INSTITUTO SUPERIOR DE AGRONOMIA

MODELOS MATEMÁTICOS E APLICAÇÕES – 2022-23

Resoluções dos Exercícios de Análise de Variância de Efeitos Fixos

1. (a) Trata-se dum delineamento a um único factor (as variedades de tomate), sendo a variável resposta Y a resistência da película (em gf). Em cada um dos $k=6$ níveis do factor há $n_c=3$ repetições (as parcelas). O número igual de repetições nas 6 situações experimentais significa que o delineamento é equilibrado. O modelo ANOVA a um factor corresponde a:

   i. A resistência Y_{ij}, na j-ésima parcela ($j=1,2,3$) associada à variedade i ($i=1,...,6$), é dada por:

   $$Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij} \ , \quad \forall\, i,j \ ,$$

   sendo μ_1 a resistência esperada da primeira variedade; $\alpha_i = \mu_i - \mu_1$ o efeito (acréscimo à resistência média da primeira variedade) da variedade i (com $\alpha_1 = 0$); e ϵ_{ij} o erro aleatório da observação Y_{ij}. Iremos (tal como o programa `R`) admitir que as variedades estão ordenadas por ordem alfabética, com os nomes de nível numéricos à cabeça, pelo que a primeira variedade acima referida é a variedade *18*.

   ii. Admite-se que os erros aleatórios são todos Normais, de média zero e variâncias homogéneas, ou seja, para qualquer i,j:

   $$\epsilon_{ij} \frown \mathcal{N}(0,\sigma^2) \ .$$

   iii. Admite-se que os erros aleatórios ϵ_{ij} são independentes.

   (b) A tabela-resumo terá apenas duas linhas (além da linha correspondente aos Totais), associadas respectivamente aos efeitos do Factor e à variabilidade Residual.

   i. Sabemos que os graus de liberdade dos efeitos do factor são $k-1=5$ e que os graus de liberdade residuais são $n-k=18-6=12$. As fórmulas para as Somas de Quadrados são dadas no formulário. A Soma de Quadrados Residual é $SQRE = \sum_{i=1}^{k}(n_i-1)s_i^2$ e, tratando-se dum delineamento equilibrado com $n_c = 3$ repetições em cada nível, tem-se $SQRE = (n_c-1)\sum_{i=1}^{k} s_i^2$. Usando as variâncias amostrais de nível dadas no enunciado, vem $SQRE = 2\times(14713.08 + 367.9434 + 5881.921 + 33132.64 + 5.414433 + 47.11163) = 108\,296.2$. É possível calcular SQF através da sua fórmula, uma vez que são disponibilizadas as médias amostrais de nível e globais. Mas é mais simples obter esse valor constatando que, numa ANOVA a um factor, se tem $SQF = SQT - SQRE$. No nosso caso $SQT = (n-1)s_y^2 = 17\times 34\,517.82 = 586\,802.9$. Logo, $SQF = 478\,506.7$. Dividindo estas Somas de Quadrados pelos graus de liberdade antes referidos obtêm-se os Quadrados Médios, e dividindo QMF por $QMRE$ obtem-se o valor calculado da estatística do teste F aos efeitos do factor. Eis a tabela-resumo:

	g.l.	SQs	Quadrados Médios	F_{calc}
Factor	5	478 506.7	$\frac{478\,506.7}{5} = 95\,701.35$	$F_{calc} = \frac{QMF}{QMRE} = \frac{95\,701.35}{9\,024.685} = 10.6044$
Residual	12	108 296.2	$\frac{108\,296.2}{12} = 9\,024.685$	–

   ii. Usando o `R`, confirmamos a tabela-resumo agora obtida:

```
> tomate.aov <- aov(res.pel ~ variedade , data=tomate)
> summary(tomate.aov)
             Df Sum Sq Mean Sq F value   Pr(>F)
```

```
variedade    5 478507   95701    10.6 0.000448
Residuals   12 108296    9025
```

(c) Eis o teste aos efeitos do factor (variedade):

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i$ vs. $H_1 : \exists \, i$ tal que $\alpha_i \neq 0$.

Estatística do Teste: $F = \frac{QMF}{QMRE} \ \frown \ F_{[k-1,n-k]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(5,12)} = 3.11$.

Conclusões: Como $F_{calc} = 10.6044 > 3.11$, rejeita-se H_0, concluindo-se que existem efeitos de variedade (ao nível $\alpha = 0.05$), o que corresponde a afirmar que existem variedades de tomate cujas películas têm resistência média diferentes de outras.

(d) O valor de prova (*p-value*) associado ao valor calculado da estatística de teste é $p = 0.000448$. Pela própria definição de *p-value*, esta é a área à direita de $F_{calc} = 10.6044$, numa distribuição $F_{[5,12]}$. Logo, seria preciso fazer um teste de hipóteses com nível de significância $\alpha = 0.000448$ (ou inferior) para que F_{calc} não pertencesse à Região Crítica e a conclusão do teste pudesse ser a de não rejeitar H_0.

(e) Tal como nas regressões lineares, a primeira coluna da matriz $\mathbf{X}$ é uma coluna de uns. No contexto duma ANOVA a um factor, as restantes colunas são variáveis indicatrizes de pertença de cada observação a um dos níveis do factor, ou seja, colunas com apenas dois valores: "1" associado a observações que pertencem ao nível do factor em causa, e "0" associado a observações associadas a outros níveis do factor. A restrição imposta no modelo ($\alpha_1 = 0$) implica que não há indicatriz do primeiro nível do factor, neste caso, o nível "18". Assim, neste caso teremos uma primeira coluna de $n = 18$ uns e cinco colunas indicatrizes dos segundo, terceiro, quarto, quinto e sexto níveis do factor ($\mathcal{I}_2$, $\mathcal{I}_3$, $\mathcal{I}_4$, $\mathcal{I}_5$ e $\mathcal{I}_6$), como se pode confirmar através do comando referido no enunciado:

```
> model.matrix(tomate.aov)
   (Intercept) variedade28 variedade29 variedade40C variedadeAce variedadeRoma
1            1           0           0           0            0             0
2            1           0           0           0            0             0
3            1           0           0           0            0             0
4            1           1           0           0            0             0
5            1           1           0           0            0             0
6            1           1           0           0            0             0
7            1           0           1           0            0             0
8            1           0           1           0            0             0
9            1           0           1           0            0             0
10           1           0           0           1            0             0
11           1           0           0           1            0             0
12           1           0           0           1            0             0
13           1           0           0           0            0             1
14           1           0           0           0            0             1
15           1           0           0           0            0             1
16           1           0           0           0            1             0
17           1           0           0           0            1             0
18           1           0           0           0            1             0
```

A ordem dos níveis do factor no `R` é, por omissão, a ordem alfabética dos nomes dos níveis. Mas essa pode não ser a ordem pela qual as observações surgem nas linhas da *data frame* com os dados. Neste exemplo, a variedade `Roma` surge como último nível (última coluna de $\mathbf{X}$), mas as observações dessa variedade não estão nas linhas finais da *data frame*, razão pela qual as duas colunas finais de $\mathbf{X}$ parecem 'trocadas'.

(f) Os valores ajustados $\hat{Y}_{ij}$, numa ANOVA a um factor, são as médias amostrais do nível a que cada observação pertence. Assim, tem-se:

```
> fitted(tomate.aov)
       1        2        3        4        5        6        7        8
560.6433 560.6433 560.6433 241.4833 241.4833 241.4833 290.9500 290.9500
       9       10       11       12       13       14       15       16
290.9500 705.7800 705.7800 705.7800 332.1067 332.1067 332.1067 377.2533
      17       18
377.2533 377.2533
```

Estas são as médias de variedade dadas no enunciado.

(g) Para comparar médias de variedade iremos utilizar o teste de Tukey. Sabemos que podemos considerar diferentes duas médias populacionais de nível, μ_i e $\mu_{i'}$, caso as respectivas médias amostrais de nível difiram mais do que o termo de comparação do teste de Tukey, ou seja, se $|\overline{y}_{i.} - \overline{y}_{i'.}| > q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}}$, onde $q_{\alpha(k,n-k)}$ indica o valor que deixa à sua direita uma região de probabilidade α, na distribuição de Tukey com parâmetros k e $n-k$. No nosso caso, $k=6$ e $n-k=12$, sendo, pelas tabelas da distribuição Tukey, $q_{0.05(6,12)}=4.75$. Como $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{9\,024.685}{3}} = 54.84732$, podemos decretar a diferença significativa entre a média amostral das resistências da variedade $40C$, $\overline{y}_{4.} = 705.8$ (que é a maior de todas), e a de qualquer outra variedade cuja média difira desta em mais de $4.75 \times 54.84732 = 260.5248$ *gf*. Ora, $705.8 - 260.5248 = 445.2752$, e apenas a variedade *18* não tem média amostral inferior a esse valor. Logo, podemos concluir que as resistências médias das variedades *28*, *29*, *Ace* e *Roma* são diferentes (inferiores) à resistência média da variedade *40C*.

```
> library(agricolae)
> tomate.aov <- aov(res.pel ~ variedade , data=tomate)
> HSD.test (tomate.aov, "variedade",console=TRUE)

Study: tomate.aov ~ "variedade"

HSD Test for res.pel

Mean Square Error:  9024.685

variedade,  means

      res.pel        std r    Min    Max
18   560.6433 121.297482 3 420.59 632.04
28   241.4833  19.181852 3 219.34 253.00
29   290.9500  76.693685 3 223.71 374.48
40C  705.7800 182.023736 3 503.51 856.39
Ace  377.2533   2.326893 3 375.18 379.77
Roma 332.1067   6.863791 3 324.82 338.45

Alpha: 0.05 ; DF Error: 12
Critical Value of Studentized Range: 4.750231

Minimun Significant Difference: 260.5375

Treatments with the same letter are not significantly different.

      res.pel groups
40C  705.7800      a
18   560.6433     ab
Ace  377.2533     bc
Roma 332.1067     bc
```

```
29   290.9500      c
28   241.4833      c
```

(h) O facto dos resíduos se encontrarem 'empilhados' em seis colunas é o reflexo natural do facto, referido na alínea anterior, de apenas haver seis diferentes valores ajustados nesta ANOVA: as seis médias amostrais de cada variedade, $\hat{y}_{ij} = \overline{y}_{i.}$ $(j = 1, 2, 3)$. Este facto ajuda a identificar as observações associadas aos resíduos de maior magnitude. Assim, por exemplo, o maior resíduo (em módulo) corresponde ao ponto no canto inferior direito. Por estar associado a uma média $\overline{y}_{i.}$ de aproximadamente 700, tem de corresponder à variedade `40C`. Por ser um resíduo negativo, tem de corresponder a uma observação com valor inferior à média dessa variedade, o que apenas acontece com a primeira das três observações desse nível. Assim, a observação a que corresponde o referido resíduo é a observação $y_{4,1} = 503.51$.

Embora o número de repetições em cada nível ($n_c = 3$) seja muito baixo, e portanto susceptível de gerar impressões enganadoras, o gráfico sugere alguma heterogeneidade nas variâncias de Y_{ij} em cada nível. Os valores das variâncias amostrais de nível indicam que há variedades com muito pouca variabilidade nas resistências observadas (como a *Ace*, com $s_5^2 = 5.414433$) e outras com uma variabilidade muito maior (como a *29*, com $s_3^2 = 5881.921$, mais de mil vezes maior).

2. (a) A variável resposta `if` é medida com base num delineamento experimental onde se cruzam dois factores: o factor `genótipo` (factor A) com $a = 6$ níveis (genótipos); e o factor `terreno` (factor B), também com $b = 6$ níveis (terrenos). Trata-se dum delineamento factorial, já que efectuaram-se observações com todas as 36 possíveis combinações genótipo/terreno e equilibrado, porque em cada uma dessas 36 células houve igual número ($n_{ij} = 1$) de observações. No entanto, como apenas foi feita uma observação em cada célula, não será possível ajustar um modelo ANOVA com efeitos de interacção. Assim, tem-se o seguinte modelo ANOVA a dois factores, sem interacção:

   i. Cada uma das $n = 36$ observações da variável resposta é representada por $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}$, $\forall\, i = 1, 2, ..., 6$, $j = 1, 2, ..., 6$, $k = 1$ (o índice k é dispensável porque não há repetições nas células), e onde
      - Y_{ij1} indica o índice de fertilidade potencial (variável `if`) para a (única) observação do genótipo i, no terreno j;
      - μ_{11} é o `if` populacional médio do genótipo 1, no terreno 1;
      - α_i indica o efeito do genótipo i, impondo-se a restrição $\alpha_1 = 0$;
      - β_j indica o efeito do terreno j, impondo-se a restrição $\beta_1 = 0$; e
      - ϵ_{ij1} indica o erro aleatório associado à observação Y_{ij1}.

   ii. $\epsilon_{ij1} \frown \mathcal{N}(0, \sigma^2)$, $\forall\, i, j$.

   iii. $\{\epsilon_{ij1}\}_{i,j}$ constituem um conjunto de variáveis aleatórias independentes.

   (b) Sabemos que os graus de liberdade associados aos efeitos de factor correspondem ao número de níveis do factor, menos um. Assim, no nosso caso, tem-se que os g.l. de factor genótipo são $a - 1 = 5$, e os do factor terreno são $b - 1 = 5$. Os graus de liberdade residuais podem ser calculados como o que falta para que a soma dê $n - 1 = 35$, ou seja, $n - (a + b - 1) = 25$, e assim se completa a primeira coluna da tabela. Tendo em conta que o Quadrado Médio Residual é, por definição, $QMRE = \frac{SQRE}{n-(a+b-1)}$, tem-se $SQRE = QMRE \times (n - (a + b - 1)) = 0.3660 \times 25 = 9.15$, e assim se completa a última linha da tabela. Os dois Quadrados Médios em falta (QMA e QMB) podem ser ambos calculados através do conhecimento dos valores calculados das duas estatística F, disponíveis na tabela. De facto, por definição,

$F_A = \frac{QMA}{QMRE}$, pelo que $QMA = F_A \times QMRE = 4.204 \times 0.3660 = 1.538664$. Por um raciocínio análogo, tem-se $QMB = F_B \times QMRE = 2.691 \times 0.3660 = 0.984906$, e assim se completa a penúltima coluna da tabela. Faltam apenas os valores das Somas de Quadrados associadas aos dois factores: SQA e SQB. Mas, por definição, tem-se $QMA = \frac{SQA}{g.l.(SQA)}$, pelo que $SQA = QMA \times (a-1) = 1.538664 \times 5 = 7.69332$. De forma inteiramente análoga, obtém-se o valor de SQB: $SQB = QMB \times (b-1) = 0.984906 \times 5 = 4.92453$. Resumindo, tem-se:

Variabilidade	g.l.	SQs	QMs	F
Genótipo (Factor A)	5	7.69332	1.538664	4.204
Terreno (Factor B)	5	4.92453	0.984906	2.691
Residual	25	9.15	0.3660	–

(c) Há dois tipos de efeitos previstos no modelo: os efeitos α_i associados ao factor A (genótipos) e os efeitos β_j associados ao factor B (terreno). Vamos efectuar os testes F correspondentes, começando pelo teste a eventuais efeitos de genótipo:

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2,3,4,5,6$ vs. $H_1 : \exists\, i = 2,3,4,5,6$ tal que $\alpha_i \neq 0$.

Estatística do teste: $F_A = \frac{QMA}{QMRE} \ \frown \ F_{(a-1\,,\,n-(a+b-1))}$, sob H_0.

Nível de significância: $\alpha = 0.01$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.01(5,25)} = 3.85$.

Conclusões: O valor da estatística do teste é dado no enunciado: $F_{A_{calc}} = 4.204$. É um valor significativo ao nível $\alpha = 0.01$ e rejeita-se H_0 a favor da hipótese de que existem efeitos de genótipo.

Agora o teste a efeitos de terreno:

Hipóteses: $H_0 : \beta_j = 0 \ , \ \forall \ j$ vs. $H_1 : \exists\, j$ tal que $\beta_j \neq 0$.

Estatística do teste: $F_B = \frac{QMB}{QMRE} \ \frown \ F_{(b-1\,,\,n-(a+b-1))}$, sob H_0.

Nível de significância: $\alpha = 0.01$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.01(5,25)} = 3.85$.

Conclusões: O valor da estatística do teste é dado no enunciado: $F_{B_{calc}} = 2.691$. É um valor não significativo ao nível $\alpha = 0.01$ e não se rejeita H_0, pelo que não há efeitos significativos de terreno.

(d) O enunciado pede para considerar o que aconteceria se, aos mesmos dados, fosse ajustado um modelo ANOVA com um único factor, o factor `genotipo`. O pedido corresponde a ignorar a existência do factor terreno (embora ele tenha sido considerado no delineamento experimental que foi efectivamente usado), tratando-se as seis observações de cada genótipo como meras repetições. Nesse caso, e como se viu nas aulas teóricas, a tabela ANOVA terá apenas duas linhas: uma correspondente ao único factor agora considerado (genótipo) e outra residual. A linha da tabela correspondente ao factor genótipo permanece inalterada quanto a graus de liberdade (na notação dos modelos a um factor tem-se $k = a = 6$, logo continua a ter-se $a - 1 = 5$ g.l. associados aos genótipos); Soma de Quadrados ($SQA = SQF = n_c \sum_{i=1}^{6} (\overline{y}_{i.} - \overline{y}_{..})^2$); e (por conseguinte) Quadrado Médio ($QMA = \frac{SQA}{a-1}$). Já quanto à nova Soma de Quadrados Residual, tem de corresponder à soma das antigas parcelas SQB e $SQRE_{2f}$ no modelo a dois factores, sem interacção, ajustado inicialmente. De facto, e como se viu nas aulas teóricas, a Soma de Quadrados Total não depende do modelo ajustado, mas apenas dos valores de Y observados ($SQT = (n-1)\, s_y^2$). No modelo a dois factores, sem interacção, essa Soma de Quadrados foi decomposta como $SQT = SQA + SQB + SQRE_{2f}$. A mesma Soma de quadrados é agora decomposta como $SQT = SQA + SQRE_{1f}$. Sendo igual o total (SQT) e a primeira parcela em cada decomposição (SQA), necessariamente se

tem $SQRE_{1f} = SQB + SQRE_{2f}$. Logo, $SQRE_{1f} = 4.92453 + 9.15 = 14.07453$. Assim, o novo Quadrado Médio Residual é $QMRE_{1f} = \frac{SQRE_{1f}}{n-a} = \frac{14.07453}{30} = 0.469151$. O valor da (única) estatística F existente no modelo a um factor será agora: $F = \frac{QMA}{QMRE_{1f}} = \frac{1.538664}{0.469151} = 3.279678$. Assim, a tabela do modelo a um único factor será:

Variabilidade	g.l.	SQs	QMs	F
Genótipo	5	7.69332	1.538664	3.279678
Residual	30	14.07453	0.469151	–

O valor calculado da estatística F terá agora de ser comparado com a fronteira duma região crítica unilateral direita numa distribuição $F_{(5,30)}$. Ao nível de significância $\alpha = 0.01$, essa fronteira será $f_{0.01(5,30)} = 3.70$. Assim, os efeitos de genótipo já não são significativos, ao nível $\alpha = 0.01$.

(e) A hipótese cujo estudo se pede é a hipótese de existirem *efeitos de interacção* entre genótipos e terrenos. Trata-se efectivamente duma hipótese possível (que seria um caso particular duma interacção genótipo × ambiente). Mas não é possível ajustar um modelo que preveja essa possibilidade (o modelo a dois factores *com* interacção) pois, como já se referiu, não existem repetições nas células.

3. (a) Trata-se dum delineamento factorial a dois factores, sendo a variável resposta Y a altura aos dois anos (em cm) dos pinheiros; o primeiro factor (A) a proveniência, com $a = 5$ níveis e o segundo factor (B) o local do ensaio (com $b = 2$ níveis). O delineamento é equilibrado, uma vez que em cada uma das $ab = 10$ células (situações experimentais) existem $n_c = 6$ observações, num total de $n = n_c ab = 60$ observações. Existem repetições nas células, logo é possível (e desejável) estudar a existência de eventuais efeitos de interacção.

O modelo ajustado é o modelo ANOVA a dois factores, com efeitos de interacção. Admite-se que os níveis de cada factor estão ordenados por ordem alfabética (que corresponde à ordem em que aparecem no enunciado). Eis o modelo:

- $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, para qualquer $i = 1, 2, 3, 4, 5$, $j = 1, 2$ e $k = 1, 2, 3, 4, 5, 6$, sendo μ_{11} a altura esperada (aos dois anos) dos pinheiros gregos em Sines; α_i o efeito principal (acréscimo à altura) associado à proveniência i (com a restrição $\alpha_1 = 0$); β_j o efeito principal (acréscimo à altura) associado a $j = 2$ (dada a restrição $\beta_1 = 0$); $(\alpha\beta)_{ij}$ o efeito de interacção, isto é, o acréscimo na altura específico da combinação da proveniência i com o local j. Dadas as restrições $(\alpha\beta)_{ij} = 0$ se $i = 1$ e/ou $j = 1$, o modelo apenas prevê efeitos de interacção nas situações experimentais correspondentes a Tavira ($j = 2$) e para proveniências diferentes da Grécia ($i > 1$). Finalmente ϵ_{ijk} é o erro aleatório da observação Y_{ijk}.
- Admite-se que os erros aleatórios são todos Normais, de média zero e variâncias homogéneas: $\epsilon_{ijk} \frown \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k.
- Admite-se que os erros aleatórios ϵ_{ijk} são independentes.

(b) Tratando-se dum modelo ANOVA factorial, a dois factores com interacção, a tabela-resumo terá de ter quatro linhas, correspondentes aos três tipos de efeitos previstos (principal de cada factor e de interacção), bem como à variabilidade residual e, opcionalmente, uma quinta linha associada à variabilidade total. A tabela terá as habituais colunas de graus de liberdade, Somas de Quadrados, Quadrados Médios e valor das estatísticas F. Vejamos como se pode preencher esta tabela.

Sabemos que, neste tipo de modelo, os graus de liberdade associados a $QMRE$ são dados por $n - ab$, onde $n = 60$ é o número total de observações e $ab = 10$ é o número de parâmetros

existentes no modelo. Assim, $g.l.(SQRE)=50$. Sabemos ainda que, para os vários tipos de efeitos, os graus de liberdade são dados pelo número de parcelas de cada tipo de efeito, após a introdução das restrições, ou seja, associado a SQA há $a-1=4$ g.l., associado a SQB há $b-1=1$ g.l., e associado a $SQAB$ há $(a-1)(b-1)=4$ graus de liberdade.

No enunciado é dada a Soma de Quadrados associada ao que foi designado factor A, tendo-se $SQA = 280.61$, donde se conclui que $QMA = \frac{SQA}{a-1} = \frac{280.61}{4} = 70.1525$. No enunciado é também dado o Quadrado Médio Residual, tendo-se $QMRE = 16.59$, donde $SQRE = QMRE \times (n-ab) = 16.59 \times 50 = 829.50$. Ora, sabemos pelo formulário que:

$$\begin{aligned} SQB &= a\, n_c \sum_{j=1}^{2} (\overline{y}_{.j.} - \overline{y}_{...})^2 \\ &= 5 \times 6 \times [(28.14 - 31.76298)^2 + (35.38 - 31.76298)^2] = 786.2645 \ . \end{aligned}$$

Donde $QMB = \frac{SQB}{b-1} = 786.2645$. O enunciado refere ainda a variância da totalidade das 60 observações, $s_y^2 = 34.49584$, donde se pode concluir que a Soma de Quadrados Total é $SQT = (n-1)\, s_y^2 = 59 \times 34.49584 = 2035.255$. Uma vez que sabemos que esta Soma de Quadrados Total se pode decompor como $SQT = SQA + SQB + SQAB + SQRE$, torna-se possível calcular $SQAB = SQT - (SQA + SQB + SQRE) = 2035.255 - (280.61 + 786.2645 + 829.50) = 138.8801$. Assim, o Quadrado Médio associado à interacção é dado por $QMAB = \frac{SQAB}{(a-1)(b-1)} = \frac{138.8801}{4} = 34.7200$.

Finalmente, os valores das estatísticas F são dados, para os três tipos de efeitos, pela razão entre o Quadrado Médio do referido tipo de efeito e $QMRE$. A tabela completa fica assim:

	g.l.	Soma de Quadrados	Quadrado Médio	F
Proveniência	4	280.61	70.1525	4.229
Local	1	786.2645	786.2645	47.394
Interacção	4	138.8801	34.7200	2.093
Residual	50	829.50	16.59	–

(c) Vai-se efectuar em pormenor o teste aos efeitos principais do Factor A (proveniência dos pinheiros), e descrever sinteticamente os testes aos efeitos principais do Factor B (local) e aos efeitos de interacção.

Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall\, i$ vs. $H_1 : \exists\, i$ tal que $\alpha_i \neq 0$.

Estatística do Teste: $F_A = \frac{QMA}{QMRE} \underset{\frown}{} F_{[a-1,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(4,50)} \approx 2.57$ (entre os valores tabelados 2.53 e 2.61).

Conclusões: Como $F_{calc} = \frac{QMA}{QMRE} = 4.229 > 2.57$, rejeita-se H_0, sendo possível concluir pela existência de efeitos principais de proveniência (ao nível $\alpha = 0.05$).

No teste aos efeitos principais do factor local do estudo, as hipóteses do teste podem ser escritas apenas como $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, uma vez que após a imposição da restrição $\beta_1 = 0$, apenas sobra um efeito deste tipo, o efeito β_2 associado a Tavira. O valor calculado da estatística de teste é muito grande ($F_{calc} = 47.394$) deixando antever a rejeição de H_0, facto que é confirmado determinando nas tabelas o limiar da região crítica unilateral direita: $f_{0.05(1,50)} \approx 4.04$ (entre os valores tabelados 4.00 e 4.08). Assim, conclui-se claramente pela existência de efeitos principais de localidade, o que neste caso significa que existe um efeito

associado à passagem do local de plantação de Sines para Tavira. Uma rápida inspecção das médias de local sugere que se trata dum maior crescimento dos pinheiros em Tavira, pelo que se deduz que β_2 terá um valor positivo.

No teste aos efeitos de interacção, com hipóteses $H_0 : (\alpha\beta)_{ij} = 0$, para todo o i e j, contra a hipótese alternativa de que existe pelo menos uma célula (i, j) onde $(\alpha\beta)_{ij} \neq 0$, o valor calculado da estatística de teste é $F_{calc} = 2.093$, inferior ao limiar da região crítica, que é (por coincidência) igual ao do teste aos efeitos do factor A, $f_{0.05(4,50)} \approx 2.57$. Logo, não se rejeita H_0 (para $\alpha = 0.05$), e conclui-se pela inexistência de efeitos significativos de interacção.

(d) Nesta alínea é pedido para verificar se o facto da maior altura média amostral de Sines (31.16, para pinheiros provenientes de Marrocos) ser menor que a mais baixa altura média amostral em Tavira (33.56, para pinheiros da segunda proveniência italiana) é uma relação que se possa estender à população. Vamos responder efectuando, como solicitado no enunciado, um teste de Tukey, e usando $\alpha = 0.05$. Ora, o termo de comparação é (como indicado no formulário e usando as tabelas da distribuição de Tukey):

$$q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(10,50)} \sqrt{\frac{16.59}{6}} = 4.68 \times 1.662829 = 7.782039 \ .$$

Ora, a diferença entre as médias amostrais das duas células referidas acima é apenas $|31.16 - 33.56| = 2.40$, logo inferior ao termo de comparação, pelo que não é uma diferença significativa (ao nível $\alpha = 0.05$). Assim, não é possível afirmar que as médias populacionais em Tavira sejam sempre maiores às de Sines, independentemente das proveniências. Alguns pares de médias populacionais podem ser consideradas diferentes (por exemplo, o crescimento médio dos pinheiros gregos em Sines e em Tavira), mas será preciso levar em conta as proveniências, e não apenas o local da realização do estudo.

4. (a) Trata-se dum delineamento factorial a dois factores: *localidade* (Factor A, com $a = 4$ níveis) e *cultivar* (Factor B, com $b = 9$ níveis). Existem $n_{ij} = 4 = n_c$ repetições em todas as $ab = 36$ situações experimentais (células), pelo que se trata dum delineamento equilibrado. Existem ao todo $n = abn_c = 144$ observações da variável resposta Y (rendimento, em kg/ha). O modelo ANOVA adequado é o modelo ANOVA a dois factores, com interacção, dado por:

i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \ \forall\, i = 1, 2, 3, 4 \ , \ j = 1, 2, ..., 9 \ , \ k = 1, 2, 3, 4,$
com $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i, onde
   - Y_{ijk} indica o rendimento na k-ésima parcela da localidade i, associada à cultivar j;
   - μ_{11} indica o rendimento médio (populacional) da cultivar *Celta*, em Elvas;
   - α_i indica o efeito principal da localidade i;
   - β_j indica o efeito principal da cultivar j;
   - $(\alpha\beta)_{ij}$ indica o efeito de interacção entre a localidade i e a cultivar j; e
   - ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk}.

ii. $\epsilon_{ijk} \frown \mathcal{N}(0, \sigma^2), \ \forall\, i, j, k.$

iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constitui um conjunto de variáveis aleatórias independentes.

(b) i. Os nove valores em falta na tabela são dados por:
   - $g.l.(SQA) = a - 1 = 3$;
   - $g.l.(SQB) = b - 1 = 8$;
   - $g.l.(SQAB) = (a - 1)(b - 1) = 3 \times 8 = 24$;

- $g.l.(SQRE) = n - ab = 144 - 36 = 108$;
- $SQB = QMB\,(b-1) = 964\,060 \times 8 = 7\,712\,480$;
- $SQAB = SQT - (SQA + SQB + SQRE) = (n-1)\,s_y^2 - 219\,628\,472 =$
  $= 143 \times 1\,714\,242 - 219\,628\,472 = 25\,508\,134$;
- $QMA = \frac{SQA}{a-1} = \frac{183\,759\,916}{3} = 61\,253\,305$;
- $QMAB = \frac{SQAB}{(a-1)(b-1)} = \frac{25\,508\,134}{24} = 1\,062\,839$;
- $F_B = \frac{QMB}{QMRE} = \frac{964\,060}{260\,704} = 3.69791$.

ii. Em qualquer modelo linear (regressão ou ANOVA), a variância dos erros aleatórios do modelo ($V[\epsilon_i] = \sigma^2$) é estimado pelo Quadrado Médio Residual. No nosso caso, a estimativa de σ^2 é dada no enunciado: $QMRE = 260\,704$. O valor muito elevado nada indica de especial, uma vez que a sua interpretação tem de levar em conta as unidades de medida dos dados, que são $(kg\,ha^{-1})^2$. De facto sabemos pelo enunciado que as unidades de medida da variável resposta são kg/ha. Sabemos que os resíduos ($e_i = y_i - \hat{y}_i$) têm as mesmas unidades de medida que a variável resposta. Sabemos que o QMRE é a Soma de Quadrados dos Resíduos a dividir pelos graus de liberdade associados, pelo que as unidades de medida do QMRE são o quadrado das unidades de medida da variável resposta. Bastava que os valores da variável resposta tivessem sido medidos em toneladas por hectare, para que o Quadrado Médio Residual viesse em $(t\ ha^{-1})^2$, ou seja, que fosse um milhão de vezes inferior ao valor acima indicado: $QMRE = 0.260704$. Mas isso não altera os dados, nem a significância de cada tipo de efeitos previsto no modelo. Assim, não é possível avaliar a estimativa de σ^2 apenas olhando para o valor absoluto de $QMRE$: é essencial ter em conta as unidades de medida associadas.

iii. Pedem-se os três testes F para cada tipo de efeitos previstos no modelo. Efectuemos em pormenor o teste à existência de efeitos de interacção entre localidade e cultivar:

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0\ ,\ \forall\ i = 2,3,4$ e $j = 2,3,...,9$ [não há interacção]
vs. $H_1 : \exists\, i = 2,3,4\ ,\ j = 2,3,...,9$ tais que $(\alpha\beta)_{ij} \neq 0$ [há interacção].

Estatística do teste: $F = \frac{QMAB}{QMRE} \underset{\frown}{} F_{[(a-1)(b-1)\,,\,n-ab]}$, sob H_0.

Nível de significância: $\alpha = 0.01$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.01(24,108)} \approx 1.97$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 4.0768$. É um valor significativo ao nível $\alpha = 0.01$, rejeitando-se H_0 a favor da hipótese alternativa de que existem efeitos de interacção entre localidade e cultivar.

No que respeita ao teste para os efeitos principais do factor *localidade*, as hipóteses em confronto são $H_0 : \alpha_i = 0\ ,\ \forall\ i = 2,3,4$ vs. $H_1 : \exists\, i = 2,3,4\ ,$ tal que $\alpha_i \neq 0$. A Região Crítica é agora dada pela rejeição de H_0 caso $F_{calc} > f_{0.01\,(3,108)} \approx 3.97$. O valor elevadíssimo da estatística calculada $F_{calc} = 234.9531$ leva à rejeição clara de H_0, concluindo-se pela existência de importantes efeitos de localidade, nos rendimentos.

Finalmente, no teste aos efeitos principais do factor *cultivar*, as hipóteses em confronto são $H_0 : \beta_j = 0\ ,\ \forall\ j = 2,3,...,9$ vs. $H_1 : \exists\, j = 2,3,...,9\ ,$ tal que $\beta_j \neq 0$. A Região Crítica é agora dada pela rejeição de H_0 caso $F_{calc} > f_{0.01\,(8,108)} \approx 2.68$. O valor da estatística calculada $F_{calc} = 3.698$ pertence à Região Crítica, levando à rejeição de H_0, concluindo-se também pela existência de efeitos de cultivar sobre os rendimentos.

Assim, conclui-se pela existência dos três tipos de efeitos, ao nível $\alpha = 0.01$, com destaque para a existência clara de efeitos de localidade.

iv. Os dois gráficos de interacção reflectem a mesma informação, embora de formas diferentes. No gráfico da esquerda, as quatro localidades definem posições no eixo horizontal. Por cima de cada localidade encontram-se nove pontos, associados às nove cultivares. A ordenada de cada um desses nove pontos é dada pelo rendimento médio das parcelas correspondentes a essa combinação de localidade e cultivar. Os segmentos de recta unem os pontos correspondentes a cada cultivar (segundo a legenda indicada no gráfico). Embora haja algum paralelismo nas nove curvas seccionalmente lineares, para as três primeiras localidades, os rendimentos na Revilheira sugerem a existência de efeitos de interacção. Por exemplo, a cultivar *TE9110* que, quer em Évora, quer em Elvas, regista o rendimento mais baixo (facto que se pode confirmar na tabela de médias dada na alínea c) tem o segundo mais elevado rendimento na Revilheira. Também a cultivar *Celta*, cujo rendimento em Benavila é o terceiro mais baixo, regista o segundo maior rendimento em Elvas. Assim, há cultivares que manifestam "preferências" ou "aversões" por diferentes localidades, reflectindo efeitos de interacção. O teste à interacção efectuado na alínea anterior confirma que esses efeitos são significativos, ao nível $\alpha = 0.01$.
O gráfico da direita dá, como se disse, uma perspectiva diferente sobre a mesma informação. Agora, são as cultivares que definem nove posições no eixo horizontal. Por cima de cada uma dessas posições (cultivares) há quatro pontos, com ordenadas dadas pelos rendimentos médios da referida cultivar, nas quatro localidades consideradas no ensaio. Segmentos de recta unem os pontos correspondentes a uma mesma localidade. Neste gráfico torna-se evidente que os rendimentos são sempre bastante superiores em Elvas (no gráfico da esquerda, esse facto reflectia-se no "pico" por cima de Elvas). Essa será a principal razão pela clara rejeição da hipótese nula no teste à existência de efeitos principais de localidade. Por outro lado, os efeitos de interacção reflectem-se na mais visível ausência de paralelismo, nomeadamente nos traços correspondentes a Évora e Revilheira, que para várias cultivares parecem ter comportamentos quase antagónicos.

v. Pede-se para discutir o efeito sobre a tabela resultante de dividir a variável resposta por mil (passando o rendimento a ser expresso em t/ha). Os graus de liberdade não são, naturalmente, afectados. O mesmo não se passa com as Somas de Quadrados. À nova variável $Y^* = Y/1000$ corresponderão novas médias de nível, de célula e global, que também resultam de dividir por mil (para ficarem em t/ha). Tendo em conta que no modelo em questão, as médias de célula definem os valores ajustados, tem-se $\hat{Y}^*_{ijk} = \hat{Y}_{ijk}/1000$. Assim, as novas Somas de Quadrados resultam de dividir as suas congéneres originais por 1000^2, ou seja, por um milhão. De facto, $SQT^* = \sum_i \sum_j \sum_k (Y^*_{ijk} - \overline{Y}^*_{...})^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \overline{Y}_{...}/1000)^2 = SQT/(1000^2)$. Também $SQRE^* = \sum_i \sum_j \sum_k (Y^*_{ijk} - \hat{Y}^*_{ijk})^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \hat{Y}_{ijk}/1000)^2 = SQRE/(1000^2)$. De forma análoga, e utilizando as fórmulas para delineamentos equilibrados,

$$SQA^* = bn_c \sum_{i=1}^{a} (\overline{Y}^*_{i..} - \overline{Y}^*_{...})^2 = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..}/1000 - \overline{Y}_{...}/1000)^2 = SQA/(1000^2)$$

$$SQB^* = an_c \sum_{j=1}^{b} (\overline{Y}^*_{.j.} - \overline{Y}^*_{...})^2 = an_c \sum_{j=1}^{b} (\overline{Y}_{.j.}/1000 - \overline{Y}_{...}/1000)^2 = SQB/(1000^2) \; .$$

Por diferença, tem igualmente de verificar-se $SQAB^* = SQAB/(1000^2)$. Assim, toda

a coluna de Somas de Quadrados na tabela será dividida por um milhão. Essa mesma transformação aplica-se à coluna de Quadrados Médios (que resulta de dividir Somas de Quadrados por graus de liberdade). Mas na coluna final, correspondente aos valores calculados das estatísticas F, o quociente de Quadrados Médios mantém-se inalterado (a transformação multiplicativa de numerador e denominador é igual). Logo, as conclusões de todos os testes (incluíndo os respectivos *p-values*) mantêm-se inalterados.

(c) O melhor rendimento observado em Elvas é o da cultivar *Trovador* ($\overline{y}_{29\,\cdot} = 5927 kg/ha$). Pede-se para usar o teste de Tukey a fim de verificar quais as cultivares cujo rendimento em Elvas não é significativamente diferente deste, ao nível $\alpha = 0.10$. O termo de comparação do teste de Tukey é, neste caso, (e utilizando o R para obter o valor da distribuição de Tukey),

$$q_{\alpha(ab,n-ab)} \sqrt{\frac{QMRE}{n_c}} \;=\; q_{0.10\,(36,108)} \sqrt{\frac{260704}{4}} \;=\; 5.24655 \times 255.2959 \;=\; 1339.423 \;.$$

Assim, os rendimentos médios considerados significativamente diferentes do da cultivar Trovador em Elvas serão os inferiores a $5927 - 1339.4 = 4587.6$. Em Elvas, apenas a cultivar *TE9110* está nessa situação. Todas as restantes têm rendimentos médios que não diferem significativamente do da cultivar *Trovador*. Este resultado reflecte a variabilidade elevada, expressa pelo $QMRE$.