# Introdução à Estatística Aplicada e Estatística Descritiva*

**Eduardo Esteves**

25 de Setembro de 2009

Departamento de Engenharia Alimentar, Instituto Superior de Engenharia da Universidade do Algarve, *Campus* da Penha, 8005-139 Faro, Portugal. E-mail: eesteves@ualg.pt; URL: http://w3.ualg.pt/~eesteves ou "página" na Tutoria Electrónica.

**Resumo**

Neste documento, que deriva dos *Apontamentos de Estatística* (para a disciplina de Métodos Estatísticos leccionada no 1º Ano do Curso de Engenharia Alimentar até ao ano lectivo de 2006/2007), pretende-se abordar alguns dos tópicos introdutórios que se consideram (mais) relevantes (*vd.* programa da disciplina). Ao longo do texto incluem-se exemplos dos assuntos em estudo para auxiliar a compreensão das matérias. Complementarmente, providenciam-se exercícios, e respectivas soluções.

## Conteúdo



## 1 Introdução

A palavra "Estatística" deriva do latim "Estate", ou Estado, e foi usada pela primeira vez em meados do século XVIII por um professor alemão, Gottfried Achenwall[1]. A sua utilização estava inicialmente relacionada com a otenção de "informação vital", como por exemplo dados demográficos, "vitais" para a governação, para o recrutamento militar ou para a cobrança de impostos. Muitas vezes é usada como sinónimo de "dados": ouvimos falar em número de candidatos ao ensino superior, percentagem do PIB aplicado na Educação, etc. Aquele termo adquiriu, durante o século XIX, o significado moderno.

No entanto, a ESTATÍSTICA é mais do que isso; diz respeito não só aos dados mas também à colheita (amostragem), RESUMO, análise e interpretação de dados com vista à avaliação objectiva da validade das conclusões acerca do "mundo real" que se obtiverem[2]. Por MÉTODOS ESTATÍSTICOS entendem-se os métodos científicos para colher, organizar, resumir, apresentar e analisar dados de modo a obter conclusões válidas.

O OBJECTIVO GERAL da disciplina Estatística Aplicada (14451014) é providenciar, aos alunos do curso de engenharia alimentar, conhecimentos básicos de estatística, teóricos e práticos, que lhes permitam analisar estatisticamente problemas relacionados com o desempenho da actividade.

*Este artigo foi preparado usando um processador de texto WYSIWYM (*What You See Is What You Mean*), LyX, (mais) adequado à preparação de documentos de índole técnico-científica (vd. `http://www.lyx.org`).

[1]Esta versão da história não é consensual *vd. Earliest Known Uses of Some of the Words of Mathematics* em `http://members.aol.com/jeff570/mathword.html` (consultado em 17/09/2007).

[2]Organização para a Cooperação e Desenvolvimento Económico (OCDE em http://stats.oecd.org/glossary/index.htm consultado em 3/10/2008) ou a American Statistical Association (AmStat em http://www.amstat.org consultado em 5/10/2008)

## 2 Conceitos básicos: População, Amostra, Variável

Antes de mais, devem apresentar-se alguns conceitos importantes e que serão necessários ao longo deste texto.

O primeiro dos conceitos de que falaremos é o de população. Simplisticamente, o que se pretende com a análise estatística é elaborar conclusões sobre um grupo de medições ou observações da variável em estudo. Ora, o conjunto de medições ou observações realizadas sobre diferentes elementos de conjuntos bem definidos e rigorosamente condicionados designa-se por POPULAÇÃO. A respectiva dimensão identifica-se por $N$.

Existem vários "tipos" de populações e podem classificar-se as populações de acordo com vários critérios. Por vezes, as populações em estudo não existem na realidade, fisicamente. Neste caso, alguns autores referem-se a populações "imaginárias", "hipotéticas" ou "potenciais". No entanto, existem classificações mais consensuais e mais vulgarizadas de "tipos" de populações. Assim, podemos falar em POPULAÇÕES FINITAS e INFINITAS. As primeiras são constituidas por um número finito de elementos.

> Por exemplo, se quisermos estudar a altura dos alunos do Instituto Superior de Engenharia (ISE): todos os alunos da EST constituem a população em estudo. Se pretendermos estudar determinada característica de uma conserva de sardinha da marca XYZ: então a população que estamos a estudar é constituida por todas as latas de conserva de sardinha produzidas por essa determinada empresa. Se, por exemplo, estivermos a estudar, em laboratório, o efeito de determinado complemento alimentar sobre a taxa de crescimento de 40 cobaias; a população de que estamos a falar não são as cobaias mas as taxas de crescimento (de todas as cobaias que, eventualmente, poderiam receber esse complemento alimentar em condições similares).

Se estivermos a estudar uma população relativamente pequena, digamos as mulheres que já atravessaram o Canal da Mancha a nado ou o número de homens que pisou a Lua, poderemos examinar toda a população porque é praticável em tempo útil obter a informação que pretenderíamos daquelas mulheres ou desses homens. Assim, quando podemos examinar toda a população (neste caso, as dita(o)s senhoras ou senhores) estamos a realizar um CENSO. Todavia, em casos particulares, efectua-se o censo de populações maiores do que aquelas. Regularmente, de 10 em 10 anos, realiza-se o censo da população portuguesa com o objectivo de obter a tal "informação vital" para a governação designadamente o número de habitantes, as idades, as profissões, se possuem electricidade, água e telefone, etc.

Vulgarmente não é possível obter informação relativa a toda uma população. De facto, se estamos a estudar uma população maior, digamos as sardinhas da costa portuguesa ou a qualidade das sardinhas enlatadas por determinado fabricante, não será possível pesar, medir ou analisar bioquimicamente todos os peixes ou o conteúdo/embalagem de todas as latas. Então, poderemos examinar uma parte dessa população, ou seja, obter uma AMOSTRA. O número de elementos/observações, isto é, o TAMANHO DA AMOSTRA[3] designa-se $n$. Uma amostra é composta por um número determinado de observações individuais, geralmente referidas por $x_i$ em que $i = 1, 2, ..., n$. Este será o modo mais viável de estudar, do ponto estatístico (e não só!), muitos problemas práticos.

Podem obter-se amostras de uma população de acordo com vários critérios. Contudo, para se elaborarem conclusões válidas, a maioria dos métodos estatísticos assume que as amostras foram obtidas de modo aleatório, ou seja, é conhecida a probabilidade com que determinado elemento da população pode ser (es)colhido e a escolha de um dado elemento não influencia a escolha de outro(s) - AMOSTRAGEM ESTATÍSTICA. Obtêm-se, assim, AMOSTRAS ALEATÓRIAS. O conjunto dessa(s) amostra(s) possíveis de obter de determinada população com base em determinado critério é designado por AMOSTRAGEM. Como vimos, no entanto, também se utiliza o termo amostragem para designar o processo de obtenção das amostras. Falaremos neste curso de AMOSTRAGEM ALEATÓRIA SIMPLES como um exemplo dos vários critérios de selecção de amostras[4] .

As populações podem ser definidas por determinados PARÂMETROS que resumem certas características (que veremos a seguir). A esses parâmetros da população são usualmente atribuidas letras gregas ($\mu$, $\sigma$, etc.) ou letras maiúsculas ($N$, $X$, etc.), para os distinguir dos parâmetros correspondentes nas amostras (designados por letras minúsculas: $\bar{x}$, $s$, $n$, $x_i$, etc.) e que, por vezes, se designam ESTATÍSTICAS (ver a seguir).

As observações individuais que compoem uma amostra podem ser QUALITATIVAS, como por exemplo a cor, o sexo ou o comportamento, etc., ou QUANTITATIVAS como por exemplo o peso, a densidade, a taxa de crescimento, etc. Os elementos da amostra descrevem ou medem determinada característica da população (por exemplo, o peso,

[3] No caso das populações finitas, quando se obtém uma amostra, a FRACÇÃO DE AMOSTRAGEM $f$ é definida por $f = \frac{n}{N}$, em que $n$ é o número de elementos da amostra ou tamanho da amostra e $N$ o número total de elementos da população. Pelo contrário, as as populações infinitas são constituidas por um número infinito (não-determinado) de elementos. Assim sendo, a fracção de amostragem é "praticamente" igual a zero, uma vez que $n$ tende para $N$.

[4] *vd.* Garson, G. D. "Sampling" In *Statnotes: Topics in Multivariate Analysis* [disponível em `http://www2.chass.ncsu.edu/garson/pa765/sampling.htm` (consultado em 18/10/2007)].

o sexo ou o comportamento). Essa característica, que é descrita ou medida pelas observações individuais designa-se por VARIÁVEL (mais adiante elaboraremos sobre outras definições de variável).

> Ou seja, se pretendemos estudar o peso dos alunos desta disciplina no presente ano lectivo (que seria a variável $X$ em estudo), poderíamos obter uma amostra de 20 alunos seleccionados aleatoriamente (tamanho da amostra seria, então, $n = 20$) e pesar cada aluno. Obteríamos uma "lista" de vinte observações individuais, geralmente referidas por $x_i$, no exemplo o peso de cada aluno ($x_1 = 68$ kg, $x_2 = 53$ kg, ..., $x_{20} = 76$ kg). Com base na média das observações (i.e. pesos) que compõe a amostra ($\bar{x} = 67,3$ kg) é possível e/ou desejável "ter uma ideia " (i.e. estimar) o peso médio de todos os alunos da disciplina ($\mu$).

Existem vários "tipos" de variáveis organizados segundo diferentes critérios. Por vezes, a variável em estudo descreve determinada qualidade ou atributo em vez de medir certa quantidade: a cor, por exemplo. Alguns autores referem-se a ATRIBUTOS para designar este "tipo" de variáveis. Contudo, podemos, para facilitar a análise e a representação, substituir esses atributos por números, isto é em vez de olhos azuis atribuir o valor 1, ou em vez de olhos castanhos considerar o valor 2, etc. Em muitos casos, no entanto, as variáveis são mensuráveis, isto é, podem medir-se ou quantificar-se de alguma forma e, portanto, podem representar-se numericamente. Nestes casos, podemos considerar dois "tipos": VARIÁVEIS DISCRETAS (e.g. contagens); e VARIÁVEIS CONTÍNUAS (e.g. medições). Nas primeiras, as observações individuais só podem assumir determinados valores, enquanto no segundo caso, a variável pode assumir um qualquer valor entre quaisquer limites observados, ou seja, é possível existir um valor entre quaisquer outros dois valores observados.

> Por exemplo, o número de folhas numa árvore só pode assumir determinados valores. É possível contar 37 folhas, mas é impossível enumerar 37,48 folhas ou 36,125 folhas nesse ramo de àrvore - variável discreta. Se medirmos a altura dos alunos desta turma é possível obter resultados de 154 cm, ou mesmo 167,3 cm, ou até de 183,92 cm (depende do equipamento usado para medir). Podemos sempre obter, pelo menos teoricamente, valores de altura dos alunos entre os valores 154 cm e 155 cm, ou 167,0 cm e 167,1 cm, ou 172,03 cm e 172,04 cm, ou 181,007 cm e 181,008 cm, etc.

Paralelamente, os dados podem ser expressos em quatro escalas diferentes: nominal (ou categórica) - os dados são classificados por categorias não-ordenadas (*e.g.* nº pessoas por cor de cabelo, ou os consumidores por sexo); ordinal - quando os dados estão classificados por categorias ordenadas (por exemplo, nº alunos por nota final de estatística); de intervalo - os dados estão expressos numa escala numérica com origem arbitrária[5] (por exemplo, temperatura em ºC); e absoluta - neste caso, os dados são expressos numa escala numérica com origem fixa[6] (e.g. temperatura em K ou comprimento em cm)[7] .

Depois da colheita de informação, da organização e resumo desses dados, de forma a apresentá-los correctamente, é geralmente intenção generalizar os resultados para toda a população. A capacidade de elaborar conclusões para toda a população a partir de características de amostras corresponde à INFERÊNCIA ESTATÍSTICA (Fig. 1).

# 3 Estatística descritiva

Na maioria dos casos, em virtude da dimensão da população em estudo, é necessário recorrer a sub-conjuntos, a amostras, para estudar uma ou várias características de uma população. Independentemente de estarmos a lidar com populações ou com amostras, desde que os dados sejam numerosos, torna-se incómodo apresentá-los todos de cada vez que isso seja necessário. À apresentação total das observações individuais (ou dos "dados"), dispostos por ordem crescente ou decrescente de grandeza, alguns autores (mais antigos) designam por lista ou rol. No entanto, é possível (e muitas vezes desejável) descrever de forma resumida a informação relativa às amostras em estudo - a caixa designada "Amostragem" na Fig. 1.

## 3.1 Distribuição de frequências

Os dados de uma amostra (ou de uma população) podem, no entanto, ser tratados de forma a simplificar a sua apresentação e "manuseamento". Um processo consiste no seu AGRUPAMENTO, isto é, na apresentação em

[5] Pode atribuir-se um significado à diferença entre os números mas não à sua razão!

[6] Para além da diferença entre valores, também a razão entre resultados tem significado! Por outras palavras, é possível interpretar o valor $2x$ como o dobro de $x$.

[7] *vd.* StatSoft Inc. *Measurement scales* [disponível em `http://www.statsoft.nl/textbook/esc.html` (consultado em 18/10/2007)] ou Wuensch, K. (2003) *Scales of measurement* [disponível em `http://core.ecu.edu/psyc/wuenschk/docs30/scales.doc` (consultado em 18/10/2007)].

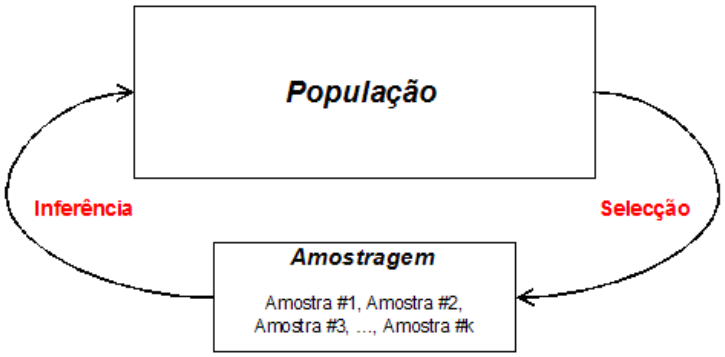


Figura 1: Esquema da relação entre os vários conceitos básicos em Estatística.

Tabela 1: Tabela de frequências (simplificada) para uma variável que pode assumir quaisquer valores entre 0 e 10. Os dados brutos obtidos são: 7, 6, 5, 7, 8, 9, 6, 7, 4, 6, 7, 10.

| Classes | Ponto-médio ($p_j$) | Frequência Absoluta ($F$) | Frequência relativa ($f$) |
|---|---|---|---|
| 3,5-4,9 | 4,2 | 1 | 0,0833 |
| 4,9-6,3 | 5,6 | 4 | 0,3333 |
| 6,3-7,7 | 7,0 | 4 | 0,3333 |
| 7,7-9,1 | 8,4 | 2 | 0,1667 |
| 9,1-10,5 | 9,8 | 1 | 0,0833 |

conjunto de todos os dados cuja grandeza é igual. Um dos modos de apresentar os dados é através de TABELAS DE FREQUÊNCIAS (Tab. 1) que abordaremos aqui; outro é elaborar TABELAS DE CONTINGÊNCIA, *i.e.* tabelas de frequências bidimensionais[8] (Tab. 2 ou 3). Enquanto no primeiro caso os dados dizem respeito a uma única variável, no segundo caso os resultados são organizados segundo dois critérios (variáveis).

Uma tabela de frequências inclui, geralmente, a seguinte informação: as classes consideradas (coluna da esquerda na Tab. 1); e as frequências propriamente ditas (nas colunas mais à direita). Opcionalmente apresentam-se os pontos médios das classes $p_j$. Vamos abordar a seguir como obter e dispôr essa informação.

**Cálculo do número de classes** Na maioria dos casos, é necessário definir arbitrariamente o número de CLASSES, ou categorias, que integram observações individuais da mesma ordem de grandeza, com que vamos elaborar a tabela de frequências.

Quando estamos a trabalhar com variáveis contínuas, um modo de resolver esta questão é recorrer à seguinte equação (fórmula de Sturges) para calcular o número de classes $k$:

$$k = I(\log_2 n) + 1$$

em que $\log_2$ é o logaritmo de base 2, $n$ é o tamanho da amostra e $I$ indica que o resultado deve ser arredondado ao número inteiro mais próximo.

> Se utilizarmos como exemplo os resultados que deram origem à Tab. 1, teríamos que para $n = 12$, logo $k = I(\log_2 12) + 1 = 4 + 1 = 5$ (uma vez que neste caso $\log_2 12 \cong 3,6$ ).

Pode-se obter o número de classes $k$ através de outra equação, talvez mais simples:

$$k = I\left(\frac{\log n}{\log 2}\right) + 1$$

neste caso, utilizam-se logarítmos de base 10. Se $n \geq 25$ é possível obter $k$ através de $k = \sqrt{n}$.

Nos casos das variáveis qualitativas (atributos) ou das variáveis discretas, o procedimento de elaboração de tabelas de frequências é ligeiramente diferente. Quando temos atibutos, podemos simplesmente definir como categorias ou classes, os diferentes atributos. Contabilizando o nº de observações/resultados por atributo preenche-se

[8] É possível, embora menos frequente, elaborar tabelas de contingência tridimensionais.

Tabela 2: Tabela de contingência relativa a duas variáveis, sexo (masculino ou feminino) e mão com que escreve (destro ou canhoto), obtida para uma amostra de 100 pessoas.

| | Masculino | Feminino | Total |
|---|---|---|---|
| Destros | 43 | 9 | 52 |
| Canhotos | 44 | 4 | 48 |
| Total | 87 | 13 | 100 |

Tabela 3: Tabela(s) de contingência relativa ao resultados dum inquérito acerca dos gostos da população (subdividida por sexo e classe etária) por diferentes tipos de cerveja ("light","normal","preta").

| 18-24 anos | Cerveja | | | |
|---|---|---|---|---|
| Sexo | "light" | "Normal" | "Preta" | Sub-Total |
| Masculino | 12 | 56 | 23 | 91 |
| Feminino | 50 | 17 | 5 | 72 |
| Sub-Total | 62 | 73 | 28 | 163 |
| 25-50 anos | Cerveja | | | |
| Sexo | "light" | "Normal" | "Preta" | Sub-Total |
| Masculino | 24 | 56 | 44 | 124 |
| Feminino | 32 | 11 | 6 | 49 |
| Sub-Total | 56 | 67 | 50 | 173 |
| >50 anos | Cerveja | | | |
| Sexo | "light" | "Normal" | "Preta" | Sub-Total |
| Masculino | 7 | 34 | 28 | 69 |
| Feminino | 18 | 11 | 2 | 31 |
| Sub-Total | 25 | 45 | 30 | 100 |

a tabela de frequências. Por outro lado, quando a variável é discreta, o processo de elaboração de tabelas de frequências é diferente e "um pouco subjectivo". A definição das classes depende tanto dos valores obtidos como do que se pretende ilustrar ou representar na tabela de frequências.

Exemplo 1: Uma empresa fabrica sete produtos congelados distintos (A a G, para simplificar) e os resultados das vendas por produto no último trimestre constam do relatório da auditoria trimestral. Neste caso (variável qualitativa - produto), cada produto constitui uma "classe" diferente e o nº de embalagens de cada produto vendidas no último trimestre corresponde à frequência absoluta nessa "classe".

Exemplo 2: Consideremos que uma variável discreta pode assumir quaisquer valores inteiros entre 0 e 20 (por exemplo, as classificações finais na disciplina de Métodos Estatísticos), e que os dados brutos obtidos são: 7, 9, 7, 10, 8, 6, 7, 6, 8, 12, 5, 10, 10, 9, 9, 8, 8, 9, 9, 11, 11 (resultados de 21 alunos num dado ano lectivo). Neste caso, o valor máximo = 12 e o valor mínimo = 5. No entanto, podemos definir vários agrupamentos diferentes consoante os objectivos: 5 classes (0 a 5; 6 a 9; 10 a 13; 14 a 17; e 18 a 20) que correspondem a "Medíocre", "Insuficiente", "Razoável", "Bom" e "Muito Bom"; 4 classes (5-6, 7-8, 9-10 e 11-12); 2 classes apenas (5-9 e 10-14); ou 20 classes (1, 2, 3, ..., 19, 20).

**Cálculo dos limites implícitos e obtenção das classes** Se estivermos a estudar características mensuráveis, isto é, no caso de variáveis contínuas, após a definição do número de classes a considerar no agrupamento dos dados ($k$), será necessário determinar que valores incluirá cada classe, ou seja entre que limites de classe serão contabilizadas as observações individuais. Um modo de determinar esses limites é recorrer aos próprios dados, daí a designação de LIMITES IMPLÍCITOS, e utilizar os valores mínimo e máximo das observações individuais. Se os valores, mínimo e máximo, das observações originais fossem, por exemplo, 4,5 e 10,5, então os limites implícitos da primeira e da última classe seriam 4,45 e 10,55 respectivamente (i.e. subtrair ao mínimo e adicionar ao máximo o valor do erro - se um resultado é indicado como nº inteiro, e.g. 4 m, o mais provável é ter sido arredondado de um qualquer valor entre 3,5 e 4,4999..., o que significa que a "medição" tem um erro de $\pm 0,5$ m; caso o resultado fosse apresentado, por exemplo, como 7,82 kg, o erro seria de $\pm 0,005$ kg, estando o peso real entre 7,815 e 7,82499...).

Tabela 4: Tabela de frequências para uma variável que pode assumir quaisquer valores entre 0 e 10. Os dados brutos obtidos são: 7, 6, 5, 7, 8, 9, 6, 7, 4, 6, 7, 10. $F$ - frequência absoluta; $F_A$ - Frequência absoluta acumulada; $f$ - frequência relativa; $f_A$ - frequência relativa acumulada.

| Classes | $p_j$ | $F$ | $f$ | $F_A$ | $f_A$ |
|---|---|---|---|---|---|
| 3,5-4,9 | 4,2 | 1 | 0,0833 | 1 | 0,0833 |
| 4,9-6,3 | 5,6 | 4 | 0,3333 | 5 | 0,4166 |
| 16,3-7,7 | 7,0 | 4 | 0,3333 | 9 | 0,7499 |
| 57,7-9,1 | 8,4 | 2 | 0,1667 | 11 | 0,9166 |
| 99,1-10,5 | 9,8 | 1 | 0,0833 | 12 | 1 |
| | | 12 | 1 | | |

Dividindo-se a amplitude dos limites implícitos pelo número de classes ($k$), calculado anteriormente, obtêm-se os (limites dos) intervalos de classe para elaboração da tabela de frequências (observe-se o exemplo seguinte).

> Consideremos que uma variável pode assumir quaisquer valores entre 0 e 10. Os dados brutos obtidos são: 7, 6, 5, 7, 8, 9, 6, 7, 4, 6, 7, 10. Se o valor máximo = 10, então limite (superior) implícito da última classe = 10,5. Se o valor mínimo = 4, então limite (inferior) implícito da primeira classe = 3,5. A amplitude (total) destes limites = 10,5 - 3,5 = 7. Uma vez que o número de classes proposto é 5, então a amplitude de cada classe será igual a 7/5 = 1,4. Ficam assim definidas as 5 classes: de 3,5 a 4,9; de 4,9 a 6,3; de 6,3 a 7,7; de 7,7 a 9,1; e de 9,1 a 10,5.

Para simplificar a tabela de frequências, para auxiliar a posterior elaboração dos histogramas (ou polígonos) de frequências ou para calcular estatísticas a partir dos dados agrupados, é útil determinar os pontos médios de cada classe, $p_j$, fazendo $p_j = (I_j + S_j)/a$, em que $I$ e $S$ são os limites de uma dada classe e $a$ é a amplitude dos intervalos de classe.

**Frequência absoluta, relativa e relativa acumulada** Depois de estabelecidas as classes, é necessário contabilizar os casos que estão incluidos em cada classe e desse modo obter a FREQUÊNCIA ABSOLUTA ($F_j$). Podemos acrescentar (e em muitos casos melhorar) a informação contida numa tabela de frequências. O cálculo da FREQUÊNCIA RELATIVA ($f_j$), de acordo com a seguinte equação:

$$f_j = \frac{F_j}{n}$$

em que $F_j$ é a frequência absoluta na classe $j$ ($j = 1, 2, ..., k$) e $n$ é número total de observações individuais (tamanho da amostra). As frequências relativas permitem comparar duas distribuições com $n$ diferente, ou seja, a partir de um segundo conjunto de dados (com $n$ igual, inferior ou superior) poderíamos preparar uma tabela de frequências relativas e desse modo comparar grosso modo (a "forma", a distribuição dos resultados) com a que acabámos de elaborar.

Mas, e se quiséssemos saber quantas observações individuais com valores entre 4,9 e 9,1 ocorreram na amostra do exemplo anterior? Ou quantas observações são maiores ou iguais a 7,7? Neste caso, podemos recorrer às FREQUÊNCIAS RELATIVAS ACUMULADAS ($f_A$) que se podem obter da soma da frequência relativa de determinada classe com a(s) frequência(s) relativa(s) das classe anteriores. Complementarmente, as frequências relativas acumuladas também são úteis para o cálculo de medidas de localização e dispersão da amostra (das quais falaremos mais adiante). Com esta informação podemos completar a tabela de frequências apresentada no início desta secção (Tab. 4).

## 3.2 Representação gráfica de distribuições de frequências

A partir de uma tabela de frequências, que apesar de muito informativa pode ser "maçadora", é possível elaborar representações gráficas, HISTOGRAMAS (Fig. 2 e Fig. 3, para variáveis contínuas e discretas, respectivamente) e POLÍGONOS DE FREQUÊNCIA (Fig. 4), mais apelativas e que permitem analisar visualmente os dados com maior facilidade.

## 3.3 Medidas de tendência central e de dispersão

Para além das tabelas de frequências e das suas representações gráficas (histogramas e polígonos de frequência), podemos descrever "resumidamente" a amostra (ou uma população) de outra forma. É possível, recorrendo a

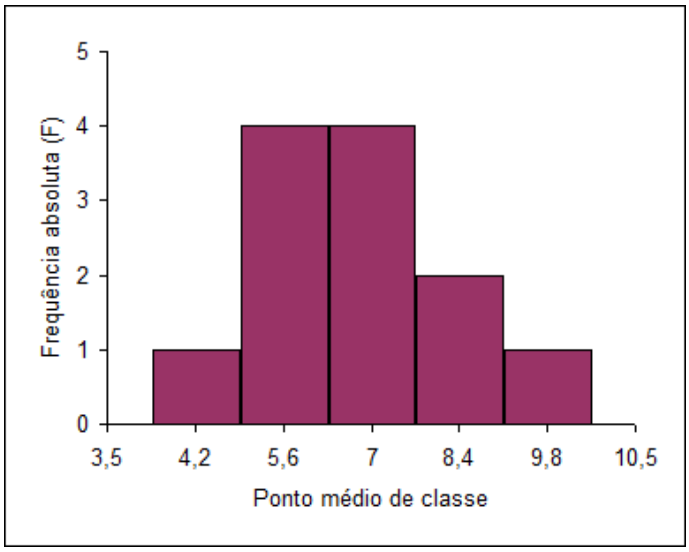


Figura 2: Histograma ("gráfico de barras") de variável contínua (da Tab. 2).

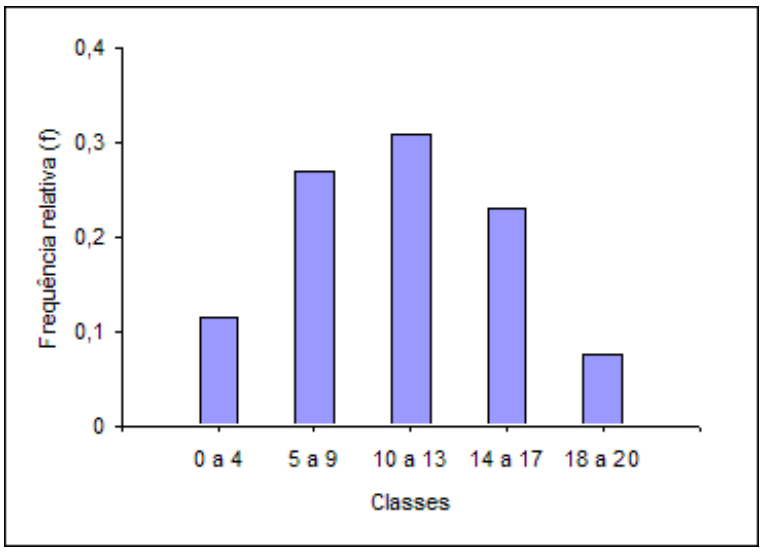


Figura 3: Histograma de frequências de variável discreta (cf. Exemplo na página 5).

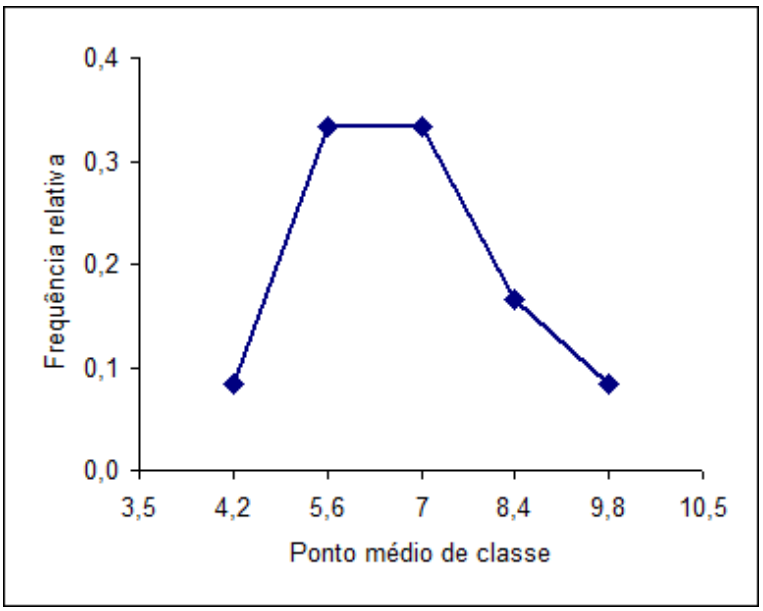


Figura 4: Polígono de frequências ("gráfico de pontos e linhas") de variável contínua (da Tab. 4).

alguns parâmetros (para populações) ou estatísticas (para amostras), caracterizar uma amostra (ou a população) em termos de localização (ou tendência central) e de dispersão. Estas medidas devem: ser objectivas; conter todas as observações; ser precisas quanto à sua interpretação; ser fáceis de calcular; e variar pouco relativamente às variações amostrais.

Geralmente, as observações de determinada característica (ou variável) de uma amostra parecem ocorrer com maior preponderância "perto" de valores "médios" ou "centrais" relativamente à amplitude observada. Assim, uma indicação da "média" da amostra (ou da população) seria expressiva e útil para a sua descrição. Existem vários PARÂMETROS ou MEDIDAS DE TENDÊNCIA CENTRAL, obtidos empriricamente da amostra ou da população, sendo que as mais comuns são a média, a mediana e a moda.

Também é importante quantificar a variabilidade (a variação) dos valores observados em torno dessa medida de tendência central. Esta informação, resumida em PARÂMETROS ou MEDIDAS DE DISPERSÃO, pode ser quantificada de diversos modos, como por exemplo através da amplitude, do intervalo inter-quartil, da variância, do desvio-padrão ou do coeficiente de variação.

As medidas que a seguir se apresentam, aplicam-se tanto a amostras como a populações. No entanto, a notação utilizada é relativamente diferente. Por exemplo, para a média e a variância populacionais usam-se as letras gregas minúsculas $\mu$ e $\sigma^2$, respectivamente. No caso de médias e variâncias amostrais utilizam-se $\bar{x}$ e $s^2$ respectivamente. Outras diferenças pontuais serão assinaladas no texto.

**Medidas de tendência central** Como se disse atrás, geralmente os valores, as medições ou as observações individuais de determinada variável numa amostra (ou população), encontram-se preponderantemente "perto" do centro da amplitude de valores. Existem várias medidas ou parâmetros de tendência central para quantificar esse valor "central", nomeadamente a MÉDIA, a MEDIANA e a MODA.

A MÉDIA ARITMÉTICA é a medida de tendência central mais usada em Estatística (e não só!) e que, em geral, se designa simplesmente por média (os autores anglófonos utilizam indiscriminadamente *mean* ou *average*). Se, numa amostra de tamanho $n$, considerarmos cada medida ou observação individual $x_i$ (em que $i = 1, 2, ...n$) da variável $X$, a média aritmética calcula-se através de:

$$\bar{x} = \frac{\sum x_i}{n}$$

sendo que $\sum$ (lê-se "sigma") indica o somatório dos elementos $x_i$. Quando se pretende calcular a média a partir de dados agrupados em tabelas de frequências com $k$ classes a média obtém-se através de:

$$\bar{x} = \frac{\sum F_j p_j}{n}$$

em que $F_j$ é a frequência absoluta e $p_j$ o ponto-médio da classe $j$. No caso de populações, a média aritmética $\mu$ (lê-se "miú") pode calcular-se de modo similar por:

$$\mu = \frac{\sum x_i}{N}$$

Admita-se que os dados amostrais obtidos são: 7, 6, 5, 7, 8, 9, 6, 7, 4, 6, 7, 10. A média (aritmética) calculada a partir daqueles dados é igual a: $\bar{x} = (\sum x_i)/n = 82/12 = 6,8$. Se recorrermos à tabela de frequências entretanto elaborada a partir daqueles dados (Tab. 4) a média será $\bar{x} = \frac{\sum F_j p_j}{n} = \frac{81,2}{12} = 6,8$. Se representarmos os dados, as observações individuais, por um histograma, a média corresponde (visualmente) ao centro de gravidade do histograma (imaginando que as barras têm peso proporcional ao tamanho), no local com barra(s) maiores (valores mais vezes observados) seria "mais pesado" (cf. Fig. 2).

Existem outras médias de uso menos frequente, designadamente a média geométrica, a média harmónica e a raíz quadrada média[9].

A MEDIANA é uma medida menos usada, apesar de em alguns casos ser mais apropriada do que a média[10]. Tipicamente, a mediana é definida como o valor, ou a observação, ou a medição, ou o caso, intermédio numa amostra arranjada por ordem de grandeza. Dito de outro modo, a mediana de um conjunto de números, ordenados por ordem de grandeza, é o valor para o qual metade dos elementos do conjunto são menores do que esse valor e outra metade são maiores do que esse valor. Podemos expressar este conceito da seguinte forma: num conjunto de valores ordenados por ordem crescente (ou decrescente, é irrelevante!), em que $i = 1, 2, ..., n$, a mediana $M$ (muitos autores utilizam a notação $\tilde{x}$) corresponde a

$$M = \tilde{x} \equiv \begin{cases} x'_{(n+1)/2} \text{ se } n\text{impar} \\ \frac{1}{2}(x'_{n/2} + x'_{1+n/2}) \text{ se } n\text{par} \end{cases}$$

em que $x'_{(n+1)/2}$ é a observação individual de ordem $(n+1)/2$. Quando $n$ é par, então $\tilde{x}$ é dado pela média aritmética dos valores de ordem $(n/2)$ e $(1 + n/2)$. No caso de dados agrupados em tabelas de frequências a mediana é dada por:

$$\tilde{x} = L + \left[\frac{\frac{n}{2} - \sum F}{F_{mediana}}\right] \cdot a$$

em que $L$ é o límite inferior da classe que contém a mediana, $n$ é o tamanho da amostra, $\sum F$ é o somatório das frequências das classes anteriores à classe que contém a mediana, $F_{mediana}$ é a frequência da classe que contém a mediana e $a$ é a amplitude dos intervalos de classe (ver exemplo seguinte). Para saber qual a classe que contém a mediana (essencial para resolver a equação anterior) deve "cruzar-se" a informação dada por $x'_{(n+1)/2}$ (independentemente do tamanho da amostra) com $F_A$.

Exemplo 1: Os dados brutos obtidos, ordenados por ordem crescente são: 4, 4, 5, 6, 8, 8, 8, 10, 10 ($n = 9$). A mediana $\tilde{x}$ é igual a 8, porque $x'_{(9+1)/2} = x'_5 = 8$.

Exemplo 2: Se os dados brutos ordenados por ordem crescente forem: 4, 5, 6, 6, 6, 7, 7, 7, 7, 8, 9, 10 ($n = 12$). $M$ = valor intermédio entre $x'_6$ e $x'_7$ que se calcula simplesmente como a média aritmética, isto é, $\tilde{x} = (7 + 7)/2 = 7$.

No caso dos mesmos dados, entretanto agrupados como na 4, $n = 12$ então $(n + 1)/2 = 6,5$, ou seja a mediana estará entre os valores de ordem 6 e 7 (i.e. $x'_6$ e $x'_7$); que está incluido na terceira classe (6,3-7,7) se observarmos a coluna de frequências acumuladas. Assim, obtém-se que $\tilde{x} = 6,7$ um valor diferente de $\tilde{x} = 7$ (obtido directamente dos dados "brutos") e ligeiramente inferior à média aritmética ($\bar{x} = 6,8$).

Na sequência do conceito de mediana, podemos ainda definir outras medidas de localização, de utilização menos comum. Um conjunto de dados organizados por ordem de grandeza, permite calcular, para além da mediana (o valor central que divide o conjunto em duas partes iguais), outros valores que dividem o conjunto em quatro, dez ou cem partes iguais, respectivamente quartis, decis ou percentis (genericamente designados por QUANTIS). Num gráfico de frequências relativas acumuladas, os quartis, decis e percentis são as abcissas cujas ordenadas correspondem à ordem $z$ (em que $z$ é o quantil pretendido). Podemos particularizar para os quatro casos de quartis (vulgarmente

[9] Nesta versão, decidi concentrar as atenções apenas na média aritmética.

[10] Para a amostra $\{1, 2, 3, 4, 50\}$ obtém-se $\bar{x} = 12$ e $s = 21,3$. Serão estas medidas representativas dos dados?

designados por $Q$) e considerando as amostras ordenadas[11] : 1) Quando $n = 4p$ (isto é, quando o tamanho da amostra é múltiplo "exacto" de quatro), o primeiro quartil é dado por $Q_1 = 1/2(x_p + x_{p+1})$, o segundo quartil é $Q_2 = \tilde{x} = 1/2(x_{2p} + x_{2p+1})$, e o terceiro quartil é $Q_3 = 1/2(x_{3p} + x_{3p+1})$; 2) Sempre que $n = 4p + 1$, então $Q_1 = 1/4(x_p) + 3/4(x_{p+1})$, $Q_2 = \tilde{x} = x_{2p+1}$, e $Q_3 = 3/4(x_{3p+1}) + 1/4(x_{3p+2})$; 3) Quando $n = 4p + 2$, logo $Q_1 = x_{p+1}$, $Q_2 = \tilde{x} = 1/2(x_{2p+1} + x_{2p+2})$, e $Q_3 = x_{3p+1}$; e 4) No caso de $n = 4p + 3$, os cálculos necessários serão $Q_1 = 3/4(x_{p+1}) + 1/4(x_{p+2})$, $Q_2 = \tilde{x} = x_{2p+2}$, e $Q_3 = 1/4(x_{3p+2}) + 3/4(x_{3p+3})$ (*cf.* exemplo seguinte). Se, porventura, os dados se encontram agrupados numa tabela de frequências, podem obter-se os quartis a partir de:

$$Q_1 = L + \left( \frac{\frac{n}{4} - F_A}{F_{Q_1}} \right) a$$

e de

$$Q_3 = L + \left( \frac{\frac{3n}{4} - F_A}{F_{Q_3}} \right) a$$

em que $L$ é o limite inferior da classe que contém o quartil, $n$ é o tamanho da amostra, $F_A$ é a frequência acumulada da classe anterior àquela que contém quartil, $F_{Q_1}$(ou $F_{Q_3}$) é a frequência da classe que contém o quartil e $a$ é a amplitude dos intervalos de classe.

> Os dados brutos obtidos, ordenados por ordem crescente são: 4, 4, 4, 5, 5, 6, 6, 8, 8, 8, 8, 10, 10, 11, 11, 12, 15 (n = 17). Neste caso estamos perante o caso 2) ou seja: o primeiro quartil será $Q_1 = 1/4(x_4) + 3/4(x_{4+1}) = 1/4(5) + 3/4(5) = 5$, o segundo quartil (ou mediana) será $Q_2 = \tilde{x} = x_{2 \cdot 4+1} = 8$ e o terceiro quartil é $Q_3 = 3/4(x_{3 \cdot 4+1}) + 1/4(x_{3 \cdot 4+2}) = 3/4(10) + 1/4(11) = 7,5 + 2,75 = 9,75$.

Na prática, obtêm-se os quantis a partir dos polígonos de frequência relativas acumuladas ou utilizando uma aplicação informática adequada, e.g. Microsoft©Excel ou OpenOffice Calc[12]. No primeiro caso, após localizar no eixo das ordenadas ($yy$) a ordem do quantil pretendido, pode-se procurar a correspondência horizontal no polígono de frequências e, depois, desenhar uma linha vertical até ao eixo das abcissas ($xx$). O ponto em que esta perpendicular intersecta o eixo dos $xx$ indica o resultado - o quantil pretendido. Procedendo de modo inverso, pode obter-se a ordem do quantil correspondente a determinado valor observado ($x_i$). Assim, se o quantil de ordem 66% de uma amostra é 27,1 cm, por exemplo, isso significa que 66% das observações, ou medições, são inferiores a 27,1 cm.

Uma terceira medida de tendência central é a MODA. A palavra moda é vulgarmente usada noutro contexto embora o seu significado estatístico não seja muito diferente daquele. A moda (ou normal segundo autores mais antigos), m, designa o valor (ou valores) que mais vezes ocorre(m) num conjunto de valores $x_i$ em que $i = 1, 2, ..., n$. Acontece que, por vezes, não é possível calcular m, pois em algumas séries de valores não existe nenhum repetido. Pelo contrário, noutros casos é possível que a série possua mais do que uma moda. No caso dos dados se encontrarem agrupados, não é possível identificar directamente a moda, mas simplesmente saber qual é a CLASSE MODAL, isto é, a classe que contém a moda.

> Os seguintes dados brutos não têm moda: 3, 5, 8, 10, 12, 15, 16. Se, no entanto, observarmos os seguintes casos, é possível determinar a moda: 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 (conjunto unimodal). Neste caso, $m = 9$. Noutro caso aínda, é possível observar duas modas: 2, 3, 4, 4, 4, 4, 5, 5, 7, 7, 7, 7, 9, 11 (conjunto polimodal). Neste caso, $m_1 = 4$ e $m_2 = 7$.

**Medidas de dispersão** É fácil constatar que as diferentes medidas de tendência central proporcionam estimativas (ligeiramente) distintas da "localização" do centro da distribuição de determinada variável numa amostra (ou população). Por outro lado, podemos verificar que a média, a mediana e a moda podem ser iguais em duas amostras que afinal podem ser substancialmente diferentes entre si. Na Fig. 5, a distribuição **a** é diferente da distribuição **b** porque os valores em b têm uma variabilidade maior do que os valores da distribuição a. Sendo assim, é necessário encontrar uma quantidade, um parâmetro análogo aos que encontrámos para a tendência central, que resuma esta variabilidade da distribuição. Existem várias medidas de dispersão para descrever numericamente essa variabilidade, nomeadamente a AMPLITUDE, o INTERVALO INTER-QUARTIL, a VARIÂNCIA, o DESVIO-PADRÃO e o COEFICIENTE DE VARIAÇÃO.

A AMPLITUDE $A$ (ou $h$) é a diferença entre o maior e o menor valor observados numa série de

[11] Não se indica, explicitamente, o facto da amostra estar ordenada (através de $x'$) para clarificar a apresentação das equações.

[12] Os quartis podem obter-se no Microsoft©Excel (PT) com =quartil(matriz;quarto) em que matriz diz respeito ao conjunto de valores (i.e. amostra) e quarto poderá ser 0, 1, 2, 3 ou 4 (respectivamente, o mínimo, o 1º quartil, a mediana, o 3º quartil e o máximo) (*vd.* menu "Ajuda" do software para lista de outras funções, e.g. percentis). No OpenOffice Calc, utilize-se a função =quartil(dados;tipo) em que dados e tipo têm o mesmo significado de matriz e quarto.

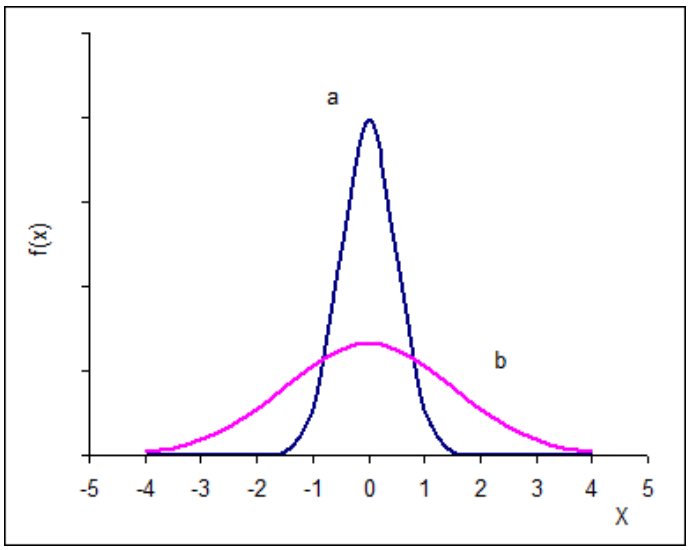


Figura 5: Representação esquemática da amplitude de duas distribuições a e b, simétricas e com igual "valor central" (neste caso, zero). No primeiro caso (distribuição **a**), $A = 3$ (*i.e.* de -1,5 a +1,5), enquanto no caso da distribuição **b**, $A = 8$ (de -4 a +4).

dados:

$$A = x_{maximo} - x_{minimo}$$

Na Fig. 5 é possível observar que as amplitudes de **a** e **b** são $A(\text{a}) = 1,5 - (-1,5) = 3$ e $A(\text{b}) = 4 - (-4) = 8$, respectivamente. A amplitude é fácil de obter e é expressa na mesma unidade da variável que estamos a estudar.

O INTERVALO INTER-QUARTIL, $IQ$, obtém-se da diferença entre o 3º e 1º quartis,

$$IQ = Q_3 - Q_1$$

No entanto, $A$ e $IQ$ são medidas "relativamente rudes" da dispersão dos dados, pois apenas consideram o valor máximo e o valor mínimo ($A$) ou o 1º e 3º quartis ($IQ$).

Outra medida de dispersão, bastante mais comum em estatística e frequentemente utilizada em análise estatística, é a VARIÂNCIA $s^2$ da amostra (ou a variância da população $\sigma^2$ - em que $\sigma$ lê-se "sigma"). Será necessário, entretanto, introduzir alguns conceitos que facilitam a compreensão do seu significado, nomeadamente os conceitos de desvio, de soma dos quadrados e de mínimos quadrados.

Poderíamos usar a informação contida na medida de tendência central (e.g. a média) e calcular a soma das diferenças entre cada valor individual $x_i$ e essa medida, (Fig. 6), para eventualmente avaliar a dispersão dos dados, isto é, calcular a soma dos desvios $D$:

$$D = \sum d_i = \sum (x_i - \bar{x})$$

Infelizmente verifica-se que $D = 0$, pois teoricamente existem tantos valores menores do que média assim como ocorrem valores maiores do que a média. Um modo de ultrapassar esta "dificuldade" é elevar ao quadrado os desvios $d_i$ e desse modo obter a **soma dos quadrados dos desvios** ou soma dos quadrados, SQ[13]:

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

Podemos refinar aínda mais esta quantidade e "ponderar" SQ pelo tamanho da amostra $n$ (ou dimensão da população $N$), e obter a média dos quadrados dos desvios dos valores individuais relativamente à média. Sendo assim, a variância da amostra $s^2$ é a média "ponderada" dos quadrados dos desvio dos valores individuais observados relativamente à média. Uma importante vantagem desta medida de dispersão é considerar todos os valores observados (e incluídos) na amostra, aliás como acontece com a média.

[13] Recorrendo ao conceito dos mínimos quadrados, demonstra-se que a SQ relativamente à média é menor do que a SQ em relação a qualquer outra medida de localização.

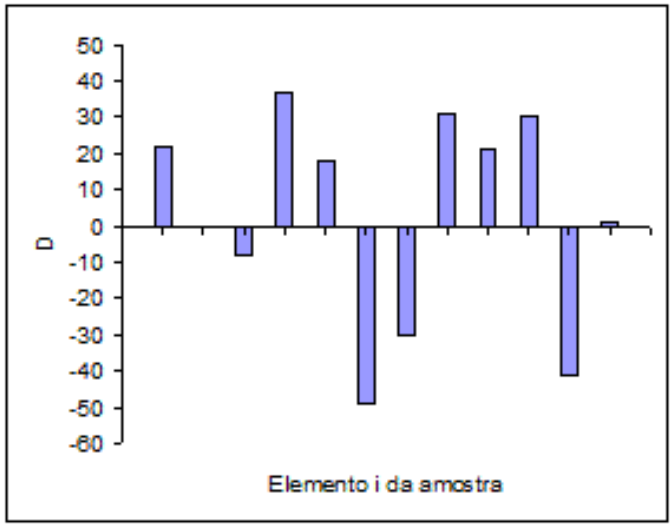


Figura 6: Representação esquemática dos desvios (D) de cada elemento $i$ da amostra relativamente à média, numa amostra de $n$=12 observações com média é igual a 26.

A VARIÂNCIA DA AMOSTRA $s^2$ expressa-se matematicamente por:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

em que $i = 1, 2, ..., n$. O denominador $n-1$, quantidade designada por **graus de liberdade** ou g.l., pretende considerar o facto de se ter usado um parâmetro da amostra (a média amostral $\bar{x}$) no cálculo. A equação apresentada permite obter uma estimativa não-enviesada (do inglês "unbiased") da variância da amostra. No caso de se pretender calcular a variância da população $\sigma^2$, então utiliza-se como denominador $N$ em vez de $n-1$, ou seja, $\sigma^2 = \sum (x_i - \bar{x})^2 / N$.

Se tivermos um conjunto grande de valores, isto é, se o tamanho da amostra for grande (n>30), então é possível calcular a variância recorrendo à seguinte expressão (simplificada):

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

Pode-se, aínda, calcular a variância numa amostra em que a informação está agrupada em tabela de frequências com $k$ classes de acordo com:

$$s^2 = \frac{n \sum F_j \cdot p_j^2 - (\sum F_j \cdot p_j)^2}{n(n-1)}$$

em que $j = 1, 2, ..., k$ e desde que $n \geq 30$. Segundo alguns autores, o desvio-padrão é a medida de dispersão mais importante em estatística paramétrica pois permite expressar a variabilidade das observações nas unidades da variável em estudo, ao contrário da variância. O DESVIO-PADRÃO da amostra, $s$, é a raíz-quadrada positiva da variância, ou seja:

$$s = \sqrt{s^2}$$

Por analogia com os parâmetros anteriores, o desvio-padrão populacional designa-se $\sigma$ e obtém-se através de $\sigma = \sqrt{\sigma^2}$. Esta medida de dispersão é expressa nas unidades dos valores observados, e antes da definição actual (de 1893) que se deve a Karl Pearson [1857-1936], designava-se por erro-médio.

> Uma queijaria regional produz queijos típicos de pequena dimensão. Obteve-se uma amostra da produção diária com as seguintes observações individuais (em g): 302, 374, 364, 318, 294, 343, 385, 348, 279, 365, 378, 357, 317, 304. O peso médio dos queijos na amostra é de 337,7 g e a variância é dada por $s^2 = [(302 - 377,7)^2 + (374 - 377,7)^2 + ... + (304 - 377,7)^2]/13 = 1208,07g$. Este valor é igual ao obtido recorrendo à equação simplificada. Tente confirmar esta afirmação? O peso médio da amostra é de 337,7 $g$ e a variância é 1208,07 $g^2$! Será melhor apresentar os resultados como $\bar{x} = 337,7g \pm s = 34,76g$!!

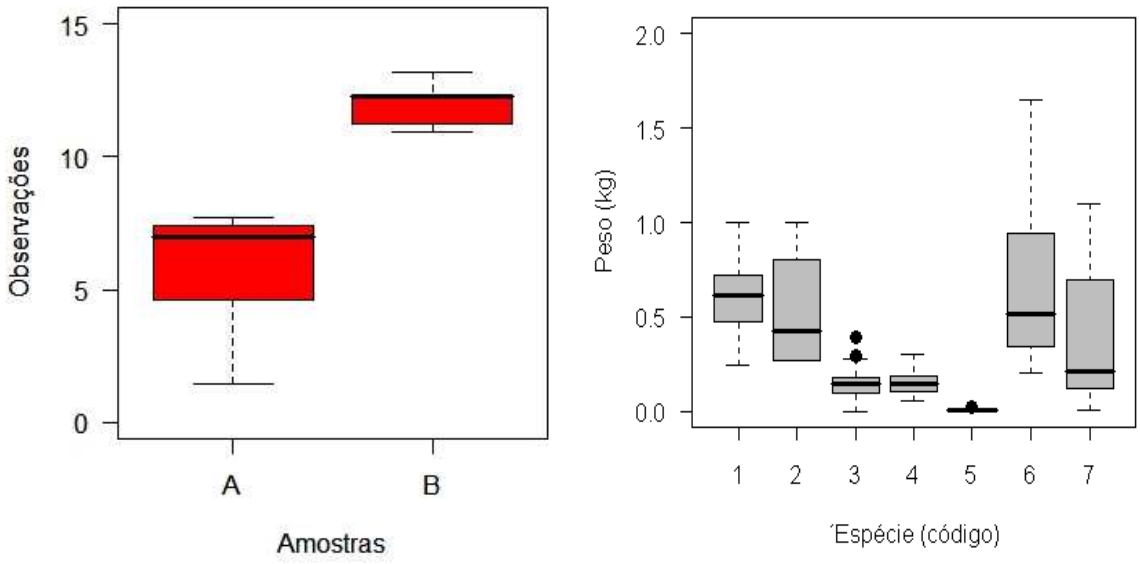


Figura 7: Gráficos de caixa-e-bigodes relativo a duas amostras (A e B), à esquerda, e para 7 espécies (à direita).

As medidas de dispersão de que falámos até agora são por vezes classificadas como medidas de dispersão absolutas pois referem-se à variabilidade numa amostra em termos concretos. Contudo, a comparação entre amostras com valores (e médias) substancialmente diferentes ou com unidades de medida diferentes, "dificulta" a sua utilização. As medidas de dispersão relativa, que resultam em números abstractos, relacionam numa mesma amostra uma medida de dispersão com uma medida de tendência central. A mais comum destas medidas é o COEFICIENTE DE VARIAÇÃO *c.v.* que é o quociente, geralmente expresso em percentagem, entre o desvio-padrão e a média, isto é:

$$c.v. = \frac{s}{\bar{x}} \cdot 100\%$$

Para resumir ou descrever a informação duma amostra (ou população) deve-se apresentar uma medida de localização e uma medida de dispersão, a média e o desvio-padrão por exemplo. É possível ilustrar estas medidas usando gráficos de caixa-e-bigodes (Fig. 7). Neste caso, representa-se a mediana (linha horizontal a negrito no interior da caixa), os $1^o$ e $3^o$ quartis (limites - base e topo - da caixa, respectivamente) e os extremos, mínimo e máximo (representados pelos bigodes).

Os tópicos abordados (amostra, tabelas de frequências e respectivas representações gráficas (histogramas e polígonos de frequências), e medidas de tendência central e de dispersão) são geralmente classificados como pertencentes ao âmbito da ESTATÍSTICA DESCRITIVA, de definição óbvia!

# 4 Exercícios

1. Na linha de enchimento de embalagens de manteiga de uma indústria de lacticínios, são retiradas periodicamente amostras para controlar o peso líquido do produto. Os pesos líquidos (em gramas) obtidos numa das amostras foram os seguintes: 256 215 276 256 260 270 280 246 234 273 214 272 293 258 229 284 218. Agrupe os dados em classes e calcule as frequências absolutas, relativas e absolutas acumuladas por classe (utilize a fórmula de Sturges e os limites implícitos para cada classe). Represente graficamente as distribuições obtidas.

2. A percentagem de àgua em salsichas do tipo frankfurt é controlada, numa determinada fábrica, retirando periodicamente amostras de salsichas antes do enlatamento. Os resultados das análises químicas a 30 salsichas foram os seguintes: 62 70 62 64 62 71 71 68 66 62 67 72 71 61 64 72 68 72 62 68 62 62 63 66 64 71 62 64 62 61 Agrupe os dados em classes e calcule as frequências absolutas e relativas por classe (utilize a fórmula de Sturges e os limites implícitos para cada classe). Represente graficamente as distribuições obtidas.

Tabela 5: Distribuição das caixas de camarão por calibre comercial numa dada amostra.

| Classificação comercial | SS | S | Q | K | T | TG |
|---|---|---|---|---|---|---|
| Nº caixas | 10 | 70 | 60 | 200 | 640 | 20 |

Tabela 6: Observações para as amostras A e B.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Amostra A | 0,9 | 1,2 | 1,4 | 1,3 | 1,3 | 1,6 | 1,4 | 1,4 | 1,2 |
| Amostra B | 1,1 | 1,5 | 1,4 | 1,4 | 1,1 | 1,3 | 1,2 | 1,3 | 1,3 |

3. Ao controlar os pesos de embalagens de certo produto, obtiveram-se os seguintes valores (em kg): 16,1 15,9 15,8 16,3 16,2 16,0 16,1 16,0 16,0 16,1 16,0 15,9 16,1 16,0 16,0 15,9 Agrupe os dados em classes e calcule as frequências relativas e relativas acumuladas por classe (utilize a fórmula de Sturges e os limites implícitos para cada classe). Represente graficamente as distribuições obtidas.

4. Foram seleccionados 18 provadores para avaliar sensorialmente o aroma de uma determinada marca de manteiga. Utilizou-se uma escala de 1 (aroma imperceptível) a 8 (aroma muito pronunciado). Represente graficamente a distribuição de frequências relativas das seguintes clasificações obtidas no teste: 7 6 7 3 6 6 7 7 6 7 7 4 5 8 6 4 6 6.

5. Numa determinada fábrica, pretende-se conhecer a distribuição de frequências, por calibre comercial, de 1000 caixas de camarão refrigerado. Num estudo efectuado obtiveram-se os resultados incluidos na Tab. 5 (os calibres comerciais estão ordenados por ordem crescente de tamanho do camarão): a) Represente graficamente a distribuição de frequências absolutas acumuladas. b) Por leitura do gráfico, indique quantas caixas existem com camarão de tamanho inferior ou igual a S; inferior ou igual a T; e superior a K.

6. Uma amostra é constituída pelos seguintes valores: 5, 1, 2, 3, 4, 5, 9, 3 e 5. a) Qual é o tamanho da amostra? b) Calcule as seguintes medidas: média, mediana, moda, menor valor (mínimo), maior valor (máximo), amplitude total, intervalo inter-quartil, variância, desvio-padrão e coeficiente de variação.

7. Duas amostras são constituidas pelos apresentados na Tab. 6. Para cada amostra calcule: a) a média e a mediana; b) a amplitude e o intervalo inter-quartil (sem agrupar os dados); c) Comente os resultados obtidos nas alíneas anteriores; d) Utilize um gráfico de caixa-e-bigodes para ilustrar os resutlados que obteve.

8. Calcule a média, a mediana, o desvio-padrão e o coeficiente de variação para: a) a tabela obtida na questão 2; b) os resultados do problema 3; c) a tabela obtida na questão 4.

9. Uma amostra de comprimentos de peixe, medidos em cm, foi resumida numa tabela de freqências (Tab. 7). a) Qual é o intervalo das classes de comprimento? E o ponto médio de cada classe? b) Calcule a média, a mediana, a variância, o desvio-padrão e o coeficiente de variação da amostra. c) Desenhe o polígono de frequências relativas acumuladas da amostra em papel milimétrico. e) Com base no polígono obtido na alínea anterior, determine o comprimento para o qual 25% dos elementos da amostra são inferiores (quantil de ordem 25% ou 1º quartil). g) Determine o quantil de ordem 50% (mediana) e compare com o resultado obtido em b). h) Marque os quantis de ordem 16% e 84%. Qual é o intervalo de comprimentos compreendido entre os quantis obtidos? i) Calcule as ordens de quantil correspondentes aos comprimentos $\bar{x} \pm s$, $\bar{x} \pm 2s$ e $\bar{x} \pm 3s$. j) Calcule a percentagem de elementos da amostra cujos comprimentos estão compreendidos entre $\bar{x} \pm 2s$.

Tabela 7: Tabela de frequências para uma dada composição de comprimentos de peixe.

| Classe de comprimento (cm) | Nº indivíduos |
|---|---|
| 10,5-12,5 | 2 |
| 12,5-14,5 | 2 |
| 14,5-16,5 | 6 |
| 16,5-18,5 | 8 |
| 18,5-20,5 | 10 |
| 20,5-22,5 | 10 |
| 22,5-24,5 | 36 |
| 24,5-26,5 | 46 |
| 26,5-28,5 | 22 |
| 28,5-30,5 | 6 |
| 30,5-32,5 | 4 |
| 32,5-34,5 | 8 |
| 34,5-36,5 | 6 |
| 36,5-38,5 | 2 |