na Figura 2.8 (página 26) usamos o comando "Histograma[True, limites, altura, True]" e mais uma vez as potencialidades do *Geogebra* na Geometria.

Já vimos algumas potencialidades do *GeoGebra*, nomeadamente, na elaboração de tabelas de frequêcia e gráficos. Vamos passar de seguida ao cálculo de medidas estatísticas, apresentando os comandos necessários.

4.1.4 Cálculo de medidas estatísticas

Depois de inserirmos os dados na folha de cálculo e de criarmos uma lista podemos calcular as medidas de localização e de dispersão colocando na janela de entrada cada um dos comandos que constam da Tabela 4.1.

Medidas estatísticas	Comando
média	Média[<Lista de Números>]
moda	Moda[<Lista de Números>]
mediana	Mediana[<Lista de Números>]
quartil 1	Q1[<Lista de Números>]
quartil 3	Q3[<Lista de Números>]
percentil	Percentil[<Lista de Números>, <Valor do Percentil>]
variância	VariânciaAmostra[<Lista de Números>]
desvio padrão	DesvioPadrãoAmostra[<Lista de Números>]

Tabela 4.1: Comandos para o cálculo de medidas estatísticas com o *GeoGebra*

Os valores das medidas estatísticas vão aparecendo na folha algébrica representados por letras, assim como a lista dos dados. Cada uma dessas letras deve ser substituída por um nome sugestivo de forma a não haver confusão. Como exemplo, apresentamos na Figura 4.5 as medidas estatísticas referentes às idades dos 25 alunos.

Uma das vantagens deste *software* no estudo da Estatística é o facto de ser dinâmico o que permite visualizar, de imediato, os efeitos, nas medidas estatísticas, da alteração de um ou mais dados da amostra. Vamos supor que substituíamos uma idade de treze anos por uma de dezoito anos. Do mesmo modo que a folha de cálculo foi alterada, também as medidas estatísticas são imediatamente atualizadas, como se pode ver na Figura 4.6. Facilmente se vê que a moda e a mediana não se alteraram, mas a média e o desvio padrão aumentaram o que confirma uma maior dispersão das idades e, por outro lado, que a moda e a mediana não são influenciadas pelos valores extremos. Refira-se que este programa também indica o valor mínimo e o valor máximo. Neste caso bastava escrever Mínimo[notas] e Máximo[notas], respetivamente.

Outras situações deste tipo irão ser exploradas neste trabalho, sendo possível beneficiar de

Figura 4.5: Medidas estatísticas para a variável idade

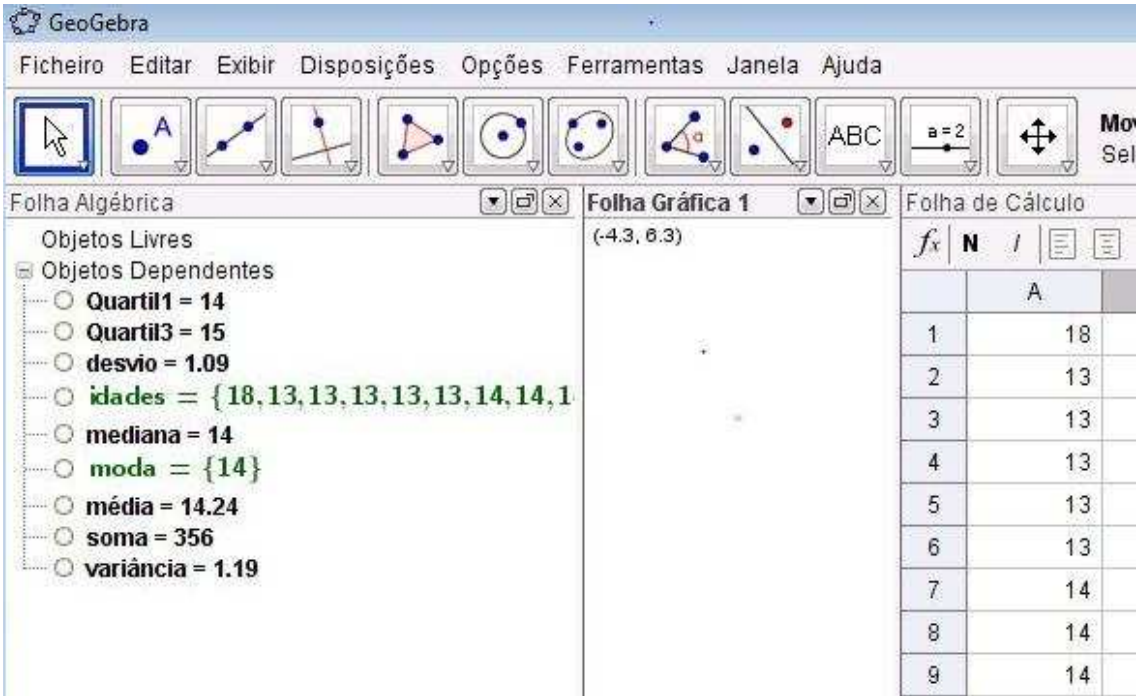


Figura 4.6: Medidas estatísticas da idade, após a alteração de um valor

outras potencialidades do *GeoGebra* tais como a *análise univariada* e a *análise multivariada*. Depois de registar os dados na folha de cálculo e de os selecionar, se clicarmos no *icon* da barra de ferramentas que apresenta um gráfico e escolhermos *análise univariada*, aparece-nos uma lista com medidas estatísticas e uma representação gráfica, entre aquelas que o *software* permite construir. Como exemplo, apresentamos na Figura 4.7 as estatísticas da variável idade e um diagrama de extremos e quartis semelhante ao diagrama já apresentado na página 37. Se alterarmos a designação do gráfico, passaremos de uns para os outros. Tudo isto é feito

automaticamente, sem qualquer comando. Mais uma vez podemos ver de uma forma simples os efeitos nas medidas estatísticas ou nos gráficos, de várias alterações que podem ser feitas na lista de dados. Outra possibilidade que o *GeoGebra* nos dá é a construção de dois diagramas de extremos e quartis empilhados, a partir de duas listas de dados, permitindo compará-las. Para isso, selecionamos as duas listas inseridas na Folha de cálculo e clicamos no *icon* da barra de ferramentas que apresenta um gráfico e escolhemos *análise multivariada*. Para além dos diagramas de extremos e quartis também surge uma tabela com as medidas estatísticas referentes aos dois conjuntos de dados.

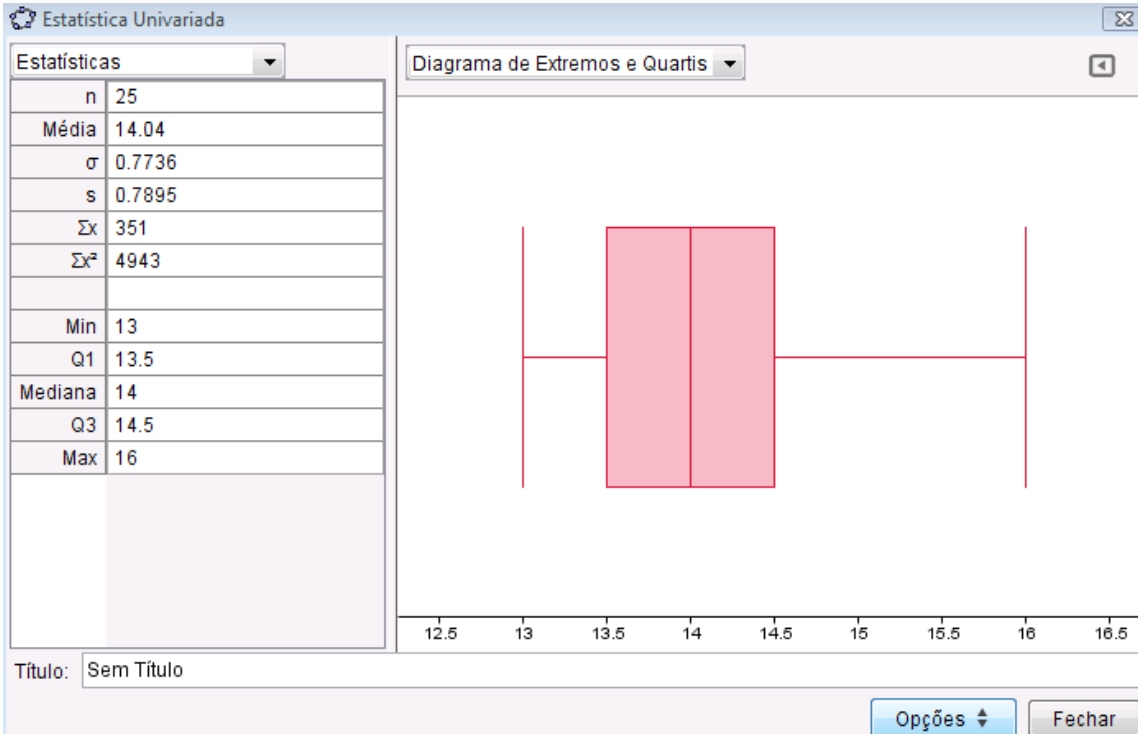


Figura 4.7: Análise univariada para a variável idade

4.1.5 Regressão linear

Figuras análogas às representadas na Figura 2.19 (página 53) podem ser construídas em sala de aula com recurso ao *software GeoGebra*, tendo a vantagem de poderem ser dinâmicas. Para este fim, no *GeoGebra*, podemos exibir a *Folha de Cálculo* e utilizar as duas primeiras colunas (*A* e *B*) para definir as coordenadas dos pontos a utilizar para a regressão linear. Com os pontos definidos podemos criar uma lista de pontos (bastará selecionar as coordenadas e, utilizando o botão do lado direito do rato, selecionar *Criar lista de pontos*, que, neste caso, cada ponto

terá duas coordenadas). Desta forma será criada a *lista1* que contém os n pontos que correspondem ao nosso conjunto de dados (x_i, y_i). Podemos de seguida determinar o coeficiente de correlação, através do comando "CoeficienteDeCorrelação[lista1]"; a reta de regressão de y em função de x de acordo com a equação (2.35), presente na página 49, recorrendo ao comando "RegressãoLinear[lista1]" e, no mesmo gráfico, representar a reta de regressão de x em função de y (cf. equação (2.41)) utilizando o comando "RegressãoLinearX[lista1]". As duas retas podem ser representadas por cores diferenciadas de forma a podermos destinguí-las facilmente. Desta forma, ao ser alterado um ou mais pontos da lista1, quer o coeficiente de correlação quer as retas serão automaticamente ajustadas à nova nuvem de pontos, o que permitirá visualizar facilmente a diferença entre as duas retas, bem como observar as alterações no coeficiente de correlação. Em sala de aula, pode-se explorar a sensibilidade do coeficiente de correlação bem como da reta ajustada a alterações de pontos, em particular à existência de pontos mais afastados (*outliers*). Caso pretendamos analisar igualmente os valores estimados para y para um dado valor de x, bastará definir um ponto A [de coordenadas $(x_0, 0)$] no eixo das abcissas, os segmentos de reta paralelos ao eixo das ordenadas que liguem o ponto A a cada uma das retas estimadas (sejam P_1 e P_2 os pontos das retas com abcissa x_0). A ordenada dos pontos P_1 e P_2 correspondem às estimativas obtidas com cada uma das retas (valores $\widehat{y}_0$ e $\widetilde{y}_0$ representados nos gráficos da Figura 2.19 da página 53). Para melhor visualizar o valor das estimativas obtidas podemos criar segmentos de retas paralelos ao eixos das abcissas que liguem P_1 e P_2 ao eixo das ordenadas (de forma análoga à apresentada nos gráficos da Figura 2.18). Com esta construção, caso alteremos a abcissa x_0 do ponto A, as estimativas obtidas serão automaticamente ajustadas, permitindo, desta forma, observar as suas diferenças quando utilizamos valores para x_0 mais próximos ou mais afastados de $\overline{x}$.

4.2 Propostas de trabalho para a sala de aula

Nesta secção vamos apresentar algumas propostas de trabalho para a sala de aula com recurso ao *GeoGebra*. Algumas destas propostas serão mais orientadas para o ensino básico e outras serão elaboradas para serem aplicadas no ensino secundário. Procurou-se aliar o uso do computador à análise de dados reais, conforme é recomendado em vários documentos tais como NCTM (2007), Branco (2000), Ponte *et al.* (2007). Para cada proposta serão definidos

objetivos que se pretendem alcançar, tendo em conta os programas atuais. Tendo em conta que o professor necessitará de uma sala com computadores para a aplicação das propostas, na impossibilidade de haver um computador para cada aluno, propomos a sua realização a pares.

4.2.1 Proposta 1 – Quantas pessoas vivem em minha casa?

Esta proposta (presente no anexo A.1) destina-se aos alunos do 1.º ciclo e tem como objetivos:

- recolher dados registando-os através de tabelas de frequências absolutas, gráficos de pontos e gráficos de barras;
- ler, explorar, interpretar e descrever tabelas e gráficos;
- identificar a moda num conjunto de dados.

Os alunos já devem ter o conceito de moda, saber o que é uma tabela de frequências absolutas e conhecer gráficos de pontos e de barras. Um dos aspetos a privilegiar nesta proposta é o facto dos dados serem significativos para os alunos. Os alunos, na própria aula, devem responder à questão "Contando contigo, quantas pessoas vivem em tua casa?". O professor pode usar uma lista dos alunos da turma que vai preenchendo no computador e depois projeta ou simplesmente escreve os nomes no quadro e as respetivas respostas. Deve ser transmitido aos alunos que antes de construírem um novo gráfico podem guardar o trabalho anterior e abrir um novo documento (havendo a possibilidade de copiar a lista de dados). Deve chamar-se a atenção de que os gráficos de pontos podem evoluir para gráficos de barras. O item em que se pede para o aluno formular questões revela-se importante para desenvolver a comunicação matemática e deve haver um momento para a apresentação e discussão das questões formuladas pelos alunos. Uma vez que o *GeoGebra* é um *software* dinâmico, o professor pode propor a alteração dos dados iniciais, fazendo questões como por exemplo "Suponham que o pai da Maria (aluna da turma) vai trabalhar para o estrangeiro. Acham que a moda se altera?" Os alunos devem dar a sua opinião, mas depois deverão confirmar, alterando a lista de dados e construindo de novo a tabela, um dos gráficos ou simplesmente usando o comando "Moda[lista1]". O professor deve ter a sensibilidade de propor alterações que façam surgir outro valor para a moda ou até mesmo duas modas. Também se deve discutir se é ou não razoável generalizar os resultados obtidos para todos os alunos da escola (de forma a começar

a transmitir a ideia de amostra como representativa da população, explorando situações de enviesamento).

4.2.2 Proposta 2 – Classificações obtidas num teste de Matemática

Esta proposta (presente no anexo A.2) destina-se aos alunos do 2.º ciclo e tem como objetivos:

- compreender e determinar a média aritmética de um conjunto de dados;
- compreender e determinar os extremos e a amplitude de um conjunto de dados;
- construir e interpretar diagramas de caule e folhas.

Os alunos já deverão ter construído diagramas de caule e folhas com papel e lápis e já devem ter os conceitos de média, extremos e amplitude. Com esta tarefa pretende-se que o aluno explore estes conceitos, nomeadamente, compreenda como pequenas alterações nos dados influencia ou não a média, a moda e a amplitude. Por outro lado, ao trabalharem com dados quantitativos com estas características, devemos questionar relativamente ao gráfico mais adequado, de entre aqueles que os alunos conhecem, permitindo que desenvolvam o espírito crítico. Uma vez que os alunos também já aprenderam a construir tabelas de frequências, poderá perguntar-se qual a tabela mais adequada para resumir estes dados. O diagrama de caule e folhas pode sugerir classes para organizar os dados numa tabela. Refira-se também a importância das alíneas que permitem estabelecer conexões com o tema "Números e operações". Dependendo das características dos alunos, o professor poderá ir mais além perguntando, por exemplo, quais seriam as alterações na média, moda, extremos e amplitude se os alunos no próximo teste subissem todos 2 valores relativamente ao anterior. Os alunos devem estabelecer as suas conjeturas e verificá-las usando novamente a folha de cálculo.

4.2.3 Proposta 3 – Meio de transporte utilizado para chegar à escola

Esta proposta (presente no anexo A.3) destina-se aos alunos do 2.º ciclo e tem como objetivos:

- construir e interpretar tabelas de frequências absolutas e relativas;
- construir e interpretar um gráfico circular.

Nesta proposta de trabalho os alunos irão construir um gráfico circular, utilizando as potencialidades do *GeoGebra*. A realização desta atividade vai permitir estabelecer conexões entre os temas Tratamento de Dados e Geometria. O aluno já deve ter construído um gráfico circular usando papel e lápis e deve agora recordar como se determinam as amplitudes dos setores que compõem o gráfico. Primeiro, os alunos constroem o gráfico circular referente aos dados fornecidos pelo professor numa tabela, mas depois deverão recolher os dados na turma e fazer o respetivo gráfico. Caso os alunos ainda não tenham trabalhado com o *GeoGebra* em geometria, o professor deverá começar por explicar-lhes os ícones da barra de ferramentas que serão utilizados.

4.2.4 Proposta 4 – Classificações internas *versus* classificações externas na disciplina de Matemática

Esta proposta (presente no anexo A.4) destina-se aos alunos do 3.º ciclo e do 10.º ano e tem como objetivos:

- escolher as medidas de localização mais adequadas para resumir a informação contida nos dados;
- compreender e determinar a mediana, os quartis e a amplitude interquartis de um conjunto de dados;
- utilizar as medidas de localização na interpretação de um conjunto de dados;
- comparar distribuições e tirar conclusões.

Os alunos devem conhecer os quartis, a amplitude interquartis e também já devem ter construído com papel e lápis um diagrama de extremos e quartis. Pretende-se rentabilizar as potencialidades do *GeoGebra* ao nível da folha de cálculo, na determinação dos quartis e na construção de diagramas de extremos e quartis, evitando assim cálculos repetitivos e rotineiros. Os conjuntos de dados apresentados têm a mesma mediana, logo esta medida não será a mais indicada para comparar estas distribuições uma vez que não evidencia as diferenças. O aluno usará as medidas de dispersão (amplitude e amplitude interquartis) assim como a observação dos diagramas para estabelecer uma comparação das distribuições. Salienta-se o facto deste tipo de questões permitir o desenvolvimento da comunicação matemática. No caso da

proposta ser desenvolvida por alunos do 10.º ano poderá ser solicitado o cálculo do desvio padrão. De modo a facilitar a realização da tarefa, o professor poderá fornecer os dados aos alunos num ficheiro.

4.2.5 Proposta 5 – Salários dos trabalhadores de uma empresa

Esta proposta (presente no anexo A.5) destina-se aos alunos do 3.º ciclo e do 10.º ano e tem como objetivos:

- compreender e determinar a média de um conjunto de dados e indicar a adequação da sua utilização num dado contexto;
- escolher as medidas de localização mais adequadas para resumir a informação contida nos dados.

Nesta proposta abordam-se as três medidas de tendência central, suas vantagens e desvantagens. Refira-se que a moda e a mediana dos salários são iguais. A razão que nos levou a elaborar muitas questões à volta do conceito da média deve-se ao facto de esta medida ser a mais utilizada e ser importante proporcionar momentos que levem os alunos a entender de que forma esta medida é influenciada pelos valores muito elevados ou muito baixos, ou simplesmente por um valor qualquer. Os alunos deverão compreender que a escolha de uma das medidas de tendência central para representar os dados depende do contexto. De modo a facilitar a tarefa o professor poderá fornecer aos alunos um ficheiro com os 72 dados. Caso isso não aconteça o professor deve alertar os alunos para a necessidade de fazer uma cópia dos dados originais (copiar, por exemplo, para a lista B) uma vez que na última questão o aluno deverá recorrer à lista inicial e não àquela que já foi alterada na questão 7. De modo a facilitar a realização da tarefa, o professor poderá fornecer os dados aos alunos num ficheiro. Algumas questões desta proposta foram inspiradas em Zawojewski (1992).

4.2.6 Proposta 6 – Comparação de duas turmas

Esta proposta (presente no anexo A.6) destina-se aos alunos do 10.º ano e tem como objetivo:

- compreender e determinar o desvio padrão de um conjunto de dados.

Os alunos já devem conhecer o conceito de desvio padrão. Nesta proposta pretende-se comparar duas turmas que apresentam a mesma média nas classificações a Matemática. Pretende-se que, inicialmente, os alunos calculem a média sem recorrer ao *GeoGebra* uma vez que os cálculos são muito simples, no entanto, depois de introduzirem os dados na folha de cálculo, poderão proceder à confirmação. Os alunos deverão compreender que as distribuições são muito diferentes por apresentarem variabilidades diferentes relativamente à média. O cálculo do desvio padrão, uma das medidas de dispersão mais utilizada, vem confirmar as diferenças. Algumas questões desta proposta foram inspiradas em Zawojewski (1992)

4.2.7 Proposta 7 – Peso e altura dos alunos de uma turma do 10.º ano

Esta proposta (presente no anexo A.7) destina-se aos alunos do 10.º ano e tem como objetivos:

- construir diagramas de dispersão;
- interpretar a reta de regressão e conhecer as suas limitações;
- distinguir a regressão de x em ordem a y da regressão de y em ordem a x.

Esta proposta prevê a recolha de dados na turma. Por vezes os alunos não gostam que o seu peso seja conhecido. Caso o professor entenda poderá levar os dados e ultrapassar essa situação. Chama-se especial atenção para o facto da maioria dos manuais escolares não explorar a representação das duas retas de regressão (a reta de regressão de y em função de x e a reta de regressão de x em função de y). Pretende-se que os alunos compreendam que são distintas, podendo ser feitas algumas experiências que consistem em alterar um ou mais pontos da lista e verificar as alterações nas retas e no coeficiente de correlação. É habitual, em alguns manuais, pedir-se estimativas de y em relação a x e de x em relação a y, usando a mesma reta. Os alunos deverão compreender que tal situação é incorreta pois as retas são distintas. Para o confirmarem bastará definir um ponto A [de coordenadas $(x_0, 0)$] no eixo das abcissas, os segmentos de reta paralelos ao eixo das ordenadas que liguem o ponto A a cada uma das retas estimadas definem os pontos P_1 e P_2 das retas com abcissa x_0. As ordenadas dos pontos P_1 e P_2 correspondem às estimativas obtidas com cada uma das retas. Esta situação pode ser explorada usando as alíneas 8 e 9 da proposta.

Capítulo 5

Conclusão

Atualmente é fundamental que o cidadão comum tenha a capacidade de compreender a informação estatística que lhe chega diariamente de várias formas. Os currículos têm-se alterado no sentido de desenvolver nos alunos esta competência. Assim, desde o reajustamento do Programa de Matemática para o ensino básico, o tema matemático "Organização e Tratamento de Dados" percorre todo o ensino básico, desde o 1.º ciclo até ao 3.º ciclo e para cada ciclo foi definido um propósito principal de ensino que se assume como a orientação mais importante para o ensino deste tema. No caso particular do terceiro ciclo pode ler-se "Desenvolver nos alunos a capacidade de compreender e de produzir informação estatística bem como de a utilizar para resolver problemas e tomar decisões informadas (...)." (Ponte *et al.*, 2007, p. 59). Acrescente-se ainda que os tópicos estudados até ao 3.º ciclo são posteriormente desenvolvidos no 10.º ano no tema "Estatística", lecionado no terceiro período.

Dada a importância desta área no ensino, ao realizarmos este trabalho, consideramos útil a elaboração de um texto, direcionado aos professores, que aposta no rigor e apresenta os conceitos fundamentais de Estatística lecionados nestes ciclos de ensino; salientamos algumas incorreções detetadas nos materiais disponíveis para o ensino da Estatística; apresentamos uma secção onde se aprofunda a regressão linear, pois encontramos um erro comum neste tema e, por fim, criamos propostas de trabalho para a sala de aula com recurso à tecnologia, permitindo a exploração dos conceitos de uma forma prática, evitando deste modo que a estatística seja reduzida à repetência de cálculos fastidiosos desprovidos de significado (cf. Carvalho (2006)).

Sabendo que o número de recursos disponíveis para o ensino da Estatística que incluam

a utilização do *GeoGebra* é escasso, situação que se confirmou aquando da análise de um conjunto de materiais disponíveis, este projeto pretende ser um contributo importante para o trabalho dos professores no ensino desta área da matemática, uma vez que nele se mostra como operacionalizar as orientações atuais e como implementar tarefas seguindo essas orientações. Sabendo que a tecnologia é crucial no tratamento de dados, recorremos ao *GeoGebra* - um *software* inovador no ensino e aprendizagem da Estatística que pode ser uma alternativa a outros *softwares*/instrumentos, por ser gratuito e dinâmico. Ao apresentarmos um conjunto de propostas de trabalho tivemos a preocupação de incluir os objetivos e escrever alguns comentários de modo a facilitar a sua aplicação na sala de aula. Refira-se que a maioria das propostas explora o facto deste *software* ser dinâmico, permitindo ao aluno alterar os dados escritos na folha de cálculo e verificar rapidamente os efeitos dessas alterações quer nos gráficos quer nas medidas estatísticas. Como refere Zawojewski "A folha de cálculo recalcula todos os dados de uma só vez e indica os resultados imediatamente. Assim, o professor pode concentrar-se no efeito de introduzir certas alterações que, de outro modo, poderiam perder-se no pântano dos cálculos individuais" (Zawojewski, 1992, p. 33).

Uma vez que este tema é apresentado nos manuais com algumas imprecisões, torna-se necessário que investigadores e professores se debrucem sobre estas questões, desenvolvendo um trabalho cada vez mais rigoroso ao nível do ensino da Estatística de modo a formar cidadãos cada vez mais ativos, interventivos e críticos.

Deste modo, espera-se que o presente trabalho possa dar um pequeno contributo para o incremento da qualidade do ensino da Estatística.

Referências Bibliográficas

APM (1998). *Matemática 2001: Diagnóstico e recomendações para o ensino e aprendizagem da Matemática*. Lisboa: APM.

BATANERO, C. & GODINO, J.D. (2003). *Estocástica y su Didáctica para Maestros*, Departamento de Didáctica de la Matemática, Facultad de Ciencias de la Educación, Universidad de Granada (disponível em `http://www.ugr.es/~jgodino/edumat-maestros`).

BRANCO, J. & MARTINS, M. (2002). Literacia Estatística, *Educação e Matemática*, n.º 69, pp. 9–13.

BROWN, P.J. (1993). *Measurement, Regression, and Calibration*, Oxford University Press.

CARVALHO, C. (2006). Desafios à educação estatística, em Ensino e Aprendizagem da Estatística, *Boletim da SPE*, Outono, pp. 7–8.

COSTA, B. & RODRIGUES, E. (2010a). *Novo Espaço-Parte 2 Matemática A 10.º Ano*, Porto Editora, Porto.

COSTA, B. & RODRIGUES, E. (2010b). *Novo Espaço-Parte 2 Matemática 7.º Ano*, Porto Editora, Porto.

DUARTE, T.O. & FILIPE, J.P. (2010). *Matemática Dez -Parte 2*, Lisboa Editora, Lisboa.

FARIA, L.; GUERREIRO, L. & ALMEIDA, P.R. (2010). *Matemática Dinâmica- Matemática 7.º Ano Parte 2*, Porto Editora, Porto.

FERNANDES, J.A. (2009). Ensino e aprendizagem da estatística - Realidades e Desafios, *Actas do XIX EIEM*, Vila Real.

INE, I.P. (2009). *Um mundo para conhecer os números*, INE-DREN-ESTP, Lisboa.

JORGE, A.M.B.; ALVES, C.B.; FONSECA, C.C.G.; BARBEDO, J. & SIMÕES, M. (2010). *Matemática A 10- Parte 3*, Areal Editores, Lisboa.

LOURA, L. (2009). Organização e Tratamento de Dados no Novo Programa de Matemática do Ensino Básico, *Educação e Matemática*, n.º 105, pp. 46–49.

LOUREIRO, C.; OLIVEIRA, F. & BRUNHEIRA, L. EDS. (2000). *Ensino e Aprendizagem da Estatística*, SPE, APM e DEIO–FCUL, Lisboa.

MAGRO, F.C.; FIDALGO, F. & LOUÇANO, P. (2010). π 7 - *Matemática 7.º ano*, ASA, Lisboa.

MARTINHO, M.H. & VISEU, F. (2009). Desenvolvimento da literacia estatística em dois manuais escolares do 7.º ano de escolaridade, *Actas do XIX EIEM*, Vila Real.

MARTINS, A., RIBEIRO, H. & SANTOS, R. (2011). Estatística no ensino secundário - um contributo para a clarificação do estudo da regressão linear simples, SPE 2011 - Programa e Resumos, pp. 283-284.

MARTINS, M.E.G.; MONTEIRO, C.; VIANA, J.P. & TURKMAN, M.A.A. (1997). *Estatística*, ME-DES, Lisboa.

MARTINS, M.E.G. & CERVEIRA, A. (1999). *Introdução às Probabilidades e à Estatística*, Universidade Aberta.

MARTINS, M.E.G. (2005). *Introdução à Probabilidade e à Estatística*, Sociedade Portuguesa de Estatística, Lisboa.

MARTINS, M.E.G.; LOURA, L. & MENDES, M. (2007). *Análise de Dados - texto de apoio para os Professores do 1.º ciclo*, ME-DGIDC, Lisboa.

MARTINS, M.E.G & PONTE, J. (2010). *Organização e Tratamento de Dados*, ME-DGIDC, Lisboa.

MONTGOMERY, D.C., PECK, E.A. & VINING, G.G. (2006). *Introduction to Linear Regression Analysis*, 4th Ed., Wiley Series in Probability and Statistics, John Wiley & Sons.

MURTEIRA, B. (1993). *Análise exploratória de dados - Estatística Descritiva*, McGraw-Hill, Lisboa.

NASCIMENTO, M. (2009). Literacia Estatística na Escola, Cidadania na Vida, *Actas do II Encontro de Probabilidades e Estatística na Escola*, Universidade do Minho, pp. 91–99.

NEGRA, C. & MARTINHO, E. (2010). *Matemática A-10.º ano Volume 2*, Santillana, Carnaxide.

NCTM (2008). *Princípios e Normas para a Matemática Escolar,* 2.ª ed., NCTM e APM, Lisboa.

NEVES, M.A.F.; GUERREIRO, L.; LEITE, A. & SILVA, J.N. (2010a). *Matemática A 10.º Ano - Estatística,* Porto Editora, Porto.

NEVES, M.A.F.; FARIA, L. & SILVA, J.N. (2010b). *Matemática– 5.º Ano - Parte 3*, Porto Editora, Porto.

NEVES, M.A.F.; SILVA, A. P.; RAPOSO, M.J. & SILVA, J.N. (2011). *Matemática– 8.º Ano - Parte 2,* Porto Editora, Porto.

OSBORNE, C.(1991). Statistical Calibration: A Review, *International Statistical Review* **59**, n.º 3, pp. 309–336.

PASSOS, I.C. & CORREIA, O.F. (2010). *Matemática em acção 7 -Parte 2*, Lisboa Editora, Lisboa.

PESTANA, D. D. & VELOSA, S. F. (2009). *Introdução à Probabilidade e à Estatística*, Vol. 1, 3.ª ed., Fundação Calouste Gulbenkian, Lisboa.

PONTE, J.P.; SERRAZINA, L.; GUIMARÃES, H. M.; BRENDA, A.; GUIMARÃES, F.; SOUSA, H.; MENEZES, L.; MARTINS, M.E. G. & OLIVEIRA, P. A. (2007). *Programa de Matemática do Ensino Básico,* ME-DGIDC, Lisboa.

REIS, E. (1998). *Estatística Descritiva,* Edições Sílabo, Lisboa.

SILVA, J.C.; FONSECA, M.G.; MARTINS, A.A.; FONSECA, C.M.C. & LOPES, I.M.C. (2001). *Programa de Matemática A - Ensino Secundário*, ME-DES, Lisboa.

ZAWOJEWSKI, J.S.; BROOKS,G.; DINKELKAMP, L.; GOLDBERG,E.D.; GOLDBERG, H.; HYDE, A.; JACKSON, T.; LANDAU, M.; MARTIN, H.; NOWAKOWSKI, J.; PAULL, S.; SHULTE, A.P.; WAGREICH, P. & WILMOT, B. (1992). *Dealing with data and chance,* The National Council of Teachers of Mathematics, Inc., Virginia (Tradução de Sónia Figueirinhas, *Lidar com dados e probabilidades - Normas para o Currículo e a Avaliação em Matemática Escolar, Colecção de Adendas, Anos de Escolaridade 5-8,* APM).

Páginas da Internet

ALEA — `www.alea.pt`

Apêndice A

Propostas de trabalho para a sala de aula

A.1 PROPOSTA 1 – Quantas pessoas vivem em minha casa?

Para resolver esta atividade é necessário, em primeiro lugar, que cada aluno da turma responda à questão: Contando contigo, quantas pessoas vivem em tua casa? Após o registo dos resultados no quadro, ficando visíveis para todos, realiza as seguintes tarefas.

1. Na janela principal do *Geogebra* começa por selecionar na barra de ferramentas *Exibir* e de seguida escolhe *Folha de Cálculo*.
2. Na coluna *A* introduz o número de pessoas que vive em casa de cada aluno.
3. De seguida, seleciona todos os elementos da lista e utiliza o botão do lado direito do rato para selecionar *Criar lista*. A lista obtida aparece na folha algébrica (lado esquerdo) designada por *lista1*.
4. Constrói uma tabela de frequências absolutas introduzindo na entrada o comando Tabela Frequências[lista1].
5. Indica a moda deste conjunto de dados.
6. Com base na tabela formula duas questões. Pede ao teu colega para responder a essas questões.
7. Constrói um gráfico de pontos introduzindo na entrada o comando GráficoPontos[lista1].
8. Constrói agora um diagrama de barras para o mesmo conjunto de dados, usando o comando DiagramaBarras[lista1,0.5], sendo 0.5 a largura das barras.
9. Que semelhanças encontras entre as duas representações gráficas? E que diferenças?

A.2 PROPOSTA 2 – Classificações obtidas num teste de Matemática

1. Começa por selecionar na barra de ferramentas *Exibir* e de seguida escolhe *Folha de Cálculo.*

2. Na coluna *A* introduz as notas obtidas num teste, na escala de 1 a 100, pelos alunos de uma turma do 6.º ano, na disciplina de Matemática.

90	50	48	44	92	41	68
82	53	62	38	81	62	43
73	44	63	88	53	73	64
42	70	75	49	59	52	53

Tabela A.1: Classificações obtidas num teste

3. De seguida, seleciona todos os elementos da lista e utiliza o botão do lado direito do rato para selecionar *Criar lista*. A lista obtida aparece na folha algébrica (lado esquerdo) designada por *lista1*.

4. Altera o nome desta lista para "notas" procedendo do seguinte modo: com o cursor colocado em *lista1*, basta utilizar o botão do lado direito do rato, clicar em *propriedades dos objetos* e escrever o nome que se pretende na janela *nome*, neste caso "notas".

5. Elabora um diagrama de caule e folhas. Para o obteres usa o comando DiagramaCaule-Folhas[notas].

   (a) Qual é a nota mais alta? E a mais baixa? Como designas estes valores?

   (b) Qual é o valor da amplitude?

   (c) Qual é a nota mais frequente? Como designas este valor?

6. Determina a média das notas dos testes através do comando Média[notas].

7. Determina a percentagem de negativas. Dá a resposta arredondada às décimas.

8. Sabendo que um aluno obtém a classificação qualitativa de Bom quando a sua nota é superior ou igual a 70 e inferior a 90, determina a percentagem de alunos que obtiveram Bom neste teste.

9. Supõe que, por lapso, o professor registou a nota da Ana (aluna desta turma) incorretamente. A nota real da Ana foi 84 e o professor escreveu 48. Responde às seguintes questões, apresentando as justificações necessárias.

   (a) A amplitude mantém-se? Porquê?

   (b) E a moda? Porquê?

   (c) A média aumentou ou diminuiu? Calcula novamente a média alterando a lista dos dados.

   (d) Representa novamente o diagrama de caule e folhas.

A.3 PROPOSTA 3 – Meio de transporte utilizado para chegar à escola

Perguntou-se a 40 alunos de uma escola qual o meio de transporte mais utilizado para irem para a escola. As respostas encontram-se na seguinte tabela:

Meio de transporte mais utilizado	Frequência absoluta
A pé	17
Autocarro	12
Carro	8
Bicicleta	3
Total	40

Tabela A.2: Meio de transporte mais utilizado pelos alunos

1. Determina a frequência relativa de cada meio de transporte.
2. Comenta a afirmação "92,5% dos alunos não selecionaram a bicicleta como meio de transporte mais utilizado para ir para a escola".
3. Constrói um gráfico circular com recurso ao *GeoGebra*, seguindo os seguintes passos:
   (a) Calcula a amplitude de cada um dos setores. Apresenta os cálculos.
   (b) Na janela gráfica marca dois pontos A e B. Desenha a circunferência de centro em A e que passa em B.
   (c) Traça o segmento de reta [AB].
   (d) Usando o icon *ângulo com uma dada amplitude*, seleciona o ponto B, o vértice do ângulo e de seguida escreve a amplitude do ângulo referente ao setor *a pé*. Une o ponto obtido com o centro da circunferência.
   (e) Repete o mesmo procedimento para os restantes meios de transporte.
   (f) Usa o icon ABC e seleciona *inserir texto*. Deste modo podes escrever em cada setor o meio de transporte assim como a percentagem que lhe corresponde.

(g) Se pretendes colorir de forma distinta os diferentes meios de transporte utiliza o comando “Sector Circular(Centro, Dois pontos)” para colorir cada um. Para selecionar a cor que pretendes, coloca o cursor num setor e clica com o botão do lado direito do rato. Depois seleciona *Propriedades dos objetos* e *Cor*.

A.4 PROPOSTA 4 – Classificações internas *versus* classificações externas na disciplina de Matemática

1. Considera as classificações internas e as classificações externas, ambas numa escala de 0 a 20, obtidas por 27 alunos do 12.º ano, na disciplina de Matemática.

2. Começa por selecionar na barra de ferramentas *Exibir* e de seguida escolhe *Folha de Cálculo*.

3. Na coluna *A* da folha de cálculo introduz as classificações internas e na coluna *B* introduz as classificações externas.

4. De seguida, seleciona a lista dos resultados da avaliação interna e utiliza o botão do lado direito do rato para selecionar *Criar lista*. A lista obtida aparece na folha algébrica (lado esquerdo) designada por *lista1*.

5. Altera o nome desta lista para "interna" procedendo do seguinte modo: com o cursor colocado em *lista1*, basta utilizar o botão do lado direito do rato, clicar em *propriedades dos objetos* e escrever o nome que se pretende na janela *nome*, neste caso "interna".

6. Determina a média, a moda e a mediana. Usa os comandos Média[interna], Moda[interna] e Mediana[interna].

7. Determina a amplitude, o 1.º quartil, o 3.º quartil e a amplitude interquartis. Usa os comandos Q1[interna] e Q3[interna].

8. Procede de modo idêntico para os resultados da avaliação externa, criando uma lista que podes designar por "externa".

9. Compara as medidas de tendência central obtidas nas duas variáveis. O que podes concluir?

10. Compara a amplitude e a amplitude interquartis obtidas. O que podes concluir?

11. Obtém, agora, os dois diagramas de extremos e quartis procedendo do seguinte modo: seleciona as duas listas inseridas na folha de cálculo e clica no *icon* da barra de ferramentas que apresenta um gráfico e escolhe "análise multivariada". Para além dos

Aluno	Avaliação interna	Avaliação externa
1	10	11
2	11	10
3	11	8
4	10	9
5	13	12
6	13	13
7	16	15
8	15	15
9	17	16
10	10	11
11	11	13
12	14	13
13	11	11
14	10	11
15	12	12
16	15	12
17	11	11
18	11	10
19	16	15
20	15	13
21	12	10
22	10	8
23	19	18
24	10	11
25	11	10
26	14	14
27	12	13

Tabela A.3: Classificações internas *versus* classificações externas

diagramas de extremos e quartis também surge uma tabela com as medidas estatísticas referentes aos dois conjuntos de dados.

12. Faz um comentário aos diagramas obtidos fazendo referência à concentração/dispersão das classificações.

A.5 PROPOSTA 5 – Salários dos trabalhadores de uma empresa

Na tabela A.4 encontram-se os salários (em euros) dos trabalhadores de uma empresa.

Tipo de emprego	Número de trabalhadores	Salário (em euros)
Presidente	1	12000
Vice-presidente	2	6000
Gerente	3	2800
Supervisor	9	1100
Operário	33	900
Funcionário de Caixa	5	750
Tesoureiro	3	700
Vendedor	12	600
Guarda	4	500

Tabela A.4: Salários dos trabalhadores de uma empresa

1. Começa por selecionar na barra de ferramentas *Exibir* e de seguida escolhe *Folha de Cálculo*.

2. Insere todos os ordenados na coluna A da folha de cálculo.

3. De seguida, seleciona a lista dos ordenados e utiliza o botão do lado direito do rato para selecionar *Criar lista*. A lista obtida aparece na folha algébrica (lado esquerdo) designada por *lista1*.

4. Altera o nome desta lista para “salario” procedendo do seguinte modo: com o cursor colocado em *lista1*, basta utilizar o botão do lado direito do rato, clicar em *propriedades dos objetos* e escrever o nome que se pretende na janela *nome*, neste caso “salario”.

5. Calcula o ordenado médio, usando o comando Média[salario]. Achas que este valor é representativo dos ordenados de todos os trabalhadores? Justifica.

6. Determina a moda e a mediana dos ordenados. Usa os comandos Moda[salario] e Mediana[salario]. Qual das três medidas representa melhor os ordenados dos trabalhadores? Justifica.

7. Em qual das três medidas de tendência central se deve basear um trabalhador desta empresa para pedir aumento de salário? Justifica.

8. Se aumentarmos para 900 euros os ordenados dos 24 trabalhadores que ganham menos que este valor, quais são os novos valores da média, moda e mediana? Altera os valores na lista e compara as 3 medidas obtidas (as medidas surgem na janela algébrica) com as já calculadas.

9. Quais foram as medidas de tendência central que se mantiveram? E as que se alteraram? Porquê?

10. Se alterasses apenas um dos 72 salários, qual era a medida de tendência central que se alterava de certeza? Porquê?

11. E qual ou quais eram as que garantidamente se manteriam? Porquê?

12. Foram contratados dois novos empregados pela empresa: um gerente de fábrica e um supervisor. Prevê se a média de salários vai aumentar, baixar, ou ficar na mesma. Explica a tua previsão. Verifica a tua conjetura recorrendo à folha de cálculo.

A.6 PROPOSTA 6 – Comparação de duas turmas

1. A tabela A.5 apresenta as notas, na escala de 1 a 5, de duas turmas do 7.º ano no final do 1.º período

Níveis a Mat.	Turma A	Turma B
1	4	0
2	5	3
3	7	20
4	5	1
5	4	1

Tabela A.5: Classificações de Matemática das turmas A e B do 7.º ano

   (a) Calcula a média para cada turma.

   (b) Se fosses aluno do 7.º ano, a que turma gostarias de pertencer? Porquê?

   (c) Começa por selecionar na barra de ferramentas *Exibir* e de seguida escolhe *Folha de Cálculo*.

   (d) Insere todas as classificações da turma A na coluna A da folha de cálculo.

   (e) De seguida, seleciona a lista das classificações e utiliza o botão do lado direito do rato para selecionar *Criar lista*. A lista obtida aparece na folha algébrica (lado esquerdo) designada por *lista1*.

   (f) Altera o nome desta lista para "turmaA" procedendo do seguinte modo: com o cursor colocado em *lista1*, basta utilizar o botão do lado direito do rato, clicar em *propriedades dos objetos* e escrever o nome que se pretende atribuir à *lista 1* , neste caso "turmaA".

   (g) Repete este procedimento para as classificações da turma B e designa a lista por "turmaB".

   (h) Confirma os cálculos efetuados na determinação da média de cada turma, utilizando os comandos "Média[turmaA]" e "Média[turmaB]".

(i) Um aluno calculou o desvio padrão dos níveis a Matemática da turma B e obteve -0,2. O que podes concluir?

(j) Calcula o desvio padrão para cada um dos conjuntos de dados. Usa o comando DesvioPadrãoAmostra[turmaA] e DesvioPadrãoAmostra[turmaB], respetivamente.

(k) Qual das turmas te parece mais homogénea nesta disciplina? Porquê?

2. Encontra um conjunto de 10 números com média 20 e desvio padrão cerca de 5. Confirma com a folha de cálculo do *GeoGebra*.

3. Encontra um conjunto de 10 números com média 20 e desvio padrão cerca de 10. Confirma com a folha de cálculo do *GeoGebra*.

A.7 PROPOSTA 7 – Peso e altura dos alunos de uma turma do 10.º ano

Para resolver esta atividade o professor deve solicitar, antecipadamente, uma lista com os pesos e as alturas dos alunos.

1. Começa por exibir a folha de cálculo e utilizar as duas primeiras colunas (*A* e *B*) para registar os pesos e as alturas dos alunos, respetivamente. Ficam assim conhecidas as coordenadas dos pontos a utilizar para a regressão linear.

2. Com os pontos definidos, cria uma lista de pontos (bastará selecionar as coordenadas e, utilizando o botão do lado direito do rato, selecionar *Criar lista de pontos*, pois, neste caso, cada ponto terá duas coordenadas). Obtém-se assim a *lista1*.

3. Determina o coeficiente de correlação, através do comando "CoeficienteDeCorrelação[lista1]";

4. Obtém a reta de regressão da altura em função do peso recorrendo ao comando "RegressãoLinear[lista1]"

5. No mesmo gráfico, representa a reta de regressão do peso em função da altura utilizando o comando "RegressãoLinearX[lista1]".

6. As retas obtidas são iguais? Podes representá-las por cores diferentes de forma a poderes distingui-las facilmente.

7. Altera um ou mais pontos da lista1 e verifica a alteração quer no coeficiente de correlação quer nas retas.

8. Considera que o Manuel só preencheu o peso (63,450kg). Qual é valor mais provável para a sua altura?

9. Considera que o João só preencheu a altura (1,68 m). Qual é valor mais provável para o seu peso?