Introdução à Estatística

2.ª Edição

Bento Murteira
Carlos Silva Ribeiro
João Andrade e Silva
Carlos Pimenta

Instituto Superior de Economia e Gestão
Universidade Técnica de Lisboa

McGraw-Hill

LISBOA • SÃO PAULO • BOGOTÁ • BUENOS AIRES • GUATEMALA
MADRID • MÉXICO • NOVA IORQUE • PANAMÁ • SAN JUAN • SANTIAGO

AUCKLAND • HAMBURGO • LONDRES • MILÃO • MONTREAL • NOVA DELI
PARIS • SINGAPURA • SYDNEY • TÓQUIO • TORONTO

Na capa apresenta-se uma variante do *Board* de Galton, que ilustra uma distribuição muito importante para a Estatística: a distribuição log-normal. Esta distribuição pode ser definida da seguinte forma: a variável aleatória X segue uma distribuição log-normal, com parâmetros μ e σ^2, se e só se, o logaritmo natural de X tem distribuição normal, $X \sim N(\mu, \sigma^2)$ Pode provar-se que

$$E(X) = \exp\{\mu + (1/2)\sigma^2\} \text{ e } \mathrm{Var}(X) = E(X)^2(e^{\sigma^2} - 1).$$

Uma particularidade desta distribuição é a sua assimetria, em que o respectivo coeficiente depende apenas de σ^2,

$$\gamma_1 = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}.$$

Nesta caso, o triângulo de agulhas é construído de forma a garantir as características da distribuição log-normal, nomeadamente a sua assimetria.

© McGraw-Hill

capítulo

1

Dados estatísticos

1.1 Introdução

A estatística tem-se difundido de forma tão rápida, que se torna cada vez mais difícil listar os ramos da actividade humana em que a sua aplicação se tem revelado fecunda ou mesmo indispensável. No levantamento que esboça, Lehmann (1988) refere que a *Encyclopedia of Statistical Sciences* apresenta, nos nove volumes iniciais que a compõem, 47 entradas correspondentes a outros tantos campos de aplicação da estatística que vão, por ordem alfabética, desde *actuarial work* até *weather modification*, passando por *econometrics... finance... management science* sem faltar *demography* e *gambling*, berços da probabilidade e da estatística.

A questão que imediatamente surge é a seguinte: qual a motivação para tão vastas aplicações? Para responder a esta pergunta deve começar por notar-se que quando alguém (indivíduo ou instituição) pretende estudar um determinado fenómeno — com a simples finalidade de o descrever, ou com objectivos mais ambiciosos, como sejam explicá-lo, interpretá-lo, fazer previsões ou tomar decisões — depara-se com uma **situação de incerteza**, que tem por consequência nunca ser possível conhecer o fenómeno de forma completamente rigorosa.

Nestas circunstâncias, começa-se normalmente por recolher os factos que pareçam importantes sobre os atributos mais relevantes do fenómeno em estudo. Estes factos devem conter informação e ser compilados de forma organizada; por isso se designam por **dados**. Em muitos casos os dados são de natureza quantitativa, isto é, são **dados numéricos**.

Muitas vezes, pela insuficiente qualidade dos dados ou pelo deficiente conhecimento das técnicas, ou, ainda, pela irrelevância para os objectivos do estudo, o tratamento estatístico não passa além da **exploração** e da **descrição** dos dados recolhidos. Neste caso, é quase sempre suficiente calcular alguns indicadores que os caracterizem sumariamente e construir alguns gráficos adequados. É este o objectivo do presente capítulo, e nele se desenvolvem alguns conceitos básicos da **estatística descritiva**. É evidente que, mesmo nesta fase, os resultados obtidos podem fundamentar algumas decisões simples.

Quando os objectivos são mais ambiciosos — **explicar**, **interpretar**, **prever** ou **decidir** — torna-se imprescindível aprofundar a análise do fenómeno, uma vez que a incerteza que o caracteriza se revela através de um dos aspectos mais significativos para a análise estatística — a **variabilidade** dos dados. Procura-se então modelar esta variabilidade, estabelecendo um **padrão** que capture os aspectos essenciais do fenómeno e consentindo que os dados observados apresentem **desvios**, maiores ou menores, em relação ao padrão. Como não é geralmente possível captar toda a variabilidade contida nos dados, o **modelo** que vai definir o padrão — e assim dar respostas aos objectivos

fixados — constitui sempre uma representação simplificada do fenómeno em estudo. Apenas se procura estabelecer uma «relação razoável» entre o padrão e os respectivos desvios. Como quase sempre existem aspectos do padrão adoptado que são desconhecidos, deve então procurar-se **estimar** o modelo, reduzindo tanto quanto possível os desvios, e **testar** a sua adequação.

Embora alguns termos tenham sido apresentados de forma intencionalmente vaga (como sejam: «incerteza», «variabilidade», «padrão», «desvios», «modelo», «relação razoável», «estimar», «testar»), eles constituem o cerne da estatística. Não querendo definir a estatística como a actividade dos estatísticos, mas sublinhando que ela é a ciência da incerteza, poderá afirmar-se que a **estatística** é um repositório de instrumentos ou procedimentos adequados para recolher, explorar, descrever e, finalmente, analisar e interpretar dados.

Uma pergunta básica que se deve fazer é a seguinte: como medir o grau de incerteza? O conceito básico que vai permitir a quantificação da incerteza é o de **probabilidade**, cuja teoria vai ser desenvolvida nos capítulos 2 a 5.

Quando a análise exploratória ou descritiva constitui apenas uma primeira etapa para o estudo do fenómeno em causa, torna-se imprescindível aprofundar a análise estatística por meio de **inferências indutivas**. Estas inferências fundamentam-se na teoria da probabilidade — supondo que os dados constituem uma **amostra** de uma **população**, ou universo, mais vasta (conceitos que vão ser estudados na secção 1.2, deste capítulo) — e procuram extrair conclusões sobre a população a partir da amostra. O estudo de alguns aspectos da **amostragem** e dos métodos da **estatística indutiva** vai ser feito nos capítulos 6 a 11, embora o capítulo 6 também possa ser considerado o capítulo-charneira entre a teoria da probabilidade e a estatística indutiva. Em particular, nos capítulos 9 (secções 9.3 e 9.4), 10 e 11 vai ser aprofundada a análise estatística das **relações** entre dois ou mais atributos. O estudo das técnicas estatísticas de **previsão** e dos métodos e dos modelos para fundamentar a tomada de **decisões** em situações de incerteza, depois de avaliadas e pesadas as opções possíveis, não vai ser abordado neste texto.

Convém esclarecer, nesta secção introdutória, se a estatística é ou não uma **teoria**. Depois da leitura dos capítulos que se seguem, o leitor vai concluir, com facilidade, que a estatística é uma ciência com um corpo teórico muito sólido, e, portanto, neste sentido, é uma teoria. Pode mesmo afirmar-se que a estatística, na forma apresentada neste livro, é uma **teoria matemática**, uma vez que é composta pelas componentes que caracterizam estas teorias: termos primitivos, axiomas, definições, teoremas e respectivas demonstrações. Mais precisamente, a estatística é um ramo da matemática aplicada e, como tal, é dotada de rigor e opera num quadro bem definido de noções, princípios e premissas.

No entanto, a preocupação fundamental da estatística reside na apreensão e na quantificação de diferentes tipos de fenómenos, mais ou menos complexos, delimitando as suas fronteiras, avaliando as suas dimensões, salientando e relacionando alguns dos seus aspectos pertinentes, auxiliando até na formulação de proposições, mas nunca procurando «explicar por si só», isto é, substituir aqueles (os especialistas) que conhecem o domínio científico em que se enquadram os fenómenos em questão. Neste sentido, importa sublinhar que a estatística é um **método** e não uma teoria.

Deve ainda referir-se que embora neste livro se esteja mais preocupado com os aspectos matemáticos introdutórios da estatística e menos com as aplicações, vão ser apresentados muitos **exemplos reais** de forma que o leitor fique com uma ideia razoável do interesse prático da estatística.

Finalmente convém relembrar que o presente capítulo se destina a apresentar as noções fundamentais de estatística descritiva.

© McGraw-Hill

1.2 Amostra. População

Chama-se **colecção de dados** (termo primitivo) a um conjunto de observações de certo(s) atributo(s), qualquer que seja a forma como foram recolhidas. Apresentam-se a seguir alguns exemplos que vão permitir clarificar esta noção.

EXEMPLO 1.1 • A colecção de dados que se apresenta no quadro 1.1 refere-se à taxa anual de nupcialidade (número de casamentos ocorrido durante o ano, referido à população média desse ano, por 1000 habitantes). ⊙

QUADRO 1.1 • Taxa anual de nupcialidade em Portugal no período 1960-2002.

1960	7.8	1969	9.0	1978	8.3	1987	7.2	1996	6.3
1961	8.8	1970	9.4	1979	8.1	1988	7.1	1997	6.5
1962	7.9	1971	9.7	1980	7.3	1989	7.4	1998	6.6
1963	7.9	1972	9.1	1981	7.7	1990	7.3	1999	6.8
1964	8.1	1973	9.9	1982	7.4	1991	7.3	2000	6.2
1965	8.4	1974	9.3	1982	7.5	1992	7.0	2001	5.7
1966	8.6	1975	10.9	1984	7.0	1993	6.8	2002	5.4
1967	8.9	1976	10.5	1985	6.8	1994	6.6		
1968	8.7	1977	9.4	1986	6.9	1995	6.6		

Fonte: INE.

EXEMPLO 1.2 • No quadro 1.2 apresenta-se uma colecção de dados referente ao volume de vendas de 25 empresas da indústria transformadora em Março de 2002. Estes dados foram obtidos a partir dos quadros de pessoal publicados pelo Ministério do Emprego e da Segurança Social. ⊙

QUADRO 1.2 • Volume de vendas de 25 empresas da indústria transformadora em Março de 2002 (em euros).

116 849.2	5 307 005.7	3 423 764.1	8 824 851.6	15 748 600.0
1 596 746.4	349 052.0	1 110 322.4	5 180 326.6	43 697 324.5
251 389.2	793 957.0	1 481 137.9	0.0	576 158.8
7 941 925.8	17 692 688.1	108 500.6	13 755 676.5	20 270 578.2
101 426 958.4	5 794 202.2	8 824 851.6	3 190 067.0	6 473 545.5

EXEMPLO 1.3 • No quadro 1.3 encontram-se os últimos três dígitos (é necessário adicionar 299 000 km/s) das 100 medidas da velocidade da luz na atmosfera aferidas por Michelson em 1879 com o método de Foucault aperfeiçoado. Note-se que a velocidade da

© McGraw-Hill

luz no vácuo universalmente aceite é de 299 792.5 km/s. Para obter estimativas deste valor, os dados de Michelson têm de ser corrigidos por um factor que depende da humidade atmosférica, da pressão e da temperatura.

Este exemplo, para além de constituir um caso histórico, mostra que os resultados variam mesmo quando as observações são repetidas aparentemente nas mesmas condições. Neste caso, a variabilidade dos dados pode ser atribuída a causas naturais difíceis de controlar ou mesmo de identificar, como sucede na maior parte dos estudos experimentais. ⊙

QUADRO 1.3 • Medidas de Michelson.

620	760	800	810	840	850	870	880	930	960
650	760	800	810	840	850	870	880	930	960
720	760	800	810	840	850	880	890	940	970
720	760	800	810	840	850	880	890	940	980
720	770	800	810	840	850	880	890	940	980
740	780	810	820	840	860	880	900	950	980
740	780	810	820	840	860	880	900	950	1000
740	790	810	830	850	860	880	910	950	1000
750	790	810	830	850	870	880	910	960	1000
760	790	810	840	850	870	880	920	960	1070

Fonte: Reiss e Thomas (1997), *Statistical Analysis of Extreme Values*, Birkhauser Verlag.

EXEMPLO 1.4 • O lançamento de uma moeda ao ar e o registo do lado que fica voltado para cima é um processo gerador de várias colecções de observações. Por exemplo, numa sequência de 20 lançamentos obteve-se

$$F, C, F, C, C, C, F, F, F, C, C, F, F, C, F, C, F, C, C, F,$$

onde F = «Face» e C = «Coroa». Convencionando que $F \equiv 0$ e $C \equiv 1$, tem-se a sequência numérica equivalente,

$$0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0. \quad ⊙$$

As **duas** categorias básicas de dados são as seguintes:

a) **Dados temporais.** Os dados são **temporais** ou cronológicos quando as observações se referem a uma mesma entidade, para vários momentos ou períodos de tempo (por exemplo: as quantidades produzidas e as quantidades de factores de produção utilizados referentes a uma determinada indústria nos últimos 10 anos; o consumo e o rendimento disponível num determinado país nos últimos 20 anos).

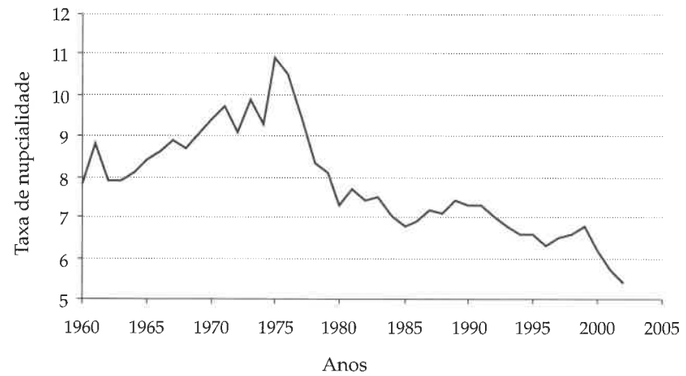
© McGraw-Hill

Uma característica essencial dos dados temporais é a sua **ordem cronológica**, podendo falar-se em **passado, presente** e **futuro**. Outro aspecto importante diz respeito à **frequência temporal das observações**, ou seja, trata-se de saber se os dados são observados numa base diária, semanal, mensal, trimestral, anual, etc.

Quando os dados são numéricos e se pretende descrever a evolução, no tempo, dos valores observados, os dados devem, como é evidente, conservar-se associados à data em que ocorreram e apresentar-se em forma de **série temporal**, dando origem a um gráfico que se chama **cronograma.** Por exemplo, se se considerar a taxa de nupcialidade em Portugal no período 1960-2002 (exemplo 1.1) tem-se o cronograma da figura 1.1, onde se pode observar a quebra significativa verificada a partir de 1975.

A estatística tem métodos próprios para o estudo de sucessões cronológicas, mas o respectivo tratamento não cabe no presente texto [ver, por exemplo, Wei (1990) ou Murteira, Muller e Turkman (1993)].

FIGURA 1.1 Cronograma das taxas de nupcialidade no período 1960-2002.



b) **Dados seccionais.** Os dados são **seccionais** quando as observações se referem a determinadas entidades (unidades estatísticas) em certo momento ou período de tempo (por exemplo: as quantidades produzidas e as quantidades de factores de produção utilizados referentes às empresas de uma certa indústria num determinado ano; as despesas em bens de consumo e as receitas das famílias consideradas numa determinada amostra e em determinado mês).

Nalguns casos pode acontecer que os dados não correspondam exactamente, para todas as entidades observadas, ao mesmo período. No entanto, se os dados se referem a períodos relativamente próximos, pode considerar-se que fazem parte do mesmo **conjunto de dados seccionais.** Por exemplo, se há observações de despesas e receitas de certas famílias realizadas num certo mês e há observações de outras famílias feitas no mês seguinte, é lícito, em muitos casos, supor que esta pequena variação temporal não afecta significativamente a análise empírica.

© McGraw-Hill

Uma característica fundamental dos dados seccionais é que **a ordem das observações é irrelevante** (pouco importa qual é a primeira família observada ou a vigésima-quinta!).

No exemplo 1.2 tem-se uma colecção de dados seccionais relativa ao volume de vendas de 25 empresas da indústria transformadora em Março de 2002.

Além das duas categorias básicas de dados apresentadas, convém referir mais duas que, de certo modo, são derivadas daquelas.

c) **Dados seccionais combinados**. Alguns conjuntos de dados têm aspectos seccionais e temporais. Diz-se que se tem um conjunto de dados seccionais combinados (*pooled cross sections*) quando se juntam vários conjuntos de dados seccionais, cada um referente a certa data. Por exemplo, em determinado ano tem-se uma amostra de 200 famílias (com as respectivas despesas e receitas), e cinco anos depois tem-se outra amostra de 250 famílias. Este tipo de dados pode ser analisado como se fosse um conjunto de dados seccionais, mas tem a vantagem de permitir testar se as famílias têm comportamentos diferentes nos dois anos considerados. Obviamente, o conjunto das 200 famílias da primeira amostra não coincide com o conjunto das 250 famílias da segunda amostra (o mais provável é que os dois conjuntos sejam disjuntos!).

d) **Dados de painel.** Em muitas situações os dados disponíveis referem-se a um conjunto fixo de entidades observadas em várias datas. Neste caso, diz-se que se tem um conjunto de **dados de painel** ou dados longitudinais (por exemplo, as observações das despesas de consumo e das receitas de um conjunto fixo de famílias ao longo de vários anos).

A **característica essencial** dos dados de painel é que o conjunto de entidades a observar é sempre o mesmo para todas as observações temporais, o que, muitas vezes, dificulta muito a sua obtenção.

Embora os dados de painel possam ser encarados como dados seccionais combinados, existem métodos adequados que, com vantagem, tiram partido da sua característica essencial (conjunto fixo de entidades a observar). As vantagens têm a ver com o facto de se dispor de **várias observações temporais** para a mesma entidade.

Como pode ver-se em Murteira (1993) é possível ir bastante longe na análise exploratória de colecções *ad hoc*. Nas secções que se seguem apresentam-se alguns métodos para efectuar este tratamento.

No entanto, o trabalho estatístico não fica em geral por aí. Quando se entra na antecâmara da inferência indutiva tem pouco sentido falar em colecções, e torna-se conveniente falar em amostras e, forçosamente, em populações ou universos.

Em estatística, **população** designa o conjunto dos elementos cujos atributos são objecto de um determinado estudo. Deste modo, as populações são os conjuntos fundamentais para efectuar análises estatísticas.

EXEMPLO 1.5 • Constituem exemplos de populações:

a) O conjunto das empresas da indústria transformadora em Março de 2007.
b) O conjunto dos cidadãos eleitores nas eleições para a Assembleia da República em Fevereiro de 2005.
c) O conjunto das temperaturas instantâneas que se registam numa certa estação meteorológica ao longo de um determinado dia. ⊙

© McGraw-Hill

A inferência estatística ocupa-se fundamentalmente das propriedades das populações. Os casos isolados podem ter interesse histórico, mas são objecto formal da estatística só quando fazem parte de um conjunto, de uma população.

Os elementos da população são em geral designados por **unidades estatísticas**: facto ou entidade elementar que é objecto de observação, qualquer que seja a sua natureza, desde que possa corresponder a uma definição precisa [um ser vivo (homem, cabeça de gado,...), uma instituição (família, empresa,...), um objecto (automóvel, tractor,...), um facto qualquer (cultura de trigo,...), etc.].

EXEMPLO 1.6 • A principal unidade estatística utilizada no Inquérito aos Orçamentos Familiares de 1989-1990 realizado pelo INE foi a unidade de alojamento. Por unidade de alojamento entende-se o local distinto e independente que, pelo modo como foi construído, reconstruído, ampliado ou transformado, se destina à habitação humana, na condição de, no momento do inquérito, não estar a ser utilizado para outros fins (hotéis, pensões, quartéis, acampamentos, etc., não são unidades de alojamento). ⊙

Para conhecer de forma completa uma população tem de analisar-se todos os seus elementos, isto é, realizar um **censo**[1] ou **indagação completa**. Exceptuando os casos em que a população tem dimensão modesta e é acessível, raramente é possível analisar todos os elementos de uma população finita (por não se dispor de orçamento, de tempo, por a observação ser destrutiva[2], etc.), e é sempre impossível observar todos os elementos de uma população infinita. Portanto, o estudo dos atributos da população tem de ser feito sobre um seu subconjunto finito, que se designa por **amostra.**

Os termos amostra e população referem-se a conjuntos de unidades estatísticas. Contudo, muitas vezes, por abuso de linguagem, estes termos também são utilizados para identificar os conjuntos de valores assumidos pelo atributo em estudo. Por exemplo, chama-se população ao conjunto das empresas da indústria transformadora (unidades estatísticas), mas também eventualmente ao conjunto dos volumes de vendas destas empresas. Naturalmente pode estabelecer-se esta simplificação só depois de se definirem claramente os objectivos do estudo.

A forma de selecção de uma amostra a partir da população, designada por **processo de amostragem**, é determinante para a qualidade das inferências que venham a fazer-se. As técnicas correntemente utilizadas para colher amostras deixam total ou parcialmente ao acaso a selecção dos elementos que virão a integrá-las. Se intervém apenas o acaso e se todos os elementos da população têm igual possibilidade de ser incluídos, a amostragem diz-se casual. No extremo oposto, quando a escolha da amostra é feita apenas por critérios arbitrários do investigador, a amostragem diz-se dirigida. Os métodos estatísticos não permitem analisar a qualidade dos resultados obtidos a partir deste último tipo de amostras. O problema da amostragem é retomado adiante.

[1] Os censos mais conhecidos são os da população (no sentido demográfico). No nosso País realizaram-se censos em 1864, 1878, 1890, 1900, 1911, 1920, 1930, 1940, 1950, 1960, 1970, 1981, 1991 e 2001.

[2] Nunca se viu uma fábrica de fusíveis ou de munições experimentar todas as unidades produzidas...

© McGraw-Hill

EXEMPLO 1.7 • Considerando as populações apresentadas no exemplo 1.5, considerem-se as seguintes amostras:

a) Uma possível amostra da população formada pelo conjunto das empresas da indústria transformadora em Março de 2002 é constituída pelas 25 empresas a que se refere o quadro 1.2, supondo que foram obtidas por escolha casual a partir dos quadros de pessoal publicados pelo Ministério do Emprego e da Segurança Social. Com base nesta amostra podem estudar-se vários atributos das empresas, como, por exemplo, o volume de vendas.
b) Uma amostra de cidadãos eleitores pode ser obtida seleccionando casualmente 1000 cidadãos à boca das urnas.
c) Considerando a população de temperaturas na estação meteorológica, pode obter-se uma amostra observando as temperaturas de 15 em 15 minutos. ⊙

Os atributos observados podem ser **quantitativos** ou **qualitativos**. Quando são **quantitativos** apresentam-se com diferentes intensidades ou valores. Quando são **qualitativos** revestem várias modalidades e registam-se numa **escala nominal**, se a ordem das modalidades não tem significado (exemplos: {homem, mulher} ou {solteiro, casado, viúvo, divorciado, separado}) ou numa **escala ordinal**, se a ordem das modalidades tem significado (exemplo: {trabalhador não especializado, semi-especializado, especializado, muito especializado}).

Qualquer que seja o atributo considerado existe sempre a possibilidade de o representar numericamente. Assim:

1. No caso de um **atributo quantitativo** os valores numéricos são obviamente as respectivas **intensidades**.
2. Quando se trata de um **atributo qualitativo**, este também pode ser representado numericamente, bastando para tal estabelecer uma correspondência qualquer entre as várias modalidades e os números inteiros.

a) No caso de **escalas nominais** tem-se, por exemplo:
   - Género — {homem, mulher}:
     {mulher} ≡ 0; {homem} ≡ 1.
   - Estado civil — {solteiro, casado, viúvo, divorciado, separado}:
     {solteiro} ≡ 1; {casado} ≡ 2; {viúvo} ≡ 3; {divorciado} ≡ 4; {separado} ≡ 5.

   Note-se que a atribuição destes valores não tem qualquer significado quantitativo, é apenas uma **codificação** numérica das várias modalidades.
b) No caso de **escalas ordinais** pode, por exemplo, fazer-se:
   - Grau de especialização dos trabalhadores — {trabalhador não especializado, semi-especializado, especializado, muito especializado}:
     {trabalhador não especializado} ≡ 1; {trabalhador semi-especializado} ≡ 2;
     {trabalhador especializado} ≡ 3; {trabalhador muito especializado} ≡ 4.

   Naturalmente numa escala ordinal a numeração é feita de forma a respeitar a **ordem** das várias modalidades, mas isso não quer dizer que, por exemplo, um trabalhador especializado corresponda a três trabalhadores não especializados.

© McGraw-Hill

EXEMPLO 1.8 • No Recenseamento Geral Agrícola de 1989, cada unidade estatística — exploração agrícola — foi observada em relação a vários atributos qualitativos, entre os quais podem salientar-se: 1) explorações segundo a natureza jurídica do produtor (sete modalidades: produtor singular autónomo, produtor singular empresário, sociedades, cooperativas agrícolas, baldios, Estado e outros entes públicos, outras entidades); 2) explorações segundo a forma de exploração (quatro modalidades: conta própria, arrendamento fixo, arrendamento de campanha e arrendamento variável). Entre os atributos quantitativos podem mencionar-se: 3) superfície agrícola utilizada (SAU); 4) número de blocos da exploração (parte da exploração inteiramente rodeada de terras, águas, etc., não pertencentes à exploração). ⊙

Qualquer que seja o atributo considerado, o seu valor numérico pode variar de elemento para elemento. Para assinalar este facto representam-se estes valores por uma **variável**, seja x. Se a amostra observada tem **dimensão** n (ou seja, n elementos), tem-se

$$x_1, x_2, \ldots, x_n, \tag{1.1}$$

onde x_i $(i = 1, 2, \ldots, n)$ é o valor do atributo na i-ésima observação.

As variáveis com que correntemente se depara são de dois tipos: **discretas** e **contínuas**. São do primeiro tipo as variáveis que podem tomar somente um número finito ou uma infinidade numerável de valores (exemplos: número de divisões por unidade de alojamento; número de acidentes por apólice de seguro; género; estado civil; grau de especialização dos trabalhadores). São do segundo tipo as variáveis que podem tomar qualquer valor dentro de um intervalo de números reais (exemplos: altura e peso de indivíduos; superfície agrícola utilizada em cada exploração; tempo de vida de uma máquina; taxas de juro).

A distinção entre variáveis discretas e variáveis contínuas é muitas vezes artificial. Com efeito, o conjunto de valores que uma variável contínua pode assumir é, muitas vezes, convertido num conjunto discreto[3], uma vez que o grau de aproximação dos instrumentos de medida é necessariamente limitado. Por exemplo, embora os pontos percentuais das taxas de juro possam ser considerados como números reais, eles são reportados, quase sempre, em dízimas (com duas casas decimais). No entanto, pode ser conveniente tratar como contínuas todas as variáveis que, pelo menos em teoria, possam assumir qualquer valor de um intervalo desde que tal não seja logicamente incompatível.

1.3 Gráfico caule-e-folhas

Quando a dimensão da amostra é pequena, a análise descritiva dos dados pode ser facilitada recorrendo a métodos semigráficos, nomeadamente os **gráficos caule-e-folhas**. A construção destes, apesar de imediata, permite que o observador, sem perda de informação, se torne mais sensível ao aspecto global dos dados e comece a explorar a existência de padrões. A simplicidade destes gráficos permite que a sua apreensão seja feita através de exemplos.

[3] Note-se, por exemplo, que até os computadores mais potentes representam o conjunto dos números reais por um conjunto discreto (finito), qualquer que seja a precisão utilizada.

© McGraw-Hill

QUADRO 1.4 • Gráfico caule-e-folhas das taxas de nupcialidade.

N.º de folhas	Caule	Folhas
2	5	47
10	6	2356668889
14	7	00123334457899
8	8	11346789
7	9	0134479
2	10	59

$n = 43$; unidade = 0.1.

EXEMPLO 1.9 • No quadro 1.4 apresenta-se o gráfico caule-e-folhas da colecção de dados sobre nupcialidade do quadro 1.1. Cada **caule — dígitos dominantes** (no caso presente, a parte inteira) — está numa linha e tem um número variável de **folhas — dígitos dominados** (no caso presente a parte decimal) –; cada observação é uma folha de determinado caule. Por exemplo, a observação 8.3 corresponde à terceira observação do quarto caule — **8 | 3**; o caule 9 tem 7 folhas — **9 | 0134479** — que correspondem às 7 observações 9.0, 9.1, 9.3, 9.4, 9.4, 9.7 e 9.9. A unidade é determinada pelo primeiro dígito dominado (no caso presente só há um) e é igual a 0.1. A primeira coluna regista o número de folhas em cada caule. ⊙

A escolha do número de linhas de um gráfico caule-e-folhas depende de dois factores: a dimensão, n, e a amplitude (diferença entre o maior e o menor dos valores) da amostra ou da colecção de dados. Embora existam na literatura várias propostas para resolver este problema [ver, por exemplo, Murteira (1993)], a decisão final deve basear-se na experiência e nos objectivos do investigador. Uma proposta referida é considerar que o número de linhas é igual à parte inteira de $2\sqrt{n}$.

EXEMPLO 1.10 • No quadro 1.5 apresentam-se as taxas aparentes (quociente entre o juro líquido e a última cotação efectuada) das 51 obrigações mais transaccionadas em 1990 na Bolsa de Valores de Lisboa.

No quadro 1.6 apresenta-se uma primeira versão do gráfico caule-e-folhas. Repare-se que cada parte inteira tem dois caules, e que se cortaram as segundas casas decimais sem arredondar; trata-se da prática recomendada para permitir identificar os valores na colecção original (quadro 1.5).

Como não é prudente que o gráfico esteja dependente dos valores invulgares ou extravagantes, a prática corrente para ladear o problema consiste em pôr de parte aqueles valores e em basear a construção do gráfico nos restantes. Os valores anormalmente pequenos ou anormalmente grandes são incluídos em linhas rotuladas «Pe» ou «Gr», convenientemente separados por vírgulas. A variante do quadro 1.7 baseia-se na retirada de três observações «pequenas» e seis observações «grandes»[4], procedimento que permite trabalhar com uma unidade de 0.01%

[4] A ideia de observações «pequenas» e observações «grandes» é precisada adiante.

© McGraw-Hill

e caracterizar melhor o padrão apresentado pelos dados. A observação mais «pequena» corresponde a uma emissão com prémio de reembolso; portanto, com uma elevada taxa de juro em função da vida média; as outras duas correspondem a obrigações com baixa taxa de juro nominal. As seis observações «grandes» correspondem a obrigações com elevada taxa de juro nominal. ⊙

QUADRO 1.5 • Taxas de juro aparentes das obrigações (%).

15.26	15.81	15.90	15.35	15.32	15.59	15.44	14.79
16.64	15.23	15.56	15.53	15.46	16.59	14.94	15.54
15.48	15.67	15.94	15.67	15.37	16.93	15.90	15.37
15.56	15.40	15.98	15.50	15.46	15.62	16.64	15.62
15.62	15.51	15.51	15.36	16.08	15.60	14.82	15.08
15.65	15.50	15.39	17.08	15.56	15.84	14.07	19.04
12.24	15.59	15.16					

Fonte: Banco de Portugal, Relatório do Conselho de Administração, Lisboa, 1991.

QUADRO 1.6 Gráfico caule-e-folhas das taxas de juro aparentes (1.ª versão).

N.º de folhas	Caule	Folhas
1	12*	2
	12–	
	13*	
	13–	
1	14*	0
3	14–	789
15	15*	012233333344444
24	15–	555555555556666666889999
1	16*	0
4	16–	5669
1	17*	0
	17–	
	18*	
	18–	
1	19*	0

$n = 51$; unidade = 0.1.

© McGraw-Hill

QUADRO 1.7
Gráfico caule-e-folhas das taxas de juro aparentes (2.ª versão).

N.º de folhas	Caule	Folhas
3	Pe	(12.24, 14.07, 14.79)
1	148	2
1	149	4
1	150	8
1	151	6
2	152	36
6	153	256779
5	154	04668
11	155	00113466699
7	156	0222577
	157	
2	158	14
4	159	0048
1	160	8
6	Gr	(16.59, 16.64, 16.64, 16.93, 17.08, 19.04)

$n = 51$; unidade = 0.01.

EXEMPLO 1.11 • No quadro 1.8 encontra-se o gráfico caule-e-folhas construído com as observações referentes aos 84 concelhos da Região Norte (dados extraídos do *Inventário Municipal*, Norte, 1992). A variável observada diz respeito ao equipamento dos centros de saúde, e mede o índice de atracção de cada concelho. O índice de atracção é o quociente entre a população que utiliza determinado equipamento e a população residente no concelho. Quando este indicador é superior a 1 conclui-se que o concelho é atractivo para o equipamento considerado, isto é, a população que utiliza o equipamento situado neste concelho é superior à sua população residente.

QUADRO 1.8
Gráfico caule-e-folhas do índice de atracção.

N.º de folhas	Caule	Folhas
	0*	
6	t	223333
26	f	44444444455555555555555555
27	s	666666666666666667777777777
13	0·	8888899999999
11	1*	00000000000
	t	
1	f	4

$n = 84$; unidade = 0.1.

© McGraw-Hill

Nesta variante, o caule correspondente a cada valor tem cinco linhas. Por exemplo, no caule correspondente ao valor 0, na linha 0* não há observações (o dígito dominado é 0 ou 1); na linha **t**, as observações 0.2 e 0.3 (o dígito dominado é 2 — **t**wo — ou 3 — **t**hree); na linha **f**, as observações 0.4 e 0.5 (o dígito dominado é 4 — **f**our — ou 5 — **f**ive); na linha **s**, as observações 0.6 e 0.7 (o dígito dominado é 6 — **s**ix — ou 7 — **s**even); finalmente, na linha 0·, as observações 0.8 e 0.9. Na análise do gráfico nota-se uma assimetria não muito pronunciada e uma grande concentração de valores em torno de 0.6 e 0.7. O valor grande, 1.4, corresponde ao concelho de S. João da Madeira, em que 100% das freguesias possuíam centros de saúde. ⊙

Em resumo, com a construção dos gráficos caule-e-folhas procura dar-se, através de uma análise exploratória extremamente simples, uma visão de conjunto da colecção de dados, salientando ou identificando propriedades, tais como:

a) Existência de valores em torno dos quais se concentram os restantes;
b) Maior ou menor dispersão dos valores;
c) Simetrias ou assimetrias;
d) Presença de valores extremos («pequenos» e «grandes»).

1.4 Distribuições de frequências. Histogramas

Apresentando-se ao investigador uma colecção numerosa de dados, os gráficos caule-e-folhas perdem boa parte do seu interesse por não ser possível analisá-los com facilidade. Torna-se então necessário recorrer às chamadas **distribuições de frequências** ou distribuições empíricas. Começa-se por analisar o caso em que os valores de determinado atributo podem ser descritos por uma variável discreta.

Considere-se uma colecção de dados $\{x_1, x_2, \ldots, x_n\}$, onde x_i representa o valor numérico do atributo observado para a unidade estatística i $(i = 1, 2, \ldots, n)$. Esta colecção pode ser analisada como tal ou constituir uma amostra de dimensão n de certa população. Seja

$$a_1, a_2, \ldots, a_j, \ldots, \tag{1.2}$$

a sucessão disposta por ordem crescente dos valores possíveis que a variável x pode assumir. O conjunto dos valores da sucessão chama-se **domínio** da variável x, e representa-se por A. Nos casos mais correntes tem-se uma sucessão de números inteiros.

Chama-se **frequência absoluta** de a_j na colecção, e representa-se por F_j, ao número de vezes que este valor é observado. Deste modo, tem-se a sucessão das frequências absolutas,

$$F_1, F_2, \ldots, F_j, \ldots, \tag{1.3}$$

verificando-se, evidentemente,

$$\sum_j F_j = n, \tag{1.4}$$

© McGraw-Hill

onde n é o número total de observações da colecção ou amostra. A **frequência relativa** do valor a_j, representada por f_j, é definida pelo quociente entre a frequência absoluta e a dimensão da colecção, ou seja, $f_j = F_j/n$.

De forma semelhante a (1.3) obtém-se a sucessão das frequências relativas,

$$f_1, f_2, \ldots, f_j, \ldots, \tag{1.5}$$

verificando-se obviamente,

$$\sum_j f_j = 1. \tag{1.6}$$

Quando se fala simplesmente em **frequência** pretende normalmente significar-se frequência relativa.

A construção de quadros de frequências para variáveis discretas não apresenta, regra geral, dificuldades, como se pode ilustrar com o seguinte exemplo.

EXEMPLO 1.12 • O quadro 1.9 contém a distribuição de frequências do número de indivíduos por agregado doméstico privado no Continente retirada do Inquérito aos Orçamentos Familiares 1989/1990 (INE, 1992). A amostra é composta por 9640 agregados.

A segunda coluna do quadro indica as frequências absolutas no conjunto dos 9640 agregados (por exemplo, há 2304 agregados domésticos privados com três indivíduos, e 848 com cinco), enquanto a terceira coluna se refere às frequências relativas (por exemplo: 23.9% dos agregados têm três indivíduos: 0.239 = 2304/9640; 8.8% dos agregados são compostos por cinco indivíduos: 0.088 = 848/9640). ⊙

QUADRO 1.9
Número de indivíduos por agregado doméstico privado.

Número de indivíduos	Frequências absolutas	Frequências relativas
1	1138	0.118
2	2748	0.285
3	2304	0.239
4	2082	0.216
5	848	0.088
≥ 6	520	0.054
Total	9640	1.000

Fonte: Inquérito aos Orçamentos Familiares 1989/1990, INE, Lisboa, 1992.

Quando a variável x é de tipo contínuo, o processo é um pouco mais elaborado, uma vez que não faz muito sentido falar em frequências absolutas (relativas) de valores isolados. Neste caso, é necessário proceder à definição de classes de valores, à contagem das frequências absolutas e ao cálculo das frequências relativas de cada classe.

© McGraw-Hill

Para classes de valores de uma variável contínua tomam-se, em regra, intervalos, correntemente designados por **intervalos de classe**. A definição das classes consiste, assim, na fixação dos limites desses intervalos ou **limites de classe**. Admita-se que o número de classes é fixado em m, e represente-se o j-ésimo intervalo de classe por I_j.

Para a arrumação ser feita sem ambiguidade é necessário que os intervalos de classe não tenham pontos em comum, isto é,

$$I_j \cap I_k = \varnothing, \quad j \neq k, \tag{1.7}$$

onde $\cap$ designa intersecção de conjuntos, e $\varnothing$, o conjunto vazio. Para não ficarem valores por classificar, o domínio A da variável deve ser igual à união de todos os intervalos,

$$A = \bigcup_{j=1}^{m} I_j, \tag{1.8}$$

onde $\cup$ representa a união de conjuntos.

Seguindo estas duas regras, que introduzem uma partição do conjunto A, podem definir-se os m intervalos de classe,

$$I_1 = (\ell_0, \ell_1], \ldots, I_j = (\ell_{j-1}, \ell_j], \ldots, I_m = (\ell_{m-1}, \ell_m], \tag{1.9}$$

abertos à esquerda e fechados à direita, a verificar

$$\ell_0 < \ell_1 < \cdots < \ell_j < \cdots < \ell_m, \ \ell_0 \leq \min(x_i) \text{ e } \ell_m \geq \max(x_i),$$

onde $\min(x_i)$ e $\max(x_i)$ representam, respectivamente, o mínimo e o máximo dos valores observados. Deste modo, um valor x pertence à j-ésima classe quando e só quando $\ell_{j-1} < x \leq \ell_j$. Por vezes, é conveniente considerar o primeiro intervalo fechado à esquerda e/ou o último intervalo aberto à direita.

A amplitude da j-ésima classe é dada por,

$$h_j = \ell_j - \ell_{j-1} \ (j = 1, 2, \ldots, m). \tag{1.10}$$

Quando não haja inconveniente e tal seja possível, devem tomar-se classes com amplitude constante, $h_j = h \ (j = 1, 2, \ldots, m)$, propriedade que pode tornar mais sugestiva a comparação das frequências das diversas classes.

Feito o apuramento dos valores pertencentes a cada classe, a frequência absoluta da classe I_j $(j = 1, 2, \ldots, m)$, F_j, ou a frequência relativa da mesma, f_j, são obtidas pelos métodos descritos anteriormente.

A construção de intervalos de classe também pode ser feita para variáveis discretas quando estas assumem um grande número de valores. No exemplo 1.12 a última classe considerada corresponde, em rigor, a um intervalo de classe.

© McGraw-Hill

EXEMPLO 1.13 • Os dados do quadro 1.10 representam o tempo de hemodiálise (HD) antes da transplantação renal efectuada em 104 doentes no Hospital de Santa Cruz[5]. Embora teoricamente se trate de uma variável contínua, as observações são medidas em meses.

No conjunto dos 104 valores observados, $\min(x_i) = 3$ e $\max(x_i) = 144$. Em face destes valores tomou-se 0 para limite inferior da primeira classe (o que é acertado, visto que o tempo de hemodiálise é não negativo) e 150 para limite superior da última classe. Formaram-se então 10 classes com a mesma amplitude $h = 15$. As classes são, portanto, em meses, (0, 15], (15, 30], ..., (135, 150]. O apuramento do número de valores em cada classe conduziu ao quadro de frequências (quadro 1.11). ⊙

QUADRO 1.10
Tempo de HD (meses) em 104 doentes transplantados.

51	35	7	17	60	50	36	64
24	60	35	24	12	55	54	70
55	42	79	18	44	96	28	19
75	34	28	19	36	36	36	54
24	60	16	32	3	60	42	10
27	65	26	52	22	26	32	21
22	48	67	37	76	69	5	30
23	144	97	49	24	19	38	55
48	84	59	103	18	43	21	30
18	50	26	18	10	22	18	72
96	28	39	136	36	58	10	108
24	36	47	84	37	10	15	30
26	93	135	16	52	21	32	30

QUADRO 1.11
Distribuição do tempo de HD (meses) em 104 doentes transplantados.

Tempo de HD Classes	Frequências absolutas	Frequências relativas
(0, 15]	9	0.087
(15, 30]	35	0.337
(30, 45]	20	0.192
(45, 60]	20	0.192
(60, 75]	7	0.067
(75, 90]	4	0.038
(90, 105]	5	0.048
(105, 120]	1	0.010
(120, 135]	1	0.010
(135, 150]	2	0.019
Total	104	1.000

[5] Dados amavelmente cedidos por Jacinto Simões e estudados em *Falências Múltiplas de Órgãos*, dissertação para doutoramento na Faculdade de Ciências Médicas de Lisboa, 1993.

© McGraw-Hill

EXEMPLO 1.14 • A Tabaqueira, SA faz um apertado controlo da qualidade dos cigarros que produz; o peso é uma das características rigorosamente acompanhadas. Com os pesos de uma amostra de 500 cigarros SG Filtro construiu-se a distribuição de frequências do quadro 1.12. ⊙

QUADRO 1.12
Distribuição do peso de 500 cigarros SG Filtro*.

Peso (mg) Classes	Frequências absolutas	Frequências relativas
(760, 780]	4	0.008
(780, 800]	43	0.086
(800, 820]	118	0.236
(820, 840]	168	0.336
(840, 860]	117	0.234
(860, 880]	39	0.078
(880, 900]	11	0.022
Total	500	1.000

* Dados amavelmente cedidos pela empresa.

EXEMPLO 1.15 • Do Inquérito à Estrutura das Explorações Agrícolas de 2003 retirou-se o quadro de frequências (ver quadro 1.13), onde, em função da superfície agrícola utilizada (SAU), se indica o número de explorações da região agrária do Alentejo. Não foram consideradas as explorações sem SAU, e teve de seguir-se o critério das classes fechadas à esquerda e abertas à direita, forma apresentada pelo inquérito.

Trata-se de exemplo típico da situação em que não é prático considerar intervalos de classe com a mesma amplitude. Por um lado, devido ao grande número de pequenas explorações, interessa discriminar, por exemplo, entre as que têm menos de 1 hectare e as que têm 1 a 5 hectares; por outro lado, é informação desnecessária saber quantas explorações têm 101 a 102 ou 102 a 105 hectares. Se fosse adoptada uma amplitude constante ou se perdia a informação referente às pequenas explorações ou, querendo reter esta, havia um número impraticável de classes. ⊙

QUADRO 1.13
SAU nas explorações de Alentejo (2003).

SAU (ha) Classes	Frequências Absolutas	Frequências Relativas
$0 < \text{SAU} < 1$	5 314	0.12249
$1 \leq \text{SAU} < 5$	17 003	0.39192
$5 \leq \text{SAU} < 20$	10 518	0.24244
$20 \leq \text{SAU} < 50$	4 016	0.09257
$\text{SAU} \geq 50$	6 533	0.15058
Total	43 384	1.00000

É corrente encontrar quadros de frequências em que a primeira e a última classes, ou só uma delas, apresentam, respectivamente, o limite inferior e o limite superior indefinidos: é o que sucede com as últimas classes dos exemplos 1.12 e 1.15. Sempre que possível deve evitar-se esta situação,

© McGraw-Hill

pois ela pode prejudicar, como vai ver-se nas secções seguintes, a realização de certos cálculos. No entanto, o problema pode não ser particularmente grave porquanto, em muitos casos, o observador dispõe da colecção original.

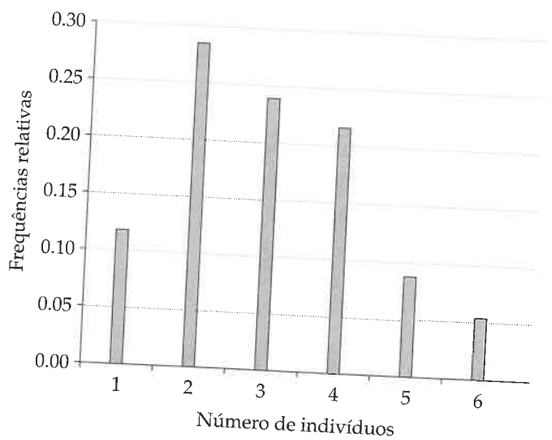
Já se disse, a propósito do número de linhas no gráfico caule-e-folhas, que há elevado grau de subjectividade na construção das diversas classes, problema que se repete para os quadros de frequências e que é, aliás, uma característica comum a muitos processos de classificação. Por um lado, quando se classificam as observações de uma variável contínua perde-se necessariamente informação, pois deixa de conhecer-se a variabilidade dentro de cada classe. Por outro lado, quando se arrumam os dados num número excessivo de classes, acontece que o número total de observações não é suficiente para pôr em evidência a eventual regularidade do fenómeno. A solução passa, portanto, por um compromisso, não sendo possível dar instruções objectivas sobre o número de classes a considerar.

A ideia de conjunto de uma distribuição de frequências pode ser melhorada com a da respectiva representação gráfica. Quando a variável é discreta recorre-se aos **diagramas de barras**, representando-se no eixo das abcissas os valores de x, e no eixo das ordenadas, as frequências (relativas ou absolutas) com que os valores foram observados.

EXEMPLO 1.16 • Na figura 1.2 faz-se a representação, em diagrama de barras, da distribuição do número de indivíduos por agregado doméstico privado (ver exemplo 1.12), supondo, para facilitar, que a última classe corresponde a seis indivíduos e não a seis indivíduos ou mais. ⊙

A representação através de frequências relativas, como se fez no exemplo 1.16, é mais sugestiva na comparação de diagramas de distribuições que tenham desigual número de observações.

FIGURA 1.2
Diagrama de barras do número de indivíduos por agregado doméstico privado.



© McGraw-Hill

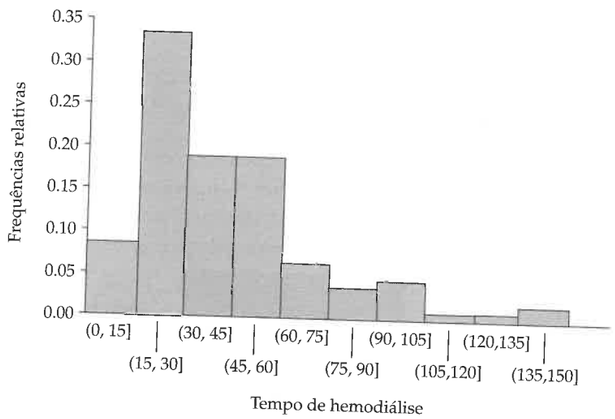
A representação gráfica de distribuições de frequências de variáveis contínuas é obtida por meio de um diagrama de áreas, o **histograma**, formado por uma sucessão de rectângulos adjacentes. Quando os intervalos de classe têm amplitude constante, h, cada rectângulo tem base h e altura igual à respectiva frequência relativa (absoluta). Assim, quando, por exemplo, se empregam frequências relativas, a área do rectângulo correspondente à classe I_j $(j = 1, 2, \ldots, m)$ é hf_j, e a soma das áreas é h. Como se pode sempre supor $h = 1$ (trata-se meramente de uma questão de escala), é indiferente dizer que a área ou a altura de cada rectângulo é igual à respectiva frequência relativa. Quando as classes têm amplitudes diferentes, h_j, os rectângulos têm base h_j e altura igual à respectiva frequência relativa (absoluta) dividida pela amplitude do intervalo de classe, f_j / h_j. Deste modo, a área do rectângulo correspondente à classe I_j $(j = 1, 2, \ldots, m)$ é a frequência relativa (absoluta), e a soma das áreas é igual a 1 (número de observações, n).

Tal como nos diagramas de barras, o recurso às frequências relativas costuma ser mais indicado porque permite comparar colecções com um número desigual de observações, na condição de os intervalos de classe serem os mesmos nas duas colecções.

EXEMPLO 1.17 • Na figura 1.3 apresenta-se o histograma da distribuição do tempo de hemodiálise. As bases dos rectângulos são os segmentos correspondentes aos intervalos (0, 15], (15, 30], (30, 45], ..., e as alturas 0.087, 0.337, 0.192, ... correspondem às frequências relativas. ⊙

EXEMPLO 1.18 • A distribuição do peso numa amostra de 500 cigarros SG Filtro é descrita pelo histograma da figura 1.4. As bases dos rectângulos são os segmentos correspondentes aos intervalos (760, 780], (780, 800], (800, 820], ..., e as alturas, 0.008, 0.086, 0.236, ... ⊙

FIGURA 1.3
Histograma da distribuição do tempo de HD (meses).



© McGraw-Hill

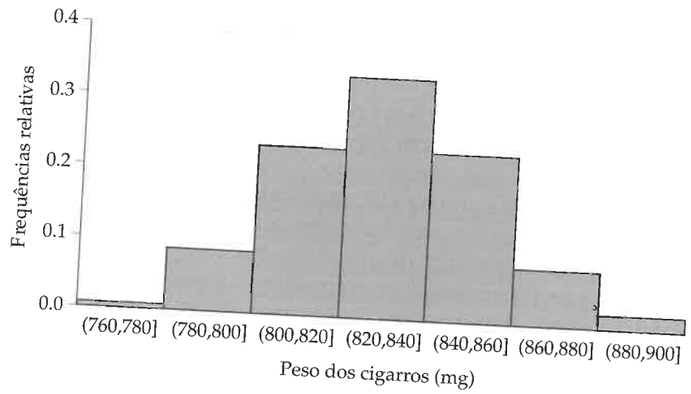


FIGURA 1.4
Histograma da distribuição do peso dos cigarros SG Filtro.

EXEMPLO 1.19 • No exemplo 1.15 apresentaram-se, em função da superfície agrícola utilizada (SAU), as frequências relativas do número de explorações da região agrária do Alentejo. Com base nestes dados, construiu-se um histograma com classes de amplitude variável (figura 1.5).

As alturas das classes encontram-se no quadro 1.14, onde se supôs que a amplitude do último intervalo de classe é de 450, uma vez que se considerou que o limite superior da última classe é 500. ⊙

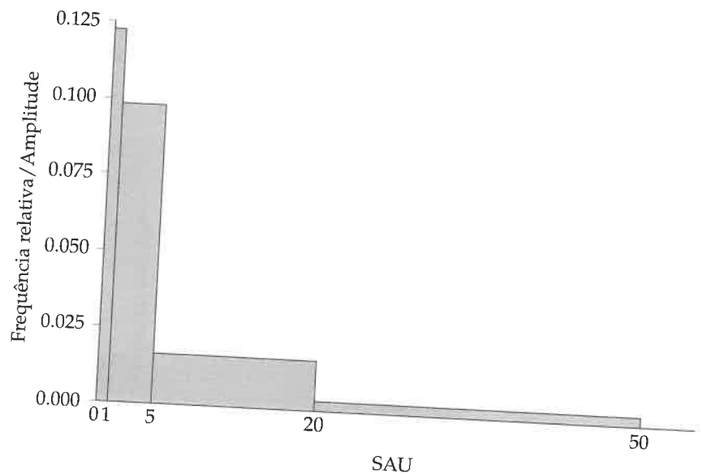


FIGURA 1.5
Histograma com rectângulos de áreas iguais às frequências relativas.

© McGraw-Hill

QUADRO 1.14
SAU nas explorações de Alentejo. Alturas dos intervalos de classe.

Classes	h_j	f_j	f_j/h_j
$0 < \text{SAU} < 1$	1	0.12249	0.12249
$1 \leq \text{SAU} < 5$	4	0.39192	0.09798
$5 \leq \text{SAU} < 20$	15	0.24244	0.01616
$20 \leq \text{SAU} < 50$	30	0.09257	0.00309
$\text{SAU} \geq 50$	450	0.15058	0.00033

1.5 Características numéricas: média e desvio padrão

As secções anteriores ocuparam-se com a descrição gráfica de colecções de dados ou amostras. Importa agora estudar a forma corrente de proceder à descrição através de características numéricas que sumariem a informação disponível.

Quando o **atributo** observado é **quantitativo**, uma das principais características numéricas, certamente conhecida do leitor, é a média. A **média** de uma colecção ou de uma amostra constituída pelos valores $\{x_1, x_2, \ldots, x_n\}$ designa-se por $\bar{x}$, e define-se pela expressão:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{1.11}$$

Sempre que possível, para não desprezar informação, a média deve ser calculada a partir dos dados originais; no entanto, se se dispõe apenas das distribuições de frequências, torna-se necessário recorrer à hipótese básica da tabulagem, que consiste em atribuir a todas as observações de uma determinada classe o valor dado pelo **ponto médio** desta, x'_j ($j = 1, 2, \ldots, m$), isto é, de acordo com (1.9) e (1.10),

$$x'_j = \ell_{j-1} + \tfrac{1}{2}h_j = \frac{\ell_{j-1} + \ell_j}{2} \,. \tag{1.12}$$

A natureza do erro que assim se comete, chamado **erro de tabulagem**, deve ser tida em conta quando se analisam os resultados. Note-se que é importante não confundir os pontos médios das m classes, $x'_1, x'_2, \ldots, x'_m$, e os n valores originais da colecção $\{x_1, x_2, \ldots, x_n\}$. Trata-se então de calcular médias ponderadas tomando as frequências relativas como pesos. Tem-se, portanto,

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} F_j x'_j = \sum_{j=1}^{m} f_j x'_j \,. \tag{1.13}$$

A média é uma medida de **localização** que, geralmente, indica um valor central da distribuição, entendido como o valor em torno do qual se distribuem os valores da colecção. Por isso, a média é, muitas vezes, usada como valor representativo de uma colecção. Na vida quotidiana ouve-se falar constantemente no rendimento médio, no preço médio, no custo médio, na média de idades, etc.

© McGraw-Hill

Da definição de média sai imediatamente,

$$\sum_{i=1}^{n}(x_i-\overline{x})=0\,, \tag{1.14}$$

propriedade que se enuncia do seguinte modo: a soma dos desvios em relação à média é nula.

Outras medidas de localização são a mediana (a estudar na secção seguinte) e a moda. Se se considerarem os dados originais, a **moda**, m_*, é o valor com maior frequência (com os dados do exemplo 1.12, verifica-se imediatamente que a moda é 2). A moda pode ser determinada qualquer que seja a natureza do atributo observado (quantitativo, qualitativo em escala nominal ou qualitativo em escala ordinal).

Se os dados estão classificados em intervalos de igual amplitude, a classe com maior frequência chama-se **classe modal**. Quando as amplitudes são variáveis, a classe modal é aquela que corresponde ao rectângulo com maior altura no histograma. A moda pode considerar-se o ponto médio da classe modal[6].

Além da localização, outro aspecto importante a considerar quando se descrevem colecções é a **dispersão**. A dispersão deve definir-se tomando em conta a posição das observações em relação a uma referência fixa que, como se afigura natural e lógico, deve ser o valor escolhido para localizar a colecção ou a distribuição empírica. Quando a medida de localização usada é a média, esta deve ser tomada para referenciar a dispersão das observações. Assim, em tal hipótese, uma medida de dispersão deve sintetizar o comportamento do conjunto de desvios em relação à média, $x_i-\overline{x}$; havendo pouca dispersão, os desvios devem ser globalmente pequenos; havendo muita dispersão os desvios devem ser globalmente grandes. Como a soma dos desvios é igual a zero — (1.14) — ocorre tomar os quadrados dos desvios para não haver compensação entre valores positivos e negativos. Assim, quando se considera a média dos quadrados dos desvios em relação à média,

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\,, \tag{1.15}$$

obtém-se a **variância** da colecção de dados ou da amostra.

Quando a amostra tem dimensão pequena utiliza-se preferencialmente a **variância corrigida**[7],

$$s'^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\,, \tag{1.16}$$

por razões que são estudadas no capítulo 6.

Sendo estas medidas valores médios de quadrados, a comparação com a média da amostra não é directamente possível. Torna-se assim mais útil recorrer à raiz quadrada positiva da variância (corrigida), que se designa por **desvio padrão** (corrigido). A definição é dada respectivamente por

$$s=+\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} \tag{1.17}$$

[6] Esta aproximação pode ser melhorada usando a fórmula de King [ver Murteira (1993)].

[7] Que alguns autores designam apenas por variância.

© McGraw-Hill

e

$$s'=+\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}\,. \tag{1.18}$$

Quando se dispõe apenas de dados classificados, as expressões anteriores são adaptadas tendo em conta a hipótese básica de tabulagem. Obtêm-se assim, respectivamente, as seguintes expressões para a variância e para a variância corrigida:

$$s^2=\frac{1}{n}\sum_{j=1}^{m}F_j(x'_j-\overline{x})^2=\sum_{j=1}^{m}f_j(x'_j-\overline{x})^2\,, \tag{1.19}$$

$$s'^2=\frac{1}{n-1}\sum_{j=1}^{m}F_j(x'_j-\overline{x})^2=\frac{n}{n-1}\sum_{j=1}^{m}f_j(x'_j-\overline{x})^2\,. \tag{1.20}$$

Os valores do desvio padrão e do desvio padrão corrigido são obtidos tomando as raízes quadradas positivas destas expressões.

Muitas vezes, o cálculo da variância pode ser feito de forma mais simplificada, notando que

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n}\sum_{i=1}^{n}(x_i^2-2x_i\overline{x}+\overline{x}^2)=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2\,. \tag{1.21}$$

Deste modo, a variância de uma colecção é igual à média dos quadrados dos valores observados menos o quadrado da média dos mesmos valores.

Para dados classificados tem-se

$$s^2=\sum_{j=1}^{m}f_j\,x'^2_j-\overline{x}^2\,. \tag{1.22}$$

O cálculo da média e da variância (ou desvio padrão) de uma amostra é um exercício simples.

EXEMPLO 1.20 • Considerando as taxas de nupcialidade apresentadas no exemplo 1.1, obtêm-se recorrendo a meios de cálculo automático (uma folha de cálculo ou uma simples máquina de calcular):

média: $\overline{x}=7.816$;
variância: $s^2=1.581$;
desvio padrão: $s=1.257$;
variância corrigida: $s'^2=1.619$;
desvio padrão corrigido: $s'=1.272$. ⊙

EXEMPLO 1.21 • Retomando o exemplo 1.14, sobre o peso dos cigarros SG Filtro, apresentam-se, no quadro 1.15, os cálculos intermédios para aplicar as fórmulas (1.13) e (1.22).

© McGraw-Hill

A partir deste quadro, obtém-se:

$$\bar{x} = \frac{415\,240}{500} = 830.48\,;$$

$$s^2 = \frac{345\,127\,600}{500} - 830.48^2 = 558.1696\,;$$

$$s = 23.6256\,;$$

Classe modal: (820, 840]. ⊙

QUADRO 1.15
Distribuição dos pesos de 500 cigarros SG Filtro (cálculos intermédios).

x'_j	F_j	$F_j x'_j$	$F_j x'^2_j$
770	4	3 080	2 371 600
790	43	33 970	26 836 300
810	118	95 580	77 419 800
830	168	139 440	115 735 200
850	117	99 450	84 532 500
870	39	33 930	29 519 100
890	11	9 790	8 713 100
Total	500	415 240	345 127 600

EXEMPLO 1.22 • Considerando os dados do quadro 1.13 (SAU nas explorações do Alentejo) e supondo que o limite superior da última classe é 500, obtém-se:

$$\bar{x} = 48.9184\,;$$
$$s^2 = 9\,149.8517\,;$$
$$s = 95.6549\,;$$
Classe modal: (0, 1).

Note-se que a classe modal corresponde a uma altura de 0.12249 do respectivo rectângulo do histograma (ver figura 1.5). Além disso, o rectângulo adjacente, referente à classe modal [1, 5), tem frequência relativa igual a 0.39192, que é superior à da classe modal. Porém, como aquela classe tem amplitude 4, o valor de f_2 / h_2 é 0.09798, inferior ao da classe modal. ⊙

Além do desvio padrão existem outras medidas de dispersão, como, por exemplo, o **desvio absoluto médio**,

$$d = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \text{ ou } d = \frac{1}{n}\sum_{j=1}^{m} F_j\,|x'_j - \bar{x}|. \tag{1.23}$$

© McGraw-Hill

O desvio padrão e o desvio absoluto médio são medidas de dispersão absoluta. Quando todas as observações têm o sinal positivo, uma medida de dispersão relativa muito usada é o **coeficiente de variação**,

$$\text{CV} = \frac{s}{\bar{x}}, \tag{1.24}$$

que, como se verifica, é independente da unidade em que se exprime a variável e permite a comparação entre duas ou mais distribuições.

1.6 Características numéricas: estatísticas de ordem

Quando o atributo é quantitativo ou qualitativo em escala ordinal, os dados da colecção $\{x_1, x_2, \ldots, x_n\}$ podem ser ordenados,

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} \text{ com } x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)},$$

e diz-se que se calcularam as **estatísticas de ordem.** Assim, $x_{(i)}$ é a estatística de ordem i, com $i = 1, 2, \ldots, n$. Duas estatísticas de ordem importantes são os **extremos**,

$$x_{(1)} = \min(x_i), \quad x_{(n)} = \max(x_i).$$

A partir das estatísticas de ordem podem ser definidas outras características numéricas da colecção. É dado particular realce à **mediana**, que é a principal concorrente da média como medida de localização. De forma aproximada, a mediana, que se representa por M, é o valor da colecção que tem 50% de observações inferiores e 50% de observações superiores. Em termos rigorosos, a mediana pode definir-se da seguinte maneira:

$$M = \begin{cases} x_{(k+1)} & \text{se } n = 2k+1, \\ (x_{(k)} + x_{(k+1)})/2 & \text{se } n = 2k. \end{cases} \tag{1.25}$$

Verifica-se então que: quando n é ímpar, a mediana é a estatística de ordem $k+1$; quando n é par, a mediana é a média simples das estatísticas de ordens k e $k+1$.

Comparando a mediana com a média, facilmente se conclui o seguinte: enquanto no cálculo da média basta alterar uma observação para que a média varie, a mediana é definida pela sua posição na sucessão ordenada de observações, isto é, a mediana ocupa o lugar central na sucessão das estatísticas de ordem (está, portanto, muito menos sujeita à influência de observações extremas ou, em linguagem mais técnica, é uma medida mais resistente do que a média).

Quando se consideram os dados originais, o cálculo da mediana é imediato após proceder à ordenação das observações.

© McGraw-Hill

EXEMPLO 1.23 • Considere-se o exemplo 1.10, referente às taxas de juro aparentes das obrigações. Ordenando as 51 observações conclui-se facilmente que a mediana, estatística de ordem 26, é igual a 15.54. Este valor também podia ser obtido, sem dificuldade, a partir do gráfico caule-e-folhas (quadro 1.7). ⊙

A mediana é um caso particular dos **quantis**. Em termos aproximados, o quantil de ordem α, $0<\alpha<1$, seja q_α, é o valor da colecção que tem αn observações inferiores e $(1-\alpha)n$ observações superiores. A mediana é o quantil de ordem $\alpha = 0.5$. Outros quantis importantes, muito utilizados na prática, são os **percentis**, obtidos quando $\alpha = 0.01, 0.02, \ldots, 0.99$, os **decis**, quando $\alpha = 0.1, 0.2, \ldots, 0.9$, o **primeiro quartil**, $q_{0.25}$, ou simplesmente Q_1, e o **terceiro quartil**, $q_{0.75}$, ou simplesmente Q_3.

Vai apresentar-se a seguir um procedimento para calcular os quantis quando se dispõe de uma colecção ordenada de dados em que, como se sabe, $x_{(1)}$ e $x_{(n)}$ são os extremos. Como existem $n-1$ intervalos unitários entre 1 e n, a ordem da observação correspondente ao quantil de ordem α é dada por $1+(n-1)\alpha$ se este valor for inteiro. Assim, fazendo $1+(n-1)\alpha = r$, o quantil de ordem α é $q_\alpha = x_{(r)}$.

Quando $1+(n-1)\alpha$ não é inteiro, seja

$$1+(n-1)\alpha = r+\varepsilon, \tag{1.26}$$

onde r é a parte inteira, e ε é a parte decimal ($0 \le \varepsilon < 1$). Neste caso, o valor $r+\varepsilon$ pode ser considerado a ordem aproximada da observação correspondente ao quantil. Na prática, o quantil de ordem α obtém-se por interpolação linear, fazendo a seguinte média ponderada entre $x_{(r)}$ e $x_{(r+1)}$:

$$q_\alpha = x_{(r)} + \varepsilon(x_{(r+1)} - x_{(r)}) = (1-\varepsilon)x_{(r)} + \varepsilon x_{(r+1)}. \tag{1.27}$$

Note-se também que o quantil de ordem α não se deve definir quando $\alpha < 1/(n+1)$ ou $\alpha > n/(n+1)$.

No caso particular dos quartis tem-se:

Quartis	Ordem aproximada
1.º quartil: Q_1	$(n+3)/4$
Mediana: M	$(n+1)/2$
3.º quartil: Q_3	$(3n+1)/4$

Alguns autores [Hogg e Tanis (2001), e Newbold, Carlson e Thorne (2007)] propõem o valor $(n+1)\alpha$ — em vez de $1+(n-1)\alpha$ — para determinar as ordens aproximadas. Neste caso, tem-se $(n+1)/4$ e $3(n+1)/4$ para os 1.º e 3.º quartis, respectivamente.

É possível definir medidas de dispersão (absoluta e relativa) a partir das estatísticas de ordem. Assim, a amplitude do intervalo de variação, $\text{AIV} = x_{(n)} - x_{(1)}$, e a amplitude do intervalo interquartis, $\text{AIQ} = Q_3 - Q_1$, são medidas de dispersão absoluta. O primeiro intervalo compreende todas as observações, o segundo intervalo compreende 50% das observações. Pode também considerar-se, quando todas as observações têm o mesmo sinal, a medida de dispersão relativa,

$$\frac{\text{AIQ}}{M}.$$

Os extremos, os quartis e a mediana permitem construir o **resumo dos 5 números**. Veja-se um exemplo.

© McGraw-Hill

EXEMPLO 1.24 • Os valores do quadro 1.16 indicam a população activa em percentagem da população residente para cada um dos 14 concelhos da Nomenclatura de Unidade Territorial III (NUT III) Alto Alentejo, subconjunto da NUT II Alentejo, no ano de 1991. A apresentação dos valores ordenados na forma que passa a indicar-se no quadro 1.17 mostra bem a capacidade descritiva dos extremos, dos quartis e da mediana. ⊙

QUADRO 1.16 • População activa em percentagem da população residente nos concelhos da NUT III Alto Alentejo.

Dados originais				Dados ordenados			
30.87	34.11	31.91	30.73	30.73	33.00	34.43	36.69
33.29	33.00	32.93	36.15	30.87	33.29	35.75	38.85
36.69	38.85	35.75	40.28	31.91	33.70	36.15	40.28
33.70	34.43			32.93	34.11		

Fonte: INE.

QUADRO 1.17 • População activa em percentagem da população residente nos concelhos da NUT III Alto Alentejo. Descrição por meio de extremos-mediana-quartis.

Min					Mediana						Max
30.73					**33.905**						40.28
	30.87				33.70	34.11					38.85
		31.91			33.29		34.43			36.69	
			32.93	33.00				35.75	36.15		
		1.º Quartil = 32.9475						**36.05 = 3.º Quartil**			

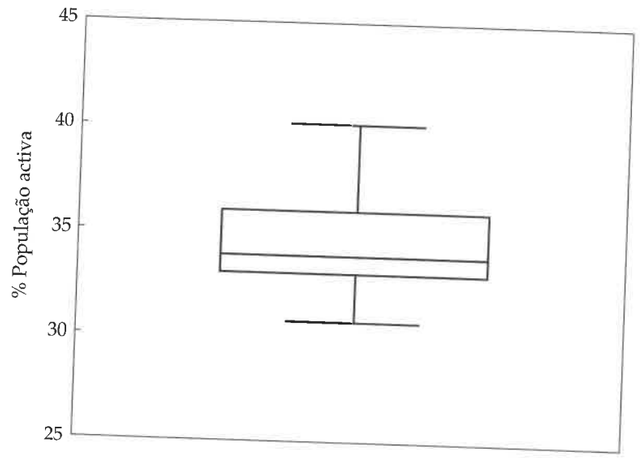
A capacidade descritiva do resumo dos 5 números aprecia-se melhor construindo um gráfico pitorescamente designado por **caixa-de-bigodes.** Na figura 1.6 apresenta-se o gráfico para os dados referidos no exemplo 1.24. A figura é suficientemente esclarecedora e indica a posição dos 5 números.

O rectângulo («a caixa») é desenhado de tal modo, que os seus lados inferior e superior correspondem aos 1.º e 3.º quartis. O segmento no seu interior refere-se à mediana. O mínimo e o máximo são representados pelos segmentos inferior e superior («os bigodes») desenhados no exterior do rectângulo. A caixa nos seus limites horizontais, isto é, entre o primeiro e o terceiro quartis, contém 50% das observações.

© McGraw-Hill

As caixas-de-bigodes prestam-se a comparações quando as variáveis se exprimem nas mesmas unidades.

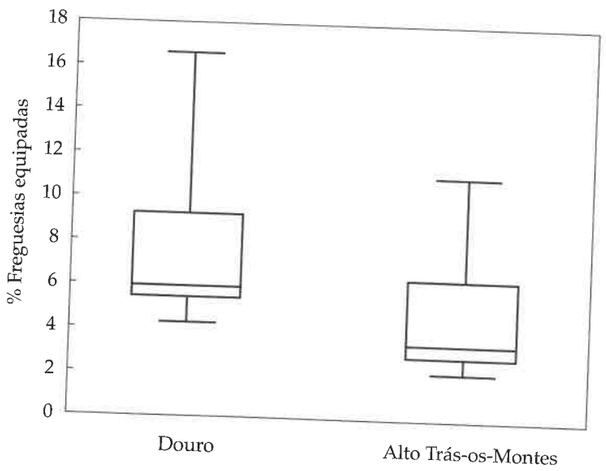
FIGURA 1.6
Caixa-de-bigodes da distribuição da população activa em percentagem da população residente nos concelhos do NUT III Alto Alentejo.



EXEMPLO 1.25 • Na figura 1.7 comparam-se as percentagens de freguesias equipadas com centros de saúde nos concelhos da NUT III — Douro e Alto Trás-os-Montes — da NUT II — Região Norte.

Como é evidente da análise da figura 1.7, a situação em Alto Trás-os-Montes é muito desfavorável quando confrontada com o que se passa no Douro. Verifica-se em ambos os casos que a mediana está muito mais próxima do 1.º quartil que do 3.º quartil. ⊙

FIGURA 1.7
Comparação da percentagem de freguesias equipadas no Douro e em Alto Trás-os-Montes.



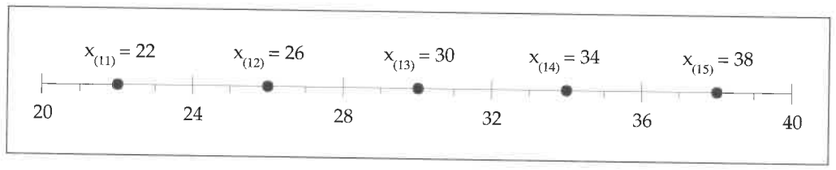
© McGraw-Hill

Quando se dispõe apenas dos dados classificados e se pretende calcular quantis (mediana, quartis, etc.) é necessário proceder de um modo adequado para estimar as estatísticas de ordem, $x_{(i)}$, já que, não se conhecendo os dados, não é possível ordená-los. O procedimento que vai propor-se permite estimar aquelas estatísticas de forma razoavelmente aproximada, e baseia-se na hipótese de que as observações classificadas em cada classe se ordenam de forma equidistante.

Para ilustrar esta ideia, vai apresentar-se um exemplo muito simples. Seja a seguinte distribuição de frequências:

Intervalos de classe	Frequências absolutas	Frequências absolutas acumuladas
(0, 20]	10	10
(20, 40]	5	15
(40, 60]	4	19

Suponha-se que se vai estimar a 14.ª estatística de ordem, $x_{(14)}$. Como facilmente se verifica, esta estatística pertence ao segundo intervalo de classe, I_2, com extremidades $\ell_1 = 20$ e, $\ell_2 = 40$, abrangendo 5 observações. Como o intervalo tem amplitude $h_2 = 20$, e se aceita a hipótese de que as observações estão ordenadas de forma equidistante, então supõe-se que as observações estão separadas por subintervalos de amplitude $h_2 / F_2 = 4$. Isto equivale a dividir o intervalo em 5 subintervalos de amplitude 4 e colocar as observações nos seus pontos médios. Assim, supõe-se que as 5 observações da classe I_2 ocupam as seguintes posições:



Pode, então, propor-se que:

$$\hat{x}_{(14)} = 34 = 20 + \left(14 - 10 - \frac{1}{2}\right)\frac{20}{5}.$$

Esta ilustração sugere a seguinte regra para estimar qualquer estatística de ordem $x_{(i)} \in I_j$:

$$\hat{x}_{(i)} = \ell_{j-1} + \left[\left(i - F^*_{j-1}\right) - \frac{1}{2}\right]\frac{h_j}{F_j}, \qquad (1.28)$$

onde: ℓ_{j-1} é o limite inferior da classe I_j; $F^*_{j-1} = F_1 + F_2 + \cdots + F_{j-1}$ é a frequência absoluta acumulada até a classe I_{j-1}; h_j é a amplitude da classe I_j; F_j é a frequência absoluta da mesma classe. Note-se que a ordem da estatística $x_{(i)}$ dentro da classe I_j é dada por $i - F^*_{j-1}$.

Como este procedimento permite estimar qualquer estatística de ordem, facilmente se podem calcular os quantis com (1.27).

© McGraw-Hill

Outro procedimento para estimar os quantis consiste em fazer uma interpolação na classe I_j a que pertence o quantil q_α. Assim, supondo que $q_\alpha \in I_j$, tem-se

$$\ell_{j-1} \le q_\alpha \le \ell_j \text{ e } f^*_{j-1} \le \alpha \le f^*_j ,$$

onde f^*_{j-1} e f^*_j são as frequências relativas acumuladas até às classes I_{j-1} e I_j, respectivamente. Então, estabelecendo a proporção,

$$\frac{f^*_j - f^*_{j-1}}{\ell_j - \ell_{j-1}} = \frac{\alpha - f^*_{j-1}}{q_\alpha - \ell_{j-1}},$$

ou, notando que $f^*_j - f^*_{j-1} = f_j$ e $\ell_j - \ell_{j-1} = h_j$,

$$\frac{f_j}{h_j} = \frac{\alpha - f^*_{j-1}}{q_\alpha - \ell_{j-1}},$$

obtém-se:

$$q_\alpha = \ell_{j-1} + \frac{\alpha - f^*_{j-1}}{f_j} h_j . \tag{1.29}$$

EXEMPLO 1.26 • Suponha-se que pretende calcular a mediana e os quartis do peso de 500 cigarros SG Filtro quando se dispõe não da colecção original, mas apenas da distribuição de frequências apresentada no quadro 1.12. No quadro 1.18 calculam-se as respectivas frequências acumuladas.

QUADRO 1.18
Distribuição do peso de 500 cigarros SG Filtro. Frequências absolutas e relativas acumuladas.

Peso (mg) Classes	Frequências absolutas acumuladas	Frequências relativas acumuladas
(760, 780]	4	0.008
(780, 800]	47	0.094
(800, 820]	165	0.330
(820, 840]	333	0.666
(840, 860]	450	0.900
(860, 880]	489	0.978
(880, 900]	500	1.000

Como $(n+1)/2 = 250.5$, a mediana é a média simples de $\hat{x}_{(250)}$ e $\hat{x}_{(251)}$. Como estes valores pertencem à classe (820, 840] (até 820 há 165 observações, e até 840 há 333 observações), tem-se:

$$\hat{x}_{(250)} = 820 + (250 - 165 - 0.5)(20/168) = 830.0595 ;$$
$$\hat{x}_{(251)} = 820 + (251 - 165 - 0.5)(20/168) = 830.1786 ;$$
$$M = (\hat{x}_{(250)} + \hat{x}_{(251)})/2 = 830.119 .$$

© McGraw-Hill

Para calcular os outros quartis procede-se de modo semelhante. Assim, para o 1.º quartil, como $(n+3)/4 = 125.75$, é necessário determinar $\hat{x}_{(125)}$ e $\hat{x}_{(126)}$. Como estes valores pertencem à classe (800, 820], obtém-se:

$$\hat{x}_{(125)} = 800 + (125 - 47 - 0.5)(20/118) = 813.1356 ;$$
$$\hat{x}_{(126)} = 800 + (126 - 47 - 0.5)(20/118) = 813.3051 ;$$
$$Q_1 = \hat{x}_{(125)} + 0.75(\hat{x}_{(126)} - \hat{x}_{(125)}) = 813.2627 ,$$

Para estimar o 3.º quartil faz-se $(3n+1)/4 = 375.25$, obtendo-se:

$$\hat{x}_{(375)} = 840 + (375 - 333 - 0.5)(20/117) = 847.094 ;$$
$$\hat{x}_{(376)} = 840 + (376 - 333 - 0.5)(20/117) = 847.265 ;$$
$$Q_3 = \hat{x}_{(375)} + 0.25(\hat{x}_{(376)} - \hat{x}_{(375)}) = 847.1368 ,$$

Utilizando (1.29) obtém-se:

$$M = 820 + \frac{0.50 - 0.33}{0.336} \times 20 = 830.119 ;$$
$$Q_1 = 800 + \frac{0.25 - 0.094}{0.236} \times 20 = 813.2203 ;$$
$$Q_3 = 840 + \frac{0.75 - 0.666}{0.234} \times 20 = 847.1795 .$$

Comparando estes resultados com os anteriores, verifica-se que o valor da mediana é praticamente igual, o 1.º quartil é ligeiramente inferior, e o 3.º quartil, ligeiramente superior. ⊙

Outra utilização dos quartis é a de fornecer critérios para determinar, de forma aproximada, a amplitude h (constante) dos intervalos de classe para classificar as observações de uma variável contínua. Um destes critérios [ver Freedman e Diaconis (1981)] estabelece o seguinte:

$$h = \frac{2 \times \text{AIQ}}{\sqrt[3]{n}} .$$

Por exemplo, no caso dos dados sobre os tempos de hemodiálise (exemplo 1.13), obtém-se:

$$h = \frac{2 \times 33}{\sqrt[3]{104}} = 14.03 .$$

No exemplo 1.13 foram consideradas amplitudes de 15 meses.

Para terminar esta secção vai apresentar-se um quadro-resumo (quadro 1.19) sobre as principais medidas de localização e dispersão. No quadro indicam-se, também, quais as medidas que se podem calcular para os vários tipos de atributos.

© McGraw-Hill

QUADRO 1.19 • Principais medidas de localização e de dispersão.

Medidas	$\bar{x}$	s	CV	m_e	M	AIQ	AIQ/M
Localização	×			×	×		
Dispersão absoluta		×				×	
Dispersão relativa			×				×
Atributo quantitativo	×	×	×	×	×	×	×
Atributo qualitativo nominal				×			
Atributo qualitativo ordinal				×	×	×	×

1.7 *Outliers*

A diferença entre os extremos (amplitude do intervalo de variação da amostra) reflecte, naturalmente, a dispersão. No entanto, é uma medida de dispersão paradigmática da falta de resistência à presença de valores que parecem muito grandes ou muito pequenos em relação às outras observações. Quando se pretende examinar a colecção para identificar estas observações, habitualmente designadas por ***outliers***, é usual recorrer a uma medida de dispersão resistente (pouco susceptível de ser «corrompida» por aquelas observações). O desvio padrão, por exemplo, pode não ser o mais adequado porque, sendo função de uma soma de quadrados, «alinha» em certa medida com os valores extremos. Uma medida de dispersão resistente é dada pela **amplitude interquartis**, $\text{AIQ} = Q_3 - Q_1$, que representa a amplitude do intervalo que compreende 50% das observações centrais da colecção, o que significa que no seu cálculo não intervêm nem as observações menores nem as maiores. Assim, a amplitude interquartis é o padrão de variação que logicamente se impõe. A regra prática consiste no seguinte: um qualquer valor da colecção, x_i, é um ***outlier* severo** quando

$$x_i < Q_1 - 3(Q_3 - Q_1) \quad \text{ou} \quad x_i > Q_3 + 3(Q_3 - Q_1), \tag{1.30}$$

e um ***outlier* moderado** quando

$$Q_1 - 3(Q_3 - Q_1) < x_i < Q_1 - 1.5(Q_3 - Q_1), \tag{1.31}$$

$$Q_3 + 1.5(Q_3 - Q_1) < x_i < Q_3 + 3(Q_3 - Q_1). \tag{1.32}$$

Os valores $Q_1 - 3(Q_3 - Q_1)$ e $Q_3 + 3(Q_3 - Q_1)$ são as chamadas **barreiras externas inferior** e **superior;** os valores $Q_1 - 1.5(Q_3 - Q_1)$ e $Q_3 + 1.5(Q_3 - Q_1)$ são as **barreiras internas inferior** e **superior**. A identificação (é ou não é?) e a interpretação (se é, porque é?) de *outliers* são tarefas complexas e altamente subjectivas que não podem aqui estudar-se. Porém, nota-se que, ao pesquisar valores com um comportamento que se afasta nitidamente do da grande maioria dos restantes valores da colecção, a explicação para a diferença encontrada resulta, muitas vezes: a) de erros humanos cometidos ao medir ou ao registar os dados (neste caso, trata-se de um *outlier* que convém eliminar do estudo); b) da própria natureza do fenómeno em estudo (neste caso, trata-se de um *outlier* que se deve manter na colecção).

A eventual presença de *outliers* obriga a modificar a caixa-de-bigodes tal como foi apresentada anteriormente. O esquema modificado consiste em calcular os quartis (1.° quartil, mediana e

© McGraw-Hill

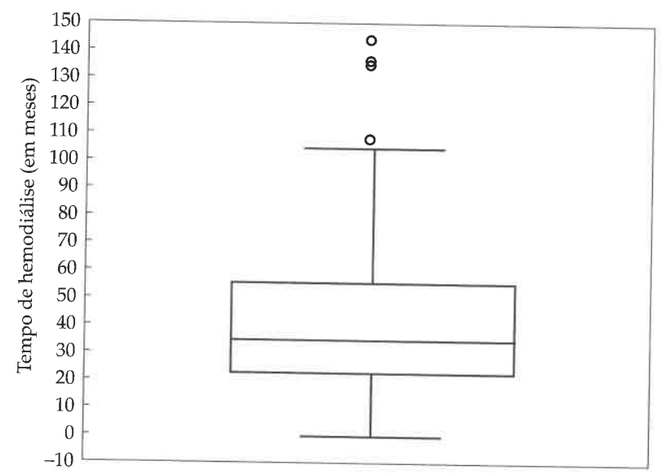
3.° quartil) tal como anteriormente, em determinar os extremos (máximo e mínimo) sem considerar os *outliers* e, a seguir, acrescentar estes valores, utilizando símbolos gráficos diferentes para distinguir os *outliers* moderados dos *outliers* severos.

EXEMPLO 1.27 • Na figura 1.8 apresenta-se a caixa-de-bigodes correspondente aos tempos de hemodiálise do quadro 1.10. Deixa-se ao leitor verificar, como exercício, que se tem $Q_1 = 22.75$ e $Q_3 = 55.75$, e, portanto, $Q_3 - Q_1 = 33$. Assim, sai:

Barreira interna inferior = 0, pois $Q_1 - 1.5(Q_3 - Q_1) = -26.75$;
Barreira externa inferior = 0, pois $Q_1 - 3(Q_3 - Q_1) = -76.25$;
Barreira interna superior = $Q_3 + 1.5(Q_3 - Q_1) = 105.25$;
Barreira externa superior = $Q_3 + 3(Q_3 - Q_1) = 154.75$;
outliers moderados: 108, 135, 136, 144.

Note-se que as barreiras inferiores não têm interesse prático já que os tempos de hemodiálise são positivos. ⊙

FIGURA 1.8
Caixa-de-bigodes dos tempos de HD.



EXEMPLO 1.28 • A figura 1.9 representa a caixa-de-bigodes correspondente às taxas de juro aparentes do quadro 1.5. Como $Q_1 = 15.37$, $Q_3 = 15.74$ e $Q_3 - Q_1 = 0.37$, vem:

Barreira interna inferior = $Q_1 - 1.5(Q_3 - Q_1) = 14.815$;
Barreira externa inferior = $Q_1 - 3(Q_3 - Q_1) = 14.26$;
Barreira interna superior = $Q_3 + 1.5(Q_3 - Q_1) = 16.295$;
Barreira externa superior = $Q_3 + 3(Q_3 - Q_1) = 16.85$;
outliers moderados: 14.79, 16,59, 16.64, 16.64;
outliers severos: 12.24, 14.07, 16.93, 17.08, 19.04. ⊙

© McGraw-Hill

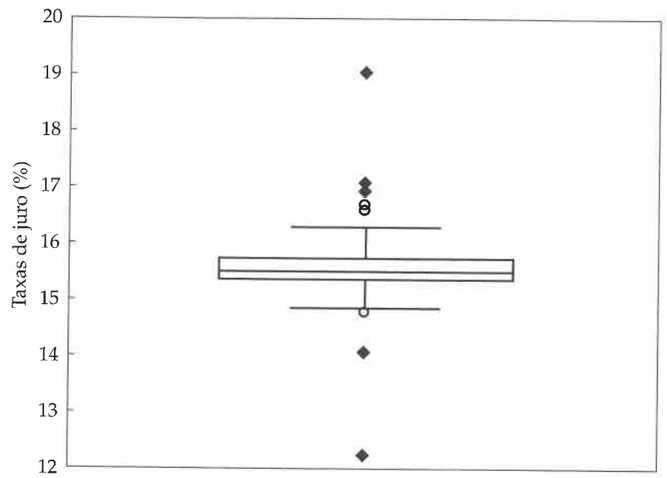


FIGURA 1.9
Caixa-de-bigodes das taxas de juro aparentes.

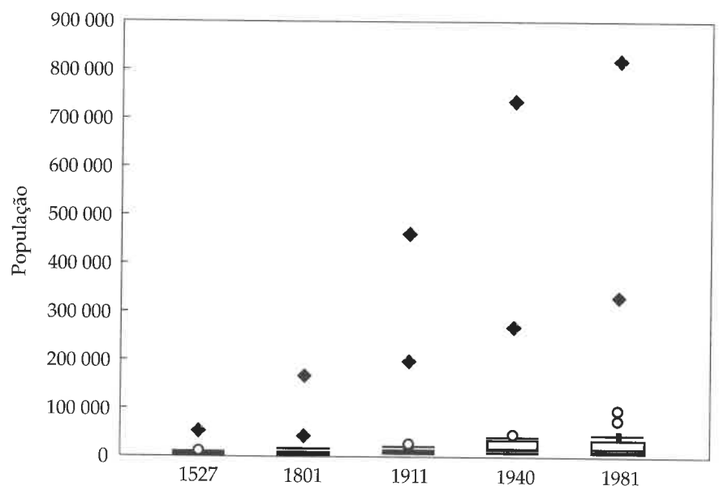


FIGURA 1.10
Caixa-de-bigodes da população dos centros urbanos de Portugal.

© McGraw-Hill

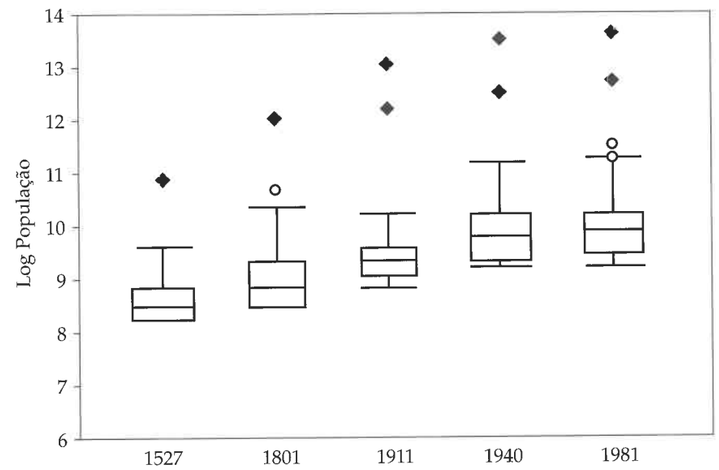


FIGURA 1.11
Caixa-de-bigodes do logaritmo da população dos centros urbanos de Portugal.

EXEMPLO 1.29 • Para ilustrar mais uma análise comparativa com caixas-de-bigodes, agora na presença de *outliers*, tomem-se os centros urbanos de Portugal nos anos de 1527, 1801, 1911, 1940 e 1981 [ver Salgueiro (1992), *A Cidade em Portugal, uma geografia urbana*, Edições Afrontamento]. A variável considerada foi a população; as caixas-de-bigodes encontram-se na figura 1.10 e são praticamente ilegíveis. A comparação é impossível, já que os *outliers* esmagam a distribuição dos pequenos centros. A prática recomendável, quando se comparam colecções com valores muito díspares, consiste em trabalhar com os logaritmos dos valores. Depois de transformar os dados, obtiveram-se, para a nova variável, as caixas-de-bigodes da figura 1.11.

A interpretação é agora mais fácil. Em 1527 apenas Lisboa se salienta como *outlier* (aliás, no Sec. XVI a população de Lisboa não se afastava muito da de Londres e era superior à de Sevilha, na data, a maior cidade de Espanha). A partir de 1801 verifica-se também a demarcação da cidade do Porto (Tratado de Methuen, expansão da cultura da vinha, etc.). Em 1981 aparecem dois novos *outliers* — Amadora e Setúbal — que, contudo, não ultrapassam as segundas barreiras. Exceptuando os referidos *outliers*, a figura 1.11 mostra uma evolução regular e equilibrada, em termos relativos, da maioria dos centros urbanos do país. ⊙

© McGraw-Hill

1.8 Correlação

Nas secções anteriores estudou-se a forma de explorar, descrever e comparar colecções de uma só variável. Viu-se como a estatística pode auxiliar o investigador a apreender, nos seus vários aspectos, o significado de uma colecção ou amostra. Quando se consideram, com o mesmo objectivo, colecções de pares de variáveis,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), \tag{1.33}$$

surge um problema novo: o estudo das relações que porventura existam entre os dois fenómenos observados. Investigar a natureza das relações gerais entre fenómenos é um dos principais propósitos da ciência, pois somente através do entendimento daquelas se podem atingir os objectivos desta, nomeadamente o aumento do conhecimento.

Cada uma das características observadas pode, como se sabe, ser de tipo qualitativo ou de tipo quantitativo. Nesta fase introdutória aborda-se apenas o caso em que ambas as características são de tipo quantitativo. Nos capítulos 9, 10 e 11 volta a tratar-se esta problemática, num quadro de hipóteses mais formalizado.

Considere-se então que se dispõe de uma colecção ou amostra referente a um **par de variáveis quantitativas**, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. A análise da colecção começa geralmente pelo estudo de cada uma das variáveis por si só, com base na metodologia abordada nas secções anteriores. Após esta primeira fase, inicia-se então o estudo das possíveis relações existentes entre as variáveis. Para tal, constrói-se um **diagrama de dispersão** onde se representam, num referencial cartesiano, os diferentes pontos da colecção. A observação deste diagrama permite, desde logo, formar uma ideia grosseira do tipo de associação que pode existir entre as variáveis.

EXEMPLO 1.30 • No quadro 1.20 apresentam-se os valores do produto interno bruto (PIB) e das importações de Portugal para o período 1979-1996, expressos em milhões de contos, a preços constantes de 1993.

A partir destes valores construiu-se o diagrama de dispersão apresentado na figura 1.12, que sugere a existência de uma relação aproximadamente linear entre estas duas variáveis. ⊙

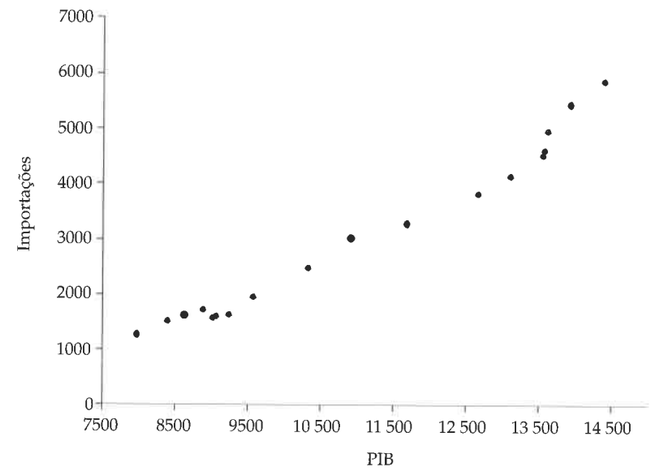
QUADRO 1.20 • PIB e importações de Portugal (1979-1996) — milhões de contos 1993.

Ano	PIB	Imp	Ano	PIB	Imp	Ano	PIB	Imp
1979	7980.0	1269.3	1985	9236.1	1645.8	1991	13100.8	4138.4
1980	8398.1	1522.1	1986	9573.8	1956.0	1992	13571.1	4593.3
1981	8639.7	1622.3	1987	10326.7	2492.0	1993	13545.9	4522.8
1982	8894.3	1717.6	1988	10919.1	3031.3	1994	13613.6	4970.6
1983	9062.8	1606.2	1989	11695.5	3288.8	1995	13926.7	5452.7
1984	9037.5	1572.9	1990	12664.8	3819.0	1996	14386.3	5861.7

Fonte: INE [10^6 contos ≡ 4 987 978.97 euros].

© McGraw-Hill

FIGURA 1.12
Diagrama de dispersão do PIB e das importações em Portugal (1979-1996).



QUADRO 1.21 • Massa monetária M_1^- (em milhões de contos) e taxa de juro (%) dos Bilhetes do Tesouro.

Mês	M_1^-	Taxa juro	Mês	M_1^-	Taxa juro	Mês	M_1^-	Taxa juro
Abr 95	3422.0	10.9	Jun 96	3843.8	7.0	Ago 97	4561.3	5.2
Mai 95	3421.4	10.7	Jul 96	4070.5	7.1	Set 97	4591.0	5.0
Jun 95	3546.9	10.2	Ago 96	4013.7	7.1	Out 97	4558.1	5.1
Jul 95	3899.6	9.8	Set 96	3917.9	7.0	Nov 97	4880.5	4.9
Ago 95	3789.3	9.7	Out 96	3862.4	6.8	Dez 97	4901.9	4.6
Set 95	3731.0	9.0	Nov 96	4042.6	6.7	Jan 98	4735.1	4.5
Out 95	3672.8	9.1	Dez 96	4301.9	6.3	Fev 98	4698.1	4.4
Nov 95	3771.9	8.8	Jan 97	4184.0	6.0	Mar 98	4840.0	4.1
Dez 95	3885.0	8.7	Fev 97	4088.3	5.9	Abr 98	4841.1	4.3
Jan 96	3760.3	8.3	Mar 97	4153.0	6.0	Mai 98	4901.2	4.2
Fev 96	3702.8	7.9	Abr 97	4148.1	5.8	Jun 98	5131.6	4.0
Mar 96	3624.1	7.7	Mai 97	4238.6	5.6	Jul 98	5494.2	3.9
Abr 96	3698.4	7.4	Jun 97	4482.6	5.5	Ago 98	5284.3	3.8
Mai 96	3725.3	7.0	Jul 97	4655.6	5.4	Set 98	5322.3	3.8

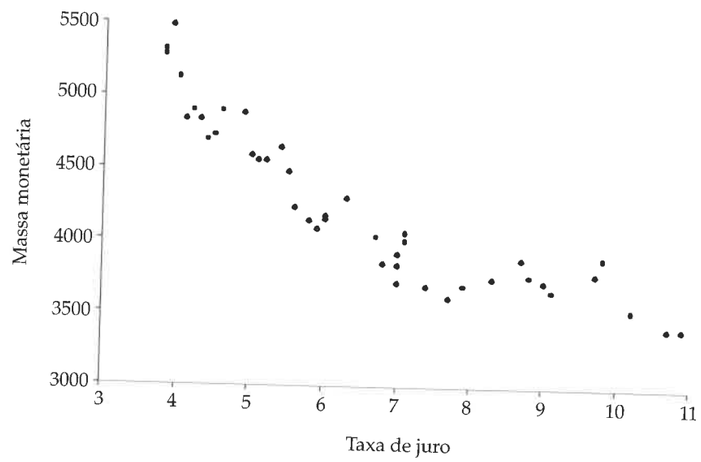
Fonte: Banco de Portugal [10^6 contos ≡ 4 987 979.97 euros].

EXEMPLO 1.31 • No quadro 1.21 apresentam-se os valores da massa monetária M_1^- (circulação monetária mais depósitos à ordem) em milhões de contos e as taxas de juro (em percentagem) dos Bilhetes do Tesouro de 181 dias a 1 ano cedidos sem recurso pelos bancos. Os dados foram extraídos do *Boletim Estatístico* do Banco de Portugal e compatibilizados, no caso do agregado M_1^-, para o período de Abril de 1995 a Setembro de 1998.

© McGraw-Hill

O diagrama de dispersão da figura 1.13 permite antever uma associação entre as variáveis, aparentemente não linear. Neste caso, a associação é de sentido inverso, isto é, quanto maior é a taxa de juro, menor é a massa monetária. Este tipo de relação entre as duas variáveis é consistente com a teoria económica. ⊙

FIGURA 1.13
Diagrama de dispersão da taxa de juro e da massa monetária.



O termo relação, por ser empregado em diversas acepções, carece de ser precisado. Quando se diz que duas variáveis estão relacionadas pode querer significar-se que a conexão é bem definida e invariável, ou pode ter-se em mente uma ligação mais indefinida e mais vaga. Por exemplo, como é bem sabido, o perímetro de uma circunferência, C, e o raio do círculo correspondente, R, estão relacionados. A relação que liga essas duas variáveis é definida e inalterável, e pode exprimir-se por meio de uma expressão matemática: a relação funcional ou função é $C = 2\pi R$. Neste caso, diz-se que se tem uma **relação determinística** entre as duas variáveis.

Quando se refere a relação que existe entre o PIB e as importações, ou entre a taxa de juro e a massa monetária, ou entre as idades dos cônjuges na data do casamento, ou entre o peso e a altura de um homem adulto, tem-se em mente um tipo de relações mais imprecisas. Nestes casos, não se tem funções, no sentido matemático, mas meras **relações estatísticas**. No caso da circunferência há uma correspondência biunívoca entre o raio e o perímetro, hipótese que não se verifica para os outros quatro exemplos. Assim: na relação entre o PIB e as importações pode muito bem suceder — e sucede — que a valores semelhantes do PIB correspondam montantes diferentes de importações mas, no entanto, parece acertado dizer que quanto maior for o PIB, maiores serão, em média, as importações; no caso da relação entre a taxa de juro e a massa monetária, pode apenas afirmar-se que os aumentos de taxa de juro implicam, em média, diminuições da massa monetária; quando se trata da relação entre as idades dos cônjuges, também é evidente que a idade do marido não determina a idade da mulher (e inversamente), mas, em média, quanto mais idoso é o marido, mais idosa é a mulher; enfim, a altura não determina o peso, mas, em média, os indivíduos mais altos são mais pesados. Fundamentalmente, é da **variação em média** que se ocupam as relações estatísticas. Entre duas variáveis ligadas por uma relação estatística diz-se que existe **correlação**. Indica-se, assim, que os fenómenos não estão indissoluvelmente ligados, mas, sim, que a intensidade de um é acompanhada **tendencialmente** pela intensidade do outro, no mesmo sentido ou em sentido inverso. Em média, a altura varia directamente com o peso; entre altura e peso há **correlação positiva**. Em média, o valor da massa monetária varia inversamente com a taxa de juro; entre massa monetária e taxa de juro há **correlação negativa**.

© McGraw-Hill

Suponha-se que se está na presença de uma relação estatística entre duas variáveis, e que se procura medir a intensidade da respectiva correlação. Tal como se definiu o conceito de variância para avaliar a dispersão de uma colecção de dados relativos a uma variável, é agora necessário dispor de um outro conceito que permita avaliar a variação conjunta de duas variáveis. Este conceito é o de **covariância** da colecção de observações de um par de variáveis, e é definido da seguinte maneira:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}). \tag{1.34}$$

Como $x_i - \overline{x}$ e $y_i - \overline{y}$ são, respectivamente, os desvios das observações das variáveis x e y em relação às suas médias, facilmente se verifica que a covariância é a média dos produtos daqueles desvios.

Pode fazer-se uma interpretação simples da expressão (1.34) a partir do respectivo diagrama de dispersão, fixando o ponto $(\overline{x}, \overline{y})$ como origem do sistema de eixos. Deste modo, o plano é dividido em 4 quadrantes, a saber:

1.º quadrante — pontos (x_i, y_i) tais que $x_i \geq \overline{x}$ e $y_i \geq \overline{y}$;
2.º quadrante — pontos (x_i, y_i) tais que $x_i \leq \overline{x}$ e $y_i \geq \overline{y}$;
3.º quadrante — pontos (x_i, y_i) tais que $x_i \leq \overline{x}$ e $y_i \leq \overline{y}$;
4.º quadrante — pontos (x_i, y_i) tais que $x_i \geq \overline{x}$ e $y_i \leq \overline{y}$.

Assim, uma covariância positiva corresponde a uma colecção onde predominam observações situadas nos 1.º e 3.º quadrantes; quando os pontos se situam maioritariamente nos 2.º e 4.º quadrantes, a covariância é negativa.

O cálculo da covariância pode ser feito de forma mais operacional através de:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - \overline{y}x_i - \overline{x}y_i + \overline{xy}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}. \tag{1.35}$$

Na avaliação do grau de associação entre duas variáveis, a covariância apresenta desde logo uma limitação importante: o valor que vai assumir depende não só da variação conjunta das variáveis como também dos valores absolutos que estas variáveis assumem. De facto, é simples exercício verificar que se se calcular a covariância entre duas variáveis medidas em centímetros se obtém um valor 10 000 vezes (100^2) superior a esta mesma covariância se as variáveis forem medidas em metros. Para obviar a esta situação e para tomar em consideração a variação de cada uma das variáveis, define-se o **coeficiente de correlação de Pearson** (ou, simplesmente, coeficiente de correlação) através de:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{1.36}$$

onde s_x e s_y representam o desvio padrão, respectivamente, das variáveis x e y.

© McGraw-Hill

Sem dificuldade se demonstra que $-1 \leq r \leq 1$, quaisquer que sejam os valores assumidos pelas variáveis x e y. Com efeito, notando que, qualquer que seja $\lambda \in \Re$,

$$\sum_{i=1}^{n} \{(y_i - \overline{y}) - \lambda(x_i - \overline{x})\}^2 \geq 0,$$

obtém-se o polinómio do segundo grau em λ, não negativo,

$$\lambda^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 - 2\lambda \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) + \sum_{i=1}^{n} (y_i - \overline{y})^2 \geq 0 .$$

Então, o seu discriminante é não positivo, ou seja,

$$\left(2\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})\right)^2 - 4\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2 \leq 0.$$

Logo,

$$\frac{\left(\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2} = r^2 \leq 1,$$

donde se conclui que $-1 \leq r \leq 1$.

Importa ainda esclarecer que o coeficiente de correlação mede o grau de **associação linear** entre as variáveis. Os casos representados na figura 1.14 ajudam a interpretar o significado do coeficiente de correlação. No primeiro caso, em que os dados se dispõem essencialmente nos 1.º e 3.º quadrantes numa mancha aproximadamente linear, obtém-se naturalmente um valor positivo elevado para o coeficiente de correlação. O segundo caso é semelhante ao primeiro, só que, estando a mancha predominantemente nos 2.º e 4.º quadrantes, o coeficiente de correlação assume um valor próximo de –1. Enquanto o sinal do coeficiente indica se a correlação é positiva ou negativa, o seu valor absoluto mede a intensidade da associação linear. Os casos 3 e 4 referem-se a situações em que o coeficiente de correlação tem um valor absoluto próximo de 0. Na primeira situação trata-se, de facto, de um caso em que as variáveis não aparentam qualquer padrão de associação, enquanto na segunda existe um padrão não linear de associação. Trata-se de sublinhar a ideia de que o coeficiente de correlação mede o grau de **associação linear** entre as variáveis.

EXEMPLO 1.32 • Retomando o exemplo 1.30, referente ao PIB e às importações de Portugal, calcula-se facilmente o coeficiente de correlação entre estas variáveis, obtendo-se $r = 0.986$. Este valor indicia, como seria de esperar à luz da teoria económica, um elevado grau de associação linear positiva entre estas variáveis. ⊙

EXEMPLO 1.33 • Considerando os dados referentes à taxa de juro e à massa monetária (exemplo 1.31), obtém-se $r = -0.902$. Este valor revela um forte grau de associação linear negativa, embora o tipo de associação não seja linear (ver figura 1.13). ⊙

© McGraw-Hill

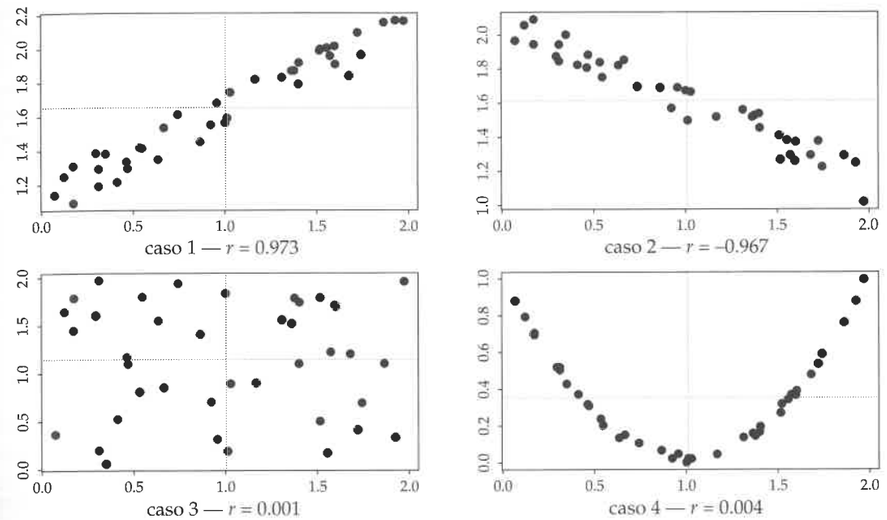


FIGURA 1.14 • Diagramas de dispersão e coeficiente de correlação.

EXEMPLO 1.34 • Uma interpretação da lei de Wagner é feita por Santos (1986) [«A Lei de Wagner e a realidade das despesas públicas», *Estudos de Economia*, vol. 4, n.º 2] nos seguintes termos: «nos países onde ocorre o processo de industrialização, à medida que aumenta a capitação do rendimento, cresce a importância do sector público relativamente ao conjunto da actividade económica». Para os 17 países europeus considerados por aquele autor apresentam-se na segunda e na terceira colunas do quadro 1.22, respectivamente, o PNB *per capita* e a proporção do PNB aplicado em despesas públicas. O diagrama de dispersão respectivo consta da figura 1.15.

Para o coeficiente de correlação obteve-se $r = 0.546$. A correlação é positiva, de acordo com a lei de Wagner, mas não muito forte, ideia que o diagrama de dispersão já sugeria. ⊙

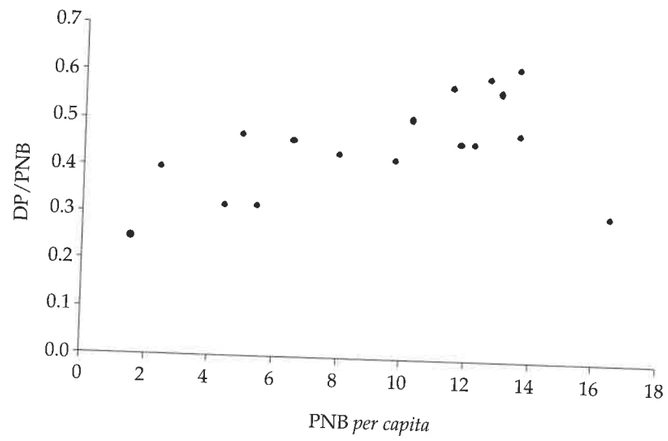
QUADRO 1.22
PNB *per capita* e proporção do PNB aplicado em despesas públicas em 17 países europeus.

Países	PNB	DP/PNB	Países	PNB	DP/PNB
Suécia	13.52	0.62	Itália	6.48	0.46
Noruega	12.65	0.60	R. Unido	7.92	0.43
Holanda	11.47	0.58	Finlândia	9.72	0.42
Dinamarca	12.95	0.57	Portugal	2.37	0.40
Áustria	10.23	0.51	Espanha	5.40	0.32
RFA	13.59	0.48	Grécia	4.38	0.32
Irlanda	4.88	0.47	Suíça	16.44	0.31
França	11.73	0.46	Turquia	1.47	0.25
Bélgica	12.18	0.46			

Fonte: Santos (1986), «A lei de Wagner e a realidade das despesas públicas», *Estudos de Economia*, vol. 4, n.º 2.

© McGraw-Hill

FIGURA 1.15
Diagrama de dispersão do PNB *per capita* e da proporção do PNB aplicado em despesas públicas em 17 países europeus.



O coeficiente de correlação de Pearson mede, como já se disse, o grau de associação linear entre duas variáveis quantitativas. Que fazer então quando a associação entre as variáveis não é linear (mesmo que aproximadamente) ou quando pelo menos uma das variáveis não é de tipo quantitativo? Nesta secção dá-se apenas uma resposta parcial a estas questões através do **coeficiente de correlação de Spearman**, r_S. Esta medida consiste em calcular o coeficiente de correlação (de Pearson), não entre as observações originais mas, sim, entre as respectivas ordens. Assim, cada par (x_i, y_i) é substituído pelo par $[\,\text{ord}(x_i), \text{ord}(y_i)\,]$, onde, por exemplo, $\text{ord}(x_i)$ representa a ordem de x_i na colecção. Uma vez feita a substituição, procede-se ao cálculo do coeficiente de correlação nos moldes habituais. Repare-se que quando se trata de variáveis ordinais não existirão problemas, uma vez que a característica fundamental de tais variáveis reside no facto de, embora sendo qualitativas, poderem ser ordenadas. Os três exemplos que se seguem ilustram o cálculo do coeficiente de correlação de Spearman.

EXEMPLO 1.35 • Considerem-se os dados do exemplo 1.34. A substituição dos dados originais pelas respectivas ordens dá lugar aos valores que constam do quadro 1.23. Deve esclarecer-se o seguinte: como a França, a Bélgica e a Itália apresentam a mesma proporção DP/PNB (igual a 0.46), foi considerada a ordem 9 (média das ordens 8, 9 e 10); como a Espanha e a Grécia apresentam uma proporção DP/PNB igual a 0.32 foi-lhes atribuída a ordem 3.5 (média das ordens 3 e 4).

O valor calculado para o coeficiente de correlação de Spearman é $r_s = 0.531$. Este valor difere pouco de $r = 0.546$ (calculado com os dados originais) uma vez que a Suíça continua a influenciar decisivamente o resultado. O valor de r_s sem a Suíça é de 0.795. ⊙

© McGraw-Hill

QUADRO 1.23
PNB *per capita* e proporção do PNB aplicado em despesas públicas em 17 países europeus (ordens).

Países	Ordem PNB	Ordem DP/PNB	Países	Ordem PNB	Ordem DP/PNB
Suécia	15	17.0	Itália	6	9.0
Noruega	13	16.0	R. Unido	7	7.0
Holanda	10	15.0	Finlândia	8	6.0
Dinamarca	14	14.0	Portugal	2	5.0
Áustria	9	13.0	Espanha	5	3.5
RFA	16	12.0	Grécia	3	3.5
Irlanda	4	11.0	Suíça	17	2.0
França	11	9.0	Turquia	1	1.0
Bélgica	12	9.0			

O exemplo 1.35 permite ver que, quando se está em presença de variáveis com uma associação linear, os valores dos coeficientes de correlação de Pearson e de Spearman não costumam diferir muito. O *outlier* constituído pela Suíça influencia ambos os coeficientes de correlação, uma vez que tem influência na ordem das observações. Em situações em que o *outlier* não influencia a ordenação das observações, o coeficiente de Spearman não é afectado por este.

EXEMPLO 1.36 • O quadro 1.24 apresenta a taxa de desemprego (em percentagem) para os quinze países com um valor mais elevado do Índice de Desenvolvimento Humano (IDH), publicado no *Relatório do Desenvolvimento Humano 1999* do Programa das Nações Unidas para o Desenvolvimento (PNUD).

QUADRO 1.24
Índice de Desenvolvimento Humano (IDH) e Taxa de Desemprego (TD) nos 15 países com maior IDH.

Ordem IDH	Países	TD	OrdemTD
1	Canadá	9.3	4
2	Noruega	4.1	13
3	EUA	5.0	11
4	Japão	3.5	15
5	Bélgica	9.0	5
6	Suécia	8.0	7
7	Austrália	8.5	6
8	Holanda	5.6	9
9	Islândia	3.8	14
10	Reino Unido	7.1	8
11	França	12.4	2
12	Suíça	4.2	12
13	Finlândia	14.5	1
14	Alemanha	9.8	3
15	Dinamarca	5.4	10

Fonte: *Relatório do Desenvolvimento Humano 1999* (PNUD), Trinova Editora, Lisboa, 1999.

© McGraw-Hill

Com o intuito de aferir se existe associação entre desenvolvimento humano e taxa de desemprego (espera-se que negativa), pode utilizar-se o coeficiente de correlação de Spearman. Note-se que não se pode recorrer ao coeficiente de correlação de Pearson, uma vez que se conhece apenas a ordem dos países em relação ao IDH, e não os valores por este assumidos.

A quarta coluna do quadro indica a ordem dos valores da taxa de desemprego (ordem decrescente). Para calcular o coeficiente de correlação de Spearman aplica-se (1.36) aos dados da primeira e quarta coluna, obtendo-se $r_s = -0.282$. ⊙

EXEMPLO 1.37 • Retome-se os dados do exemplo 1.31, referentes à massa monetária M_1^- e às taxas de juro dos Bilhetes do Tesouro. Como a associação entre as variáveis começa a afastar-se de um padrão linear — veja-se a figura 1.13, que sugere uma exponencial negativa ou uma função inversa —, é natural esperar-se que o valor absoluto do coeficiente de correlação de Spearman seja superior ao de Pearson. Considerando as ordens dos valores do quadro 1.21, obtém-se $r_s = -0.945$, valor este que confirma a expectativa referida (recorde-se que $r = -0.902$). ⊙

Antes de concluir a presente secção torna-se importante sublinhar que **não se deve confundir correlação com causalidade**. De facto, um valor elevado, em termos absolutos, para o coeficiente de correlação indica apenas que, na colecção, existe uma associação aproximadamente linear entre as duas variáveis, não sendo lícito inferir qualquer relação de causa e efeito. Por exemplo, quando se estabelece que o coeficiente de correlação entre o PIB e as importações é de 0.986, não se pode concluir, sem mais, que as importações dependem do PIB, ou inversamente. Pode até dar-se o caso de duas variáveis estarem altamente correlacionadas e, no entanto, não se conseguir descortinar qualquer relação interessante entre elas. Diz-se então que se tem uma **correlação espúria**.

EXEMPLO 1.38 • Sabendo que Mário Soares tinha 63 anos em Dezembro 1987 e que foi Presidente da República de 1987 a 1996, calcula-se sem dificuldade o coeficiente de correlação entre a idade do Presidente e o PIB (ver quadro 1.20) naquele período, obtendo-se $r = 0.95$. Note-se que este valor seria exactamente o mesmo, qualquer que fosse o cidadão considerado. A explicação para esta situação reside no facto de o PIB e a idade dependerem claramente de uma terceira variável, o tempo. ⊙

Os manuais de estatística estão cheios de exemplos bem ilustrativos de que a dependência indirecta conduz à errada interpretação do coeficiente de correlação e a conclusões ridículas: (a) a forte correlação positiva entre o número de prédios destruídos em cada fogo urbano e o número de viaturas de bombeiros utilizadas no combate ao mesmo, pode fazer supor que o problema se resolve deixando de recorrer a estes ...; (b) a forte correlação positiva entre o número anual de casos de insolação e a colheita de trigo pode levar a concluir que é a produção de trigo que causa insolações, ou que são as insolações que causam boas colheitas de trigo; o que na verdade se verifica é que os verões quentes causam insolações e boas colheitas de trigo.

© McGraw-Hill

Palavras-chave

Amostra
Amplitude de classe
Amplitude de variação (empírica)
Amplitude interquartis (empírica)
Associação linear
Atributo
Atributo qualitativo
Atributo quantitativo
Barreiras (internas, externas)
Caixa-de-bigodes
Causalidade
Censo
Classe modal (empírica)
Coeficiente de correlação (empírico)
Coeficiente de correlação de Spearman
Coeficiente de variação (empírico)
Colecção de dados
Correlação
Correlação espúria
Covariância (empírica)
Cronograma
Dados (numéricos)
Dados classificados
Dados de painel
Dados seccionais
Dados temporais
Decil (empírico)
Desvio absoluto médio (empírico)
Desvio padrão (empírico)
Desvio padrão corrigido (empírico)
Diagrama de barras
Diagrama de dispersão
Distribuição de frequências
Erro de tabulagem
Escala nominal
Escala ordinal
Estatística
Estatística de ordem (empírica)
Estatística descritiva
Estatística indutiva
Extremo (empírico)
Frequência absoluta
Frequência relativa
Gráfico caule-e-folhas
Histograma
Incerteza
Intervalos de classe
Limites de classe
Máximo (empírico)
Média (empírica)
Mediana (empírica)
Medida de dispersão
Medida de localização
Mínimo (empírico)
Moda (empírica)
Modelo
Outlier
Percentil (empírico)
Ponto médio de classe
População ou universo
Probabilidade
Processo de amostragem
Quadro de frequências
Quantil (empírico)
Quartil (empírico)
Relação determinística
Relação estatística
Resumo dos 5 números
Série temporal
Unidade estatística
Variabilidade
Variância (empírica)
Variância corrigida (empírica)
Variável
Variável contínua
Variável discreta

© McGraw-Hill

Exercícios

1. Apresente três exemplos de populações, e, para cada uma delas, indique dois atributos plausíveis para as respectivas unidades estatísticas.
2. Enumere uma possível amostra de dimensão 3 das populações seguintes:
   a) Jornais diários publicados em Portugal.
   b) Alunos da sua escola.
   c) Estações de rádio portuguesas.
   d) Clubes de futebol que participam na 1.ª Liga Profissional portuguesa.
3. Dê dois exemplos de variáveis discretas e dois exemplos de variáveis contínuas.
4. Dê dois exemplos de atributos quantitativos e quatro exemplos de atributos qualitativos (dois nominais e dois ordinais). Para cada um dos exemplos apresentados indique as respectivas unidades estatísticas.
5. Caracterize a natureza de cada uma das variáveis seguintes (quantitativa discreta, quantitativa contínua, qualitativa ordinal, qualitativa nominal).
   a) Número de *e-mails* enviados diariamente por um aluno no último mês.
   b) Custo dos livros de estudo comprados por um estudante da sua universidade no presente ano lectivo.
   c) Valor mensal da sua conta de telemóvel no último ano.
   d) Categorias existentes na carreira docente do ensino superior público.
   e) Distrito de naturalidade de um estudante da sua universidade.
6. A biblioteca de determinado estabelecimento de ensino superior realizou um inquérito por amostragem aos seus utilizadores, para identificar aspectos do seu funcionamento que deveriam ser melhorados. Caracterize a natureza de cada variável referida nas alíneas seguintes.
   a) Pergunta: «A biblioteca possui os livros necessários para o bom acompanhamento das disciplinas do curso?» Existem 4 respostas possíveis: 1) concorda plenamente; 2) concorda; 3) discorda; 4) discorda totalmente.
   b) Pergunta: «Pensa que o horário de abertura da biblioteca deveria ser alargado?» As respostas possíveis são as seguintes: sim; não; não sabe.
   c) Pergunta: «Quantas vezes utilizou a biblioteca na semana passada?» A resposta é um número a ser escrito por quem responde ao questionário.
7. Os dados que se seguem correspondem ao consumo (litros/100 km) em circuito urbano (U), extra-urbano (EU) e combinado (C) de 54 modelos de veículos ligeiros de passageiros movidos a gasolina e com cilindrada entre 1300 e 1400 cm^3 dos existentes no mercado dos veículos novos em Portugal. Foram retirados, em Novembro de 2006, da página da internet da revista *Guia do Automóvel*.

Marca/Modelo	U	EU	C	Marca/Modelo	U	EU	C
Chevrolet Kalos 1.4	8.6	6.1	7	Opel Meriva	8.3	5.3	6.4
Chevrolet Aveo 1.4 LT	8.9	5.4	6.7	Opel Astra Cosmos	8.2	5.2	6.3
Chevrolet Aveo 1.4 LT Aut	9.6	5.6	7	Opel Astra EasyTronic	7.7	4.9	5.9
Chevrolet Lacetti 1.4	9.3	5.9	7.1	Peugeot 1007 1.4i	8.4	5.3	6.4
Citroen C2 1.4i VTR	7.9	4.9	6	Peugeot 206 SW	8.5	5.2	6.4

© McGraw-Hill

Marca/Modelo	U	EU	C	Marca/Modelo	U	EU	C
Citroen C3 1.4i SX	8.2	4.9	6.1	Peugeot 207 Premium	8.5	5.2	6.4
Citroen Pluriel C3 1.4i	8.9	5.5	6.8	Renault Clio III 1.4	8.7	5.4	6.6
Fiat Stilo 1.4 16v	8.5	5.7	6.7	Renault Megane II Fase II 1.4	9.2	5.6	6.9
Fiat Stilo 1.4 16v Sport	8.2	5.5	6.5	Renault Scenic II	9.7	5.9	7.2
Ford Fiesta 14i	8.3	5.2	6.3	Rover 25 1.4	9.1	5.5	6.8
Ford Fusion 1.4i	8.4	5.5	6.5	Rover Streetwise 1.4	6.4	6.1	7.3
Ford Focus 1.4i	8.7	5.4	6.6	Rover 45 1.4	9.3	5.7	7
Honda Jazz 1.4 ES	7.3	5.1	5.8	Seat Ibiza 1.4 100cv	8.9	5.3	6.6
Honda Jazz 1.4 ES CVT	7.2	5.1	5.9	Seat Ibiza 1.4 75cv CxAut	10.2	5.9	7.5
Honda Civic 1.4	7.6	4.9	5.9	Seat Ibiza 1.4 85cv	8.5	5.2	6.4
Hyundai Getz 1.4	7.5	5.1	6	Seat Cordoba 1.4	8.5	5.2	6.4
Hyundai Accent	8	5.1	6.2	Seat Leon 1.4	9.6	5.7	7.1
Kia Rio 1.4 Ex	7.9	5.1	6.2	Seat Altea 1.4 Rebel	9.9	5.9	7.4
Lancia Musa 1.4 16v Oro	8.5	5.5	6.6	Seat Altea 1.4 Rebel XL	9.7	5.8	7.3
Mazda 2 MZ-CD 1.4	5.9	4.1	4.7	Seat Toledo 1.4	9.9	5.9	7.4
Hyundai ZR 105	9.1	5.5	6.8	Skoda Fabia 1.4 Ambiente	8.8	5.2	6.4
Hyundai ZS 105	9.3	5.7	7	Skoda Roomster 1.4	8.9	5.6	6.8
Nissan Micra 1.4	7.9	5.4	6.3	Skoda Octavia 1.4 Ambiente	9.6	5.6	7
Nissan Note 1.4	8.1	5.5	6.4	Toyota Corolla 1.4 VVT i	8.4	5.7	6.7
Opel Corsa 1.4	8.1	5.1	6.2	Volkswagen Polo 1.4 Aut	10.4	6	7.6
Opel Corsa GTC 1.4	7.8	4.8	5.9	Volkswagen Golf 1.4	9.4	5.5	6.9
Opel Tigra 1.4	8.1	5	6.1	Volkswagen Golf Plus 1.4	9.8	5.8	7.2

   a) Utilizando como unidade para cada caule 0.5 litros/100 km, construa um gráfico caule-e-folhas para o consumo em circuito urbano.
   b) Construa uma caixa-de-bigodes para o consumo dos veículos em circuito urbano.
   c) Determinado técnico defende que, para se obter o consumo médio em circuito urbano dos veículos ligeiros de passageiros (novos) em circulação com cilindrada entre 1300 e 1400 cm^3, basta calcular a média simples das observações. Critique a ideia do técnico.
   d) Compare os 3 tipos de consumo (urbano, extra-urbano e combinado) recorrendo a um resumo dos 5 números e a caixas-de-bigodes.
8. Os dados apresentados referem-se à taxa anual de mobilidade em 40 bairros residenciais de determinada cidade americana (n.º de agregados que deixaram o bairro num determinado ano a dividir pelo n.º total de agregados que residem no bairro).

0.10	0.15	0.16	0.18	0.19	0.21	0.23	0.26
0.12	0.15	0.16	0.18	0.20	0.22	0.24	0.26
0.12	0.15	0.17	0.19	0.20	0.22	0.25	0.28
0.13	0.16	0.17	0.19	0.21	0.22	0.25	0.30
0.14	0.16	0.18	0.19	0.21	0.22	0.26	0.36

© McGraw-Hill

a) Construa um gráfico caule-e-folhas.
b) Construa um resumo dos 5 números para esta colecção.
c) Identifique eventuais *outliers* e construa uma caixa-de-bigodes.
d) Calcule a média e a mediana da colecção.
e) Suponha que o valor 0.36 resulta de um erro no registo da informação, sendo o valor correcto de 0.26. Será que este erro tem impacto nas medidas que calculou na alínea anterior.

9. O quadro que se segue apresenta os valores unitários (euros/m^2) de avaliação bancária da habitação no 3.º trimestre de 2006 nas diferentes sub-regiões de Portugal identificadas de acordo com a classificação NUTS III.

Valores unitários (euros/m^2) de avaliação bancária de habitação por natureza dos alojamentos				
Região	**Sub-Região**	**Habitação**	**Apartamentos**	**Moradias**
Continente		1217.32	1275.68	1133.19
Norte	Alto Trás-os-Montes	893.71	955.19	857.35
Norte	Ave	959.53	915.32	989.71
Norte	Cávado	1027.55	1033.24	1023.68
Norte	Douro	945.98	954.26	942.67
Norte	Entre Douro e Vouga	932.40	918.51	947.81
Norte	Grande Porto	1183.99	1168.61	1235.97
Norte	Minho-Lima	1043.68	1027.89	1050.73
Norte	Tâmega	929.00	871.54	960.20
Centro	Baixo Mondego	1281.37	1443.84	1152.37
Centro	Baixo Vouga	1057.10	1105.96	1014.31
Centro	Beira Interior Norte	957.96	1081.78	901.47
Centro	Beira Interior Sul	1007.12	956.11	1062.84
Centro	Cova da Beira	892.17	868.17	910.40
Centro	Dão-Lafões	1043.03	1082.71	1026.23
Centro	Pinhal Interior Norte	992.17	899.53	1034.48
Centro	Pinhal Interior Sul	918.22	922.55	916.21
Centro	Pinhal Litoral	1143.48	1097.35	1186.88
Centro	Serra da Estrela	785.69	791.71	781.80
Lisboa e Vale do Tejo	Grande Lisboa	1596.47	1594.27	1619.29
Lisboa e Vale do Tejo	Lezíria do Tejo	1149.47	1080.15	1200.48
Lisboa e Vale do Tejo	Médio Tejo	1095.29	1086.73	1102.74
Lisboa e Vale do Tejo	Oeste	1280.37	1292.90	1269.74
Lisboa e Vale do Tejo	Península de Setúbal	1347.99	1290.27	1565.59
Alentejo	Alentejo Central	1321.90	1328.88	1317.94
Alentejo	Alentejo Litoral	1296.28	1238.30	1355.55
Alentejo	Alto Alentejo	1105.62	1008.14	1172.14
Alentejo	Baixo Alentejo	1116.04	1267.66	1072.26
Algarve	Algarve	1468.25	1453.76	1520.30

© McGraw-Hill

a) Para o valor agregado referente à habitação construa um gráfico caule-e-folhas, tomando como unidade para os caules a centena de euros/m^2.
b) Para o valor agregado referente à habitação construa uma caixa-de-bigodes e identifique eventuais *outliers*.
c) Faz sentido calcular os valores para o continente por média dos valores nas diferentes sub-regiões? Justifique.
d) Calcule os coeficientes de correlação de Pearson e de Spearman entre os valores de avaliação dos apartamentos e das moradias. Comente e compare os resultados.

10. Os dados que se seguem referem-se à precipitação caída em Lisboa (em mm) no mês de Janeiro dos anos de 1980 a 2006.

1980	53.5	1986	52.2	1992	61.5	1998	62.3	2004	129.5
1981	11.6	1987	130.0	1993	19.0	1999	107.5	2005	0.0
1982	112.8	1988	107.7	1994	79.7	2000	17.7	2006	57.7
1983	6.2	1989	62.5	1995	57.3	2001	195.5		
1984	40.6	1990	63.1	1996	394.0	2002	72.9		
1985	224.6	1991	64.9	1997	136.6	2003	131.0		

a) Calcule a média, a variância e o coeficiente de variação para estas observações.
b) Calcule a variância corrigida e o desvio padrão corrigido.
c) Apresente o resumo dos 5 números.
d) Calcule os primeiro e terceiro decis.
e) Construa uma caixa-de-bigodes para esta informação e identifique eventuais *outliers*.
f) Esboce um cronograma com esta informação.
g) Comente as várias representações gráficas a que procedeu, realçando vantagens e desvantagens de cada uma delas.

11. As observações que se seguem consistem no rendimento, em percentagem, obtido pelos fundos Caixagest nos últimos 12 meses (a data de referência é 21 de Novembro de 2006). Estes 24 fundos são agrupados em 6 categorias (Liquidez, Aforro, Diversificação, Capitalização, Valor Acrescentado e Fundos Postal) em função dos objectivos pretendidos por cada um deles.

Liquidez		**Valor Acrescentado**	
Curto prazo	1.18	Acções Europa	21.39
Tesouraria	1.48	Acções EUA	–1.43
Moeda	1.98	Acções Portugal	26.47
Aforro		Acções Oriente	16.71
Renda mensal	1.94	Acções Japão	–11.68
Rendimento	1.68	Acções emergentes	13.49
Diversificação		PPA	28.90
Estratégia conservadora	3.45	Energias renováveis	1.52
Estratégia moderada	3.24	**Fundos Postal**	
Estratégia moderada II	3.24	Postal tesouraria	1.33

© McGraw-Hill

(Continua)

(Continuação)

Estratégia agressiva	8.08	Postal rendimento	1.14
Estratégias alternativas	3.08	Postal gestão global	2.41
Capitalização		Postal capitalização	1.31
Obrigações Euro	0.80	Postal acções	26.24

a) Construa uma caixa-de-bigodes para o conjunto dos fundos e identifique eventuais *outliers*.
b) Compare, por meio de caixas-de-bigodes, os resultados obtidos pelos fundos de Valor Acrescentado e Postal. Critique a metodologia utilizada.

12. Nalgumas situações particulares, a média pode constituir uma medida de localização «curiosa». Construa uma situação em que 19 das 20 observações de que dispõe se situam acima da média.

13. Considere uma colecção de observações $\{x_1, x_2, \ldots, x_n\}$.
a) Suponha que se adiciona a mesma constante, c, a cada uma das observações. Obtenha a média, a mediana, a variância e os quartis da colecção referente aos dados transformados em função das mesmas características da colecção original.
b) Admita que os dados originais eram multiplicados pela constante $c \neq 0$. Obtenha as mesmas características numéricas referidas na alínea anterior.

14. Sejam $\overline{x}_n$ e s'^2_n a média e a variância corrigida, respectivamente, de uma colecção com n observações $\{x_1, x_2, \ldots, x_n\}$. Designe-se por $\overline{x}_{n+1}$ e s'^2_{n+1} a média e a variância corrigida, respectivamente, da colecção que se obtém quando se acrescenta a observação x_{n+1} à colecção inicial.
a) Obtenha $\overline{x}_{n+1}$ como combinação linear de $\overline{x}_n$ e de x_{n+1}.
b) Mostre que,

$$n s'^2_{n+1} = (n-1)s'^2_n + \frac{n}{n+1}(x_{n+1} - \overline{x}_n)^2.$$

Verifica-se assim que s'^2_{n+1} pode ser obtido a partir de s'^2_n, x_{n+1} e $\overline{x}_n$.

15. O número de sinistros participados por cada apólice do ramo automóvel é uma informação de grande importância para que as seguradoras possam construir a sua tarifa. Lemaire (1995) [*Bonus-Malus Systems in Automobile Insurance*, Kluwer Academic Publishers] apresenta a seguinte informação no que se refere à sinistralidade em Responsabilidade Civil de uma seguradora belga.

N.º de sinistros	N.º de apólices
0	96 978
1	9 240
2	704
3	43
4	9
>4	0
Total	106 974

© McGraw-Hill

a) Calcule a média e a variância do número de sinistros por apólice.
b) Apresente a informação sob a forma de um diagrama de barras.

16. O quadro que se segue apresenta o número de golos por jogo da 1.ª Liga Profissional de Futebol portuguesa na época 1999/2000.

N.º de golos	0	1	2	3	4	5	6	7	8
N.º de jogos	27	67	80	69	29	19	12	2	1

a) Represente graficamente esta informação da forma que entender mais adequada.
b) Qual o número médio de golos por jogo naquela época? E o desvio padrão corrigido?
c) Será que se pode considerar os jogos com mais de 6 golos como sendo *outliers*?

17. Recolhida informação semelhante à do exercício anterior mas referente à época de 2004/2005, observou-se que:

N.º de golos	0	1	2	3	4	5	6	7
N.º de jogos	34	67	77	59	35	23	9	2

a) Qual o número médio de golos por jogo naquela época? E o desvio padrão corrigido?
b) Apresente um resumo dos 5 números e identifique eventuais *outliers*.
c) Compare os resultados de 2004/2005 com os de 1999/2000 recorrendo a uma análise numérica e gráfica.

18. Construa um exemplo para o qual a medida de localização adequada seja:
a) a média;
b) a mediana;
c) a moda.

19. Um estudo sobre os atrasos nos voos europeus durante o Verão de 2006, realizado em determinado aeroporto, conduziu aos seguintes resultados:

Atraso	(0, 10]	(10, 20]	(20, 30]	(30, 40]	(40, 50]	(50, 60]
N.º de voos	29	23	17	14	11	6

a) Construa um histograma.
b) Calcule a média, a variância e o desvio padrão referentes aos minutos de atraso dos voos.
c) Obtenha a mediana.
d) Determine o intervalo interquartis e a respectiva amplitude.
e) Determine a classe modal.

© McGraw-Hill

20. Considere a precipitação (em mm) caída no Porto nos anos de 1980 a 2006 e tenha presente a informação dada no exercício 10.

1980	118.2	1986	171.3	1992	99.3	1998	149.1	2004	126.8
1981	3.6	1987	107.7	1993	54.1	1999	101.7	2005	9.6
1982	94.4	1988	217.1	1994	232.1	2000	26.8	2006	62.3
1983	16.4	1989	35.5	1995	165.8	2001	403.4		
1984	279.0	1990	164.5	1996	331.4	2002	157.5		
1985	203.3	1991	164.2	1997	162.5	2003	332.2		

a) Compare a pluviosidade no Porto e em Lisboa no mês de Janeiro recorrendo às medidas de localização e de dispersão que conhece.
b) Efectue a referida comparação com base nas caixas-de-bigodes.
c) Construa um diagrama de dispersão para a pluviosidade em Lisboa e no Porto. Com base na figura que acabou de traçar, parece existir correlação significativa entre as variáveis?
d) Calcule o coeficiente de correlação de Pearson e de Spearman, e refira o que está a avaliar com cada um deles.

21. Os dados que se seguem referem-se ao valor do índice PSI (o índice mais representativo da Bolsa de Valores de Lisboa — Euronext Lisboa) em dias úteis consecutivos contados a partir de 6 de Outubro de 2006.

Dia	PSI20	Dia	PSI20	Dia	PSI20	Dia	PSI20	Dia	PSI20
1	10385.89	8	10421.12	15	10564.79	22	10589.82	29	10631.37
2	10412.59	9	10460.42	16	10569.64	23	10598.71	30	10622.91
3	10431.87	10	10442.92	17	10539.46	24	10599.66	31	10604.38
4	10427.72	11	10463.99	18	10515.57	25	10607.99	32	10580.02
5	10451.13	12	10503.55	19	10508.49	26	10601.69	33	10628.80
6	10455.77	13	10471.12	20	10499.59	27	10604.43	34	10611.28
7	10434.00	14	10564.68	21	10504.49	28	10588.03	35	10617.78

a) Recorrendo a meios computacionais adequados construa um cronograma com estes valores.
b) Uma medida muito utilizada para avaliar a evolução da bolsa consiste em considerar médias móveis. Para construir a série das médias móveis (de amplitude 5, por exemplo) a partir da série original, substitui-se cada valor desta pela média considerando o valor do dia e os 4 valores anteriores. Procura-se assim remover (ou pelo menos atenuar) as flutuações «especulativas» de muito curto prazo. De forma semelhante, constrói-se a série dos desvios padrões corrigidos móveis para avaliar (embora de forma algo «simplista») a evolução da volatilidade das cotações. Calcule, recorrendo a uma folha de cálculo, as séries das médias e dos desvios padrões corrigidos móveis (de amplitude 5) para o período em análise.

© McGraw-Hill

c) Repita o procedimento, considerando agora uma amplitude 15 (o valor do dia e os 14 valores anteriores). Comente os resultados obtidos.
d) Construa os cronogramas correspondentes às duas alíneas anteriores. Compare os cronogramas das médias móveis com o cronograma obtido na alínea a).

22. Uma amostra de 20 analistas financeiros foi inquirida sobre qual o rendimento por acção mais provável para determinada empresa. As respostas foram as seguintes:

29.39%	5.83%	6.31%	13.42%	8.79%	7.22%	14.23%
12.69%	6.46%	12.59%	0.13%	12.21%	12.33%	27.93%
–1.40%	25.37%	14.02%	11.78%	14.56%	5.62%	

a) Construa um resumo dos 5 números.
b) Calcule a média e o desvio padrão.
c) Com base no intervalo interquartis obtenha um intervalo de variação para o rendimento de expectativa.

23. A procura de água mineral engarrafada aumenta, como é bem sabido, nos meses de verão. Procura saber-se se, numa determinada empresa do sector, a máquina enchimento das garrafas de 1.5 litros está bem calibrada. Para tal, o departamento de produção recolheu uma amostra de 50 garrafas cujos pesos, em gramas, são os seguintes:

1503.50	1501.55	1503.37	1513.31	1497.78
1498.61	1496.55	1503.15	1496.94	1500.76
1506.22	1495.77	1511.71	1507.69	1497.39
1511.38	1500.11	1504.57	1509.51	1503.19
1510.99	1501.13	1504.07	1514.59	1504.84
1513.67	1494.41	1502.43	1504.58	1505.14
1494.08	1502.16	1514.86	1502.38	1503.39
1503.83	1502.98	1509.33	1508.38	1515.97
1510.48	1505.67	1516.88	1503.09	1496.29
1499.57	1503.17	1501.73	1508.79	1501.32

O trabalho consiste em analisar a amostra, utilizando o EXCEL, e elaborar um pequeno relatório executivo para a administração da empresa (lembra-se que a água pesa um quilograma por litro, e suponha-se que o invólucro de plástico pesa 10 gramas) recorrendo a medidas de localização e dispersão e a representações gráficas adequadas.

24. Num *site* para resolução de problemas de *Sudoku online* são referidas as medianas dos tempos gastos pelos internautas para resolver problemas de dificuldade semelhante. Refira porque razão a mediana parece ser, numa situação desta natureza, um indicador de localização mais fiável do que a média.

© McGraw-Hill

25. Dê dois exemplos de relações determinísticas.

26. Dê dois exemplos de correlações espúrias.

27. O quadro que se apresenta abaixo refere o valor das acções (em euros) das SAD do Sporting Clube de Portugal (SCP) e do Futebol Clube do Porto (FCP) de 3 de Janeiro a 27 de Março de 2000.

Dia	SCP	FCP	Dia	SCP	FCP	Dia	SCP	FCP
3-Jan	4.17	4.35	31-Jan	4.04	4.11	28-Fev	6.60	5.99
4-Jan	4.15	4.35	1-Fev	4.04	4.25	29-Fev	6.39	5.84
5-Jan	4.20	4.30	2-Fev	4.15	4.18	1-Mar	6.25	5.75
6-Jan	4.13	4.40	3-Fev	4.10	4.20	2-Mar	6.14	5.40
7-Jan	4.23	4.42	4-Fev	4.20	4.22	3-Mar	6.07	5.10
10-Jan	4.15	4.40	7-Fev	4.19	4.14	6-Mar	5.95	5.30
11-Jan	4.16	4.37	8-Fev	4.08	4.20	7-Mar	5.95	5.30
12-Jan	4.17	4.37	9-Fev	4.20	4.25	8-Mar	5.60	5.20
13-Jan	4.00	4.43	10-Fev	4.15	4.21	9-Mar	5.75	5.30
14-Jan	4.15	4.36	11-Fev	4.50	4.84	10-Mar	5.80	5.40
17-Jan	4.13	4.36	14-Fev	5.02	4.84	13-Mar	5.80	5.50
18-Jan	4.10	4.35	15-Fev	5.04	4.62	14-Mar	5.57	5.65
19-Jan	4.08	4.29	16-Fev	4.70	4.74	15-Mar	5.63	5.50
20-Jan	4.10	4.35	17-Fev	4.77	4.73	16-Mar	5.52	5.27
21-Jan	4.11	4.31	18-Fev	4.77	4.97	17-Mar	5.59	5.50
24-Jan	4.04	4.30	21-Fev	5.00	5.00	20-Mar	5.84	5.49
25-Jan	4.19	4.30	22-Fev	5.00	4.95	21-Mar	5.89	5.56
26-Jan	4.14	4.28	23-Fev	5.25	5.19	22-Mar	5.90	5.51
27-Jan	4.12	4.29	24-Fev	5.33	5.13	23-Mar	5.70	5.23
28-Jan	4.10	4.28	25-Fev	6.50	5.40	24-Mar	5.85	5.24

a) Construa um histograma para o valor das acções da SAD do SCP utilizando as seguintes classes: (0, 1.5], (1.5, 3], (3, 4.5], (4.5, 6] e (6, $+\infty$). Parece-lhe que tal histograma traduz o comportamento do activo em questão?

b) Construa um histograma para o valor das acções da SAD do FCP com base nas seguintes classes: (4, 4.2], (4.2, 4.5], (4.5, 5], (5, 6] e (6, $+\infty$).

c) Trace o diagrama de dispersão do valor destes dois activos financeiros. Calcule a covariância e o coeficiente de correlação entre estas variáveis.

d) Construa, na mesma figura, os cronogramas referentes a estes dois activos financeiros. Será que a análise gráfica permite confirmar a conclusão tirada na alínea anterior?

28. O quadro que segue diz respeito aos países europeus com «desenvolvimento humano elevado» e foi construído com base em informação contida no *Relatório do Desenvolvimento Humano 1999*, publicado para o Programa das Nações Unidas para o Desenvolvimento (PNUD, Trinova

© McGraw-Hill

Editora, Lisboa). Para estes países, registaram-se os valores assumidos para algumas variáveis: valor do índice de desenvolvimento humano (IDH), n.º de televisores por 1000 habitantes (TEL), produto interno bruto (PIB) real *per capita* em milhares de dólares — valores corrigidos para atender à paridade dos poderes de compra —, e % de mulheres no parlamento (MUL).

a) Construa um gráfico caule-e-folhas para a variável PIB.

b) Das variáveis TEL, PIB e MUL, qual lhe parece dever apresentar maior correlação com a variável IDH? Verifique a sua «intuição» calculando os respectivos coeficientes de correlação.

c) Parece-lhe que a avaliação da associação entre variáveis se alteraria substancialmente se se recorresse ao coeficiente de correlação de Spearman em vez do coeficiente de Pearson? Valide a sua teoria para uma das situações.

País	TEL	PIB	IDH	MUL	País	TEL	PIB	IDH	MUL
Suécia	476	19.790	0.923	42.7	Rep. Checa	406	10.510	0.833	13.9
Dinamarca	533	23.690	0.905	37.4	Irlanda	469	20.710	0.900	13.7
Noruega	709	24.450	0.927	36.4	Portugal	367	14.270	0.858	13.0
Finlândia	605	20.150	0.913	33.5	Polónia	418	6.520	0.802	12.9
Holanda	495	21.110	0.921	31.6	Eslováquia	384	7.910	0.813	12.7
Alemanha	493	21.260	0.906	29.8	R. Unido	612	20.730	0.918	12.3
Islândia	447	22.497	0.919	25.4	Itália	436	20.290	0.900	10.0
Áustria	496	22.070	0.904	24.7	Malta	497	13.180	0.850	9.2
Suíça	493	25.240	0.914	20.3	França	598	22.030	0.918	9.1
Luxemburgo	628	30.863	0.902	20.0	Eslovénia	375	11.800	0.845	7.8
Espanha	509	15.930	0.894	19.9	Grécia	442	12.769	0.867	6.3
Bélgica	464	22.750	0.923	15.8	Chipre	146	14.201	0.870	5.4

29. O quadro seguinte apresenta a duração, o tipo de chamada (L = local; LOP = local para outros operadores; M = móvel; INT = internacional) e o preço das chamadas efectuadas por determinado cliente da Portugal Telecom em Novembro de 2006.

Tipo chamada	Duração (seg)	Valor (euros)	Tipo chamada	Duração (seg)	Valor (euros)
L	55	0.041	L	49	0.07
M	60	0.248	L	200	0.131
LOP	116	0.096	M	53	0.248
M	70	0.496	L	58	0.07
L	91	0.083	L	27	0.07
M	72	0.496	M	394	1.735
L	27	0.041	L	395	0.216

(Continua)

© McGraw-Hill

(Continuação)

Tipo chamada	Duração (seg)	Valor (euros)	Tipo chamada	Duração (seg)	Valor (euros)
LOP	39	0.07	M	171	0.744
M	26	0.248	INT	5	0.1
M	68	0.496	INT	502	2.992
LOP	20	0.07	M	27	0.248
M	20	0.248	L	137	0.081
L	97	0.075	M	190	0.992
LOP	36	0.07	M	62	0.496
M	33	0.248	L	85	0.074
LOP	93	0.085	M	41	0.248

a) Calcule o coeficiente de correlação (de Pearson) entre a duração e o preço das chamadas.
b) Dentro de certas condições, as chamadas são facturadas ao segundo. Refira porque razão não se obtém um valor mais elevado para o coeficiente de correlação.
c) Calcule o coeficiente de correlação de Spearman, e refira se tem interesse neste contexto.

30. O quadro que se segue apresenta o número de acessos ao Diário da República Electrónico (DRE) nos dia úteis entre 2 de Outubro de 2006 e 15 de Novembro do mesmo ano. Naturalmente que, tratando-se de um sistema que funciona via *internet*, o acesso também é possível nos restantes dias, mas origina, como é fácil de perceber, valores bem menores.

Obs	Dia	N.º acessos DRE	Obs	Dia	N.º acessos DRE	Obs	Dia	N.º acessos DRE
1	2-Out	36250	11	17-Out	37616	21	31-Out	36116
2	3-Out	36734	12	18-Out	35771	22	2-Nov	37423
3	4-Out	35197	13	19-Out	35646	23	3-Nov	38546
4	6-Out	29996	14	20-Out	40198	24	6-Nov	39689
5	9-Out	41625	15	23-Out	35603	25	7-Nov	38303
6	10-Out	38703	16	24-Out	36844	26	8-Nov	38446
7	11-Out	50505	17	25-Out	39189	27	9-Nov	35118
8	12-Out	34782	18	26-Out	36920	28	10-Nov	32361
9	13-Out	36669	19	27-Out	46780	29	13-Nov	38233
10	16-Out	37404	20	30-Out	40350	30	14-Nov	37254
						31	15-Nov	39089

a) Calcule a média e a variância associadas a esta colecção de dados.
b) Trace um cronograma para o número de acessos/dia ao DRE, admitindo que as observações eram consecutivas. Será que o gráfico lhe sugere a existência de alguma tendência linear?
c) Confirme a sua resposta à alínea anterior utilizando o coeficiente de correlação entre o número da observação e o número de acessos/dia.
d) Construa uma caixa-de-bigodes referente ao número de acessos/dia e identifique eventuais *outliers*.

© McGraw-Hill

capítulo

Probabilidade

2.1 Introdução

No capítulo 1 sublinhou-se que um dos propósitos da estatística é o de estudar as características de determinada população ou universo a partir da observação de uma amostra. Tratando-se de uma inferência indutiva (da amostra para a população), não é possível extrair conclusões sem que a estas esteja associado um determinado grau de **incerteza.** Para tratar a incerteza é necessário recorrer à teoria da probabilidade, único método capaz de a quantificar. A linguagem das probabilidades é, assim, uma ferramenta insubstituível para a inferência estatística.

O objecto da teoria da probabilidade é o estudo de certos fenómenos observáveis, influenciados pelo acaso, ou seja, o estudo dos **fenómenos aleatórios.** Como a natureza destes fenómenos é muito variada (físicos, económicos, sociais, etc.), o objecto enunciado peca por ser excessivamente ambicioso. É mais razoável considerar que a teoria da probabilidade é a teoria matemática que se ocupa dos métodos de análise que são comuns ao estudo dos fenómenos aleatórios (termo primitivo da teoria), seja qual for o campo a que pertençam.

Os fenómenos aleatórios vão estudar-se fazendo apelo à noção de «experiência» aleatória. Quando se fala aqui em «experiência» deve entender-se como tal qualquer processo ou conjunto de circunstâncias capaz de produzir resultados observáveis. Trata-se, portanto, de uma ideia muito mais ampla do que a correntemente associada ao termo experiência. Quando o processo está sujeito à influência de factores casuais ou contingentes e conduz a resultados incertos fala-se em experiência aleatória. Mais precisamente:

DEFINIÇÃO 2.1 • EXPERIÊNCIA ALEATÓRIA • Diz-se que uma experiência é aleatória se apresentar as seguintes características:

a) O conjunto dos resultados possíveis é fixado antecipadamente;
b) O resultado da experiência nunca pode ser previsto de forma exacta, mesmo que se desenvolvam todos os esforços para manter sob controlo as circunstâncias relevantes para o resultado.

Ao contrário do que é habitual, não se refere como característica de uma experiência aleatória a possibilidade de a repetir um grande número de vezes em condições semelhantes. Com efeito, a definição apresentada é suficientemente geral para permitir considerar experiências aleatórias que se podem realizar apenas uma vez (situações únicas).

Soluções dos exercícios

Capítulo 1

5. a) quantitativa discreta; b) quantitativa contínua; c) quantitativa contínua; d) qualitativa ordinal; e) qualitativa nominal.
6. a) qualitativa ordinal; b) qualitativa nominal. c) quantitativa discreta.
7. d) Urbano: (5.9, 8.1, 8.5, 9.275, 10.4); Extra urbano: (4.1, 5.2, 5.45, 5.7, 6.1); Combinado: (4.7, 6.3, 6.6, 7, 7.6).
8. b) (0.1, 0.16, 0.19, 0.2225, 0.36); c) *outlier* moderado: 0.36; d) 0.1985, 0.19; e) só sobre a média.
9. b) (785.69, 954.965, 1050.39, 1208.085, 1596.47), *outlier* moderado: 1596.47; c) não, só se for ponderada; d) 0.8430, 0.8046.
10. a) 90.8111, 6397.6691, 0.8808; b) 6643.7333, 81.5091; c) (0, 52.85, 63.1, 121.15, 394); d) 15.26, 56.54; e) *outlier* moderado: 224.6, *outlier* severo: 394.0.
11. a) (–11.68, 1.325, 2.195, 9.4325, 28.9); *outliers* moderados: –11.68, 26.24, 26.47, 28.9; b) (–11.68, 0.7825, 15.1, 22.685, 28.9), (1.14, 1.31, 1.33, 2.41, 26.24).
15. a) 0.1011, 0.1074.
16. b) 2.4150, 1.5747; c) sim (*outliers* moderados).
17. a) 2.3562, 1.5829; b) (0, 1, 2, 3, 7); *outliers* moderados: 7.
19. b) 22.3, 243.71, 15.6112; c) 19.1304; d) (8.6207, 34.2857), 25.665; e) (0, 10].
20. b) Porto (3.6, 78.35, 149.1, 187.3, 403.4); d) 0.6567, 0.6148.
22. a) (–1.4, 6.4225, 12.27, 14.0725, 29.39); b) 11.974, 7.9026; c) (6.4225, 14.0725).
27. c) 0.4348, 0.9601.
28. b) 0.5524, 0.8861, 0.5576; c) 0.5779, 0.7718, 0.6043.
29. a) 0.7742; c) 0.3925.
30. a) 37850.3226, 13399033.5088; c) 0.0271; d) (29996, 35943.5, 37404, 38896, 50505); *outlier* moderado: 46780; *outlier* severo: 50505.