

ADELAIDE FREITAS • JOÃO PEDRO CRUZ

BIO

estatística

Síntese Teórica e Exercícios

Com exercícios resolvidos e propostos
Questões de escolha múltipla



EDIÇÕES SÍLABO

Bioestatística

Síntese Teórica e Exercícios

Com exercícios resolvidos e propostos

Questões de escolha múltipla

ADELAIDE FREITAS

JOÃO PEDRO CRUZ

EDIÇÕES SÍLABO

É expressamente proibido reproduzir, no todo ou em parte, sob qualquer forma ou meio gráfico, eletrónico ou mecânico, inclusive fotocópia, este livro.

As transgressões serão passíveis das penalizações previstas na legislação em vigor.

Não participe ou encoraje a pirataria eletrónica de materiais protegidos.

O seu apoio aos direitos dos autores será apreciado.

Visite a Sílabo na rede

www.silabo.pt

FICHA TÉCNICA:

Título: Bioestatística – Síntese Teórica e Exercícios

Autores: Adelaide Freitas, João Pedro Cruz

© Edições Sílabo, Lda.

Capa: Pedro Mota

1ª Edição – Lisboa, janeiro de 2020

Impressão e acabamentos: VASP-DPS

Depósito Legal: 465988/20

ISBN: 978-989-561-039-6



Editor: Manuel Robalo

R. Cidade de Manchester, 2

1170-100 Lisboa

Tel.: 218130345

e-mail: silabo@silabo.pt

www.silabo.pt

Conteúdo

Prefácio	7
Síntese teórica	9
1.1 Estatística descritiva	9
1.1.1 Classificação das variáveis	9
1.1.2 Análise preliminar de dados numéricos	10
1.1.3 Organização da amostra	10
1.1.4 Descrição numérica da amostra	10
1.1.5 Descrição gráfica da amostra	13
1.2 Variáveis aleatórias	14
1.2.1 Variáveis aleatórias discretas	14
1.2.2 Variáveis aleatórias contínuas	16
1.2.3 Propriedades da esperança e da variância	17
1.2.4 Teorema do limite central	17
1.2.5 Distribuição da soma (exata e aproximada)	18
1.3 Inferência estatística	19
1.3.1 Características amostrais	19
1.3.2 Estimação pontual	20
1.3.3 Estimação intervalar e testes de hipóteses	21
1.4 Teste de ajustamento do qui-quadrado	23
1.5 Avaliar a normalidade dos dados	24
1.6 Inferência envolvendo um parâmetro	25
1.6.1 Inferência sobre a média μ desconhecida de uma população $X \sim N(\mu, \sigma^2)$ com variância conhecida	25
1.6.2 Inferência sobre a média μ desconhecida de uma população $X \sim N(\mu, \sigma^2)$ com variância desconhecida	26
1.6.3 Inferência sobre a média μ desconhecida de uma população X que não segue uma distribuição normal ou tem distribuição desconhecida	26
1.6.4 Inferência sobre a variância σ^2 de uma população $X \sim N(\mu, \sigma^2)$ com média desconhecida	27
1.6.5 Inferência sobre a proporção de sucesso p de uma população $X \sim \text{Bernoulli}(p)$	27
1.7 Inferência envolvendo dois parâmetros	28
1.7.1 Amostras emparelhadas	28

1.7.2	Inferência sobre a diferença de médias de duas populações normais, $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$, emparelhadas	28
1.7.3	Amostras independentes	29
1.7.4	Inferência sobre a diferença de médias de duas populações normais $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ independentes com variâncias σ_X^2 e σ_Y^2 conhecidas	29
1.7.5	Inferência sobre a diferença de médias de duas populações normais $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ independentes com variâncias σ_X^2 e σ_Y^2 desconhecidas mas iguais	30
1.7.6	Inferência sobre a diferença de médias de duas populações normais $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ independentes com variâncias σ_X^2 e σ_Y^2 desconhecidas e diferentes	31
1.8	Inferência envolvendo várias médias: ANOVA a 1 fator	32
1.8.1	ANOVA a 1 fator sobre populações independentes - efeitos fixos	33
1.8.2	ANOVA a 1 fator sobre populações independentes - efeitos aleatórios	35
1.9	Correlação	36
1.9.1	Coefficiente de correlação amostral de Pearson	36
1.9.2	Inferência sobre o coeficiente de correlação populacional de Pearson	37
1.10	Regressão linear simples	37
1.10.1	Inferência na regressão	38
1.10.2	Avaliação da qualidade da regressão	40
1.11	Análise de dados qualitativos	40
1.11.1	Teste de independência do qui-quadrado	41
1.11.2	Teste de homogeneidade do qui-quadrado	42
Exercícios resolvidos		43
2.1	Estatística descritiva	43
2.2	Variáveis aleatórias	48
2.3	Avaliar a normalidade dos dados	56
2.4	Inferência estatística envolvendo um ou dois parâmetros	57
2.5	ANOVA a um fator	69
2.6	Correlação e regressão linear simples	76
2.7	Análise de dados qualitativos	83
Exercícios propostos		87
3.1	Enunciados	87
3.2	Soluções	123
Bibliografia		125

Prefácio

O presente livro é fruto do trabalho desenvolvido pelos autores nos últimos anos na lecionação da unidade curricular de Bioestatística na Universidade de Aveiro. Apresenta um vasto conjunto de questões de escolha múltipla, com resolução (Capítulo 2) ou solução (Capítulo 3), precedido por uma exposição teórica essencial de conceitos e técnicas estatísticas (Capítulo 1), respondendo assim à necessidade dos estudantes por mais instrumentos que lhes permitam treinar e, desse modo, aprofundar os seus conhecimentos em Estatística ao nível de cursos de licenciatura do ensino superior ou politécnico nas áreas das Ciências da Vida.

Para as resoluções de cada exercício do Capítulo 2, são detalhados os comandos a usar, quer numa calculadora gráfica (escolheu-se a TI-84 para ilustração), quer no software estatístico R.

Parte do presente trabalho foi suportado pelo Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA, Universidade de Aveiro) através da Fundação para a Ciência e a Tecnologia (FCT), referências UIDB/04106/2020 e UIDP/04106/2020.

Aveiro, janeiro/2020

Adelaide Freitas e João Pedro Cruz.

Síntese teórica

1.1 Estatística descritiva

1.1.1 Classificação das variáveis

No estudo de uma população, uma característica mensurável ou um atributo (categórico) de interesse chama-se variável (estatística). Uma variável pode ser classificada quanto ao:

tipo : quantitativa ou qualitativa.

Quantitativa (ou numérica): assume valores intrinsecamente numéricos. Estas podem ainda ser classificadas quanto à sua natureza:

- discreta: assume um conjunto finito ou infinito numerável (que se consegue contar) de possibilidades.

Ex: 'Número de bactérias distintas por m³ de água contaminada'.

- contínua: assume valores em intervalos reais.

Ex: 'Tempo de vida de uma bactéria *Helicobacter pylori*'.

Qualitativa (ou categórica): assume valores intrinsecamente não-numéricos.

Ex: 'Ocorre/Não ocorre infecção pela bactéria *E. coli*'.

escala de medida : escala de razão, escala intervalar, ordinal, nominal.

- Escala de razões: os valores numéricos possuem ordem e as diferenças têm significado. O valor zero representa ausência da característica e os múltiplos de valores possuem significado.

Ex: 'Altura de um eucalipto'; 'Número de filhos por ninhada'.

- Escala intervalar: os valores numéricos possuem ordem e as diferenças têm significado. O zero não está bem definido.

Ex: 'Temperatura (em graus °C)'; 'Altitude'.

- Escala ordinal: os atributos (ou códigos numéricos atribuídos) possuem uma ordem intrínseca.

Ex: 'Grau de risco' (baixo, médio, alto); 'Pigmentação' (fraca, moderada, forte, total).

- Escala nominal: os atributos (ou códigos numéricos atribuídos) não possuem uma ordem intrínseca.

Ex: ‘Tipo de sangue’ (A, B, O, AB); ‘Identidade de Género’ (Feminino, Masculino, Transgénero, Outro).

1.1.2 Análise preliminar de dados numéricos

Genericamente uma amostra (estatística) univariada é constituída por n observações numéricas ou n observações categóricas.

Por vezes, nas Ciências da Vida, é frequente usar o termo ‘amostra’ para designar uma observação (unidade amostral).

Ex: ‘Na nascente de rio foram recolhidas 10 amostras de 1ml de água’.

A análise preliminar de uma amostra de dados numéricos poderá passar por três fases:

- Organização da amostra
- Descrição numérica da amostra
- Descrição gráfica da amostra

1.1.3 Organização da amostra

Uma amostra numérica poderá ser referida de várias formas. As formas mais usadas são:

- pelas suas n observações: x_1, x_2, \dots, x_n ;
- pelas suas n observações colocadas por ordem crescente; nesse caso, a amostra será denotada do seguinte modo: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$; e
- pelos seus k valores distintos e ordenados, $x_1^* < x_2^* < \dots < x_k^*$, com frequências absolutas n_1, n_2, \dots, n_k , respetivamente.

A organização de dados de uma amostra passa pela construção de tabelas de frequências (absolutas ou relativas, acumuladas ou não).

Ex: Seja a amostra de 10 observações: 2, 1, 0, 2, 1, 5, 1, 5, 2, 1. A tabela de frequências absolutas associada é:

Observação	$x_1^* = 0$	$x_2^* = 1$	$x_3^* = 2$	$x_4^* = 5$
Frequência	$n_1 = 1$	$n_2 = 4$	$n_3 = 3$	$n_4 = 2$

1.1.4 Descrição numérica da amostra

É efetuada através do cálculo de medidas estatísticas, nomeadamente:

- medidas de localização central;
- medidas de localização relativa;
- medidas de dispersão;
- medidas de forma.

Medidas de localização central

1. Média (\bar{x})

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^k x_i^* n_i}{n}.$$

2. Mediana (m_e)

Dizer que m_e é a mediana de uma amostra significa que pelo menos 50% das observações da amostra são menores ou iguais ao valor de m_e e pelo menos 50% das observações são maiores ou iguais ao valor de m_e . Uma expressão para o cálculo do valor da mediana é:

$$m_e = \begin{cases} x_{(\lfloor n/2+1 \rfloor)} & , \text{ se } n \text{ é ímpar} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & , \text{ se } n \text{ é par} \end{cases}$$

onde $\lfloor x \rfloor$ denota a parte inteira de x . Ex: $\lfloor 8.5 \rfloor = 8$.

3. Moda (m_o)

$$\bar{m}_o = x_i^* \text{ tal que } n_i = \max_{j=1, \dots, k} n_j .$$

Medidas de localização relativa

1. Mínimo e máximo (min e max)

$$\begin{aligned} \min &= x_1^* = x_{(1)} ; \\ \max &= x_k^* = x_{(n)} . \end{aligned}$$

2. Quantil de ordem p (Q_p).

Dizer que Q_p é o quantil de ordem p significa que pelo menos $p \times 100\%$ das observações da amostra são menores ou iguais ao valor de Q_p e pelo menos $(1 - p) \times 100\%$ das observações são maiores ou iguais ao valor de Q_p . Uma expressão para o cálculo do valor de Q_p é:

$$Q_p = \begin{cases} x_{(\lfloor np+1 \rfloor)} & , \text{ se } np \text{ não é inteiro} \\ \frac{x_{(np)} + x_{(np+1)}}{2} & , \text{ se } np \text{ é inteiro} \end{cases} .$$

Ex: $Q_{1/4}$ = quantil de ordem $1/4 = 1^\circ$ quantil.

Medidas de dispersão

1. Amplitude amostral

$$\max - \min = x_k^* - x_1^* = x_{(n)} - x_{(1)} .$$

2. *Distância interquartil (H)*

$$H = Q_{3/4} - Q_{1/4}.$$

3. *Variância (s^2)*

A simplificação da expressão de cálculo da variância depende de como estão organizados os dados.

Dados não ordenados:

$$\begin{aligned} s^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \\ &= \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2. \end{aligned}$$

Dados ordenados tabelados com frequências dos distintos valores:

$$s^2 = \frac{\sum_{i=1}^k (x_i^* - \bar{x})^2 n_i}{n} = \frac{\sum_{i=1}^k n_i x_i^{*2}}{n} - \bar{x}^2.$$

4. *Variância corrigida ($s_c^2 = \frac{n}{n-1} s^2$)*

$$\begin{aligned} s_c^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \\ &= \frac{\sum_{i=1}^n n_i x_i^2}{n-1} - \frac{n}{n-1} \bar{x}^2 \\ &= \frac{\sum_{i=1}^k (x_i^* - \bar{x})^2 n_i}{n-1} = \frac{\sum_{i=1}^k n_i x_i^{*2}}{n-1} - \frac{n}{n-1} \bar{x}^2. \end{aligned}$$

5. *Desvio padrão (s)*

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

6. *Desvio padrão corrigido (s_c)*

$$s_c = \sqrt{s_c^2} = \sqrt{\frac{n}{n-1}} s.$$




Medidas de forma

Para descrever a forma da curva da distribuição, seja discreta ou contínua, em termos de simetria ou assimetria, poderá considerar-se:

1. uma *medida de simetria* designada por g_3 e dada por:

$$g_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n s^3} = \frac{\sum_{i=1}^k (x_i^* - \bar{x})^3 n_i}{n s^3}.$$

2. a relação entre as medidas de localização moda (m_o), mediana (m_e) e média (\bar{x}), conforme a tabela seguinte:

assimetria	forma da distribuição	sinal de g_3	relação
positiva		> 0	$m_o \leq m_e \leq \bar{x}$
nula (simétrica)		$= 0$	$m_o = m_e = \bar{x}$
negativa		< 0	$\bar{x} \leq m_e \leq m_o$

1.1.5 Descrição gráfica da amostra

Os gráficos mais comuns para visualizar a distribuição dos dados de uma amostra são:

1. *Caixa de bigodes* (ou ‘boxplot’). Trata-se de uma representação que envolve cinco medidas sumárias (mínimo e máximo, primeiro e terceiro quartis e a mediana) conforme Figura 1.1.

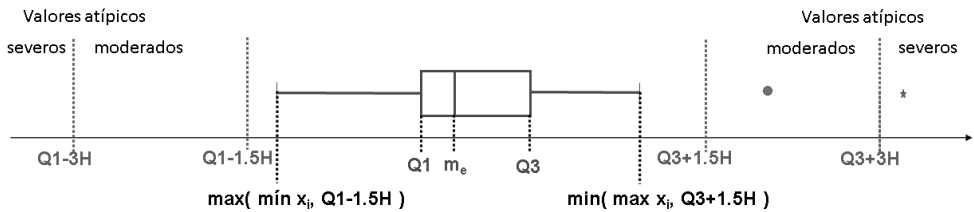


Figura 1.1: Esquema de construção de uma caixa de bigodes, onde Q1 e Q3 denotam $Q_{1/4}$ e $Q_{3/4}$, respetivamente.

Identificação de valores atípicos:

- moderados: observações da amostra com valores entre $Q_{1/4} - 3H$ e $Q_{1/4} - 1.5H$ ou entre $Q_{3/4} + 1.5H$ e $Q_{3/4} + 3H$;
- severos: observações da amostra com valores inferiores a $Q_{1/4} - 3H$ ou superiores a $Q_{3/4} + 3H$.

2. *Diagrama de Barras*. Gráfico de linhas verticais (ou horizontais), sem área, refletindo a forma da distribuição de frequências dos dados discretos da amostra. Cada linha tem altura igual à frequência do valor assinalado na base.

Ex: Fig 1.2 (a).

3. *Histograma*. Gráfico de retângulos justapostos na vertical (ou na horizontal), cuja área individual é proporcional à frequência do intervalo de valores assinalado na base do retângulo.

Ex: Fig 1.2 (b).

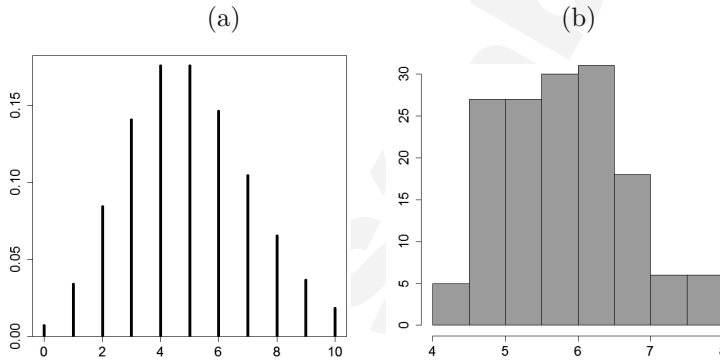


Figura 1.2: Exemplo de um (a) gráfico de barras, (b) histograma.

1.2 Variáveis aleatórias

Dito de forma simplificada, uma função que a cada acontecimento do espaço dos possíveis, associado a uma experiência aleatória, faz corresponder um número real chama-se *variável aleatória* (v.a.).

Em termos gerais, as v.a. podem ser discretas (se o domínio de valores possíveis é finito ou infinito numerável) ou contínuas (se o domínio de valores possíveis for infinito não numerável, por exemplo, dado por um intervalo real).

1.2.1 Variáveis aleatórias discretas

A distribuição de probabilidade de uma v.a. discreta pode ser caracterizada pela

- *função massa de probabilidade* (fmp)

$$f(x) = P(X = x), \quad \text{com } \sum_x f(x) = 1 ;$$

- *função de distribuição* (fd)

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t) .$$

No caso de v.a. discretas, a fd satisfaz as seguintes propriedades:

- (1) $P(X = a) = \text{salto da fd F no ponto } a = F(a^+) - F(a^-)$;
- (2) $P(a < X \leq b) = F(b) - F(a)$.

A esperança (ou média) e a variância de uma distribuição discreta são determinadas por:

$$\text{Esperança: } E[X] = \sum_x x f(x) \quad ;$$

$$\text{Variância: } \text{Var}[X] = \sum_x (x - E[X])^2 f(x) \quad .$$

No que se segue, seja p a probabilidade de sucesso de um evento. As seguintes distribuições discretas são usuais:

1. *Bernoulli*, $X \sim \text{Bernoulli}(p)$

$X=0$ se insucesso, $X=1$ se sucesso
com fmp

$$f(x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\},$$

média $E[X] = p$ e variância $\text{Var}[X] = p(1-p)$.

2. *Binomial*, $X \sim B(n, p)$

X = 'número de sucessos em n provas independentes'
com fmp

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n,$$

média $E[X] = np$ e variância $\text{Var}[X] = np(1-p)$.

3. *Geométrica*, $X \sim \text{Geom}(p)$

X = 'número de provas até ao primeiro sucesso'
com fmp

$$f(x) = P(X = x) = (1-p)^{x-1} p, \quad x = 1, 2, \dots,$$

média $E[X] = 1/p$ e variância $\text{Var}[X] = (1-p)/p^2$.

O mesmo nome 'distribuição geométrica' também é usado para modelar a v.a.

Y = 'número de insucessos até ao primeiro sucesso'.

Verifica-se que $Y = X - 1$. A v.a. Y tem fmp

$$f(y) = P(Y = y) = (1-p)^y p, \quad y = 0, 1, 2, \dots,$$

média $E[Y] = (1-p)/p$ e variância $\text{Var}[Y] = (1-p)/p^2$.

4. *Poisson*, $X \sim \text{Poisson}(\lambda)$

A distribuição de Poisson é frequentemente usada para modelar

X = 'número de ocorrências por unidade de tempo (ou espaço)'

com fmp

$$f(x) = e^{-\lambda} \lambda^x / x!, \quad x = 0, 1, 2, \dots,$$

média $E[X] = \lambda$ e variância $\text{Var}[X] = \lambda$.

1.2.2 Variáveis aleatórias contínuas

A distribuição de probabilidade de uma v.a. contínua pode ser caracterizada pela

- *função densidade de probabilidade* (fdp)
Qualquer função f tal que $f(x) \geq 0$, $\forall x \in \mathbb{R}$, e

$$\int_{-\infty}^{+\infty} f(x)dx = 1;$$

- *função de distribuição* (fd)

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt.$$

No caso de v.a. contínuas, a fd satisfaz as seguintes propriedades:

- (1) $P(X = a) = F(a^+) - F(a^-) = 0$, $\forall a \in \mathbb{R}$;
- (2) $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = F(b) - F(a)$.

A esperança (ou média), a variância e o quantil de ordem p de uma distribuição contínua são determinados por:

$$\text{Esperança: } E[X] = \int_{-\infty}^{+\infty} xf(x)dx \quad ;$$

$$\text{Variância: } \text{Var}[X] = \int_{-\infty}^{+\infty} (x - E[X])^2 f(x)dx \quad ;$$

$$\text{Quantil de ordem } p: \text{ valor } x_p \text{ tal que } F(x_p) = p \quad (0 < p < 1) \quad .$$

As seguintes distribuições contínuas são usuais:

1. *Uniforme*, $X \sim \text{Unif}(a, b)$

X com fdp

$$f(x) = \frac{1}{b-a}, \quad x \in [a, b],$$

média $E[X] = (a + b)/2$ e variância $\text{Var}[X] = (b - a)^2/12$.

2. *Exponencial*, $X \sim \text{Exp}(\lambda)$

X com fdp

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0,$$

média $E[X] = 1/\lambda$ e variância $\text{Var}[X] = 1/\lambda^2$.

3. *Normal*, $X \sim N(\mu, \sigma^2)$

X com fdp

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma > 0,$$

média $E[X] = \mu$ e variância $\text{Var}[X] = \sigma^2$.



ADELAIDE FREITAS é licenciada em Matemática Investigação (1987) e doutora em Matemática (1998) pela Universidade de Aveiro (UA). É mestre em Probabilidades e Estatística (1991) pela Faculdade de Ciências da Universidade de Lisboa. Atualmente é Professora Auxiliar no Departamento de Matemática (UA), onde tem lecionado disciplinas na área da Estatística a várias licenciaturas, mestrados e doutoramentos; é também membro do Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA-UA).



JÃO PEDRO CRUZ é licenciado em Eng. Informática pela Universidade de Coimbra e mestre em Engenharia Eletrotécnica e de Computadores pelo Instituto Superior Técnico (1999). É doutor em Matemática pela Universidade de Aveiro (2005), onde é Professor Auxiliar no Departamento de Matemática e tem lecionado diversas disciplinas na área da Estatística e da Análise Numérica. É membro do Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA-UA).

Este livro apresenta uma exposição teórica, sucinta mas rigorosa, de conceitos e procedimentos estatísticos, ilustrada com um amplo conjunto de exercícios que abrangem um largo espectro de tópicos habitualmente utilizados e necessários em cursos de licenciatura nas áreas das Ciências da Vida. Os exercícios – com resolução detalhada ou solução final – estão na forma de questões de escolha múltipla, sendo alguns extraídos de provas de avaliação da unidade curricular de Bioestatística lecionada pelos autores na Universidade de Aveiro. As resoluções apresentadas são acompanhadas de instruções claras para uma calculadora gráfica e descrição ilustrada de comandos para o *software* estatístico R.

ISBN 978-989-561-039-6



9 789895 610396