Capítulo 3 Introdução à Inferência Estatística

Exercícios de Distribuições de Amostragem e Estimadores

3.1. De uma população normal de média 8 e desvio padrão 4, extraíu-se uma amostra aleatória de dimensão 100.

a) Qual a probabilidade de a diferença, em valor absoluto, entre a média da amostra e a média da população ser superior a 0.5?

b) Se a população não fosse normal, qual seria essa probabilidade? Justifique.

3.2. O peso médio dos indivíduos duma certa espécie de bivalves é 31g e o respectivo desvio padrão é 2.4g. Recolhe-se uma amostra aleatória de 100 indivíduos desta espécie.

a) Qual a probabilidade, aproximada, de a média da amostra ser inferior a 30g?

b) Qual a probabilidade, aproximada, de o peso total da amostra ser superior a 3150g?

3.3. Mediu-se o comprimento (em cm) de 25 coelhos adultos de uma dada raça escolhidos aleatoriamente. A média da amostra foi 56.2 cm e o desvio padrão 4 cm. Admita que a população dos comprimentos dos coelhos adultos daquela raça é bem modelada por uma distribuição normal.

a) Indique estimativas para a média e o desvio padrão da população.

b) Usando as estimativas da alínea anterior como os parâmetros da população, determine a probabilidade de o comprimento de um coelho adulto daquela raça, escolhido ao acaso, ser superior a 57 cm.

c) Chama-se erro padrão da média amostral a uma estimativa do desvio padrão da v.a. média da amostra. Com base nos resultados obtidos, determine o erro padrão da média amostral?

3.4. Seja X uma variável aleatória com distribuição de Bernoulli de parâmetro p desconhecido e (X_1, X_2, X_3) uma amostra aleatória associada a esta variável.

a) Quais as propriedades da amostra aleatória (X_1, X_2, X_3)?

b) Sugerem-se os dois estimadores para p, $\hat{P}_1 = \overline{X}$ e $\hat{P}_2 = (X_1 + 2X_2 + X_3)/4$. Determine o valor esperado de cada um dos estimadores? Interprete.

c) Analise os estimadores quanto à sua variância. Qual dos dois estimadores considera que deve ser escolhido para estimar p? Justifique.

d) Procedeu-se à recolha de uma amostra, nas condições do enunciado, tendo-se obtido os valores $(0, 1, 0)$. Obtenha duas estimativas para p.

3.5. Depois de fabricado e embalado, a actividade de um certo adubo segue aproximadamente uma distribuição normal com $\mu = 120$ dias e $\sigma = 40$ dias.

Pretende-se enviar um lote de embalagens do referido adubo de modo que a actividade média amostral $(\overline{X})$ seja superior a 118 dias com probabilidade de 0.95. Qual o tamanho do lote a enviar?

3.6. Seja $(X_1, ..., X_n)$ uma amostra aleatória de tamanho n proveniente d uma população X com distribuição $U(0, 1)$. Calcule a probabilidade, aproximada, de $\overline{X}$ ser superior 0.9.

3.7. Uma amostra aleatória de dimensão 50, $(X_1, ..., X_{50})$, é extraída de uma população de Poisson com $\lambda = 10$. Recorra à distribuição normal para calcular um valor aproximado de $P(\overline{X} > 11)$.

3.8. Seja $X_1, X_2, \ldots, X_n$ uma amostra de variáveis aleatórias independentes e com a mesma distribuição, de valor médio μ e variância σ^2. Considere os seguintes estimadores de μ e σ^2, respectivamente,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{e} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

a) Prove que $E[\overline{X}] = \mu$ e $Var[\overline{X}] = \sigma^2/n$.

b) Prove que

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu)^2$$

c) Usando os resultados das alíneas anteriores, mostre que $E[S^2] = \sigma^2$.

3.9. Considere a amostra aleatória $X_1, X_2, \ldots, X_{100}$, retirada de uma população X para a qual se tem $E[X] = 0.5$ e $Var[X] = 0.0625$. Considere a variável aleatória $S = \sum_{i=1}^{100} X_i$.

a) Defina "amostra aleatória".

b) Determine o valor esperado e a variância de S. Justifique convenientemente.

c) Determine, aproximadamente, $P[S < 45]$.

Exercícios de Intervalos de Confiança e Testes de Hipóteses

3.10. Seja X uma população com distribuição normal, de média μ e desvio padrão σ=2. Uma amostra aleatória de dimensão 25 foi extraída desta população, tendo-se obtido $\overline{x} = 78.3$.

a) Calcule o intervalo de confiança a 99% para μ.

b) Qual o erro máximo cometido (a 99% de confiança) ao estimar μ por $\overline{x}$ =78.3?

c) Qual deverá ser a dimensão da amostra para que o erro máximo cometido, a 99% de confiança, ao estimar μ por $\overline{x}$, não exceda 0.1?

3.11. Para avaliar a tensão máxima suportada por uma barra de aço testaram-se n barras tendo-se obtido $\overline{x}_n = 20$ e para extremo superior do intervalo de confiança a 95% para o verdadeiro médio da tensão obteve-se 21.7. Sabendo que se admite que a tensão suportada por uma barra de aço é uma v.a. normal com desvio padrão $\sigma = 3$, determine o extremo inferior do intervalo de confiança e a dimensão n da amostra.

3.12. Uma empresa de peixe congelado está a ser investigada com o objectivo de se verificar se cada embalagem pesa de facto 1 Kg em média. Numa amostra aleatória de 100 embalagens de peixe registou-se o peso (Kg) de cada embalagem, x_i ($i = 1, \ldots, 100$), tendo-se obtido os seguintes valores:

$$\sum_{i=1}^{100} x_i = 95.9 \text{ Kg} \qquad \sum_{i=1}^{100} x_i^2 = 93.12 \text{ Kg}^2$$

a) Indique uma estimativa para a média e para a variância do peso de uma embalagem de peixe.

b) O que pode dizer sobre o resultado da investigação? Justifique convenientemente, especificando as hipóteses que necessitou de considerar.

c) Determine, explicitando as hipóteses necessárias, um intervalo a 95% de confiança para a variância do peso de uma embalagem de peixe.

3.13. O grau de acidez do azeite produzido em certa região supõe-se ter distribuição normal. Uma amostra da produção de dimensão 25 conduziu a uma acidez média de 1 grau e a um desvio padrão de 0.33 graus.

a) Face a estes valores, alguém sugeriu o intervalo]0.815, 1.185[para o valor esperado da acidez. Qual o nível de confiança associado a este intervalo?

b) Para o nível de confiança determinado na alínea anterior, qual o erro máximo cometido ao estimar o verdadeiro valor médio da acidez por 1 grau?

3.14. Numa amostra de 16 elementos, que se supõe ter sido retirada de uma população com distribuição normal, o desvio padrão obtido foi de 5.2

a) Foi calculado um intervalo de confiança para a média populacional tendo-se obtido $]24.2297; 29.7703[$. Indique justificando:

i) Qual a média da amostra;

ii) Qual o grau de confiança do intervalo calculado.

b) Qual o intervalo de confiança a 95% para a variância populacional?

3.15. Num laboratório foi recentemente desenvolvido um método de análise mais rápido e barato do que o método convencional. Realizaram-se testes com o objectivo de comparar a precisão do novo método com a do convencional, repetindo-se sucessivamente a mesma análise. No seguinte quadro encontram-se cálculos efectuados com os resultados dos testes, em unidades adequadas (x designa o resultado do teste pela aplicação de cada um dos métodos):

	nº de observações	$\sum_i (x_i - \overline{x})^2$
Método convencional	24	30.5
Método novo	30	55.7

Sendo a precisão de um método medida pela variância dos resultados da aplicação desse método, poder-se-á concluir que os dois métodos têm precisões distintas? Que hipóteses é necessário admitir para responder a esta questão?

3.16. Cada um de dois fabricantes de máquinas de lavar tem uma rede de vendas nacional, COM UM REPRESENTANTE POR CIDADE. Seleccionaram-se aleatoriamente dez representantes de cada fabricante. A média de máquinas vendidas pelos 10 representantes do fabricante A (durante um ano) foi $\overline{x}_1 = 84$ com um desvio padrão $s_1 = 8$. A média de máquinas vendidas pelos 10 representantes do fabricante B foi $\overline{x}_2 = 77$, com um desvio padrão s_2 =10 (no mesmo ano). Responda às seguintes alíneas, explicitando em cada caso qualquer hipótese adicional que seja necessária.

a) Será que a variância das vendas dos representantes de cada fabricante é igual?

b) Comente a veracidade da seguinte afirmação: ”os representantes do fabricante A vendem em média tantas máquinas de lavar como os representantes do fabricante B”.

c) Suponha agora que foram seleccionadas aleatoriamente 10 cidades do país onde os resultados das vendas (durante o mesmo ano) foram as seguintes:

Fab. A	76	60	85	58	91	75	82	64	79	90
Fab. B	81	52	87	70	86	77	90	63	85	85

Comente neste caso a afirmação da alínea b).

d) Independentemente dos resultados concretos obtidos nas alíneas b) e c), se tivesse que decidir qual o processo estatístico a utilizar para aferir da validade da afirmação da alínea b), qual escolheria: a selecção aleatória de 10 representantes de cada fabricante ou a selecção aleatória de dez cidades? Justifique a sua resposta.

3.17. Numa experiência agronómica pretende-se avaliar o crescimento total de uma certa espécie de plantas (expresso em peso seco) relativamente a dois regimes de fertilização A e B.

Ao fim de determinado tempo procedeu-se a medições, tendo-se obtido os seguintes resultados:

A	5.44	5.36	5.60	6.46	6.75	6.03	4.15	4.44
B	5.12	3.80	4.96	6.43	5.03	5.08	3.22	4.42

a) Numa experiência anterior (com um elevado número de plantas da mesma cultivar) relativa ao tratamento A, obteve-se uma variância de 0.42. Verifique se os dados actuais são consistentes com esse valor. Diga, justificando, se haveria alguma(s) hipótese(s) necessária(s) à resolução do problema.

b) Verifique se os dois regimes de fertilização A e B evidenciam diferenças significativas no que respeita ao crescimento das plantas. Explicite as hipóteses necessárias à resolução do problema.

3.18. Num estudo sobre o número de folhas por planta de tabaco, obtiveram-se os seguintes dados:

n º de folhas	17	18	19	20	21	22	23	24
n º de plantas	3	22	44	42	22	10	6	1

a) Indique a unidade estatística e a variável em estudo.

b) Determine a média e a mediana da amostra. Compare-as e comente.

c) Com base nesta amostra poder-se-á afirmar que 90% das plantas de tabaco têm 21 ou menos folhas? Justifique.

3.19. Duas marcas de chocolate (A e B) produzem chocolate rotulado "70% de cacau". Seleccionaram-se aleatoriamente 15 chocolates de cada uma das marcas, que foram analisados quanto ao teor de cacau. Os resultados foram introduzidos no *software* R. Utilize, sempre que possível, os resultados apresentados abaixo para responder às seguintes questões (alguns cálculos apresentados são desnecessários ou inadequados).

a) Conjectura-se que as variabilidades do teor de cacau dos chocolates das duas marcas são diferentes. Será esta conjectura compatível com os dados obtidos? Explicite e valide os pressupostos necessários à resolução do problema.

b) Indique estimativas para o valor esperado do teor de cacau dos chocolates de cada uma das marcas.

c) Os dados recolhidos evidenciam que o teor médio de cacau do chocolate da marca A é significativamente superior ao da marca B? Explicite e valide os pressupostos necessários à resolução do problema.

d) O produtor de chocolate da marca A afirma que o teor médio de cacau do seu chocolate é superior a 70%. O que pode dizer sobre a afirmação do produtor? Justifique convenientemente.

   Especifique as hipóteses nula e alternativa de um teste de hipóteses que permita averiguar a validade desta afirmação.

```
> A
 [1] 72.8 69.6 70.0 73.3 69.8 71.1 71.0 71.0 71.0 76.3 71.3 75.4 72.1 70.4 68.6
> B
 [1] 69.8 71.3 67.3 69.4 67.7 67.9 70.7 70.5 70.2 71.4 71.6 73.1 70.6 68.5 70.2

> var(A)              > var(B)                > var(A-B)
[1] 4.483143          [1] 2.616952            [1] 3.643810


> shapiro.test(A)                          > shapiro.test(B)
        Shapiro-Wilk normality test                Shapiro-Wilk normality test
data:  A                                   data:  B
W = 0.9066, p-value = 0.1199               W = 0.9613, p-value = 0.7144

> shapiro.test(A-B)
        Shapiro-Wilk normality test
data:  A - B
W = 0.9754, p-value = 0.9283


> var.test(A,B)
        F test to compare two variances
data:  A and B
F = 1.7131, num df = 14, denom df = 14, p-value = 0.3254
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5751437 5.1026658
sample estimates:
ratio of variances
```

```
          1.713116


> t.test(A,B,paired=T,alternative="less")
        Paired t-test
data:  A and B
t = 3.1787, df = 14, p-value = 0.9967
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.434763
sample estimates:
mean of the differences
               1.566667


> t.test(A,B,paired=T,alternative="greater")
        Paired t-test
data:  A and B
t = 3.1787, df = 14, p-value = 0.003349
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6985701       Inf
sample estimates:
mean of the differences
               1.566667


> t.test(A,B,paired=F,var.equal=T,alternative="less")
        Two Sample t-test
data:  A and B
t = 2.2771, df = 28, p-value = 0.9847
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.737039
sample estimates:
mean of x mean of y
 71.58000  70.01333


> t.test(A,B,paired=F,var.equal=T,alternative="greater")
        Two Sample t-test
data:  A and B
t = 2.2771, df = 28, p-value = 0.01531
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
```

```
 0.3962939       Inf
sample estimates:
mean of x mean of y
 71.58000  70.01333
```

3.20. Para decidir se deveria ou não lançar um novo produto no mercado, uma empresa de bens alimentares fez um inquérito em 10 supermercados do Sul e 20 do Norte do país, acerca do número de unidades X do referido produto que estes esperam poder vender semanalmente. Obtiveram-se os seguintes resultados:

	$\sum x_i$	$\sum x_i^2$
Sul	1000	102550
Norte	1200	75950

a) Determine estimativas para o número médio e para a variância das unidades que os supermercados do Sul e do Norte esperam vender.

b) "Embora a média de vendas difira significativamente no Norte e Sul, a variabilidade das vendas é praticamente igual". Critique fundamentadamente a afirmação anterior, justificando.

3.21. A fim de investigar os efeitos de ambientes nitrosos e de ambientes fosfatados no desenvolvimento de colónias de bactérias, contaminam-se 10 plaquetas envolvidas em cada um daqueles ambientes com as bactérias em estudo, e deixa-se incubar durante 24 horas. Após esse tempo, procede-se à contagem do número de colónias de bactérias em cada plaqueta, tendo-se obtido os seguintes resultados:

Ambiente nitroso	60	47	12	29	51	46	49	74	63	101
Ambiente fosfatado	8	46	21	13	58	33	20	46	31	38

a) Investigue a hipótese de o tipo de ambiente não influir no desenvolvimento das colónias de bactérias.

b) Que hipótese(s) foi necessário considerar para poder resolver a alínea a)?

3.22. Numa Estação Florestal estudam-se problemas de intercepção de precipitação. Nesse sentido em 12 dias de chuva (suponha que se trata de uma amostra aleatória de dias com precipitação) são colocados dois udómetros para medir a quantidade de precipitação: um numa zona desarborizada e outro sob as copas das árvores. As leituras da quantidade de água em cada udómetro (medidas em cm de altura) deram os seguintes valores (cada coluna corresponde a um dia):

zona descoberta	5.87	1.30	2.34	2.82	5.89	9.09	1.93	9.27	4.65	4.35	5.00	8.43
sob coberto	4.96	1.14	2.24	2.26	4.75	7.83	1.86	8.85	4.17	3.65	4.08	7.99

a) Estime a precipitação média na zona desarborizada e na zona sob coberto.

b) Será admíssivel supor que a 95% de confiança as precipitações médias nos dois casos são iguais? Explicite as hipóteses necessárias à resolução desta questão.

c) Tendo em conta que, pela própria natureza do problema, o nível de precipitação sob coberto não deverá ser superior ao nível da correspondente precipitação a descoberto, indique um procedimento estatístico mais adequado para avaliar se existem diferenças significativas nos dois casos.

3.23. Pretende-se testar se a proporção de ulmeiros afectados pela grafiose é idêntica em duas zonas A e B. Na zona A foi recolhida uma amostra aleatória de 30 ulmeiros e verificou-se que 20 estavam afectados pela grafiose. Na zona B recolheu-se uma amostra de 35 ulmeiros e verificou-se que 27 estavam afectados pela grafiose. Que conclusão se pode tirar ao nível de significância de 0.05?

3.24. Diga, justificando, se cada uma das afirmações seguintes é verdadeira ou falsa:

a) Num teste de hipóteses $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ em que se obteve um *p-value* $= 0.034$, não se rejeita a hipótese de a média populacional ser nula com um nível de significância de 0.01.

b) Se $]2.15, 3.24[$ é um intervalo de confiança a 95% para σ^2 então a probabilidade de σ^2 pertencer a este intervalo é igual a 0.95.

3.25. Suponha que o rendimento de um pé de tomateiro expresso em kg é uma variável aleatória com distribuição normal de valor médio 1 kg. Numa parte da produção foi utilizado um novo fertilizante. Observada uma amostra de 10 pés de tomateiro da parte da produção em que foi utilizado o novo fertilizante obtiveram-se os seguintes resultados:

1.375 1.223 1.773 1.752 0.779 1.407 1.068 1.633 1.201 1.042

expressos em kg. Que decisão se deverá tomar perante estes resultados, face ao novo fertilizante?

3.26. Um processo de determinação do conteúdo de enxofre de um determinado produto forneceu os seguintes resultados:

1.12 1.10 1.08 1.06 1.08 1.14 1.10 1.11 1.14

a) Encontre um intervalo de confiança a 95% para o verdadeiro conteúdo médio de enxofre, apresentando as hipóteses a fazer sobre a população que achar convenientes.

b) De acordo com certas normas, o conteúdo de enxofre não deve ter um desvio padrão superior a 0.02. Acha que a amostra recolhida permite afirmar que o produto está dentro das normas? Justifique convenientemente a resposta.

c) De uma outra fábrica que produz idêntico produto recolheu-se uma mostra de 20 elementos que proporcionou um desvio padrão de 0.085. Pode concluir que a variabilidade do conteúdo de enxofre é, nesta fábrica, três vezes superior à da outra?

3.27. É desencadeado um programa de controlo da poluição de um rio em que são efectuadas medições, antes de lançar a campanha antipoluição e um ano após. As medições são combinações de vários índices; quanto maior for o valor resultante maior é a poluição. Obtiveram-se os seguintes resultados:

Ponto de controlo	1	2	3	4	5	6	7	8	9	10
Antes da campanha	68	88	101	82	96	74	65	74	52	99
Um ano após	67	87	90	76	98	69	68	65	59	70

Será que a campanha antipoluição reduziu de facto a poluição? Explicite e verifique todas as hipóteses necessárias à resolução do problema, justificando.

3.28. Para comparar dois tipos de máquina de ceifar (segadeiras) quanto à sua eficiência, foram seleccionadas ao acaso 9 searas tendo sido cada uma dividida em dois lotes. Em cada seara uma das segadeiras foi atribuída ao acaso a um dos lotes, ficando a outra para o outro lote.

A eficiência é avaliada num intervalo de valores de 0 (eficiência mínima) a 10 (eficência máxima). Os valores registados relativos à eficiência de cada segadeira foram introduzidos no *software* R. Responda às seguintes questões utilizando os resultados de alguns dos comandos apresentados abaixo.

a) Indique uma estimativa da eficiência média de cada uma das segadeiras.

b) Calculou-se um intervalo a 95% de confiança para a variância da eficiência da segadeira 1, tendo-se obtido $]0.5144; 4.1381[$. Qual o valor observado para a variância da segadeira 1?

c) Será que a segadeira 1 é mais eficiente do que a segadeira 2? Justifique convenientemente indicando e validando as condições necessárias à resolução.

```
> # seg1 designa os valores observados de eficiência da segadeira 1
> # seg2 designa os valores observados de eficiência da segadeira 2

> mean(seg1)   > mean(seg2)     > var(seg2)       > var(seg1-seg2)
[1] 7.4       [1] 6.488889     [1] 0.5861111      [1] 0.6761111


> shapiro.test(seg1)
        Shapiro-Wilk normality test
data:  seg1
W = 0.8848, p-value = 0.1762

> shapiro.test(seg2)
        Shapiro-Wilk normality test
data:  seg2
W = 0.9105, p-value = 0.3193
```

```
> shapiro.test(seg1-seg2)
        Shapiro-Wilk normality test
data:  seg1 - seg2
W = 0.9648, p-value = 0.8472


> t.test(seg1,seg2,paired=T,alternative="two.sided")
        Paired t-test
data:  seg1 and seg2
t = 3.3242, df = 8, p-value = 0.01047
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2790663 1.5431559
sample estimates:
mean of the differences
              0.9111111


> t.test(seg1,seg2,paired=T,alternative="greater")
        Paired t-test
data:  seg1 and seg2
t = 3.3242, df = 8, p-value = 0.005237
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.4014339       Inf
sample estimates:
mean of the differences
              0.9111111


> t.test(seg1,seg2,paired=T,alternative="less")
        Paired t-test
data:  seg1 and seg2
t = 3.3242, df = 8, p-value = 0.9948
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.420788
sample estimates:
mean of the differences
              0.9111111

> t.test(seg1,seg2,paired=F,alternative="two.sided")
        Welch Two Sample t-test
```

```
data:  seg1 and seg2
t = 2.088, df = 14.548, p-value = 0.05482
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02147036  1.84369258
sample estimates:
mean of x mean of y
 7.400000  6.488889


> t.test(seg1,seg2,paired=F,alternative="greater")
        Welch Two Sample t-test
data:  seg1 and seg2
t = 2.088, df = 14.548, p-value = 0.02741
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1446037       Inf
sample estimates:
mean of x mean of y
 7.400000  6.488889


> t.test(seg1,seg2,paired=F,alternative="less")
        Welch Two Sample t-test
data:  seg1 and seg2
t = 2.088, df = 14.548, p-value = 0.9726
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.677619
sample estimates:
mean of x mean of y
 7.400000  6.488889
```

3.29. Uma máquina de ensacar está regulada para encher sacos com 16 kg de açúcar. Para controlar o seu funcionamento escolheram-se, aleatoriamente, 15 sacos, que foram pesados. Os valores obtidos foram introduzidos no *software* R. Responda às seguintes questões utilizando, sempre que possível, os resultados dos comandos apresentados.

a) Indique uma estimativa para o peso médio e para a variância do peso de cada saco.

b) Foi calculado um intervalo de confiança para a variância do peso de cada saco, σ^2, tendo-se obtido $]0.01368\ ;\ 0.04931[$.

   i) Qual o estimador que foi utilizado na construção do intervalo?

ii) Determine o nível de confiança do intervalo dado.

c) Que conclusão pode tirar quanto à regulação da máquina? Justifique convenientemente, indicando e validando os pressupostos necessários à sua resposta.

```
> peso
 [1] 16.1 15.8 15.9 16.1 15.8 16.2 16.0 15.9 16.0 15.7 15.8 15.7 16.0 16.0 15.8

> mean(peso)                    > shapiro.test(peso)
[1] 15.92                                Shapiro-Wilk normality test
                                  data:  peso
> var(peso)                      W = 0.9377, p-value = 0.3544
[1] 0.02314286
```

3.30. Um enólogo pretende avaliar a acidez total de um vinho. Para isso selecciona aleatoriamente 20 garrafas de vinho na adega e analisa o seu conteúdo através do método clássico e de um dispositivo de titulação automática. Alguns resultados das análises, em g/l, foram:

garrafa	1	2	3	···	18	19	20
método clássico	4.8	3.4	2.5	···	2.9	5.4	2.1
titulação automática	6.1	5.1	2.1	···	4.5	3.9	1.5

Os dados foram introduzidos no *software* R. Abaixo apresentam-se resultados de comandos, alguns inadequados. Responda às seguintes questões utilizando os resultados apresentados abaixo.

a) De acordo com a legislação em vigor um vinho de mesa deverá ter uma acidez total superior a 3.5 g/l. Com base nos resultados das análises efectuadas pelo método clássico, o enólogo poderá concluir que o seu vinho cumpre os requisitos de acidez impostos pela legislação? Explicite e valide os pressupostos necessários à resolução do problema.

b) Com base nos valores obtidos poder-se-á concluir que os dois métodos de análise da acidez total do vinho têm resultados significativamente diferentes? Explicite e valide os pressupostos necessários à resolução do problema.

```
> classico
 [1] 4.8 3.4 2.5 3.8 4.3 3.6 3.5 3.5 4.0 3.6 6.3 3.0 3.1 3.7 2.8 5.1 4.0 2.9 5.4 2.1
> automatico
 [1] 6.1 5.1 2.1 4.9 6.6 4.0 4.5 1.0 4.7 3.5 8.2 3.9 3.6 4.5 3.5 6.3 4.7 4.5 3.9 1.5

> var(classico)       > var(automatico)       > var(classico-automatico)
[1] 1.040105           [1] 2.887868              [1] 1.326605
```

```
> shapiro.test(classico)
        Shapiro-Wilk normality test
data:  classico
W = 0.951, p-value = 0.3827
```

```
> shapiro.test(automatico)
        Shapiro-Wilk normality test
data:  automatico
W = 0.9625, p-value = 0.5959
```

```
> shapiro.test(classico-automatico)
        Shapiro-Wilk normality test
data:  classico - automatico
W = 0.9163, p-value = 0.08413


> t.test(classico,mu=3.5)
        One Sample t-test
data:  classico
t = 1.184, df = 19, p-value = 0.2510
alternative hypothesis: true mean is not equal to 3.5
95 percent confidence interval:
 3.292693 4.247307
sample estimates:
mean of x
     3.77

> t.test(classico,mu=3.5,alternative="greater")
        One Sample t-test
data:  classico
t = 1.184, df = 19, p-value = 0.1255
alternative hypothesis: true mean is greater than 3.5
95 percent confidence interval:
 3.375677      Inf
sample estimates:
mean of x
     3.77

> t.test(classico,mu=3.5,alternative="less")
        One Sample t-test
data:  classico
t = 1.184, df = 19, p-value = 0.8745
alternative hypothesis: true mean is less than 3.5
95 percent confidence interval:
     -Inf 4.164323
sample estimates:
mean of x
     3.77
```

```
> t.test(classico,automatico,paired=FALSE, var.equal=TRUE)
        Two Sample t-test
data:  classico and automatico
t = -1.32, df = 38, p-value = 0.1947
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.4821486  0.3121486
sample estimates:
mean of x mean of y
    3.770     4.355

> t.test(classico,automatico,paired=TRUE)
        Paired t-test
data:  classico and automatico
t = -2.2714, df = 19, p-value = 0.03493
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.12405128 -0.04594872
sample estimates:
mean of the differences
                 -0.585
```

3.31. Uma associação de defesa de consumidores pretende publicar um estudo comparativo dos preços de dois supermercados. Para isso seleccionou 20 produtos e registou os preços de cada um nos supermercados A e B (em cêntimos):

Produto	1	2	3	4	5	6	···	15	16	17	18	19	20
Super. A	202	201	560	253	384	332	···	624	851	742	501	476	765
Super. B	185	187	516	239	349	295	···	613	851	731	546	490	762

Utilize os resultados apresentados abaixo para responder às seguintes questões.

a) Determine uma estimativa para a diferença dos preços médios nos dois supermercados.

b) Com base nos dados recolhidos, o que pode afirmar relativamente aos preços médios praticados pelos dois supermercados? Explicite e valide as hipóteses necessárias à resolução do problema.

```
> A
 [1] 202 201 560 253 384 332 549 722 153 676 804 535 472 335 624 851 742 501 476 765
> B
 [1] 185 187 516 239 349 295 552 667 132 676 745 529 460 316 613 851 731 546 490 762

> mean(A)      > mean(B)          > var(A)           > var(B)           > var(A-B)
[1] 506.85     [1] 492.05         [1] 46116.77       [1] 46454.68       [1] 568.379


> shapiro.test(A)                                    > shapiro.test(B)
  Shapiro-Wilk normality test                         Shapiro-Wilk ormality test
data:  A                                             data:  B
W = 0.957, p-value = 0.4864                          W = 0.9552, p-value = 0.4521

> shapiro.test(A-B)
    Shapiro-Wilk normality test
data:  A - B
W = 0.9514, p-value = 0.3889
```

Exercícios de Revisão de Inferência Estatística

R3.1. Recolheu-se uma amostra aleatória de dimensão 5 de uma população normal. Determine a probabilidade de o desvio padrão da amostra ser inferior ao desvio padrão da população.

R3.2. Seja X uma variável aleatória com distribuição binomial, $X \cap B(n, p)$. Considere os estimadores para p assim definidos $\widehat{P_1} = \frac{X+1}{n+1}$ $\widehat{P_2} = \frac{X}{n+1}$

a) Determine, em função dos parâmetros da distribuição de X, o valor médio e a variância dos estimadores $\widehat{P}_1$ e $\widehat{P}_2$. Qual deles preferiria? Justifique.

b) Suponha que numa amostra de dimensão 20, se observou a ocorrência de 5 sucessos. Determine uma estimativa para p.

R3.3. A quantidade (em *ppm*) de um poluente no solo de uma certa região é uma v.a. X que se admite ter distribuição normal com valor médio μ e variância σ^2, desconhecidos.

Recolheu-se uma amostra de dimensão 10 tendo-se obtido o desvio padrão de 1.3943 *ppm*. Com base na amostra construiu-se o seguinte intervalo de confiança para μ

$$]22.447 \ ; \ 25.313[$$

a) Indique estimativas do valor médio e da variância da quantidade daquele poluente no solo.

b) Qual o grau de confiança daquele intervalo?

c) Construa um intervalo a 95% de confiança para σ^2.

R3.4. Num certo processo quimico é muito importante que uma dada solução tenha um pH de exactamente 8.20. O método utilizado na determinação do pH fornece medições que se admite terem distribuição normal de valor médio igual ao verdadeiro valor do pH da solução e desvio padrão de 0.02.

Para avaliar o pH de uma solução, efectuaram-se 10 medições independentes tendo-se obtido os seguintes valores:

8.18 8.16 8.17 8.22 8.19 8.17 8.15 8.21 8.16 8.18

a) Indique uma estimativa do valor médio do pH da solução.

b) Com base nestas 10 medições, o que pode concluir relativamente à utilização desta solução no referido processo químico?

c) Pretende-se efectuar um novo conjunto de medições para diminuir o erro máximo cometido na estimativa do verdadeiro valor do pH da solução. Mantendo-se todas as condições referidas acima, qual deverá ser o tamanho da amostra para que aquele erro máximo não exceda 0.01, a 95% de confiança?

R3.5. Faz-se uma experiência para saber se dois regimes alimentares A e B produzem o mesmo aumento de peso nos animais, durante um período de tempo fixado. Tomam-se 20 animais e de entre eles 10 ao acaso aos quais é dado o alimento A. Aos outros 10 é dado o alimento B. Os aumentos de peso (expressos em kg) no mesmo intervalo de tempo são os seguintes:

Regime A	-2.0	0.0	4.2	6.3	9.6	4.3	10.2	11.0	12.4	13.1
Regime B	4.0	6.0	8.0	11.3	12.3	14.4	14.5	14.7	14.7	16.0

Diga se existe diferença significativa entre os dois regimes alimentares, justificando convenientemente todas as hipóteses necessárias à resolução do problema.

R3.6. Um estudo pretende comparar um tipo de semente melhorada com o tipo de semente usado anteriormente. A nova semente passará a ser utilizada se, em média, o crescimento das plantas após 20 dias fôr superior ao das obtidas das 'velhas' sementes. São criadas 15 diferentes situações laboratoriais, variando temperatura e humidade. Em cada situação semeia-se uma semente de cada tipo e obtêm-se os seguintes resultados para o crescimento (em cm) das plantas após 20 dias :

Situação	1	2	3	4	5	6	7	8
Sementes melhoradas	3.46	3.48	2.74	2.83	4.00	4.95	2.24	6.92
sementes tradicionais	3.18	3.67	2.92	3.10	4.10	4.86	2.21	6.91

Situação	9	10	11	12	13	14	15
sementes melhoradas	6.57	6.18	8.30	3.44	4.47	7.59	3.87
sementes tradicionais	6.83	6.19	8.05	3.46	4.18	7.43	3.85

Deverá passar a usar-se as sementes melhoradas? Responda justificando e explicitando quaisquer hipóteses adicionais que seja necessário impôr.

R3.7. Num estudo sobre a incidência de certa doença numa população de insectos, um grupo de biólogos registou ao longo de um ano o número de insectos contaminados em cada amostra de 5 insectos, tendo para tal recolhido 100 amostras. Os resultados obtidos foram:

Num. de insectos contaminados	0	1	2	3	4	5
Num. de amostras	9	26	34	22	8	1

a) Construa uma tabela de frequências absolutas, relativas e relativas acumuladas para o número de insectos contaminados por amostra.

b) Determine a média e a mediana do número de insectos contaminados por amostra.

c) Considerando agora a totalidade de insectos observados como uma amostra aleatória da população de insectos:

   i) Determine uma estimativa da verdadeira proporção de insectos contaminados.

ii) Qual é o erro máximo associado à estimativa obtida na alínea anterior, com uma confiança de 95% ?

R3.8. O dono de uma ervanária produz um chá relativamente ao qual afirma que é 90% eficaz para curar dores de cabeça. Num inquérito feito a 250 pessoas, 198 concordaram que o chá cura as dores de cabeça. Acha que o resultado do inquérito é compatível com a pretensão do produtor ?

R3.9. Um investigador pretende estudar a incidência a nível nacional, de uma doença que ataca os pinheiros. Observações efectuadas através do país resultaram em 1233 casos de pinheiros afectados (a nível nacional) num total de 4250 observações.

a) Estime a percentagem de pinheiros afectados a nível nacional.

b) Determine um intervalo a 95% de confiança para a verdadeira proporção de pinheiros afectados.

R3.10. Um agricultor pretende experimentar dois tipos (I e II) de semeador de milho para comparar a produtividade das duas máquinas.

Para isso, num campo agrícola marcaram-se aleatoriamente 10 parcelas de igual área sendo cada uma delas dividida em duas secções iguais. Em cada parcela sorteou-se a atribuição de um dos tipos de semeador a uma das secções.

A produtividade (em unidades adequadas) de cada um dos semeadores foi registada, e os dados foram introduzidos no *software* R. Utilizando os resultados apresentados abaixo (alguns desnecessários), poderá o agricultor admitir que a produtividade esperada das duas máquinas é igual? Justifique convenientemente a sua resposta.

```
>x<-c(5.6, 8.4, 8.0, 8.6, 6.4, 7.8, 6.2, 7.7, 8.0, 7.7)
observações para o semeador tipo I.
>y<-c(6.0, 7.4, 7.3, 7.5, 6.4, 6.0, 5.5, 6.6, 5.6, 6.1)
observações para o semeador tipo II.


>mean(x)        >mean(y)         >mean(x-y)
[1] 7.44         [1] 6.44          [1] 1

>var(x)         >var(y)          >var(x-y)
[1] 1.018222     [1] 0.5448889     [1] 0.68


> shapiro.test(x)

        Shapiro-Wilk normality test
```

```
data:  x
W = 0.8731, p-value = 0.1087

> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.9026, p-value = 0.2341

> shapiro.test(x-y)

        Shapiro-Wilk normality test

data:  x - y
W = 0.9765, p-value = 0.9435
```

Soluções de alguns Exercícios

3.1. a) 0.2112

b) $\simeq 0.2112$

3.2. a) Como $n = 100$ é "grande", pelo Teorema Limite Central, verifica-se para a média amostral, $\overline{X} = \sum_{i=1}^{100} X_i/n$, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Longrightarrow \frac{\overline{X}-31}{0.24} \sim N(0,1)$. Logo, $P[\overline{X} < 30] \simeq \Phi(\frac{30-31}{0.24}) = \Phi(-4.17) = 1 - \Phi(4.17) = 1 - 1 = 0$

b) Seja $S_{100} = \sum_{i=1}^{100} X_i$ o peso total da amostra aleatória. Pelo Teorema Limite Central tem-se

$$\frac{S_{100} - 100\mu}{\sigma\sqrt{100}} \sim N(0,1) \Longrightarrow \frac{S_{100} - 3100}{24} \sim N(0,1).$$

Então, $P[S_{100} > 3150] = 1 - P[S_{100} \leq 3150] \simeq 1 - \Phi(\frac{3150-3100}{24}) = 1 - \Phi(2.08) = 1 - 0.9812 = 0.0188$.

3.3. a) $\overline{x} = 56.2$ cm e $s = 4$ cm

b) 0.42074

c) 4/5=0.8 cm.

3.4. b) p

d) $\hat{p}_1 = 1/3$ e $\hat{p}_2 = 1/2$

3.5. 1082

3.6. $\approx 1 - \Phi(0.4\sqrt{12n})$, para n elevado

3.7. ≈ 0.01267

3.10. a) $]77.27,\ 79.33[$

b) 1.03

c) 2654

3.11. 18.3 e $n = 12$

3.12. a) $\overline{x} = 0.959$ Kg e $s^2 = 0.0116$ Kg^2

c) $]0.00886,\ 0.0155[$

3.13. a) 99%

b) 0.185

3.14. a) i) 27 ii) 0.95

b) $]14.76,\ 64.77[$

3.18. a) Unidade estatística: planta de tabaco; variável: nº de folhas por planta

b) $\overline{x} = 2967/150 = 19.78$ e $\tilde{x} = 20$

3.20. a) $\overline{x}_\mathrm{S} = 100$, $\overline{x}_\mathrm{N} = 60$, $s^2_\mathrm{S} = 288.33$ e $s^2_\mathrm{N} = 207.89$

3.24. a) Verdadeira

b) Falsa

R3.1. 0.7127

R3.2. a) $E[\widehat{P_1}] = \frac{np+1}{n+1}$; $E[\widehat{P_2}] = \frac{np}{n+1}$;

$Var[\widehat{P_1}] = Var[\widehat{P_2}] = \frac{npq}{(n+1)^2}$.

$\widehat{P_2}$ é o que apresenta menor EQM portanto é o melhor.

b) Usando o melhor estimador, $\widehat{P_2}$, temos uma estimativa para p, $\hat{p}_2 = 5/21$.

R3.3. a) Estimativa do valor médio é $\overline{x} = 23.88\ ppm$ e de σ^2 é $s^2 = 1.9441\ ppm^2$

b) O grau de confiança é 99%.

R3.4. a) $\overline{x} = 8.179$

c) $n \geq 16$

R3.7. a)

x_i	Freq. Abs. n_i	Freq. Rel. f_i	Freq. Rel Acum. F_i
0	9	0.09	0.09
1	26	0.26	0.35
2	34	0.34	0.69
3	22	0.22	0.91
4	8	0.08	0.99
5	1	0.01	1.00

b) $\overline{x} = 197/100$ e $\tilde{x} = 2$

c) i) $\hat{p} = 197/500 = 0.394$ ii) 0.0428

R3.9. a) 29.01%

b) $]0.2765,\ 0.3037[$