
INSTITUTO SUPERIOR DE AGRONOMIA
ESTATÍSTICA E DELINEAMENTO – 2011/12
Algumas resoluções de Exercícios de Análise de Variância

4. A descrição do delineamento feita no enunciado geral aponta para um delineamento a um factor (*Grupo*, ou seja, momento do processamento em que se efectuavam as medições), com $k = 4$ níveis. O delineamento é equilibrado uma vez que há igual número ($n_i = 9 = n_c$) de observações em cada nível do factor, perfazendo um total de $n = k n_c = 36$ observações. A variável resposta Y é, neste caso, o conteúdo de zinco. Embora não pedido no enunciado, o modelo ANOVA associado a este delineamento é o seguinte:

- (i) $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall i = 1, 2, 3, 4$, $j = 1, 2, \dots, 9$, com $\alpha_1 = 0$, onde
- Y_{ij} indica a concentração de zinco no j -ésimo lote de feijão do Grupo (passo do processamento) i ;
 - μ_1 indica o conteúdo médio (populacional) de zinco antes de qualquer tratamento (primeiro Grupo);
 - α_i indica o efeito principal do Grupo (passo do processamento) i ; e
 - ϵ_{ij} indica o erro aleatório associado à observação Y_{ij} .
- (ii) $\epsilon_{ij} \cap \mathcal{N}(0, \sigma^2)$, $\forall i, j$.
- (iii) $\{\epsilon_{ij}\}_{i,j}$ constituem um conjunto de variáveis aleatórias independentes.

(a) Vamos construir a tabela com o auxílio do R, uma vez que os dados estão disponíveis na *data frame* zinco, com os valores da variável resposta na coluna `concentracao` e os grupos no factor `grupo` (alternativamente, podem sempre usar-se as fórmulas disponíveis no formulário para *SQF* e *SQRE* em delineamentos a um factor, sabendo-se também que os graus de liberdade associados ao Factor são $k - 1 = 3$ e os residuais $n - k = 32$):

```
> summary(aov(concentracao ~ grupo, data=zinco))
              Df Sum Sq Mean Sq F value    Pr(>F)
grupo          3 20.597   6.8656   6.3343 0.001703 **
Residuals     32 34.684   1.0839
```

(b) O teste F desta ANOVA diz respeito à possível existência de efeitos do Factor, ou seja,

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

Estatística do teste: $F = \frac{QMF}{QMRE} \cap F_{(k-1, n-k)}$, sob H_0 .

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,32)} \approx 2.90$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 6.3343$.

É um valor significativo ao nível $\alpha = 0.05$ e rejeita-se H_0 a favor da hipótese de que existem efeitos do Factor, ou seja, que as concentrações médias de zinco não são iguais em todos os passos do tratamento.

(c) Pedem-se para comparar as médias amostrais de grupos, a fim de determinar quais as que são significativas, ou seja, que levam a concluir que as correspondentes médias populacionais de grupos são diferentes. Vamos responder utilizando intervalos de confiança de Tukey (ver acetatos 302 e 305 das aulas teóricas), que serão construídos com o auxílio do R:

```
> TukeyHSD(aov(concentracao ~ grupo, data=zinco))
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = concentracao ~ grupo, data = zinco)
```

```
$grupo
      diff      lwr      upr      p adj
2-1  0.61888889 -0.71080385  1.9485816  0.5936756
3-1  0.00555556 -1.32413719  1.3352483  0.9999995
4-1  1.85555556  0.52586281  3.1852483  0.0034379
3-2 -0.61333333 -1.94302608  0.7163594  0.6006631
4-2  1.23666667 -0.09302608  2.5663594  0.0757933
4-3  1.85000000  0.52030726  3.1796927  0.0035442
```

O intervalo a 95% de confiança para a diferenças das médias do segundo e primeiro níveis, $\mu_2 - \mu_1$ é] -0.71080385, 1.9485816 [. Este intervalo inclui o valor zero, que é assim um valor admissível para $\mu_2 - \mu_1$. Logo, não há razões para concluir que antes do branqueamento (Grupo 2, correspondente ao segundo passo do processamento) o teor médio de zinco seja diferente do que é antes do início do processamento (Grupo 1). Da mesma forma, o intervalo a 95% de confiança para $\mu_3 - \mu_1$ é (arredondando os extremos a três casas decimais)] -1.324, 1.335 [. Também neste caso, admite-se $\mu_3 = \mu_1$. Já no que respeita à comparação entre as médias dos Grupos mais extremos (μ_1 e μ_4), o intervalo a 95% de confiança é] 0.526, 3.185 [. Assim, o valor zero não é um valor admissível (a 95% de confiança) para $\mu_4 - \mu_1$, pelo que concluímos que $\mu_4 \neq \mu_1$. Analisando os restantes intervalos de confiança, resulta que μ_3 e μ_2 podem ser considerados iguais, tal como μ_4 e μ_2 (por pouco), mas conclui-se que $\mu_4 \neq \mu_3$ (a 95% de confiança), uma vez que o intervalo de confiança para $\mu_4 - \mu_3$ não contém o valor zero. Estas conclusões (que não permitem uma arrumação das médias de grupo em compartimentos estanques) são por vezes sintetizadas ordenando as médias amostrais de grupo por ordem crescente e sublinhando subconjuntos de médias cujas médias não diferem significativamente. No nosso caso:

```
> model.tables(aov(concentracao ~ grupo, data=zinco), type="means")
```

```
Tables of means
```

```
Grand mean
```

```
2.847778
```

```
grupo
```

```
grupo
```

```
      1      2      3      4
2.228 2.847 2.233 4.083
```

Assim, ordenando as médias e tendo em conta as conclusões acima referidas, teríamos:

$\bar{y}_1.$	$\bar{y}_3.$	$\bar{y}_2.$	$\bar{y}_4.$
2.228	2.233	2.847	4.083

Uma forma alternativa de representar as conclusões consiste em utilizar letras iguais para indicar os subconjuntos de médias que não diferem significativamente. No nosso caso, poderíamos escrever:

$\bar{y}_1.$	$\bar{y}_3.$	$\bar{y}_2.$	$\bar{y}_4.$
2.228 ^a	2.233 ^a	2.847 ^{ab}	4.083 ^b

Alternativamente, é possível efectuar o teste de Tukey. O termo de comparação para qualquer diferença de médias amostrais é, neste caso (com $\alpha = 0.05$):

$$q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(4,32)} \sqrt{\frac{1.0839}{9}} = 3.83 \times 0.3470351 = 1.329144 .$$

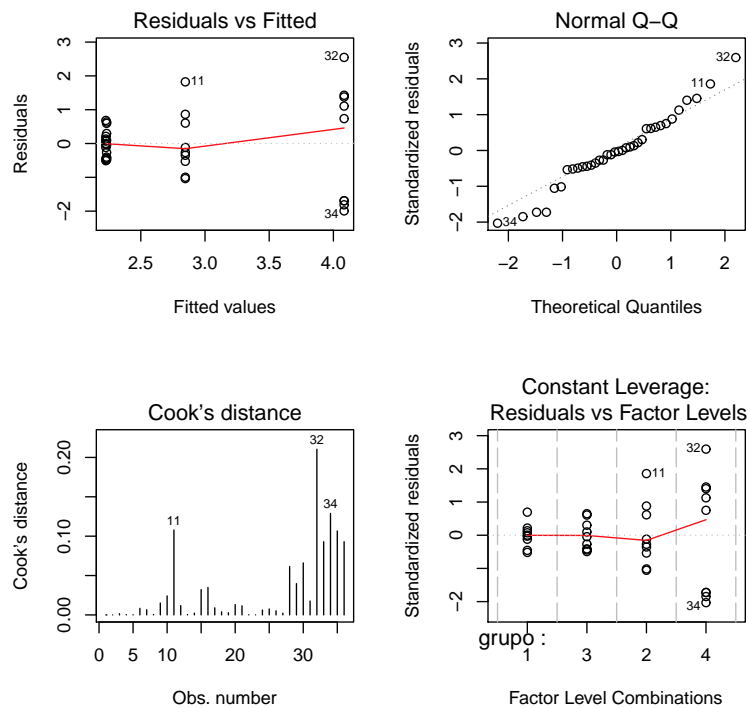
Assim, duas médias populacionais de nível consideram-se diferentes caso as correspondentes médias amostrais difiram em mais do que 1.329 unidades. As conclusões são as acima referidas.

- (d) Como em qualquer modelo linear, o resíduo é a diferença entre cada valor observado da variável resposta e o correspondente valor ajustado pelo modelo, ou seja, e usando a notação da ANOVA a 1 Factor, $e_{ij} = y_{ij} - \hat{y}_{ij}$. Sabe-se que, num modelo ANOVA a um factor, o valor ajustado dum dada observação corresponde à média amostral das observações no mesmo nível do factor: $\hat{y}_{ij} = \bar{y}_{i.}$. Assim, todas as observações do primeiro grupo têm valor ajustado igual a $\hat{y}_{1j} = \bar{y}_{1.} = 2.228$. O resíduo da primeira observação do primeiro grupo será $e_{11} = 2.23 - 2.228 = 0.002$ e o da segunda observação desse grupo é $e_{12} = 2.20 - 2.228 = -0.028$. De forma análoga, os valores ajustados de qualquer observação no segundo grupo são dados por $\hat{y}_{2j} = \bar{y}_{2.} = 2.847$. O resíduo da terceira observação do segundo grupo é assim $e_{23} = y_{23} - \bar{y}_{2.} = 3.45 - 2.847 = 0.603$. Para calcular a totalidade dos resíduos podemos recorrer ao R (arredondando a três casas decimais):

```
> round(residuals(aov(concentracao ~ grupo, data=zinco)),d=3)
      1      2      3      4      5      6      7      8      9     10     11     12
0.002 -0.028  0.212 -0.118  0.072 -0.508 -0.448  0.132  0.682  0.863  1.823  0.603
     13     14     15     16     17     18     19     20     21     22     23     24
-0.117 -0.267 -0.997 -1.037 -0.527 -0.347  0.297  0.637  0.597  0.097 -0.043 -0.433
     25     26     27     28     29     30     31     32     33     34     35     36
-0.483 -0.403 -0.263  1.377  1.107  1.427  0.737  2.547 -1.693 -1.993 -1.813 -1.693
```

Com o auxílio do R, podemos construir os quatro gráficos de resíduos já considerados no estudo dos modelos de Regressão Linear, produzidos pelo comando:

```
plot(aov(concentracao ~ grupo, data=zinco), which=c(1,2,4,5))
```



O gráfico do canto superior esquerdo é o gráfico de resíduos usuais (no eixo vertical) vs. valores ajustados da variável resposta (eixo horizontal). O facto de os resíduos surgirem “empilhados” em colunas é característico numa ANOVA a um factor e resulta do já referido facto de todas as observações dum dado nível terem o mesmo valor ajustado $\hat{y}_{ij} = \bar{y}_{i.}$, logo, a mesma coordenada no eixo horizontal. Neste caso, à primeira vista apenas se observam

três colunas, e não as quatro que seriam de esperar, tendo em conta que o nosso factor Grupos tem quatro níveis. A grande proximidade das médias amostrais do primeiro e terceiro níveis significa que há, na realidade, duas colunas de resíduos quase sobrepostas na parte esquerda do gráfico. A principal conclusão a extrair deste gráfico é que pode haver problemas com a hipótese de homogeneidade das variâncias, uma vez que a variabilidade no segundo grupo, e sobretudo no quarto grupo (identificável pelo facto de corresponder à maior média, logo corresponder à coluna que surge mais à direita no gráfico) parece bastante maior que a dos outros dois níveis. Será conveniente efectuar um teste de Bartlett à homogeneidade das variâncias. A utilização desse teste parece adequada, uma vez que o *qq-plot* (no canto superior direito) não indicia problemas graves com a Normalidade, dada a disposição aproximadamente linear dos pontos. Do gráfico no canto inferior esquerdo resulta que nenhuma observação tem uma distância de Cook desmesurada, pelo que não há observações excessivamente influentes, embora $D_{32} > 0.2$. O gráfico do canto inferior direito costuma ter no eixo horizontal os valores do efeito alavanca (h_{ii} , ou *leverages*) de cada observação. No entanto, para delineamentos equilibrados a um factor, o R substitui esse eixo por uma simples indicação dos diferentes níveis do factor (ordenados por ordem crescente das médias \bar{y}_i). A razão dessa opção reside no seguinte facto: é possível mostrar que o efeito alavanca de qualquer observação y_{ij} numa ANOVA a um factor é dada por $\frac{1}{n_i}$, onde n_i indica o número de observações no nível i da observação. Em delineamentos equilibrados, esse valor é igual para todas as observações (no nosso caso, todas teriam efeito alavanca igual a $\frac{1}{9}$), pelo que o gráfico tradicional seria de pouca utilidade, empilhando todos os resíduos numa única coluna. O gráfico alternativo produzido pelo R quando os delineamentos são equilibrados fica assim semelhante ao do canto superior esquerdo, embora sem qualquer efeito de escala no eixo horizontal e com os resíduos (internamente) standardizados no eixo vertical, em vez dos resíduos usuais.

As observações número 11, 32 e sobretudo 34 surgem em todos os gráficos de resíduos como observações extremas, o que merece alguma atenção. No quadro com os valores observados, disponível no enunciado, verifica-se que a observação 11 (a segunda observação do Grupo 2, com valor $y_{22} = 4.67$) é claramente superior às restantes observações do segundo Grupo. O mesmo se passa com a observação 32, a que corresponde o maior valor (6.63) de toda a tabela. Já a observação 34 está na situação oposta. O seu valor (2.09) é pequeno em qualquer grupo, mas sobretudo no quarto Grupo, a que corresponde. Aliás, neste Grupo há uma subdivisão das suas 9 observações em dois subconjuntos: o das cinco primeiras observações, todas superiores a 4.80, e o das quatro últimas observações, todas inferiores a 2.40. Esta divisão clara (que mereceria uma atenção especial no estudo em causa) contribui para que este quarto grupo tenha uma variabilidade de resíduos que, no primeiro e último gráficos, se evidencia claramente como maior do que os restantes grupos.

Completemos a alínea efectuando o teste de Bartlett. O facto de haver $n_c = 9 > 5$ repetições em cada nível do factor significa que a aproximação assintótica da distribuição da respectiva estatística de teste pode ser considerada válida. Indicando por σ_i^2 a variância populacional no nível i do factor (no nosso caso, para cada passo do procedimento), temos:

Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ vs. $H_1 : \exists i, j$ tais que $\sigma_i^2 \neq \sigma_j^2$.

Estatística do teste:
$$K^2 = \frac{(n-k) \ln QMRE - \sum_{i=1}^k (n_i-1) \ln S_i^2}{C} \sim \chi_{k-1}^2, \text{ sob } H_0,$$

onde $C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \frac{1}{n_i-1} - \frac{1}{n-k} \right]$ e S_i^2 representa a variância amostral do nível i .

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $K_{calc}^2 > \chi_{0.05(3)}^2 = 7.815$.

Conclusões: Para calcular o valor da estatística do teste vamos recorrer ao R (note-se que o comando `aov` não é utilizado para invocar o teste de Bartlett):

```
> bartlett.test(concentracao ~ grupo, data=zinco)
Bartlett test of homogeneity of variances
data:  concentracao by grupo
Bartlett's K-squared = 23.1873, df = 3, p-value = 3.691e-05
```

O valor calculado, $K_{calc}^2 = 23.1873$, é claramente significativo (o que também se desprende do *p-value* muito baixo), pelo que se confirmam as dúvidas sobre a existência de maior variabilidade nalguns momentos do processamento. Assim, há dúvidas sobre a validade da ANOVA efectuada.

NOTA: Nesta situação, podem existir transformações estabilizadoras da variância (por exemplo, uma logaritmização da variável resposta) que permitam ultrapassar este problema. Mas pode ser preferível utilizar uma variante não-paramétrica da ANOVA a um factor, nomeadamente o Teste de Kruskal-Wallis (não incluído no programa desta UC).

- (e) A nova descrição da experiência corresponderia a uma ANOVA a dois factores, sendo o segundo factor constituído pelos únicos nove diferentes lotes de feijão agora existentes, a que correspondem as linhas da tabela. Repare-se que, a ser verdadeira esta nova descrição do delineamento, as nove observações de cada nível do factor Grupo não são independentes das nove observações de cada um dos outros Grupos. Pelo contrário, é natural que observações do mesmo lote de feijão, efectuadas em diferentes Grupos (diferentes passos do tratamento) estejam correlacionadas entre si. Refira-se ainda que na descrição da experiência dada nesta alínea, cada nível do novo (segundo) factor constitui aquilo a que, na tradição da Análise de Variância, se designa por *bloco*. Esta designação surge historicamente associada a factores cuja inclusão na experiência resulta, não tanto de se pretender estudar directamente o seu efeito sobre a variável resposta, mas sobretudo de saber que constituem uma fonte de heterogeneidade das unidades experimentais, associada a variabilidade na variável resposta. Pretende-se incorporar essa heterogeneidade no modelo, controlando-a e podendo assim filtrar a variabilidade nos valores da variável resposta que lhe está associada.

A *data frame zinco* já está preparada para ser analisada como um delineamento a dois factores, havendo uma terceira coluna da *data frame* designada `blocos`, que corresponde ao novo factor. Refira-se que o novo factor tem $b = 9$ níveis, e que nas 36 células agora existentes não há repetições de observações (ou seja, $n_c = 1$). Logo, independentemente de ser desejável, não é possível incluir efeitos de interacção no modelo. A experiência pode ser descrita por um modelo a dois factores, sem interacção:

- i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}$, $\forall i = 1, 2, 3, 4$, $j = 1, 2, \dots, 9$, $k = 1$ (o índice k é dispensável porque não há repetições nas células), com $\alpha_1 = 0$ e $\beta_1 = 0$, e onde
 - Y_{ijk} indica a observação ($k = 1$ sempre) do Grupo i , associado ao lote de feijão j ;
 - μ_{11} é a concentração média populacional pré-processamento (Grupo 1) do lote 1;
 - α_i indica o efeito principal do lote i ;
 - β_j indica o efeito principal do Grupo j ; e
 - ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk} .
- ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$.
- iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constituem um conjunto de variáveis aleatórias independentes.

A tabela-resumo correspondente a este novo modelo é:

```
> summary(aov(concentracao ~ grupo + blocos, data=zinco))
              Df Sum Sq Mean Sq F value    Pr(>F)
grupo          3  20.597   6.8656   9.7361 0.0002183 ***
blocos         8  17.760   2.2200   3.1482 0.0139310 *
Residuals     24  16.924   0.7052
```

Repare-se que:

- Existe uma nova linha na tabela (em comparação com a tabela do modelo a um factor), correspondente ao novo factor.
- Os graus de liberdade, Soma de Quadrados e Quadrado Médio do Factor *Grupo* são idênticos aos da tabela-resumo do modelo a um factor. Este facto significa que os antigos graus de liberdade e Soma de Quadrados Residual foram agora decompostos na soma de duas parcelas: uma que é atribuída aos efeitos do novo Factor (*blocos*, i.e., lote de feijão) e outra que permanece residual, ou seja, não explicada pelo novo modelo. Concretamente, os 32 graus de liberdade residuais do modelo a um factor são agora repartidos em $b - 1 = 8$ associados aos $b = 9$ níveis do novo factor B mais $n - (a + b - 1) = 24$ (aqui $a = k = 4$, os níveis do Factor *Grupo*) que permanecem associados ao residual no novo modelo a dois factores, sem interacção (que tem $a + b - 1$ parâmetros). Quanto às Somas de Quadrados, a antiga $SQRE_A = 34.684$ é repartida nas duas novas parcelas $SQB = 17.760$ e $SQRE_{A+B} = 16.924$. Ou seja, o novo Factor explica um pouco mais de metade da variabilidade que permanecia inexplicada (residual) no modelo só com o Factor A (*Grupo*).
- Na estatística F aos efeitos do Factor *Grupo*, o numerador QMF (agora QMA , na notação para modelos a dois factores) fica igual, enquanto que o denominador $QMRE$ sofre uma dupla transformação: o seu numerador $SQRE$ é menor do que no modelo a um factor (pois $SQRE_A = SQRE_{A+B} + SQB$), mas também o seu denominador é menor (pois $g.l.(SQRE_{A+B}) = n - (a + b - 1) < n - a = g.l.(SQRE_A)$).
- No exemplo em questão, o novo $QMRE$ do modelo com dois factores é mais baixo: 0.7052 (em vez de 1.0839 no modelo só com o Factor *Grupo*). A estatística F no teste aos efeitos do Factor *Grupo* (que, recorde-se, continua a ter o mesmo numerador) cresce assim para $F_a = 9.7361$ (no modelo apenas com esse factor era $F = 6.3343$). A rejeição da hipótese de inexistência de efeitos do Factor *Grupo* ($H_0 : \alpha_i = 0, \forall i$) torna-se mais clara (o p -value é agora apenas $p = 0.0002$, quando era $p = 0.0017$ no modelo apenas com esse factor). Em geral, se o novo denominador da estatística F para o teste aos efeitos do Factor A é maior ou menor do que antes depende da relação entre a redução na Soma de Quadrados Residual e a redução nos correspondentes graus de liberdade. Mas, caso existam realmente efeitos do novo factor, a nova Soma de Quadrados Residual $SQRE_{A+B}$ será bastante inferior à antiga e também $QMRE_{A+B}$ será menor, o que aumenta a estatística F , que tende assim a ser mais significativa.

10. (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. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall i = 1, 2, 3, 4$, $j = 1, 2, \dots, 9$, $k = 1, 2, 3, 4$,
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 interação entre a localidade i e a cultivar j ; e
 - ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk} .
- ii. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$.
- iii. $\{\epsilon_{ijk}\}_{i,j,k}$ constituem 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. Pedem-se os três testes F para cada tipo de efeitos previstos no modelo. Efectuemos em pormenor o teste à existência de efeitos de interação entre localidade e cultivar:

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

Estatística do teste: $F = \frac{QMAB}{QMRE} \cap F_{[(a-1)(b-1), n-ab]}$, 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 interação 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 $H_0 : \beta_j = 0$, $\forall j = 2, 3, \dots, 9$ vs. $H_1 : \exists j = 2, 3, \dots, 9$, 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.

iii. 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 regista o rendimento mais baixo em Elvas (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 Elvas e Revilheira, que para várias cultivares parecem ter comportamentos quase antagónicos.

iv. 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, $SQT^* = \sum_i \sum_j \sum_k (Y_{ijk}^* - \bar{Y}_{...}^*)^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \bar{Y}_{...}/1000)^2 = SQT/(1000^2)$. 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,

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

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

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 (incluindo os respectivos p -values) mantêm-se inalterados.

- (c) O melhor rendimento observado em Elvas é o da cultivar *Trovador* ($\bar{y}_{29} = 5927\text{kg/ha}$). 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 $QMRE$.

11. (a) Trata-se dum delineamento factorial a dois factores: *Temperatura de conservação* (Factor A), com $a = 2$ níveis, e *Tempo de armazenamento* (Factor B), com $b = 4$ níveis. Para modelar a variável resposta Y (*alterações no conteúdo em taninos das polpas de sapoti*), utiliza-se um modelo ANOVA a dois factores, com interacção. É possível estudar a interacção devido à presença de repetições nas $2 \times 4 = 8$ células. Sempre que possível, é desejável considerar este modelo para delineamentos factoriais a dois factores, deixando que sejam os dados a sugerir se se deve admitir a existência desse tipo de efeitos. O delineamento é equilibrado, uma vez que todas as células têm o mesmo número de repetições: $n_{ij} = 4 = n_c$ ($\forall i, j$), para um total de $n = 8 \times 4 = 32$ observações. O modelo é dado por:

i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall i = 1, 2$, $j = 1, 2, 3, 4$, $k = 1, 2, 3, 4$,

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 a k -ésima observação (repetição) na célula definida pelo nível i do Factor A e o nível j do Factor B;
- μ_{11} indica a média (populacional) das observações na célula (1, 1), ou seja, com temperatura alta e 0 dias de armazenamento;
- α_i indica o efeito do nível i do Factor A (*Temperatura*);
- β_j indica o efeito do nível j do Factor B (*Tempo de armazenamento*);
- $(\alpha\beta)_{ij}$ indica o efeito de interacção na célula (i, j) ; e
- ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk} .

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

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

- (b) A tabela-resumo desta ANOVA terá três linhas associadas a cada tipo de efeitos previsto no modelo (ou seja, efeitos principais do Factor A, efeitos principais do Factor B e efeitos de interacção) e ainda uma linha para o residual (podendo também incluir-se a linha associada à variabilidade Total). Como em qualquer modelo ANOVA, a tabela-resumo tem as seguintes colunas: Somas de Quadrados, graus de liberdade correspondentes, Quadrados Médios e estatísticas F . Os graus de liberdade são dados por:

- Factor A: $a - 1 = 1$;
- Factor B: $b - 1 = 3$;
- Interação: $(a - 1)(b - 1) = 3$;
- Residual: $n - ab = 32 - 8 = 24$.

Para calcular as Somas de Quadrados, registamos que no enunciado é dada a Soma de Quadrados Residual $SQRE = 20.72$. É igualmente dado o Quadrado Médio do Factor B, e multiplicando pelos respectivos graus de liberdade obtém-se $SQB = QMB(b - 1) = 96.01 \times 3 = 288.03$. A Soma de Quadrados Total também pode ser calculada facilmente, uma vez que no enunciado á dada a variância da totalidade das observações de Y , $s_y^2 = 47.83222$, e $SQT = (n - 1) s_y^2 = 31 \times 47.83222 = 1482.799$. Assim, faltam as duas Somas de Quadrados relativas aos efeitos principais do factor A (SQA) e aos efeitos de interação ($SQAB$). Utilizando a expressão para SQA , no caso de delineamentos equilibrados (disponível no formulário) e os valores das médias de nível do factor A e da média geral (disponíveis no enunciado), tem-se $SQA = bn_c \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 = 16 [(24.681 - 22.14375)^2 + (19.606 - 22.14375)^2] = 16 \times 12.87781 = 206.045$. A última Soma de Quadrados em falta ($SQAB$) pode ser calculada a partir das restantes quatro: $SQAB = SQT - (SQA + SQB + SQRE) = 1482.799 - (206.045 + 288.03 + 20.72) = 968.004$. Assim,

Variacão	g.l.	SQs	QMs	F_{calc}
Factor A	1	206.045	$QMA = \frac{SQA}{a-1} = 206.045$	$F = \frac{QMA}{QMRE} = 238.6622$
Factor B	3	288.03	$QMB = \frac{SQB}{b-1} = 96.01$	$F = \frac{QMB}{QMRE} = 111.2085$
Interação	3	968.004	$QMAB = \frac{SQAB}{(a-1)(b-1)} = 322.668$	$F = \frac{QMAB}{QMRE} = 373.7467$
Residual	24	20.72	$QMRE = \frac{SQRE}{n-ab} = 0.8633333$	–
Total	31	1482.799	–	–

- (c) De acordo com o modelo, a influência do Factor B nos valores da variável resposta pode resultar de dois tipos de efeitos: os efeitos principais do Factor B (os β_j) ou os efeitos de interação (os $(\alpha\beta)_{ij}$). Efectuaremos estes dois testes, começando pelo dos efeitos de interação. Neste exemplo, e como o Factor A apenas tem dois níveis, o índice i nos efeitos de interação apenas toma o valor $i = 2$.

Hipóteses: $H_0 : (\alpha\beta)_{2j} = 0, \forall j = 2, 3, 4$ vs. $H_1 : \exists j = 2, 3, 4$ tal que $(\alpha\beta)_{2j} \neq 0$.

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,24)} = 3.01$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 373.7467$. É um valor claramente significativo e rejeita-se H_0 a favor da hipótese alternativa de que existem efeitos de interação.

Já é possível responder afirmativamente: o Factor B tem efeitos sobre os valores médios de Y . No entanto, efectuaremos também o teste aos efeitos principais do Factor B:

Hipóteses: $H_0 : \beta_j = 0, \forall j = 2, 3, 4$ vs. $H_1 : \exists j = 2, 3, 4$ tal que $\beta_j \neq 0$.

Estatística do teste: $F = \frac{QMB}{QMRE} \cap F_{(b-1, n-ab)}$, sob H_0 .

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,24)} = 3.01$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 111.2085$. É um valor claramente significativo e rejeita-se H_0 a favor da hipótese de que existem efeitos principais do Factor B.

Assim, quer pela via dos efeitos principais, quer pela via dos efeitos de interacção, o Factor B (*tempo de armazenamento*) afecta os conteúdos médios de taninos nos sapotis.

- (d) Os dois gráficos de interacção apresentam a mesma informação, embora de forma diferente. Nos dois gráficos, os segmentos de recta unem oito pontos, associados às oito células definidas pelo nosso delineamento. Em ambos os casos, no eixo vertical encontram-se valores da variável resposta Y . Os valores médios de Y em cada célula definem a coordenada y dos oito pontos. No eixo horizontal indicam-se os níveis de um dos factores.

No gráfico da esquerda é o Factor B que define o eixo horizontal, e por cima de cada um dos seus quatro níveis existem dois pontos, correspondentes às duas células associada a esse nível do Factor B. Os segmentos de recta de cada tipo unem os pontos referentes ao mesmo nível do Factor A. Assim, a tracejado estão os segmentos que unem as médias de célula nas quais o Factor A está no nível $i = 1$ (*alta*), enquanto que as linhas contínuas unem as médias de célula em que o Factor A tem nível $i = 2$ (*baixa*). O facto dessas duas curvas seccionalmente lineares estarem longe de qualquer paralelismo sugere a existência de efeitos de interacção, confirmando o resultado do respectivo teste, efectuado na alínea anterior.

No gráfico da direita é o Factor A que define o eixo horizontal, e por cima de cada um dos seus dois níveis encontram-se quatro pontos, correspondentes às médias das quatro células associadas a esse nível do Factor A. Os dois pontos correspondentes a um mesmo nível no Factor B são unidos por segmentos de recta, à semelhança do que acontece no gráfico anterior. Mais uma vez, há uma forte indicação de efeitos de interacção, sobretudo resultante das células associadas ao tempo de armazenamento 0, cujo comportamento é substancialmente diferente dos que correspondem aos restantes níveis do Factor B.

- (e) A afirmação do investigador é que as médias populacionais das quatro células em que $i = 1$ não diferem entre si. Vamos estudar esta afirmação comparando as quatro médias amostrais dessas células através dum teste de Tukey. O termo de comparação para qualquer diferença de médias de nível, utilizando um nível global de significância $\alpha = 0.05$, é dado por

$$q_{\alpha(ab, n-ab)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(8,24)} \sqrt{\frac{0.8633333}{4}} = 4.68 \times 0.4645787 = 2.174228 .$$

Assim, devemos concluir pela diferença das médias populacionais de duas quaisquer células, caso as respectivas médias amostrais difiram em mais do que 2.174228 unidades. Uma análise das médias de célula disponíveis no enunciado mostra que, para temperaturas de armazenamento altas ($i = 1$), os pares de médias das células com tempos de armazenamento superiores a 0 (ou seja, para $j = 2, 3, 4$) diferem sempre, entre si, por menos do que esse termo de comparação (as médias são 26.85, 25.97 e 26.40). No entanto, a média da célula (1, 1), correspondente a tempo de armazenamento nulo, tem média 19.50, que difere em mais do que 2.174228 unidades das médias amostrais das células (1, 2), (1, 3) e (1, 4). Assim, devemos rejeitar a afirmação do investigador, ao nível $\alpha = 0.05$.

12. A *data frame* referida no enunciado deste Exercício contém mais dados do que aqueles que são necessários para responder às perguntas feitas.

- (a) Trata-se dum delineamento factorial a dois factores: Fibra (Factor A, com $a = 2$ níveis) e Enzima (Factor B, com $b = 2$ níveis). Em cada uma destas $ab = 4$ células há $n_c = 12$

repetições, pelo que se trata dum delineamento equilibrado. A variável resposta é *CEL*, o Coeficiente de Utilização Digestiva da celulose. Representando por Y_{ijk} a k -ésima observação desta variável resposta *CEL*, correspondente ao nível i de Fibra e j de Enzima, tem-se o seguinte modelo ANOVA a dois factores, com interacção:

- i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall i = 1, 2$, $j = 1, 2$, $k = 1, 2, \dots, 12$,
com $\alpha_1 = 0$, $\beta_1 = 0$ e $(\alpha\beta)_{ij} = 0$ se i ou j tomarem o valor 1. Neste caso concreto, e tendo em conta que cada factor tem apenas dois níveis, só existe um efeito de cada tipo: α_2 , β_2 e $(\alpha\beta)_{22}$. Na equação,

- μ_{11} indica o CUD médio (populacional) para a celulose, na célula (1, 1);
- α_i indica o efeito do nível i do Factor A (*Fibra*);
- β_j indica o efeito do nível j do Factor B (*Enzima*);
- $(\alpha\beta)_{ij}$ indica o efeito de interacção na célula (i, j) ; e
- ϵ_{ijk} indica o erro aleatório associado à observação Y_{ijk} .

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

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

- (b) Pedem-se a realização dos três testes F , associados a cada tipo de efeitos previstos no modelo. Tendo em conta que os dados estão disponibilizados na *data frame* `leitoeos`, vamos construir a tabela-resumo da ANOVA com o auxílio do R:

```
> summary(aov(CEL ~ Fibra*Enzima, data=leitoeos))
              Df Sum Sq Mean Sq F value Pr(>F)
Fibra           1  0.0239  0.02385   1.450 0.23500
Enzima          1  0.1376  0.13760   8.364 0.00593 **
Fibra:Enzima    1  0.0257  0.02567   1.560 0.21824
Residuals      44  0.7239  0.01645
```

Eis os três testes (escrevendo as hipóteses da forma especial que resulta de terem-se apenas dois níveis em cada factor), começando pelo teste ao efeito de interacção:

Hipóteses: $H_0 : (\alpha\beta)_{22} = 0$ vs. $H_1 : (\alpha\beta)_{22} \neq 0$.

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(1,44)} \approx 4.06$.

Conclusões: O valor da estatística do teste foi já calculado: $F_{calc} = 1.560 < 4.06$, pelo que não se rejeita H_0 , não havendo motivo para admitir a existência de efeitos de interacção.

O teste ao efeito do Factor A é análogo:

Hipóteses: $H_0 : \alpha_2 = 0$ vs. $H_1 : \alpha_2 \neq 0$.

Estatística do teste: $F = \frac{QMA}{QMRE} \cap F_{[a-1, n-ab]}$, sob H_0 .

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(1,44)} \approx 4.06$.

Conclusões: O valor da estatística do teste é dado na tabela-resumo: $F_{calc} = 1.450 < 4.06$, pelo que não se rejeita H_0 , não havendo motivo para admitir a existência de efeitos de fibra na digestibilidade.

Finalmente, o teste ao efeito da presença de enzimas nas dietas:

Hipóteses: $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$.

Estatística do teste: $F = \frac{QMB}{QMRE} \cap F_{[b-1, n-ab]}$, sob H_0 .

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(1,44)} \approx 4.06$.

Conclusões: O valor da estatística do teste é calculado: $F_{calc} = 8.364 > 4.06$, pelo que se rejeita H_0 , concluindo-se pela existência de efeitos associados à presença de enzimas no alimento.

Assim, a adição de enzimas introduz alterações na digestibilidade média dos alimentos, não havendo efeitos associados ao factor Fibra.

- (c) Repare-se que as conclusões da alínea anterior permitem responder à pergunta através duma via alternativa à utilização de testes de Tukey. Uma vez que apenas há efeitos do factor B, e este só tem dois níveis, conclui-se que as médias de célula apenas diferem entre si caso pertençam a diferentes níveis do factor Enzima. De facto, recorde-se que $\mu_{21} = \mu_{11} + \alpha_2$, pelo que ao se admitir que $\alpha_2 = 0$, está-se a admitir que $\mu_{21} = \mu_{11}$. De igual modo, $\mu_{12} = \mu_{11} + \beta_2$, pelo que ao rejeitar-se a hipótese $\beta_2 = 0$, se está a concluir que $\mu_{12} \neq \mu_{11}$. Finalmente, $\mu_{22} = \mu_{11} + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$. Uma vez que se admite $\alpha_2 = 0$ e $(\alpha\beta)_{22} = 0$, admite-se $\mu_{22} = \mu_{11} + \beta_2 = \mu_{12}$.

No entanto, efectuaremos os teste de Tukey, como pedido no enunciado. O facto de a teoria subjacente a testes de Tukey e testes F da ANOVA não ser idêntica pode fazer surgir alguma discrepância nas respectivas conclusões. O termo de comparação do teste de Tukey, utilizando um nível de significância global $\alpha = 0.05$, é dado por

$$q_{\alpha(ab, n-ab)} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(4,44)} \sqrt{\frac{0.01645}{12}} \approx 3.78 \times 0.03702477 = 0.1399536 .$$

Ora, as quatro médias amostrais de célula podem ser obtidas, no R, por meio do comando

```
> model.tables(aov(CEL ~ Fibra*Enzima, data=leitoeis), type="means")
Tables of means
Grand mean
0.413125

Fibra
Fibra
  1      2
0.4354 0.3908

Enzima
Enzima
  1      2
0.3596 0.4667

Fibra:Enzima
  Enzima
Fibra 1      2
      1 0.4050 0.4658
      2 0.3142 0.4675
```

As médias de célula são indicadas na tabela final. Dos seis possíveis pares de médias de células, apenas em dois casos as médias de célula diferem por mais do que o termo de comparação: $|\bar{Y}_{21} - \bar{Y}_{12}| = 0.1516 > 0.1400$ e $|\bar{Y}_{21} - \bar{Y}_{22}| = 0.1533 > 0.1400$. Logo, e ordenando as quatro médias de célula por ordem crescente, tem-se:

$\bar{y}_{21.}$	$\bar{y}_{11.}$	$\bar{y}_{12.}$	$\bar{y}_{22.}$
0.3142	0.4050	0.4675	0.4658

As conclusões não são inteiramente coerentes com as conclusões obtidas através dos testes F , uma vez que não se conclui que μ_{11} seja diferente das duas médias de célula associadas ao nível 2 do factor *Enzima*.

- (d) Neste caso, o teste de Bartlett compara as variâncias de célula. A hipótese nula afirma que as quatro variâncias populacionais de célula são iguais (como se admite no modelo), enquanto que a hipótese alternativa afirma que, para algum par de células, as correspondentes variâncias populacionais diferem:

$$H_0 : \sigma_{11}^2 = \sigma_{12}^2 = \sigma_{21}^2 = \sigma_{22}^2 \quad vs. \quad H_1 : \exists i, j, i', j' \text{ tais que } \sigma_{ij}^2 \neq \sigma_{i'j'}^2 .$$

A estatística deste teste tem uma forma pouco amigável (dada no acetato 384 das aulas teóricas). Para calcular o seu valor, utilizaremos o comando `bartlett.test` do R. No entanto, este comando (na sua actual versão) apenas admite uma variável de classificação das diferentes categorias cujas variâncias se deseja comparar. Isto significa que será necessário criar uma única variável, cujos valores identificam as $ab = 4$ células do delineamento. Isso pode ser feito através do seguinte comando do R que irá “colar” os nomes dos níveis de cada factor, utilizando um “0” como símbolo separador:

```
> celulas <- paste(leitoes$Fibra, leitoes$Enzima, sep = "0")
> celulas
[1] "101" "101" "101" "101" "101" "101" "102" "102" "102" "102" "102" "102"
[13] "201" "201" "201" "201" "201" "201" "202" "202" "202" "202" "202" "202"
[25] "101" "101" "101" "101" "101" "101" "102" "102" "102" "102" "102" "102"
[37] "201" "201" "201" "201" "201" "201" "202" "202" "202" "202" "202" "202"
```

O cálculo da estatística do teste de Bartlett pode ser pedido assim:

```
> bartlett.test(CEL ~ celulas, data=leitoes)
```

```
Bartlett test of homogeneity of variances
data: CEL by celulas
Bartlett's K-squared = 15.7157, df = 3, p-value = 0.001297
```

O valor calculado da estatística é $K_{cal}^2 = 15.7157$. Comparado com o valor fronteira da Região Crítica ao nível de significância $\alpha = 0.05$, que é $\chi_{0.05(3)}^2 = 7.81473$ (os graus de liberdade são, neste caso, $ab - 1 = 4 - 1 = 3$), temos uma rejeição de H_0 e a opção pela hipótese alternativa, correspondente à existência de variâncias heterogéneas. Esta conclusão, que também se pode justificar com base no valor de prova (*p-value*, $p = 0.001297$) dado na listagem produzida pelo R, significa que as conclusões dos testes acima efectuados podem não ser válidas, uma vez que um dos pressupostos do modelo (variâncias homogéneas) é questionável (veja-se a *Nota* no final da resolução da alínea d) do Exercício 4).

13. Continuando a considerar os dados do Exercício 12, temos:

- (a) Para o modelo a dois factores, com interacção,
- i. A matriz \mathbf{X} tem 48 linhas (uma para cada observação) e quatro colunas: uma primeira coluna de uns; uma segunda coluna dada pela indicatriz de pertença ao segundo nível do factor Fibra; uma terceira coluna dada pela indicatriz de pertença ao segundo nível

do factor Enzima; uma quarta e última coluna dada pela indicatriz de pertença à célula (2,2). Essa estrutura pode ser confirmada com o auxílio do comando:

```
> model.matrix(aov(CEL ~ Fibra*Enzima, data=leitoe))
```

- ii. Para construir a matriz de projecção ortogonal $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, precisamos de conhecer os seguintes comandos do R:

- a função `t`, que transpõe uma matriz que seja passada como argumento – por exemplo, `t(A)` calcula a transposta duma matriz `A` (previamente definida);
- a função `solve`, que inverte uma matriz que seja passada como argumento – por exemplo, `solve(A)` calcula a inversa da matriz `A` (caso exista);
- o operador `%%` que efectua a multiplicação matricial de duas matrizes, que surjam antes e depois do símbolo do operador. Por exemplo, o produto `AB` (por essa ordem) de duas matrizes `A` e `B` (já definidas), obtém-se escrevendo `A %% B`.

Assim, a matriz \mathbf{H} pode obter-se da seguinte forma:

```
> X <- model.matrix(aov(CEL ~ Fibra*Enzima, data=leitoe))
```

```
> H <- X %% solve(t(X) %% X) %% t(X)
```

- iii. Utilizando a matriz \mathbf{H} construída na alínea anterior, os valores ajustados de Y resultam do produto $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$, que no R pode ser obtido da seguinte forma (por razões de espaço, o resultado do comando apenas é reproduzido parcialmente):

```
> H %% leitoe$CEL
```

```
  [,1]
1 0.4050000
2 0.4050000
3 0.4050000
4 0.4050000
5 0.4050000
6 0.4050000
7 0.4658333
8 0.4658333
...
47 0.4675000
48 0.4675000
```

Sabemos que estes valores ajustados correspondem às médias amostrais das células onde cada observação foi efectuada. Este facto pode ser confirmado comparando os valores acima obtidos com a tabela das médias obtida na alínea c) do Exercício 12.

NOTA: A forma mais fácil de obter os valores ajustados de Y no R seria, naturalmente, através da utilização do comando `fitted`, aplicado ao ajustamento do modelo ANOVA:

```
> fitted(aov(CEL ~ Fibra*Enzima, data=leitoe))
```

- iv. Tendo em conta que os resíduos se definem como $E_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$, podemos calcular a Soma de Quadrados Residual da seguinte forma:

```
> sum((leitoe$CEL-H %% leitoe$CEL)^2)
```

```
[1] 0.7239083
```

Este valor de *SQRE* corresponde ao que foi obtido na tabela-resumo da ANOVA, calculada no Exercício 12b).

- (b) Vamos repetir os comandos da alínea anterior, mas tendo agora por base o modelo ANOVA a dois factores, *sem* efeitos de interacção:

```
> X <- model.matrix(aov(CEL ~ Fibra+Enzima, data=leitoe))
```

```
> H <- X %% solve(t(X) %% X) %% t(X)
```

```
> sum((leitoes$CEL-H %*% leitoes$CEL)^2)
[1] 0.7495771
```

- (c) Para o modelo apenas com o Factor *Enzima*, a Soma de Quadrados Residual resulta dos comandos:

```
> X <- model.matrix(aov(CEL ~ Enzima, data=leitoes))
> H <- X %*% solve(t(X) %*% X) %*% t(X)
> sum((leitoes$CEL-H %*% leitoes$CEL)^2)
[1] 0.7734292
```

Para calcular a Soma de Quadrados do Factor (SQF , correspondente à Soma SQR nos modelos de Regressão) neste modelo a um Factor, recordamos que, por definição, é dado pela soma, ao longo de todas as observações, do quadrado da diferença entre cada Y ajustado e a média global de todas as observações: $SQF = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (\hat{Y}_{ijk} - \bar{Y}_{...})^2$. Esta Soma de Quadrados pode assim ser calculada no R da seguinte forma:

```
> sum((H %*% leitoes$CEL-mean(leitoes$CEL))^2)
[1] 0.1376021
```

- (d) Por analogia com o que foi feito na alínea anterior, temos, num modelo a um Factor, só com o Factor *Fibra*:

```
> X <- model.matrix(aov(CEL ~ Fibra, data=leitoes))
> H <- X %*% solve(t(X) %*% X) %*% t(X)
> sum((leitoes$CEL-H %*% leitoes$CEL)^2)
[1] 0.8871792
> sum((H %*% leitoes$CEL-mean(leitoes$CEL))^2)
[1] 0.02385208
```

- (e) Recordando as definições das várias Somas de Quadrados numa Análise de Variância num modelo a dois factores, com interacção, observamos que:

- $SQRE$ é a Soma de Quadrados Residual calculada na alínea a): $SQRE_{A*B} = 0.7239083$.
- a Soma de Quadrados associada aos efeitos de interacção é, por definição, a diferença das Somas de Quadrados Residuais dos modelos sem, e com, interacção: $SQAB = SQRE_{A+B} - SQRE_{A*B} = 0.7495771 - 0.7239083 = 0.0256688$.
- a Soma de Quadrados associada aos efeitos do Factor B (*Enzima*) é, por definição, a diferença das Somas de Quadrados Residuais do modelo com o único factor *Fibra* (Factor A), e do modelo a dois factores, sem interacção: $SQB = SQRE_A - SQRE_{A+B} = 0.8871792 - 0.7495771 = 0.1376021$
- Finalmente, a Soma de Quadrados associada ao Factor A (*Fibra*) é definido como a Soma de Quadrados do ajustamento (SQF) no modelo com apenas esse factor: $SQA = SQF_A = 0.02385208$.

Verificamos que se trata dos valores indicados na tabela-resumo do Exercício 12b).

Uma vez que o delineamento é equilibrado, seria possível calcular os valores de SQA e SQB trocando a ordem de exclusão dos efeitos desses factores do modelo. Assim, SQA poderia ser definida como a diferença entre a Soma de Quadrados Residual do modelo com o único Factor *Enzima* (Factor B) e a Soma de Quadrados Residual do modelo a dois factores, sem interacção: $SQA = SQRE_B - SQRE_{A+B} = 0.7734292 - 0.7495771 = 0.0238521$. A Soma

de Quadrados associada ao Factor B seria agora a Soma de Quadrados do ajustamento (SQF) do modelo apenas com o factor B (*Enzima*): $SQB = SQF_B = 0.1376021$. Esta alternativa produz os mesmos valores para SQA e SQB do que a opção anterior, reflectindo a total simetria do papel de ambos os factores no estudo do modelo. De novo, previne-se que se trata duma característica de delineamentos *equilibrados*. Caso o delineamento não fosse equilibrado, uma ou outra opção produziriam valores diferentes para SQA e para SQB . Trata-se de mais uma razão que aconselha a utilização de delineamentos equilibrados.

15. (a) Pedese para mostrar que a soma dos n_i resíduos e_{ij} , correspondentes ao nível i do Factor ($i = 1, 2, \dots, k$), numa ANOVA a 1 Factor, é nula. Sabemos que, neste tipo de delineamento, os valores ajustados de cada observação correspondem à média amostral das n_i observações no nível i do Factor em que essa observação foi efectuada. Assim,

$$\sum_{j=1}^{n_i} e_{ij} = \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij}) = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.}) = 0,$$

uma vez que se trata duma soma de desvios dum conjunto de observações em relação à sua média (ou seja, do tipo $\sum_{i=1}^n (x_i - \bar{x})$, estudada no Exercício 3a da Regressão Linear Simples) que tem sempre soma zero.

- (b) Trata-se duma situação análoga à da alínea anterior. Num modelo ANOVA a dois factores, com efeitos de interacção, sabemos que os valores ajustados \hat{y}_{ijk} correspondem às médias $\bar{y}_{ij.}$ das observações da célula da referida observação. Assim, a soma dos resíduos das n_{ij} observações efectuadas na célula (i, j) é dada por:

$$\sum_{k=1}^{n_{ij}} e_{ijk} = \sum_{k=1}^{n_{ij}} (y_{ijk} - \hat{y}_{ijk}) = \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij.}) = 0.$$

16. Está-se no contexto dum modelo ANOVA a 1 Factor, onde as observações Y_{ij} constituem n variáveis aleatórias independentes, todas com distribuição $Y_{ij} \cap \mathcal{N}(\mu_1 + \alpha_i, \sigma^2)$.

- (a) Sabemos que neste modelo, os estimadores dos parâmetros μ_1 e $\alpha_i = \mu_i - \mu_1$ são dados pelas correspondentes quantidades amostrais.

- o estimador da média populacional do primeiro nível, μ_1 , é dado pela média amostral das observações desse nível, $\bar{Y}_{1.} = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}$. Mas, como é sabido (ver apontamentos da UC de Estatística, dos primeiros ciclos do ISA), a média \bar{X} duma amostra aleatória $\{X_i\}_{i=1}^n$ de n variáveis aleatórias com distribuição $\mathcal{N}(\mu, \sigma^2)$, tem distribuição $\bar{X} \cap \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Assim, e tendo em conta que $\alpha_1 = 0$, tem-se $Y_{1j} \cap \mathcal{N}(\mu_1, \sigma^2)$ e $\hat{\mu}_1 = \bar{Y}_{1.} \cap \mathcal{N}(\mu_1, \frac{\sigma^2}{n_1})$, como se quer mostrar.
- O estimador de $\alpha_i = \mu_i - \mu_1$, para $i > 1$, é dado pela correspondente diferença de médias amostrais, $\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{1.}$. Viu-se na alínea anterior que a segunda parcela tem distribuição $\mathcal{N}(\mu_1, \frac{\sigma^2}{n_1})$. Por um raciocínio análogo, a primeira parcela tem distribuição $\bar{Y}_{i.} \cap \mathcal{N}(\mu_1 + \alpha_i, \frac{\sigma^2}{n_i})$. As duas parcelas são independentes, uma vez que as parcelas que entram para o cálculo da média $\bar{Y}_{1.}$ são diferentes das que entram no cálculo da média $\bar{Y}_{i.}$. Logo, essa diferença de duas variáveis aleatórias Normais independentes tem distribuição Normal. Os parâmetros dessa distribuição são: $E[\hat{\alpha}_i] = E[\bar{Y}_{i.} - \bar{Y}_{1.}] =$

$$E[\bar{Y}_{i.}] - E[\bar{Y}_{1.}] = (\mu_1 + \alpha_i) - \mu_1 = \alpha_i; \text{ e } V[\hat{\alpha}_i] = V[\bar{Y}_{i.} - \bar{Y}_{1.}] = V[\bar{Y}_{i.}] + V[\bar{Y}_{1.}] - 2 \underbrace{Cov[\bar{Y}_{i.}, \bar{Y}_{1.}]}_{=0} = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_1} \text{ (a covariância é nula, tendo em conta a independência de duas$$

médias de nível diferentes). Logo, $\hat{\alpha}_i \cap \mathcal{N}\left(\alpha_i, \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_1}\right)\right)$, como se queria mostrar.

(b) Consideremos primeiro o caso de μ_1 .

- Da distribuição de $\hat{\mu}_1$ obtida na alínea anterior vem: $Z = \frac{\hat{\mu}_1 - \mu_1}{\sqrt{\sigma^2/n_1}} \cap \mathcal{N}(0, 1)$;
- Sabemos que, para qualquer modelo linear, a razão entre a Soma de Quadrados Residual e a variância comum a todos os erros aleatórios, σ^2 , tem distribuição χ^2 com os graus de liberdade associados a *SQRE*. No contexto duma ANOVA a um factor, tem-se assim: $W = \frac{SQRE}{\sigma^2} \cap \chi_{n-k}^2$;
- Em qualquer Modelo Linear, *SQRE* é independente dos parâmetros estimados, logo W e Z são independentes.
- Como sabemos da UC de Estatística dos primeiros ciclos do ISA, uma distribuição t -Student surge de tomar o quociente duma Normal reduzida e a raiz quadrada dum χ^2 (independente da Normal) sobre os seus graus de liberdade (esta última constante é também o parâmetro da distribuição t -Student). Logo, $\frac{Z}{\sqrt{W/(n-k)}} = \frac{\hat{\mu}_1 - \mu_1}{\sqrt{\frac{QMRE}{n_1}}} \cap t_{n-k}$.

Este último resultado é o ponto de partida para a construção dum intervalo a $(1 - \alpha) \times 100\%$ de confiança para o parâmetro μ_1 . Designando (como de costume) por $t_{\alpha/2(n-k)}$ o valor que, numa distribuição t -Student com $n - k$ graus de liberdade, deixa à sua direita uma região de probabilidade $\frac{\alpha}{2}$, temos

$$\begin{aligned} P \left[-t_{\alpha/2(n-k)} < \frac{\hat{\mu}_1 - \mu_1}{\sqrt{\frac{QMRE}{n_1}}} < t_{\alpha/2(n-k)} \right] &= 1 - \alpha \\ \Leftrightarrow P \left[-t_{\alpha/2(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} < \hat{\mu}_1 - \mu_1 < t_{\alpha/2(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} \right] &= 1 - \alpha \\ \Leftrightarrow P \left[t_{\alpha/2(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} > \mu_1 - \hat{\mu}_1 > -t_{\alpha/2(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} \right] &= 1 - \alpha \\ \Leftrightarrow P \left[\hat{\mu}_1 - t_{\alpha/2(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} < \mu_1 < \hat{\mu}_1 + t_{\alpha/2(n-k)} \cdot \sqrt{\frac{QMRE}{n_1}} \right] &= 1 - \alpha \end{aligned}$$

Calculando os extremos deste intervalo de probabilidade para a nossa amostra (e recordando que $\hat{\mu}_1 = \bar{Y}_{1.}$) obtemos o intervalo de confiança referido no enunciado.

Para obter um intervalo de confiança para α_i , segue-se um raciocínio em tudo análogo ao acabado de referir, mas partindo da distribuição para $\hat{\alpha}_i$ obtida na alínea anterior. Agora,

- $Z = \frac{\hat{\alpha}_i - \alpha_i}{\sqrt{\sigma^2\left(\frac{1}{n_i} + \frac{1}{n_1}\right)}} \cap \mathcal{N}(0, 1)$;
- Tomando à mesma $W = \frac{SQRE}{\sigma^2} \cap \chi_{n-k}^2$ e repetindo o raciocínio anterior, obtém-se $\frac{Z}{\sqrt{W/(n-k)}} = \frac{\hat{\alpha}_i - \alpha_i}{\sqrt{QMRE\left(\frac{1}{n_i} + \frac{1}{n_1}\right)}} \cap t_{n-k}$.

A dedução do intervalo de confiança para α_i é também em tudo análoga ao que foi feita no caso de μ_1 , substituindo μ_1 por α_i , $\hat{\mu}_1$ por $\hat{\alpha}_i$ e $\sqrt{\frac{QMRE}{n_1}}$ por $\sqrt{QMRE\left(\frac{1}{n_i} + \frac{1}{n_1}\right)}$.