INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2015/16

Resoluções de (quase todos os) Exercícios de Análise de Variância

1. Pretende-se modelar a variável resposta numérica `concentracao`, tendo como variável explicativa apenas os quatro diferentes laboratórios.

   (a) Estamos perante um delineamento a um factor (`laboratorio`), com $k = 4$ níveis (os 4 laboratórios). Para cada nível há $n_i = 6$ observações, e sendo este número igual para todos os laboratórios estamos perante um delineamento equilibrado. O Modelo ANOVA a 1 factor correspondente é:

   i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}, \ \forall\, i = 1, 2, 3, 4 \ , \ j = 1, 2, ..., 6 \ , \quad$ com $\alpha_1 = 0$, onde
      - Y_{ij} indica a concentração do produto químico para a j-ésima repetição observada no i-ésimo laboratório;
      - μ_1 indica a concentração média no primeiro laboratório ($i = 1$);
      - α_i indica o efeito (acréscimo em relação à média do primeiro laboratório) associado ao i-ésimo laboratório; 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. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

   (b) O quadro-resumo da ANOVA tem a seguinte estrutura:

Fonte	g.l.	SQ	QM	F_{calc}
Factor	$k-1$	$SQF = \sum_{i=1}^{k} n_i \cdot (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$QMF = \frac{SQF}{k-1}$	$\frac{QMF}{QMRE}$
Resíduos	$n-k$	$SQRE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2$	$QMRE = \frac{SQRE}{n-k}$	
Total	$n-1$	$SQT = (n-1)\, s_y^2$	–	–

   No nosso caso, $k = 4$; $n_i = 6 = n_c$ $(\forall\, i)$; $n = n_c \times k = 24$;

   $SQRE = (n_c - 1) \sum_{i=1}^{k} s_i^2 = 5 \times (4.1507 + 19.4750 + 1.1200 + 1.5867) = 131.662$; e

$$\begin{aligned} SQF &= n_c \cdot \sum_{i=1}^{k} (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2 \\ &= 6\,[(52.3333 - 49.5375)^2 + (49.35 - 49.5375)^2 + (46.7 - 49.5375)^2 + (49.7667 - 49.5375)^2] \\ &= 95.73356. \end{aligned}$$

   Alternativamente, seria possível calcular SQT a partir da variância amostral da totalidade das observações da variável resposta ($SQT = (n-1)\, s_y^2 = 23 \times 9.8868 = 227.3964$) e subtrair-lhe uma das Somas de Quadrados anteriores para obter a outra. As pequenas diferenças nos valores obtidos por estas duas abordagens resultam dos erros de arredondamento nos valores das médias e variâncias de nível dados no enunciado.

Assim, $QMRE = \frac{SQRE}{n-k} = \frac{131.662}{20} = 6.5831$ e $QMF = \frac{SQF}{k-1} = \frac{95.73356}{3} = 31.91119$.
Finalmente, $F_{calc} = \frac{QMF}{QMRE} = \frac{31.91119}{6.5831} = 4.847441$.
Os valores obtidos podem ser confirmados (a menos de erros resultantes dos arredondamentos com que são apresentadas no enunciado as médias e variâncias de nível), utilizando os dados disponíveis na *data frame* `toxicos` e os comandos do `R`.

```
> summary(aov(concentracao ~ laboratorio, data=toxicos))
            Df Sum Sq Mean Sq F value Pr(>F)
laboratorio  3  95.73   31.91   4.848 0.0108 *
Residuals   20 131.66    6.58
```

(c) Pede-se um teste F, que neste contexto significa perguntar se se deve admitir a igualdade das concentrações médias nos quatro laboratórios (H_0) ou se opta pela hipótese alternativa (H_1). Mais concretamente:

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 $F_{calc} > f_{0.05(3,20)} = 3.10$.
Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 4.848$. 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, de que será necessário verificar a uniformidade dos protocolos de análise dos laboratórios.

(d) A alteração do nível de significância não implica alterações nos dois primeiros passos do teste. Quanto aos restantes,

Nível de significância: $\alpha = 0.01$.
Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.01(3,20)} = 4.94$.
Conclusões: O valor da estatística do teste não depende do nível de significância e continua a ser $F_{calc} = 4.848$. Mas ao nível $\alpha = 0.01$ este já não é um valor da região crítica, pelo que a esse nível não se rejeita a hipótese nula. O facto de a conclusão mudar entre os níveis de significância $\alpha = 0.05$ e $\alpha = 0.01$ significa que o valor de prova (*p-value*) do valor $F_{calc} = 4.848$ estará entre esses dois níveis, isto é, $0.01 < p < 0.05$, facto que se confirma no quadro-resumo produzido pelo `R`.

(e) A matriz do modelo, $\mathbf{X}$, será constituída por quatro colunas: uma coluna de $n = 24$ uns e as colunas indicatrizes dos segundo, terceiro e quarto níveis do factor($\mathcal{I}_2$, $\mathcal{I}_3$ e $\mathcal{I}_4$), como se pode confirmar através do comando referido no enunciado:

```
> toxicos.aov <- aov(concentracao ~ laboratorio, data=toxicos)
> model.matrix(toxicos.aov)
   (Intercept) laboratorio2 laboratorio3 laboratorio4
1            1            0            0            0
2            1            0            0            0
3            1            0            0            0
4            1            0            0            0
5            1            0            0            0
6            1            0            0            0
7            1            1            0            0
8            1            1            0            0
9            1            1            0            0
10           1            1            0            0
11           1            1            0            0
```

```
12          1           1           0           0
13          1           0           1           0
14          1           0           1           0
15          1           0           1           0
16          1           0           1           0
17          1           0           1           0
18          1           0           1           0
19          1           0           0           1
20          1           0           0           1
21          1           0           0           1
22          1           0           0           1
23          1           0           0           1
24          1           0           0           1
```

(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(toxicos.aov)
       1        2        3        4        5        6        7        8
52.33333 52.33333 52.33333 52.33333 52.33333 52.33333 49.35000 49.35000
       9       10       11       12       13       14       15       16
49.35000 49.35000 49.35000 49.35000 46.70000 46.70000 46.70000 46.70000
      17       18       19       20       21       22       23       24
46.70000 46.70000 49.76667 49.76667 49.76667 49.76667 49.76667 49.76667
```

As médias aqui indicadas são as que também eram dadas (arredondadas a duas casas decimais) na penúltima linha da tabela do enunciado.

(g) O facto dos resíduos se encontrarem empilhados em quatro colunas é o reflexo natural do facto, referida na alínea anterior, que há apenas quatro diferentes valores ajustados nesta ANOVA: as quatro médias amostrais de cada laboratório. Assim, apenas há quatro diferentes valores no eixo horizontal, a que correspondem os valores ajustados $\hat{y}_{ij} = \overline{y}_{i.}$. Do gráfico não parecem surgir indicações de grandes diferenças na variância dos resíduos em cada nível, excepção feita para a segunda coluna, onde surge um resíduo atípico, de valor inferior a -8. É possível identificar o laboratório a que se refere essa observação, uma vez que a média amostral correspondente excede de pouco o valor 49: trata-se do laboratório 2 (cuja média amostral é 49.35). Em particular, trata-se da segunda observação nesse laboratório, cujo valor (40.5) é inferior em mais de oito unidades ao valor médio do laboratório. Assim, a observação a que corresponde o referido resíduo é a observação y_{22}.

2. Neste exercício sobre os grãos de café em Angola, não existem os dados originais, sendo apenas conhecida a tabela do enunciado, com as médias e variâncias amostrais de cada região.

(a) A variável resposta Y é a percentagem do peso total de grãos sem defeito. Para explicar eventuais diferenças nos valores médios populacionais desta variável, apenas se dispõe de um factor: o factor região, com $k = 6$ níveis (as seis regiões indicadas no enunciado). O modelo ANOVA correspondente é assim o modelo a um factor, semelhante ao do primeiro exercício, mas em que agora o factor tem $k = 6$ níveis, existindo ao todo $n = 66$ observações repartidas de forma equilibrada pelos seis níveis: $n_i = 11$, para qualquer $i = 1, 2, ..., 6$.

i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall\, i = 1, 2, 3, 4, 5, 6$, $j = 1, 2, ..., 11$, com $\alpha_1 = 0$, onde

- Y_{ij} indica a percentagem do peso de grãos sem defeito, no j-ésimo lote observado na região i;
- μ_1 indica a percentagem média de peso de grãos sem defeito na primeira região ($i = 1$) que, na ordem da tabela, é a região de Cabinda;

- α_i indica o efeito (acréscimo em relação à média do Cabinda) da região 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. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

(b) Comecemos pelo cálculo das Somas de Quadrados. Tendo em conta as fórmulas vistas nas aulas teóricas e os valores dados no enunciado, temos:

$$SQRE = (n_c - 1) \sum_{i=1}^{6} s_i^2 = 10 \times (48.1636 + ... + 454.1161) = 18326.71 \ ;$$

$$SQF = n_c \sum_{i=1}^{6} (\overline{y}_{i.} - \overline{y}_{..})^2 = 11 * ((44.19 - 53.25667)^2 + ... + (42.11 - 53.25667)^2)$$
$$= 4068.939 \ ,$$

sendo necessário, para obter SQF, calcular primeiro a média geral da totalidade das $n = 66$ observações, que (uma vez que o delineamento é equilibrado) é a média simples das $k = 6$ médias regionais: $\overline{y}_{..} = (44.19 + 58.87 + ... + 42.11)/6 = 53.25667$. Logo, tem-se a seguinte tabela-resumo:

Fonte	g.l.	SQ	QM	F_{calc}
Factor	$k-1 = 5$	$SQF = 4068.939$	$QMF = \frac{SQF}{k-1} = 813.7878$	$\frac{QMF}{QMRE} = 2.6643$
Resíduos	$n-k = 60$	$SQRE = 18326.71$	$QMRE = \frac{SQRE}{n-k} = 305.4451$	

(c) Neste caso, e uma vez que não são conhecidas as variáveis originais, apenas é possível calcular a variância da totalidade das $n = 66$ observações recorrendo à decomposição da Soma de Quadrados Total correspondente a esta ANOVA:

$$s_y^2 = \frac{SQT}{n-1} = \frac{SQF + SQRE}{n-1} = \frac{4068.939 + 18326.71}{65} = \frac{22395.65}{65} = 344.55 \ .$$

Repare-se que este valor *não é* a média das variâncias amostrais de nível.

(d) Embora se possa escrever as hipóteses do teste com base nos efeitos α_i do factor (como se fez no exercício anterior), nas ANOVAs a um único factor é equivalente formular as hipóteses em termos das médias populacionais (valores esperados das observações $E[Y_{ij}] = \mu_i = \mu_1 + \alpha_i$) em cada nível do factor. Eis o teste com $\alpha = 0.05$:

Hipóteses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ vs. $H_1 : \exists\, i, i'$ tal que $\mu_i \neq \mu_{i'}$.

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 $F_{calc} > f_{0.05(5,60)} = 2.37$.

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 2.664$. 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, de que a percentagem média dos pesos de grãos defeituosos não é igual em todas as regiões.

No caso de se utilizar um nível de significância $\alpha = 0.01$, apenas muda a fronteira da região crítica, que passa a ser $f_{0.01(5,60)} = 3.34$. Assim, a estatística calculada (que continua a ser $F_{calc} = 2.664$) já não é significativa a este novo nível de significância, não sendo agora possível rejeitar a hipótese de iguais médias populacionais nas seis regiões.

O valor de prova associado à estatística calculada é (tendo em conta a natureza unilateral direita do teste) $P[F_{(5,60)} > F_{calc}] = P[F_{(5,60)} > 2.664]$. Não é possível obter este valor nas tabelas (embora já saibamos que ele se encontra entre 5% e 1%, uma vez que a conclusão dos testes muda para esses níveis de significância), mas pode calcular-se essa probabilidade com o auxílio do **R**:

```
> 1-pf(2.664, 5, 60)
[1] 0.03063001
```

Assim, tem-se $p = 0.03063$.

(e) Sabemos que duas médias de nível μ_i e $\mu_{i'}$ devem ser consideradas diferentes caso as respectivas médias amostrais difiram (em módulo) por mais do que o termo de comparação $q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}}$, onde $q_{\alpha(k,n-k)}$ corresponde ao valor que deixa à sua direita uma região de probabilidade α numa distribuição de Tukey de parâmetros k e $n-k$, e n_c indica o número comum de observações em cada nível do factor (o resultado que sustenta o teste de Tukey parte do pressuposto que o delineamento é equilibrado). No nosso caso tem-se $k = 6$ e $n = 66$. Trabalhando (como pedido no enunciado) com $\alpha = 0.05$, e recorrendo às tabelas da distribuição de Tukey (tabelas específicas, disponíveis na página *web* da disciplina), tem-se $q_{0.05(6,60)} = 4.16$. Um valor mais preciso pode ser obtido através do comando `qtukey` do **R**:

```
> qtukey(0.95, 6, 60)
[1] 4.163161
```

Sabemos pela alínea (b) que $QMRE = 305.4451$ e também que $n_c = 11$. Logo, o termo de comparação é dado por $q_{\alpha(k,n-k)}\sqrt{\frac{QMRE}{n_c}} = 4.16 \times \sqrt{\frac{305.4451}{11}} = 21.9212$. Trata-se dum valor elevado e por uma inspecção simples das médias amostrais de nível vemos que a maior diferença entre médias amostrais de nível é a diferença entre a média do Libolo e de Amboim: $|61.96 - 42.11| = 19.85 < 21.9212$. Assim, nenhum par de médias tem diferença maior que o termo de comparação, pelo que se admite a igualdade de todos os pares de médias (logo, a igualdade de todas as médias). Este resultado é contraditório com o resultado do teste F ao nível $\alpha = 0.05$, o que pode acontecer quando se usam duas ferramentes baseadas em teoria diferente (testes F e de Tukey). No entanto, as várias conclusões desses testes estão próximas da fronteira, pelo que a discrepância de resultados não é assim tão surpreendente.

(f) É pedido um teste de Bartlett à homogeneidade das variâncias de nível. Representando por σ_i^2 a variância populacional na região i, tem-se:

Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2$ vs. $H_1 : \exists\, i, j$ tais que $\sigma_i^2 \neq \sigma_j^2$.

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

Pode admitir-se a validade da distribuição assintótica, uma vez que em todos os níveis do factor há mais do que 5 observações (tem-se $n_i = 11$ para todos os níveis).

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $K^2_{calc} > \chi^2_{0.05(5)} = 11.0705$.

Conclusões: O valor calculado, $K^2_{calc} = 15.53$, é superior à fronteira da região crítica, logo significativo ao nível $\alpha = 0.05$. Rejeita-se H_0, o que significa concluir que há variabilidades diferentes nas várias regiões, ou por outras palavras, que as diferenças observadas nas variâncias amostrais de região são significativas ao nível $\alpha = 0.05$. Assim, podem-se levantar dúvidas relativamente à validade da ANOVA efectuada.
Nota: Havendo problemas com os pressupostos do modelo, pode ser preferível utilizar uma variante não-paramétrica da ANOVA a um factor: o teste de Kruskal-Wallis (não incluído no programa desta disciplina).

3. A variável resposta Y é, neste caso, a variação de massa (coluna `variacao.massa` na *data frame*). Existem ao todo $n = 50$ observações.

   (a) Para estudar este problema através duma ANOVA, ignora-se os valores numéricos das concentrações de dióxido de carbono, tratando cada diferente concentração apenas como um diferente tratamento. Assim, o factor CO_2 terá $k = 5$ níveis, havendo $(n_i = 10 = n_c)$ observações para cada concentração de CO_2 (nível do factor). O modelo ANOVA associado a este delineamento é o seguinte:

      i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}, \ \forall\, i = 1, 2, 3, 4, 5 \ , \ j = 1, 2, ..., 10 \ , \quad$ com $\alpha_1 = 0$, onde
         - Y_{ij} indica a variação de massa para a j-ésima repetição associada à i-ésima concentração de CO_2;
         - μ_1 indica o variação de massa média (populacional) na ausência de CO_2 $(i = 1)$;
         - α_i indica o efeito (acréscimo em relação à média populacional do primeiro nível) da i-ésima concentração de dióxido de carbono, isto é, $\alpha_i = \mu_i - \mu_1$; 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. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.

   (b) Vamos construir a tabela-resumo da ANOVA com o auxílio do R, uma vez que os dados estão disponíveis na *data frame* `CO2`, com os valores da variável resposta na coluna `variacao.massa` e os diferentes níveis de CO_2 no factor `CO2.factor` (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 $k - 1 = 4$ e os residuais $n - k = 45$):

```
> summary(aov(variacao.massa ~ CO2.factor, data=CO2))
            Df Sum Sq Mean Sq F value Pr(>F)
CO2.factor   4  11274  2818.6   101.6 <2e-16 ***
Residuals   45   1248    27.7
```

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

   Hipóteses: $H_0 : \alpha_i = 0 \ , \ \forall \ i = 2, 3, 4, 5$ vs. $H_1 : \exists\, i = 2, 3, 4, 5$ 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 $F_{calc} > f_{0.05(4,45)} \ \approx \ 2.58$.
   Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 101.6$. É um valor claramente 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 de CO_2 estão associadas a diferentes variações médias na massa das culturas do *Pseudomonas fragi*.

(c) Pede-se para comparar as médias amostrais de grupos, a fim de determinar quais as que são significativamente diferentes, ou seja, que levam a concluir que as correspondentes médias populacionais de nível são diferentes. Vamos responder através de intervalos de confiança de Tukey. Sabemos que o intervalo para a diferença de médias populacionais de qualquer par (i, j) de níveis, ou seja, para $\mu_i - \mu_j$, tem a seguinte expressão:

$$\left] \ (\overline{y}_{i.} - \overline{y}_{j.}) \ - \ q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}} \ , \ (\overline{y}_{i.} - \overline{y}_{j.}) \ + \ q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}} \ \right[\ .$$

A semi-amplitude destes intervalos é sempre a mesma, qualquer que seja o par de níveis considerado. No nosso caso, tem-se $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{27.7}{10}} = 1.664332$. Por outro lado, o valor que na distribuição de Tukey com os parâmetros $k = 5$ e $n - k = 45$ deixa à sua direita uma gama de valores de probabilidade $\alpha = 0.05$ é $q_{0.05(5,45)} \approx 4.02$. Assim, a semi-amplitude comum a todos os intervalos é $4.02 \times 1.664332 = 6.691$.

No caso do par de níveis $(1, 2)$, pode calcular-se a média amostral a partir dos dados indicados no enunciado: $\overline{y}_{1.} = 59.14$. De forma análoga, a média amostral no segundo nível é: $\overline{y}_{2.} = 46.04$. Assim, o intervalo a 95% de confiança para a diferenças das médias do segundo e primeiro níveis, $\mu_1 - \mu_2$, é $]\,(59.14 - 46.04) - 6.691, 13.10 + 6.691\,[=]\,6.409\,,\ 19.791\,[$. Este intervalo não inclui o valor zero, que não é assim um valor admissível para $\mu_1 - \mu_2$. Logo, rejeita-se a igualdade das variações médias na massa dos *Pseudomonas*, para as duas primeiras concentrações de dióxido de carbono.

Para construir os restantes intervalos de confiança, utilizar-se-á o comando `TukeyHSD` do `R`. Repare-se que, por convenção, o `R` opta por considerar ICs para diferenças $\mu_i - \mu_j$ onde $i > j$, pelo que o intervalo correspondente ao que se acabou de calcular será o intervalo para a diferença $\mu_2 - \mu_1$, com a correspondente alteração de sinais. Repare-se ainda no problema dos erros de arredondamento, que resultam também da utilização nos cálculos anteriores do valor de $QMRE$ na tabela-resumo (arredondado a uma casa decimal: 27.7).

```
> TukeyHSD(aov(variacao.massa ~ CO2.factor, data=CO2))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = variacao.massa ~ CO2.factor, data = CO2)
$CO2.factor
             diff      lwr        upr     p adj
0.083-0    -13.10 -19.7921  -6.407896 0.0000133
0.29-0     -22.69 -29.3821 -15.997896 0.0000000
0.5-0      -33.67 -40.3621 -26.977896 0.0000000
0.86-0     -42.70 -49.3921 -36.007896 0.0000000
0.29-0.083  -9.59 -16.2821  -2.897896 0.0016698
0.5-0.083  -20.57 -27.2621 -13.877896 0.0000000
0.86-0.083 -29.60 -36.2921 -22.907896 0.0000000
0.5-0.29   -10.98 -17.6721  -4.287896 0.0002615
0.86-0.29  -20.01 -26.7021 -13.317896 0.0000000
0.86-0.5    -9.03 -15.7221  -2.337896 0.0034105
```

Todas as restantes comparações de pares de médias de nível (ao todo há $C_2^5 = 10$ pares de níveis) produzem resultados semelhantes: nenhum intervalo de confiança para $\mu_i - \mu_j$ contém o valor zero. Assim, conclui-se que a variação média de massa é sempre diferente nas cinco concentrações de CO_2 estudadas. As cinco médias amostrais de nível, que estão na base desta conclusão, podem ser obtidas através do seguinte comando do `R`:

```
> CO2.aov <- aov(variacao.massa ~ CO2.factor, data=CO2)
```

```
> model.tables(CO2.aov, type="means")
Tables of means
Grand mean
36.708
 CO2.factor
CO2.factor
    0 0.083  0.29   0.5  0.86
59.14 46.04 36.45 25.47 16.44
```

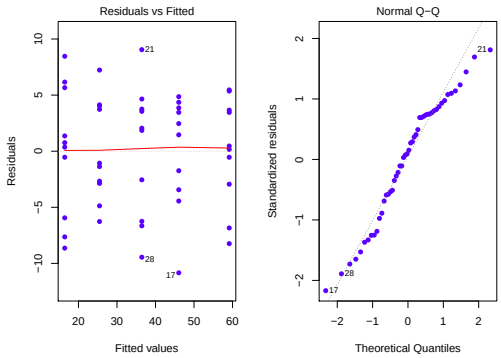
Neste caso pode afirmar-se que as diferenças entre estas médias amostrais são significativas, ou seja, permitem (ao nível de confiança global 95% que é, por omissão, usado pelo R na construção dos intervalos de confiança de Tukey) afirmar que reflectem diferenças nas correspondentes médias populacionais de nível.

(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 duma dada observação corresponde à média amostral das observações no mesmo nível do factor: $\hat{y}_{ij} = \overline{y}_{i.}$. Assim, todas as observações do primeiro grupo têm valor ajustado igual a $\hat{y}_{1j} = \overline{y}_{1.} = 59.14$. O resíduo da primeira observação do primeiro grupo será $e_{11} = 62.6 - 59.14 = 3.46$ e o da segunda observação desse grupo é $e_{12} = 59.6 - 59.14 = 0.46$. De forma análoga, os valores ajustados de qualquer observação no segundo grupo são dados por $\hat{y}_{2j} = \overline{y}_{2.} = 46.04$. O resíduo da terceira observação do segundo grupo é assim $e_{23} = y_{23} - \overline{y}_{2.} = 47.5 - 46.04 = 1.46$. Para calcular a totalidade dos resíduos podemos recorrer ao R (arredondando a três casas decimais):

```
> round(residuals(CO2.aov), d=3)
    1     2     3     4     5     6     7     8     9    10    11    12    13
 3.46  0.46  5.36  0.16 -0.54  5.46 -8.24 -2.94 -6.84  3.66  4.86 -1.74  1.46
   14    15    16    17    18    19    20    21    22    23    24    25    26
 3.46  2.46  4.36 -10.84  3.86 -3.44 -4.44  9.05  4.65 -6.65  1.85  3.75  2.05
   27    28    29    30    31    32    33    34    35    36    37    38    39
-6.25 -9.45  3.55 -2.55  4.03 -2.67 -6.27 -4.87  3.73 -1.37 -2.87  7.23 -1.07
   40    41    42    43    44    45    46    47    48    49    50
 4.13  8.46  0.76 -8.64 -5.94  1.36  5.66  6.16  0.36 -0.54 -7.64
```

Com o auxílio do R, podemos obter os dois gráficos de resíduos já considerados no estudo dos modelos de Regressão Linear, através do comando:

```
> plot(CO2.aov, which=c(1,2), pch=16, col="blue")
```



O gráfico da esquerda 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 $\hat{y}_{ij} = \overline{y}_{i.}$, logo, a mesma coordenada no eixo horizontal. Neste caso, observam-se $k = 5$ colunas. Não parece existir problema com a hipótese de homogeneidade das variâncias, uma vez que a variabilidade dos resíduos não parece diferir muito nos cinco níveis do factor. Será, no entanto, conveniente efectuar um teste de Bartlett à homogeneidade das variâncias para confirmar esta conclusão. A utilização desse teste parece adequada, uma vez que o *qq-plot* (gráfico à direita) não indicia problemas graves com a Normalidade, dada a disposição aproximadamente linear dos pontos.

Os restantes diagnósticos que foram considerados aquando do estudo da regressão (distâncias de Cook, efeito alavanca) são geralmente de menor utilidade no contexto duma ANOVA. Em relação às distâncias de Cook, por exemplo, sabe-se de antemão qual o efeito de retirar uma observação: além de desequilibrar um delineamento equilibrado, afectará a média das observações no mesmo nível do factor (ou seja, os valores ajustados $\hat{y}$ nesse nível). Assim valores elevados da distância de Cook correspondem a observações atípicas (*outliers*) no seio dum dado nível. Mas para identificar tais observações, basta o gráfico usual de resíduos contra $\hat{y}$, não sendo necessário um diagnóstico específico. Em relação aos efeitos alavanca, 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 $\frac{1}{10}$). O gráfico obtido no `R` com a opção `which=5` tinha, na regressão linear, os valores do efeito alavanca (h_{ii}, ou *leverages*) de cada observação no eixo horizontal. No entanto, para ANOVAs com 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 $\overline{y}_{i.}$), uma vez que um gráfico análogo ao construído na regressão linear apenas empilharia todos os resíduos numa única coluna. O gráfico alternativo produzido pelo R quando os delineamentos são equilibrados fica assim semelhante ao primeiro gráfico de resíduos, embora sem qualquer efeito de escala no eixo horizontal e com os resíduos (internamente) estandardizados no eixo vertical, em vez dos resíduos usuais.

Completemos a alínea efectuando o teste de Bartlett. O facto de haver $n_c = 10 > 5$ 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 concentração de CO_2), temos:

Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$ vs. $H_1 : \exists\, i, j$ tais que $\sigma_i^2 \neq \sigma_j^2$.

Estatística do teste: $K^2 = \dfrac{(n-k)\ln QMRE - \sum_{i=1}^{k}(n_i-1)\ln S_i^2}{C} \sim \chi^2_{k-1}$, 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 $K^2_{calc} > \chi^2_{0.05(4)} = 9.488$.

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(variacao.massa ~ CO2.factor, data=CO2)
Bartlett test of homogeneity of variances
```

```
data:  variacao.massa by CO2.factor
Bartlett's K-squared = 1.0701, df = 4, p-value = 0.899
```

O valor calculado, $K^2_{calc} = 1.0701$, é claramente não significativo (o que também se depreende do *p-value* muito elevado). Não se rejeita H_0, pelo que se conclui pela inexistência de variabilidades diferentes nos vários tratamentos.

Assim, não parece haver problemas com a validade da ANOVA efectuada.

(e) Nesta alínea pede-se para aproveitar os valores das concentrações de CO_2 utilizadas, e tratar essa variável preditora como uma variável numérica, estudando a regressão linear simples de `variacao.massa` sobre `CO2.numerico`.

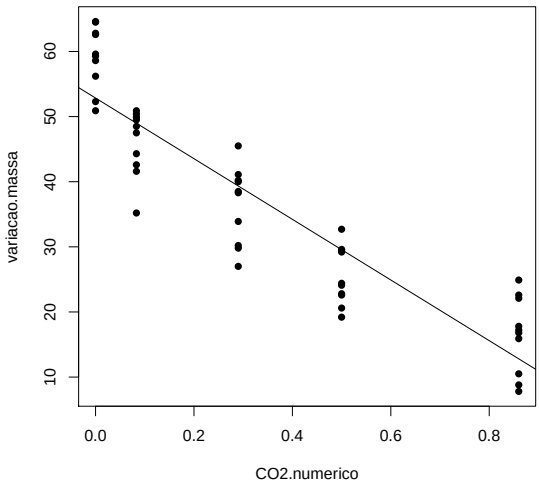
i. O gráfico pedido pode ser construído com o seguinte comando do `R`. O resultado é mostrado na alínea seguinte.

```
> plot(variacao.massa ~ CO2.numerico, data=CO2, pch=16)
```

ii. A regressão linear pedida é dada por:

```
> CO2.lm <- lm(variacao.massa ~ CO2.numerico, data=CO2)
> summary(CO2.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    52.849      1.408   37.52   <2e-16 ***
CO2.numerico  -46.569      3.030  -15.37   <2e-16 ***
---
Residual standard error: 6.637 on 48 degrees of freedom
Multiple R-squared: 0.8312,Adjusted R-squared: 0.8276
F-statistic: 236.3 on 1 and 48 DF,  p-value: < 2.2e-16
```

A nuvem de pontos pedida na alínea anterior, já com a recta de regressão (traçada com o comando `abline(CO2.lm)`) é:



Apesar de alguma tendência para uma relação curvilinear, uma regressão linear simples pode constituir uma modelação aproximada da relação entre concentrações de dióxido de carbono e variação na massa das culturas de *Pseudomonas fragi* (repare-se como seria impossível tirar esta ilação se o número de níveis fosse mais pequeno, *e.g.*, $k = 3$). O valor do coeficiente de determinação é claramente significativo ($p < 2.2 \times 10^{-16}$) e bastante elevado ($R^2 = 0.8312$), explicando mais de 83% da variabilidade total observada na variável resposta.

iii. Os testes F de ajustamento global do contexto regressão linear simples e do contexto ANOVA a um factor, não são os mesmos. Como se viu nas aulas teóricas, a ANOVA a um factor pode ser vista como uma espécie de regressão linear múltipla em que as variáveis preditoras são as indicatrizes dos níveis (excepto o primeiro) do factor. Assim, a informação disponível para prever os valores da variável resposta é, no caso da regressão considerada nesta alínea, a variável `CO2.numerico`, com valores numéricos diferentes em cada nível (mas repetidos para as observações dum mesmo nível). No caso da ANOVA a um factor, é o conjunto das indicatrizes de nível e o vector dos n uns. Sendo diferente a informação preditora, serão diferentes os valores ajustados e os valores dos respectivos F_{calc} e coeficientes de determinação. Em relação a este último, e embora não seja hábito utilizá-lo no contexto duma ANOVA a um factor, o seu valor é aqui $R^2 = 0.9003$, superior ao que se obteve na regressão ($R^2 = 0.8312$), como se pode constatar através do ajustamento obtido utilizando simultaneamente o comando `lm` e o factor preditor `CO2.factor`:

```
> summary(lm(variacao.massa ~ CO2.factor, data=CO2))
(...)
Residual standard error: 5.266 on 45 degrees of freedom
Multiple R-squared: 0.9003,Adjusted R-squared: 0.8915
F-statistic: 101.6 on 4 and 45 DF,  p-value: < 2.2e-16
```

Repare-se como o valor da estatística calculada, $F_{calc} = 101.6$, é o que foi obtido usando o comando `aov`.

Um comentário final: o modelo ANOVA não permite, ao contrário da regressão, fazer previsões sobre as variações de massa com concentrações de CO_2 não observadas na experiência, uma vez que os níveis do factor CO_2 não têm escala (são apenas categorias diferentes).

4. (a) A descrição da experiência corresponde a um delineamento factorial a dois factores, sendo o primeiro factor constituído pelas fases do processamento e o segundo factor constituido pelos diferentes lotes. Refira-se que na descrição da experiência dada nesta alínea, cada nível do 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. Neste caso, é natural supôr que a diferentes lotes de feijão correspondam diferentes concentrações de zinco, independentemente de qualquer tratamento a que sejam submetidos[1].

A *data frame* `zinco` tem três colunas: a variável resposta (`concentracao`), o factor com $a = 4$ níveis, cujos efeitos se pretende realmente estudar (`fase`) e o factor/bloco (`lote`), com $b = 9$ níveis, introduzido para controlar a heterogeneidade das unidades experimentais (lotes de feijão). Nas 36 células deste delineamento não há repetições de observações (ou

[1]Seria mais adequado supôr que ao factor `lotes` correspondem *efeitos aleatórios*, expressão usada para designar o contexto em que os níveis do factor analisados não são os únicos de interesse, mas apenas uma amostra aleatória dum número muito maior de níveis. Neste caso, não é de crer que haja interesse em estudar apenas *aqueles* nove lotes usados na experiência. Mais realista será supôr que constituem uma amostra aleatória duma infinidade de potenciais lotes de feijão. Assim, seria mais adequado associar efeitos aleatórios aos lotes, continuando a associar efeitos fixos às fases do processamento (aqui sim, existe real interesse em estudar *aqueles* quatro momentos do processamento). Um modelo onde se misturam efeitos fixos e efeitos aleatórios é conhecido por *modelo misto*, mas ultrapassa o programa desta disciplina.

seja, $n_c = 1$). Logo, independentemente de ser desejável, não é possível incluir efeitos de interacção no modelo. Utilizar-se-á um modelo a dois factores, sem interacção:

i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}$, $\forall\, i = 1, 2, 3, 4$, $j = 1, 2, ..., 9$, $k = 1$ (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 concentração de zinco da fase i, associada ao lote de feijão j;
   - μ_{11} é a concentração esperada de zinco no início do processamento, para o lote 1;
   - α_i indica o efeito da fase i;
   - β_j indica o efeito do lote 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) Recorrendo ao `R`, obtém-se a tabela-resumo correspondente a este modelo:

```
> summary(aov(concentracao ~ fase + lote, data=zinco))
            Df Sum Sq Mean Sq F value   Pr(>F)
fase         3  20.60   6.866   9.736 0.000218 ***
lote         8  17.76   2.220   3.148 0.013931 *
Residuals   24  16.92   0.705
```

Repare-se que (em comparação com a tabela do modelo a um factor) existe uma nova linha na tabela, correspondente ao novo factor. Os graus de liberdade associados a cada factor são o número de níveis desse factor, menos 1 (como reflexo da imposição das restrições $\alpha_1 = 0$ e $\beta_1 = 0$), o que neste caso significa $a - 1 = 3$ e $b - 1 = 8$ graus de liberdade. Os graus de liberdade associados ao residual são, como de costume, o número de observações menos o número de parâmetros no modelo, ou seja, $n - (a + b - 1) = 36 - (4 + 9 - 1) = 24$. Uma vez que o delineamento é equilibrado, com uma única repetição por célula ($n_c = 1$) é possível utilizar as fórmulas constantes dos acetatos das aulas teóricas (e também do formulário, uma vez que as expressões para SQA e SQB são iguais às do modelo *com* interacção, no caso de delineamentos equilibrados) para calcular as restantes quantidades da tabela. Para tal, será útil dispor das concentrações médias em cada fase e de cada lote:

```
> model.tables(aov(concentracao ~  fase + lote, data=zinco), type="means")
Tables of means
Grand mean
2.847778
 fase
fase
    1     2     3     4
2.228 2.847 2.233 4.083
 lote
lote
    1     2     3     4     5     6     7     8     9
3.483 3.733 3.558 2.998 3.425 1.940 1.858 2.195 2.443
```

Assim, e como $n_c = 1$, temos: $SQA = b\, n_c \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2 = 9 \times ((2.228 - 2.847778)^2 + (2.847 - 2.847778)^2 + (2.233 - 2.847778)^2 + (4.083 - 2.847778)^2) = 20.59066$, e $SQB = a\, n_c \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2 = 4 \times ((3.483 - 2.847778)^2 + (3.733 - 2.847778)^2 + ... + (2.443 - 2.847778)^2) = 17.76391$. Para obter a Soma de Quadrados residual, basta recordar que a Soma de Quadrados Total é o numerador da variância de todas as $n = 36$ observações. Sabendo que esta variância é:

```
> var(zinco$concentracao)
[1] 1.579458
```

pode-se deduzir que $SQT = (n-1)\,s_y^2 = 35 \times 1.579458 = 55.28102$. Logo, $SQRE = SQT - (SQA + SQB) = 55.28102 - (20.59066 + 17.76391) = 16.92645$. Os restantes valores da tabela resultam da aplicação directa das suas definições.

(c) Nesta fase apenas é pedido o teste à existência de efeitos do factor A (fases do processamento). Este teste F é indicado de seguida.

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: $F = \frac{QMA}{QMRE} \ \cap\ F_{(a-1\,,\,n-(a+b-1))}$, 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: $F_{calc} = 9.736$. 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 diferentes fases do processamento têm efeito sobre as concentrações médias de zinco.

(d) Nesta alínea, diz-se que foi ajustado um modelo apenas a um factor, o factor fases de processamento, ignorando a existência do factor (blocos) lote. O resultado obtido será:

```
> summary(aov(concentracao ~ fase , data=zinco))
            Df Sum Sq Mean Sq F value Pr(>F)
fase         3  20.60   6.866   6.334 0.0017 **
Residuals   32  34.68   1.084
```

Registem-se os seguintes factos, relativos à comparação desta tabela-resumo e da tabela-resumo do modelo a dois factores, sem interacção, ajustado nas alíneas anteriores:

- Existe uma linha comum nas duas tabelas, correspondente ao factor `fase`, e os graus de liberdade, Soma de Quadrados e Quadrado Médio do factor `fase` são idênticos aos da tabela-resumo do modelo a dois factores.
- Uma vez que a Soma de Quadrados Total é igual nos dois casos (já que $SQT = (n-1)\,s_y^2 = 35 \times 1.5795 = 55.28$ não depende do modelo ajustado) este facto tem de significar que a Soma de Quadrados Residual é aqui a soma das parcelas SQB e $SQRE$ do modelo a dois factores sem interacção. De facto, verifica-se que $SQRE_A = 34.68 = 17.76 + 16.92 = SQB + SQRE_{A+B}$. Ou seja, a não existência neste modelo de efeitos do factor B implica que a variabilidade que lhe poderia ser imputada (SQB) vai acabar por ser variabilidade residual, isto é, vai contribuir para aumentar o valor de $SQRE_A$. Neste exemplo, ao factor `lote` corresponde cerca de metade da variabilidade que é considerada residual (não explicada pelo modelo) no modelo apenas com o factor `fase`.
- Mas os graus de liberdade associados ao residual também são diferentes nos dois casos. E, mais uma vez, os graus de liberdade associados ao residual, neste modelo a um só factor, correspondem à soma dos graus de liberdade residuais e associados ao outro factor, no modelo a dois factores: $32 = 8 + 24$. Isto não acontece por acaso. Também no caso dos graus de liberdade dos modelos lineares, a soma de todas as parcelas é constante (e igual a $n-1$). Logo, a não existência, no modelo ajustado nesta alínea, de efeitos do factor `lote` significa que os graus de liberdade residuais (tal como a soma de quadrados residual) também aumentam.
- Na estatística F aos efeitos do factor `fase`, o numerador QMF (QMA, na notação para modelos a dois factores) fica igual, enquanto que o denominador $QMRE$ sofre

uma dupla transformação: o seu numerador $SQRE$ é maior do que no modelo a dois factores (pois $SQRE_A = SQRE_{A+B} + SQB$), mas também o seu denominador é maior (pois $g.l.(SQRE_{A+B}) = n-(a+b-1) < n-a = g.l.(SQRE_A)$). Assim, se a estatística F é maior, ou menor, dependerá da dimensão relativa destes aumentos do numerador e denominador.

- No exemplo em questão, o $QMRE$ do modelo com dois factores é mais baixo: 0.7052 (em vez de 1.0839 no modelo só com o factor `fase`). A estatística F no teste aos efeitos do factor `fase` (que, recorde-se, continua a ter o mesmo numerador) era $F_A = 9.7361$ no modelo a dois factores e no modelo a um factor é agora $F = 6.3343$). A rejeição da hipótese de inexistência de efeitos do Factor *fase* ($H_0 : \alpha_i = 0, \forall\, i$) era mais clara no modelo a dois factores, e embora neste caso não se altere qualitativamente a conclusão para os níveis de significância usuais, poderia dar-se esse caso.
- Caso existam realmente efeitos do novo factor, a Soma de Quadrados Residual do modelo a dois factores sem interacção, $SQRE_{A+B}$, será bastante inferior à do modelo a um factor e também $QMRE_{A+B}$ será menor, pelo que aumenta a estatística F, que tende assim a ser mais significativa. Pelo contrário, se a parcela SQB fôr relativamente pequena, pode acontecer a situação contrária, e a estatística F tornar-se menor, afastando-se assim das regiões críticas.

Conclusão: caso existam realmente efeitos dum factor adicional, que torna as unidades experimentais muito heterogeneas, a inclusão desse factor no delineamento e no modelo ANOVA contribuirá para evidenciar eventuais efeitos do outro factor, que realmente se pretende estudar. Mas no caso de ao factor adicional não corresponderem realmente efeitos importantes, a sua inclusão no delineamento e no modelo poderá até contribuir para camuflar eventuais efeitos do factor no qual estamos realmente interessados.

5. Trata-se dum delineamento factorial a dois factores (`terreno` e `variedade`), mas com uma única observação em cada célula (em cada terreno, apenas há uma parcela com cada variedade). Logo, só é possível ajustar um modelo a dois factores sem interacção, tal como no exercício 4.

   (a) A tabela-resumo correspondente é:

```
> summary(aov(rend ~ variedade + terreno, data=terrenos))
            Df Sum Sq Mean Sq F value  Pr(>F)
variedade    3  1.799  0.5997   6.145 0.00175 **
terreno     12  2.407  0.2006   2.056 0.04737 *
Residuals   36  3.513  0.0976
```

   Desta tabela depreende-se que, aos níveis de significância usuais, deve considerar-se a existência de efeitos do factor variedade:

   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: $F = \frac{QMA}{QMRE} \ \cap\ F_{(a-1\,,\,n-(a+b-1))}$, sob H_0.
   Nível de significância: $\alpha = 0.05$.
   Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,36)} \ \approx\ 2.87$.
   Conclusões: $F_{calc} = 6.145$, um valor significativo mesmo ao nível $\alpha = 0.005$. Logo, rejeita-se H_0 a favor da hipótese de que existem efeitos do factor. Assim, é de concluir que diferentes variedades estejam associadas a diferentes rendimentos médios.

   (b) Um teste aos efeitos do factor `terreno` permite tirar a conclusão que os efeitos deste factor são menos importantes que os efeitos do factor `variedade`, embora ao nível de significância $\alpha = 0.05$ sejam (por pouco) significativos. Assim,

Hipóteses: $H_0 : \beta_j = 0 \ , \ \forall \ j = 2, .., 13$ vs. $H_1 : \exists\, j = 2, ..., 13$ tal que $\beta_j \neq 0$.
Estatística do teste: $F = \frac{QMB}{QMRE} \ \cap \ F_{(b-1\,,\,n-(a+b-1))}$, sob H_0.
Nível de significância: $\alpha = 0.05$.
Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(12,36)} \ \approx \ 2.04$.
Conclusões: $F_{calc} = 2.056$, um valor significativo (por muito pouco) ao nível $\alpha = 0.05$. Logo, rejeita-se H_0 a favor da hipótese de que existem efeitos do factor `terreno`.

NOTA: Num caso como este, em que a conclusão dependende do nível de significância usado, é especialmente importante que eventuais fontes de variabilidade, exteriores ao factor sob estudo, mas que afectem a variável resposta, sejam tidas em conta, de forma a reduzir a variabilidade não explicada pelo modelo, isto é, o valor de $QMRE$.

(c) É pedido o valor ajustado da (única) observação de Y na célula $(1, 1)$, ou seja, pede-se o valor de $\hat{y}_{111}$. Sabemos, a partir dos acetatos das aulas teóricas, que $\hat{y}_{ijk} = \overline{y}_{i..} + \overline{y}_{.j.} - \overline{y}_{...}$, ou seja, que qualquer valor ajustado numa célula genérica (i, j) é dado pela soma das médias de todas as observações no nível i do factor A e de todas as observações no nível j do factor B, menos a média global da totalidade das n observações de Y. No nosso caso temos no enunciado a média das observações da variedade A, ou seja, $\overline{y}_{1..} = 1.556$, admitindo que o factor A é o factor variedade. A média das quatro observações associadas ao terreno I é $\overline{y}_{.1.} = (1.800 + 2.457 + 0.722 + 0.789)/4 = 1.4420$. Finalmente, a média global de todas as observações (que pode ser calculada directamente a partir das $n = 52$ observações, ou como a média das quatro médias de variedade - embora neste último caso com um pequeno erro de arredondamento) é $\overline{y}_{...} = 1.358308$. Logo, o valor ajustado pedido é $\hat{y}_{111} = 1.556 + 1.4420 - 1.358308 = 1.639692$. Assinale-se que este valor ajustado não é (ao contrário do que se poderia supôr com base no modelo ANOVA a um factor) a média das observações da célula respectiva (neste caso o único valor observado nessa célula, $y_{111} = 1.800$). Tal relação apenas será verdadeira num modelo ANOVA a 2 factores, mas com efeitos de interacção. Os valores aqui indicados podem ser obtidos no `R` com o auxílio dos comandos `model.tables` (com a opção `type="means"`) e `fitted`, como indicado de seguida.

```
> model.tables(aov(rend ~ terreno + variedade, data=terrenos), type="means")
Tables of means
Grand mean
1.358308
 terreno
terreno
     I     II    III     IV     IX      V     VI    VII   VIII      X     XI
1.4420 1.5995 1.3395 1.2665 1.0360 1.7643 1.4678 1.3795 1.4033 0.9458 1.4213
   XII   XIII
1.1190 1.4738
 variedade
variedade
     A      B      C      D
1.5560 1.5322 1.1669 1.1782

> fitted(aov(rend ~ terreno + variedade, data=terrenos))
        1         2         3         4         5         6         7         8
1.6396923 1.7971923 1.5371923 1.4641923 1.9619423 1.6654423 1.5771923 1.6009423
        9        10        11        12        13        14        15        16
1.2336923 1.1434423 1.6189423 1.3166923 1.6714423 1.6158462 1.7733462 1.5133462
[...]
```

6. FALTA

7. Trata-se dum delineamento factorial a dois factores, o factor A (Fósforo), com $a = 3$ níveis (Baixa, Média e Elevada dosagem de adubação) e o Factor B (Potássio), igualmente com $b = 3$ níveis (Baixa, Média e Elevada dosagem de adubação). O delineamento é equilibrado, uma vez que em cada uma das $ab = 9$ situações experimentais (células) há igual número de observações $n_{ij} = n_c = 3$. Havendo repetições nas células, é possível estudar o modelo ANOVA a 2 factores, com interacção. A equação de base deste modelo é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall\, i = 1, 2, 3$, $j = 1, 2, 3$, $k = 1, 2, 3$, onde Y_{ijk} indica o rendimento obtido na k-ésima repetição da adubação correspondente à célula que cruza o nível i do fósforo e o nível j do potássio. Impõem-se as restrições $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ para qualquer j, e $(\alpha\beta)_{i1} = 0$ para qualquer i.

   (a) A tabela-resumo é dada no enunciado, mas com seis valores omissos. Os graus de liberdade do factor A (fósforo) são $a-1 = 2$. Os graus de liberdade associados aos efeitos de interacção são $(a-1)(b-1) = 4$. O Quadrado Médio associado ao factor B (potássio) é $QMB = \frac{SQB}{b-1} = \frac{18.7563}{2} = 9.37815$. O Quadrado Médio Residual é $QMRE = \frac{SQRE}{n-ab} = \frac{2.59333}{18} = 0.1440739$. O valor da estatística F para o teste aos efeitos principais do factor A é $F_A = \frac{QMA}{QMRE} = \frac{1.121481}{0.1440739} = 7.784068$. Finalmente, o valor da estatística F no teste aos efeitos principais do factor B é $F_B = \frac{QMB}{QMRE} = \frac{9.37815}{0.1440739} = 65.09264$.

   (b) Há três tipos de efeitos: principais do factor fósforo, associados às parcelas α_i; principais do factor potássio, associados às parcelas β_j; e de interacção entre os dois tipos de adubação, associados às parcelas $(\alpha\beta)_{ij}$. Existe um teste F para testar hipóteses associadas a cada um destes tipos de efeitos. Em concreto:

   Teste à interacção. As hipóteses são:

   $$H_0 : \quad (\alpha\beta)_{ij} = 0 \ , \ \forall\, i, j \qquad vs. \qquad H_1 : \quad \exists\, i, j \ tal \ que \quad (\alpha\beta)_{ij} \neq 0 \ .$$

   Teste aos efeitos principais do factor A. As hipóteses são:

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

   Teste aos efeitos principais do factor B. As hipóteses são:

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

   Para cada um destes testes, as estatísticas F são definidas como $F = \frac{QMxx}{QMRE}$, onde $QMxx$ indica o quadrado médio associado ao respectivo tipo de efeitos. As distribuições destas estatísticas de teste, caso seja verdadeira cada uma das hipóteses nulas, são F com graus de liberdade dados pelos g.l. dos quadrados médios no numerador e denominador, respectivamente, da estatística correspondente. Todas as regiões críticas são unilaterais direitas. Assim, e tendo em conta os valores da tabela-resumo e utilizando o nível de significância $\alpha = 0.05$, tem-se que se rejeitam as hipóteses nulas dos três testes. De facto, rejeita-se a inexistência de efeitos de interacção, uma vez que $F_{AB_{calc}} = 3.36504 > f_{0.05(4,18)} = 2.927744$. Rejeita-se a inexistência de efeitos principais do factor fósforo uma vez que $F_{A_{calc}} = 7.784068 > f_{0.05(2,18)} = 3.554557$. Finalmente, rejeita-se clarissimamente a inexistência de efeitos principais do factor potássio já que $F_{B_{calc}} = 65.09264 > f_{0.05(2,18)} = 3.554557$. Assim, conclui-se pela existência dos três tipos de efeitos. Estas conclusões poderiam também ser obtidas directamente a partir dos valores

de prova (*p-values*) correspondentes às três estatísticas de teste, disponíveis no enunciado. O valor de prova mais elevado, no caso do teste aos efeitos de interacção ($p = 0.03187154$) indica que, ao nível de significância $\alpha = 0.01$, a conclusão já seria a não rejeição da hipótese nula, isto é, não seria possível concluir pela existência de efeitos de interacção. Já a existência de efeitos principais do factor potássio está associado a um *p-value* da ordem de 10^{-8}.

(c) O problema pode ser respondido através da comparação dos rendimentos esperados em cada uma das duas células indicadas. Dada a natureza do problema, pode utilizar-se um teste de Tukey na resposta. A diferença entre as médias amostrais de célula será considerada significativa caso exceda, em módulo, o termo de comparação do teste de Tukey: $q_{\alpha(ab\,,\,n-ab)}\sqrt{\frac{QMRE}{n_c}}$. Utilizando o nível de significância $\alpha = 0.05$ tem-se, pelas tabelas da distribuição de Tukey, $q_{0.05(9,18)} = 4.96$, logo o termo de comparação é 1.08696. Ora, as células cuja comparação é pedida são as células $(1,3)$ e $(2,3)$, cujas médias amostrais são $\overline{y}_{13} = 6.733$ e $\overline{y}_{23} = 7.6$. Uma vez que $|6.733 - 7.6| = 0.867 < 1.08696$, não se rejeita a igualdade dos rendimentos esperados nestas duas combinações de adubação. Assim, não se pode concluir pela existência dum rendimento significativamente superior (ao nível $\alpha = 0.05$) quando a elevada dosagem de potássio se faz acompanhar por uma dosagem média na adubação à base de fósforo (ou seja, a média amostral mais elevada na célula $(2,3)$ não pode ser considerada estatisticamente significativa ao nível $\alpha = 0.05$).

(d) Nesta alínea pede-se para considerar-se o modelo sem efeitos de interacção, ou seja, cuja equação de base é $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk},\ \forall\, i,j,k$, e com as restrições $\alpha_1 = \beta_1 = 0$. O facto de o modelo não prever efeitos de interacção significa que a respectiva Soma de Quadrados (indicada no enunciado) passa a englobar a Soma de Quadrados Residual (uma vez que já não corresponde a efeitos previstos pelo modelo). Tem-se agora $SQRE = 2.59333 + 1.93926 = 4.53259$. Os graus de liberdade sofrem uma transformação análoga (este modelo tem agora menos $(a-1)(b-1)$ parâmetros do que anterior, pelo que os graus de liberdade residuais aumentam nesse montante). Assim, $g.l.(SQRE) = 18 + 4 = 22$. Logo o novo Quadrado Médio Residual vem: $QMRE = \frac{4.53259}{22} = 0.2060268$. As somas de quadrados, graus de liberdade e quadrados médios associados aos efeitos principais de cada factor permanecem iguais (são calculados de forma análoga) pelo que a tabela-resumo é agora a seguinte:

variação	g.l.	SQs	QMs	F_{calc}
fosforo	2	2.24296	1.121481	5.443374
potassio	2	18.75630	9.37815	45.51908
residual	22	4.53259	0.2060268	–

Para identificar os valores de prova (*p-values*) dos novos valores das estatísticas F sobrantes, é necessário ter em conta os novos valores dos graus de liberdade residuais. Tem-se:

```
> 1-pf(5.443374, 2, 22)
[1] 0.01200658
> 1-pf(45.51908, 2, 22)
[1] 1.517658e-08
```

Assim, os dois valores calculados das estatísticas continuam a ser significativos ao nível $\alpha = 0.05$. No entanto, os efeitos do factor fósforo já não seriam considerados significativos ao nível $\alpha = 0.01$. Este exemplo ilustra o perigo de ignorar a existência de efeitos que realmente existam (neste caso, ignorar os efeitos de interacção): pode ajudar a camuflar a

existência de outros tipos de efeitos, mesmo dos que são previstos no modelo, através do inflacionamento da variabilidade residual ($QMRE$).

8. (a) Trata-se dum delineamento factorial, a dois factores: tempo de exposição, com $a = 3$ níveis, e temperatura, também com $b = 3$ níveis. O delineamento é equilibrado, com $n_c = 3$ repetições em cada uma das $ab = 9$ células, para um total de $n = ab\,n_c = 27$ observações. Havendo repetições nas células, é possível ajustar um modelo a dois factores, com interacção, cuja equação de base é:

   i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ \ \forall\, i = 1,2,3\ ,\ j = 1,2,3,\ k = 1,2,3,$
   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 absorção na k-ésima repetição da situação experimental dada pela combinação do tempo de exposição i com a temperatura j.
      - μ_{11} indica a absorção média (populacional) na célula definida pelo tempo de exposição E_1 com a temperatura T_1;
      - α_i indica o efeito principal do tempo de exposição i;
      - β_j indica o efeito principal da temperatura j;
      - $(\alpha\beta)_{ij}$ indica o efeito de interacção entre o tempo de exposição i e a temperatura 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. $\{\epsilon_{ijk}\}_{i,j,k}$ constitui um conjunto de variáveis aleatórias independentes.

   (b) Há três tipos de efeitos: principais do factor A, associados às parcelas α_i; principais do factor B, associados às parcelas β_j; e de interacção, associados às parcelas $(\alpha\beta)_{ij}$. Existe um teste F para testar hipóteses associadas a cada um destes tipos de efeitos. Em concreto:

   Teste à interacção. As hipóteses são:

$$H_0 :\quad (\alpha\beta)_{ij} = 0\ ,\ \forall\, i, j \qquad vs. \qquad H_1 :\quad \exists\, i, j \ tal\ que\ \ (\alpha\beta)_{ij} \neq 0\ .$$

   Teste aos efeitos principais do factor A. As hipóteses são:

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

   Teste aos efeitos principais do factor B. As hipóteses são:

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

   (c) Para efectuar os três testes F pedidos, vamos construir a tabela-resumo da ANOVA, utilizando para o efeito os comandos do `R`:

```
> absorcao.aov <- aov(abs ~ exposicao * temperatura, data=absorcao)
> summary(absorcao.aov)
                     Df Sum Sq Mean Sq F value   Pr(>F)
exposicao             2   2113  1056.3   63.59 6.92e-09 ***
temperatura           2   3674  1836.8  110.58 7.75e-11 ***
exposicao:temperatura 4   2704   676.1   40.70 8.74e-09 ***
Residuals            18    299    16.6
```

   Repare-se na utilização do asterisco para indicar, na fórmula que define o modelo usado, que se pretende utilizar um modelo a dois factores *com* interacção.

Teste à interacção .

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

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 $F_{calc} > f_{0.05(4,18)} = 2.93$.

Conclusões: O valor da estatística do teste é $F_{calc} = 40.70$. É um valor claramente significativo ao nível $\alpha = 0.05$, rejeitando-se H_0 a favor da hipótese alternativa de que existem efeitos de interacção entre tempo de exposição e temperatura.

Teste aos efeitos principais do factor A .

Hipóteses: $H_0 : \alpha_i = 0\ ,\ \forall\ i = 2,3$ [não há efeitos principais de A] vs. $H_1 : (\alpha_2 \neq 0) \vee (\alpha_3 \neq 0)$ [há efeitos principais de A].

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 $F_{calc} > f_{0.05(2,18)} = 3.55$.

Conclusões: O valor da estatística do teste é $F_{calc} = 63.59$. É um valor claramente significativo ao nível $\alpha = 0.05$, rejeitando-se H_0 a favor da hipótese alternativa de que existem efeitos principais de tempo de exposição.

Teste aos efeitos principais do factor B .

Hipóteses: $H_0 : \beta_j = 0\ ,\ \forall\ i = 2,3$ [não há efeitos principais de B] vs. $H_1 : (\beta_2 \neq 0) \vee (\beta_3 \neq 0)$ [há efeitos principais de B].

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 $F_{calc} > f_{0.05(2,18)} = 3.55$.

Conclusões: O valor da estatística do teste é $F_{calc} = 110.58$. É um valor ainda mais claramente significativo do que o da estatística calculada no teste aos efeitos principais do outro factor. Rejeita-se H_0 a favor da hipótese alternativa de que existem efeitos principais de temperatura.

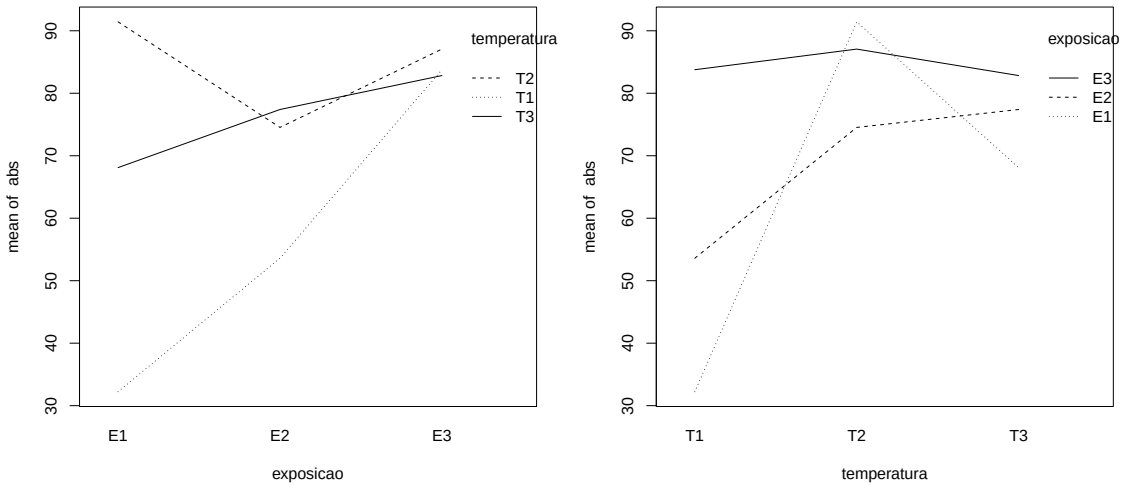
Assim, conclui-se que (ao nível $\alpha = 0.05$) existem os três tipos de efeitos previstos no modelo e todos contribuem para formar o valor esperado duma observação.

(d) Os dois gráficos de interacção podem ser criados, no `R`, com o comando `interaction.plot`:

```
> attach(absorcao)
> interaction.plot(response=abs, x.factor=exposicao, trace.factor=temperatura)
> interaction.plot(response=abs, trace.factor=exposicao, x.factor=temperatura)
> detach(absorcao)
```

NOTA: Os comandos `attach` e `detach`, que se complementam, visam tornar as colunas da *data frame* `absorcao` momentaneamente disponíveis às pesquisas efectuadas para encontrar as variáveis de nome `abs`, `exposicao` e `temperatura`, utilizadas no comando `interaction.plot`. É necessário usar o `attach`, uma vez que o comando que constrói os gráficos de interacção não dispõe dum argumento do tipo `data`. Alternativamente, poderia escrever-se o nome das três variáveis por extenso (como, por exemplo, `absorcao$temperatura`).

Os resultados são os gráficos a seguir reproduzidos:



Os dois gráficos contêm a mesma informação, mas expressa de modo diferente (pois trocam o factor que define o eixo horizontal e o que define os pontos que serão unidos pelos segmentos de recta). A existência de interacção é expressa pela ausência de "paralelismo" das curvas seccionalmente lineares, que é particularmente evidente no gráfico da direita. Por seu lado, a existência de efeitos principais de cada factor é particularmente visível nos gráficos em que esse factor define as curvas seccionalmente lineares. Assim, no gráfico da esquerda é visível que (independentemente das variações associadas aos níveis do outro factor), há temperaturas (T1) que parecem estar associados a valores de absorção média (legíveis no eixo vertical) globalmente inferiores aos de outras temperaturas. De forma análoga, no gráfico da direita constata-se que o tempo de exposição E_2 tem valores de absorção média globalmente inferiores aos do tempo de exposição E_3. Que essas diferenças são muito significativas foi o que se constatou nos testes da alínea anterior.

(e) A menor absorção média amostral verifica-se na célula associada ao tempo de exposição E_1 e temperatura T_1: $\overline{y}_{11.} = 32.23$. No outro extremo, a maior absorção média amostral observa-se quando ao mesmo tempo de exposição (E_1) se associa a temperatura T_2: $\overline{y}_{12.} = 91.43$.

(f) A pergunta pode reformular-se da seguinte forma: tendo sido visto que o maior valor médio amostral se observa na célula $(1, 2)$, quer-se saber se é possível inferir daí que, também a nível populacional, essa situação experimental gera um valor médio superior ao das restantes células. Para isso, vamos utilizar testes de Tukey e comparar a média amostral $\overline{y}_{12.}$ com as das restantes células. Sabemos que, nos testes de Tukey (aplicados a um delineamento factorial a dois factores), qualquer hipótese relativa à igualdade de duas médias populacionais de célula, $\mu_{ij} = \mu_{i'j'}$ deve ser rejeitada, caso a correspondente diferença de médias amostrais exceda, em módulo, o termo de comparação $q_{\alpha(ab,n-ab)} \cdot \sqrt{\frac{QMRE}{n_c}}$. No nosso caso, temos $ab = 9$ e $n = 27$, pelo que, ao nível $\alpha = 0.05$, o quantil de ordem $1-\alpha$ na distribução de Tukey com parâmetros $ab = 9$ e $n - ab = 18$ é $q_{0.05(9,18)} = 4.96$. Este valor pode ser obtido através das tabelas de Tukey (disponíveis na página *web* da disciplina) ou através do seguinte comando do `R`:

```
> qtukey(0.95,9,18)
[1] 4.955209
```

Temos ainda $n_c = 3$ e $QMRE = 16.6$. Logo, o valor do termo de comparação é $4.955 \times \sqrt{16.6/3} = 11.656$. Ora, há algumas médias amostrais de célula que diferem por menos do

que este valor da média amostral da célula $(1,2)$. Por exemplo, $|\overline{y}_{33.} - \overline{y}_{12.}| = 8.6$, o que significa que não se rejeita a igualdade das médias populacionais μ_{33} e μ_{12}. Aliás, qualquer média amostral superior a $\overline{y}_{12.} - 11.656 = 79.77$ não é significativamente diferente da média da célula $(1,2)$. Assim, as médias populacionais das três situações experimentais onde surge o tempo de exposição E_3 não se podem considerar diferentes da média populacional μ_{12}, enquanto que as médias populacionais nas restantes cinco células devem ser consideradas diferentes de μ_{12}.

(g) Para considerar as nove diferentes situações experimentais como nove níveis dum único factor, será necessário substituir as duas colunas que na *data frame* `absorcao` indicavam os dois factores (isto é, as colunas `exposicao` e `temperatura`) por um único factor (ao qual dar-se-á o nome `sit.exp`), indicando as nove situações experimentais. Para criar essa nova variável, utilizar-se-á o comando `paste` do `R`, que permite "colar" os valores de cada um dos factores originais, utilizando um ponto como símbolo de separação. O vector assim produzido será transformado em factor através do comando `as.factor`:

```
> sit.exp <- as.factor(paste(absorcao$exposicao, absorcao$temperatura, sep="."))
> sit.exp
 [1] E1.T1 E1.T1 E1.T1 E2.T1 E2.T1 E2.T1 E3.T1 E3.T1 E3.T1 E1.T2 E1.T2 E1.T2 E2.T2
[14] E2.T2 E2.T2 E3.T2 E3.T2 E3.T2 E1.T3 E1.T3 E1.T3 E2.T3 E2.T3 E2.T3 E3.T3 E3.T3
[27] E3.T3
Levels: E1.T1 E1.T2 E1.T3 E2.T1 E2.T2 E2.T3 E3.T1 E3.T2 E3.T3
```

Seguidamente, utiliza-se esse novo factor como preditor numa ANOVA a um factor:

```
> summary(aov(absorcao$abs ~ sit.exp))
            Df Sum Sq Mean Sq F value   Pr(>F)
sit.exp      8   8491  1061.3   63.89 1.22e-11 ***
Residuals   18    299    16.6
```

Em resposta directa à pergunta feita, é evidente pelo *p-value* baixíssimo que se deve considerar que as médias nas nove situações experimentais não são iguais (o que é, aliás, coerente com o que se viu acima). Note-se, no entanto, que a Soma de Quadrados Residual neste novo modelo é igual à que se havia obtido no modelo a dois factores com interacção. O que significa que a soma de quadrados associada ao único factor agora existente tem de ser equivalente às Somas de Quadrados SQA, SQB e $SQRE$ nesse modelo a dois factores com interacção. Esta constatação reforça a ideia que a diferença entre os dois modelos não reside tanto na capacidade explicativa global, que é a mesma, mas na forma como é (dois factores), ou não (um único factor), possível atribuir essa variabilidade explicada a várias causas (interacção, factor A, factor B).

9. (a) A troca de ordem dos factores no comando do `R` não têm efeito sobre o ajustamento do modelo a dois factores com interacção (além de trocar a ordem das duas primeiras linhas da tabela-resumo), como se pode constatar comparando o ajustamento obtido na alínea 8c com o que se obtém trocando a ordem dos factores:

```
> summary(aov(abs ~ temperatura * exposicao, data=absorcao))
                     Df Sum Sq Mean Sq F value   Pr(>F)
temperatura           2   3674  1836.8  110.58 7.75e-11 ***
exposicao             2   2113  1056.3   63.59 6.92e-09 ***
temperatura:exposicao 4   2704   676.1   40.70 8.74e-09 ***
Residuals            18    299    16.6
```

No entanto, esta invariância depende do facto de se estar a trabalhar com um delineamento equilibrado, como se verá na alínea seguinte.

(b) Retirando a primeira e as duas últimas observações, passamos a ter um delineamento análogo, mas não equilibrado. Repare-se nas tabelas-resumo obtidas agora, trocando a ordem dos factores:

```
> summary(aov(abs ~ exposicao * temperatura, data=absorcao[-c(1,26,27),]))
                     Df Sum Sq Mean Sq F value   Pr(>F)
exposicao             2 1445.6   722.8   43.92 5.36e-07 ***
temperatura           2 3032.3  1516.2   92.14 3.76e-09 ***
exposicao:temperatura 4 2444.6   611.1   37.14 1.29e-07 ***
Residuals            15  246.8    16.5
---
> summary(aov(abs ~ temperatura * exposicao , data=absorcao[-c(1,26,27),]))
                     Df Sum Sq Mean Sq F value   Pr(>F)
temperatura           2 2700.4  1350.2   82.05 8.36e-09 ***
exposicao             2 1777.5   888.8   54.01 1.40e-07 ***
temperatura:exposicao 4 2444.6   611.1   37.14 1.29e-07 ***
Residuals            15  246.8    16.5
```

Como se pode constatar, embora as linhas associadas à interacção e ao residual tenham idênticas somas de quadrados, graus de liberdade e quadrados médios nos dois casos, já as linhas associadas ao efeito principal de cada factor são diferentes nos dois casos. Este problema foi já referido nas aulas teóricas, no final da discussão sobre o modelo a dois factores, sem interacção.

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. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$, $\forall\, i = 1,2,3,4\ ,\ j = 1,2,...,9\ ,\ k = 1,2,3,4$, 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. $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, $\forall\, i,j,k$.

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

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, $SQT^* = \sum_i \sum_j \sum_k (Y^*_{ijk} - \overline{Y}^*_{...})^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \overline{Y}_{...}/1000)^2 = SQT/(1000^2)$. 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* ($\overline{y}_{29\,.} = 5927 kg/ha$). 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 $QMRE$.

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: $b - 1 = 3$;
- Interacção: $(a - 1)(b - 1) = 3$;
- 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 interacção ($SQAB$). 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 $SQA = bn_c \sum_{i=1}^{a} (\overline{Y}_{i..} - \overline{Y}_{...})^2 = 16\,[(24.681 - 22.14375)^2 + (19.606 - 22.14375)^2] = 16 \times 12.87781 = 206.045$. 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,

Variação	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	$QMB = \frac{SQB}{b-1} = 96.01$	$F = \frac{QMB}{QMRE} = 111.2085$
Interacção	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 interacção (os $(\alpha\beta)_{ij}$). Efectuaremos estes dois testes, começando pelo dos efeitos de interacção. Neste exemplo, e como o Factor A apenas tem dois níveis, o índice i nos efeitos de interacção 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 interacção.

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. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ \ \forall\, i = 1, 2\ ,\ j = 1, 2\ ,\ k = 1, 2, .., 12,$
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 principal do nível i do Factor A (*Fibra*);
   - β_j indica o efeito principal 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) Pede-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* `leitoes`, vamos construir a tabela-resumo da ANOVA com o auxílio do R:

```
> summary(aov(CEL ~ Fibra*Enzima, data=leitoes))
             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 principais 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 principais 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=leitoes), 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: $|\overline{Y}_{21.} - \overline{Y}_{12.}| = 0.1516 > 0.1400$ e $|\overline{Y}_{21.} - \overline{Y}_{22.}| = 0.1533 > 0.1400$. Logo, e ordenando as quatro médias de célula por ordem crescente, tem-se:

$\overline{y}_{21.}$	$\overline{y}_{11.}$	$\overline{y}_{12.}$	$\overline{y}_{22.}$
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:

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

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 = ".")
> celulas
 [1] "1.1" "1.1" "1.1" "1.1" "1.1" "1.1" "1.2" "1.2" "1.2" "1.2" "1.2" "1.2"
[13] "2.1" "2.1" "2.1" "2.1" "2.1" "2.1" "2.2" "2.2" "2.2" "2.2" "2.2" "2.2"
[25] "1.1" "1.1" "1.1" "1.1" "1.1" "1.1" "1.2" "1.2" "1.2" "1.2" "1.2" "1.2"
[37] "2.1" "2.1" "2.1" "2.1" "2.1" "2.1" "2.2" "2.2" "2.2" "2.2" "2.2" "2.2"
```

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 $K^2_{cal} = 15.7157$. Comparado com o valor fronteira da Região Crítica ao nível de significância $\alpha = 0.05$, que é $\chi^2_{0.05\,(3)} = 7.81473$ (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.

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=leitoes))
        ```

        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=leitoes))
        > H <- X %*% solve(t(X) %*% X) %*% t(X)
        ```

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

        ```
        > H %*% leitoes$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=leitoes))
```

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((leitoes$CEL-H %*% leitoes$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=leitoes))
> 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: $SQF = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} \left(\hat{Y}_{ijk} - \overline{Y}_{...} \right)^2$. 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 duma 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.

14. Os dados deste exercício encontram-se na *data frame* `TabRegua`. Para modelar a variável-resposta rendimento, existem dois factores: o local e ano. Mas não se trata dum delineamento factorial: os anos observados em cada local não são os mesmos.

(a) Para se tratar dum delineamento factorial, cada um dos $a = 2$ locais, Tabuaço e Régua, teria de ter sido observado em todos os anos analisados. No entanto, não se dispõem de dados para o Tabuaço em 2000 e 2002, nem para a Régua em 2003. Assim, os níveis do factor `ano` dependem das localidades, isto é, dos níveis do factor `local`. Tem-se uma hierarquia na definição dos factores, ou seja, está-se perante um *delineamento hierarquizado*. O modelo correspondente (recordando que o `R` ordena os níveis de um factor por ordem alfabética, pelo que a Régua será o primeiro nível do factor `local` e o Tabuaço o segundo) :

i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk},\ \forall i{=}1,2\ ,\ j{=}1,2,3{=}b_1$ $(se\ i{=}1)$ $ou\ j{=}1,2{=}b_2$ $(se\ i{=}2)\ ,\ k{=}1,2,..,8$, com $\alpha_1 = 0$ e $\beta_{1(i)} = 0,\ \forall i$. Neste caso concreto, só existem os efeitos α_2, $\beta_{2(1)}$, $\beta_{3(1)}$ e $\beta_{2(2)}$. Na equação,

- μ_{11} indica o rendimento médio populacional na Régua em 1999;
- α_2 indica o efeito do local Tabuaço;
- $\beta_{2(1)}$ indica o efeito do ano 2000 na Régua;
- $\beta_{3(1)}$ indica o efeito do ano 2002 na Régua;
- $\beta_{2(2)}$ indica o efeito do ano 2003 no Tabuaço;
- 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.

O delineamento é equilibrado, pois nas $b_1 + b_2 = 5$ situações experimentais há sempre $n_c = 8$ observações, para um total de $n = 40$ observações.

(b) Neste tipo de delineamentos há dois tipos de efeitos: o do factor dominante e o do factor subordinado. Para cada tipo de efeitos há um teste F, semelhante ao de anteriores modelos ANOVA. Para construir a tabela-resumo desta ANOVA a dois factores hierarquizados, utiliza-se, na fórmula do comando `lm` o símbolo "/", que indica uma relação de hierarquia entre factores. Atenção que, neste tipo de delineamentos, é importante distinguir o factor dominante e o factor subordinado (que vem após o símbolo "/"):

```
> TabRegua.aov <- aov(rend ~ local/ano, data=TabRegua)
> summary(TabRegua.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
local        1  0.418  0.4175   2.215 0.1456
local:ano    3  4.885  1.6282   8.638 0.0002 ***
Residuals   35  6.597  0.1885
```

Assim, tem-se um primeiro teste à existência de efeitos de ano (o factor subordinado):

Hipóteses: $H_0 : \beta_{2(1)} = \beta_{3(1)} = \beta_{2(2)} = 0$ vs. $H_1 : (\beta_{2(1)} \neq 0) \vee (\beta_{3(1)} \neq 0) \vee (\beta_{2(2)} \neq 0)$.

Estatística do teste: $F = \frac{QMB(A)}{QMRE} \ \cap \ F_{[(b_1-1)+(b_2-1)\,,\,n-(b_1+b_2)]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(3,35)} \approx 2.88$.

Conclusões: O valor da estatística do teste foi já calculado: $F_{calc} = 8.638 > 2.88$, pelo que se rejeita H_0, havendo motivo para admitir a existência de efeitos de anos (subordinados a local).

E também um teste à existência de efeitos do factor `local`, neste caso ao único efeito de local previsto no modelo (α_2):

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

Estatística do teste: $F = \frac{QMA}{QMRE} \ \cap \ F_{[a-1\,,\,n-(b_1+b_2)]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(1,35)} \approx 4.12$.

Conclusões: O valor da estatística do teste é dado na tabela-resumo: $F_{calc} = 2.215 < 4.12$, pelo que não se rejeita H_0, não havendo motivo para admitir a existência de efeitos de local.

(c) Vamos utilizar os testes de Tukey para comparar as cinco situações experimentais do nosso problema. De entre as cinco médias populacionais existentes (μ_{11}, μ_{12}, μ_{13}, μ_{21} e μ_{22}), devemos considerar um qualquer par delas diferentes se as respectivas médias amostrais diferirem mais do que o termo de comparação $q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}}$, onde $k = b_1 + b_2$ indica

o número total de situações experimentais. Ora, pelas tabelas da distribuição de Tukey, $q_{0.05(5,35)} = 4.07$. Tem-se ainda $\sqrt{\frac{0.1885}{8}} = 0.1535008$, pelo que o termo de comparação é 0.624715. Por outro lado, as cinco médias de situação experimental são dadas pelo comando `model.tables` (com a opção `type=''means''`):

```
> model.tables(TabRegua.aov, type="means")
Tables of means
Grand mean
0.685625
 local
    Regua Tabuaco
    0.769  0.5605
rep 24.000 16.0000
 local:ano
        ano
local     1999  2000  2002  2003
  Regua   0.269 0.687 1.352
  rep     8.000 8.000 8.000 0.000
  Tabuaco 0.646             0.475
  rep     8.000 0.000 0.000 8.000
```

(a organização da tabela das médias de local/ano ilustra bem, com os seus espaços em branco, que não estamos perante um delineamento factorial).

Ordenando as médias de situação experimental por ordem crescente, verifica-se que nenhum par que envolva as quatro médias amostrais mais pequenas é significativamente diferente (ao nível $\alpha = 0.05$), enquanto que a média $\overline{y}_{13}$ (Régua em 2002) é significativamente diferente de todas as outras:

$\overline{y}_{11.}$	$\overline{y}_{22.}$	$\overline{y}_{21.}$	$\overline{y}_{12.}$	$\overline{y}_{13.}$
0.269	0.475	0.646	0.687	1.352

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:

$\overline{y}_{11.}$	$\overline{y}_{22.}$	$\overline{y}_{21.}$	$\overline{y}_{12.}$	$\overline{y}_{13.}$
$0.269^{\,a}$	$0.475^{\,a}$	$0.646^{\,a}$	$0.687^{\,a}$	$1.352^{\,b}$

(d) Para efectuar um teste de Bartlett, começamos por designar por σ^2_{ij} a variância populacional da localidade i, ano j (naquela localidade), para testar a igualdade de todas as $k = b_1 + b_2 = 5$ variâncias populacionais (H_0) contra a diferença de pelo menos um par delas (H_1):

Hipóteses: $H_0 : \sigma^2_{11} = \sigma^2_{12} = \sigma^2_{13} = \sigma^2_{21} = \sigma^2_{22}$ vs. $H_1 : \exists\, i, i', j, j'$ tais que $\sigma^2_{ij} \neq \sigma^2_{i'j'}$.

Estatística do teste: $K^2 = \dfrac{(n-k)\ln QMRE - \sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b_i}(n_c-1)\ln S^2_{ij}}{C} \sim \chi^2_{k-1}$, sob H_0,

onde $C = 1 + \frac{1}{3(k-1)}\left[\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b_i}\frac{1}{n_{ij}-1} - \frac{1}{n-k}\right]$ e S^2_{ij} representa a variância amostral da situação experimental (i, j).

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $K^2_{calc} > \chi^2_{0.05(4)} = 9.488$.

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), mas criando primeiro um factor com cinco níveis, correspondendo às cinco situações experimentais resultantes de associar os locais e anos estudados:

```
> loc.ano <- as.factor(paste(TabRegua$local, TabRegua$ano, sep="."))
> loc.ano
 [1] Tabuaco.1999 Tabuaco.1999 Tabuaco.1999 Tabuaco.1999 Tabuaco.1999
 [6] Tabuaco.1999 Tabuaco.1999 Tabuaco.1999 Tabuaco.2003 Tabuaco.2003
[11] Tabuaco.2003 Tabuaco.2003 Tabuaco.2003 Tabuaco.2003 Tabuaco.2003
[16] Tabuaco.2003 Regua.1999   Regua.1999   Regua.1999   Regua.1999
[21] Regua.1999   Regua.1999   Regua.1999   Regua.1999   Regua.2000
[26] Regua.2000   Regua.2000   Regua.2000   Regua.2000   Regua.2000
[31] Regua.2000   Regua.2000   Regua.2002   Regua.2002   Regua.2002
[36] Regua.2002   Regua.2002   Regua.2002   Regua.2002   Regua.2002
Levels: Regua.1999 Regua.2000 Regua.2002 Tabuaco.1999 Tabuaco.2003

> bartlett.test(TabRegua$rend ~ loc.ano)
Bartlett test of homogeneity of variances
data:  TabRegua$rend by loc.ano
Bartlett's K-squared = 6.7258, df = 4, p-value = 0.1511
```

Assim, não se rejeita a hipótese de variâncias populacionais iguais nas cinco situações experimentais, o que está de acordo com o pressuposto de homogeneidade de variâncias do modelo ANOVA.

15. Esta pergunta saiu no exame de segunda chamada do ano lectivo 2012-13.

(a) Trata-se dum delineamento a dois factores – o factor `Local` (factor A) e o factor `Ano` (factor B) – mas *hierarquizado*, uma vez que os anos observados numa localidade diferem dos anos observados na outra localidade. Assim, o factor A (`Local`) tem $a = 2$ níveis (`Elvas` e `Braga`, pela ordem da listagem do enunciado) e constitui o factor dominante: o significado desses níveis é imediato, sem referência ao outro factor. O factor subordinado (factor B, `Ano`), tem $b_1 = 2$ níveis no primeiro nível do factor A (os anos `2000` e `2004` do estudo em Elvas) e $b_2 = 3$ níveis no segundo nível do factor A (os anos de `2007` a `2009` observados em Braga). O delineamento é equilibrado, pois há $n_c = 4$ repetições em cada uma das $b_1 + b_2 = 5$ situações experimentais. Tem-se assim um total de $n = n_c\,(\sum_{i=1}^{2} b_i) = 4 \times 5 = 20$ observações. O modelo correspondente a este delineamento é:

- $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, $\forall\, i, j, k$, onde Y_{ijk} indica o peso do k-ésimo bolbo no local i, no ano j ($i = 1, 2$; $j = 1, 2$ se $i = 1$ e $j = 1, 2, 3$ se $i = 2$; e $k = 1, 2, 3, 4$). Impõem-se as restrições $\alpha_1 = 0$, $\beta_{1(i)} = 0$ para $i = 1$ e $i = 2$. Com estas restrições, os parâmetros têm a seguinte interpretação:
  - μ_{11} é o peso médio populacional dos bolbos de Elvas, no ano `2000`;
  - α_2 é o efeito do `Local Braga`; e
  - $\beta_{j(i)}$ ($j > 1$) é o efeito do ano j, no local i.

  A parcela ϵ_{ijk} representa o erro aleatório associado à observação Y_{ijk}, e representa a variabilidade não explicada pelos efeitos previstos no modelo.
- $\epsilon_{ijk} \cap \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k.
- Os erros aleatórios ϵ_{ijk} são independentes.

(b) Há dois testes F neste contexto, correspondentes aos dois tipos de efeitos previstos neste modelo: efeito de localidade e efeitos de ano dentro das localidades. Comecemos pelo teste aos efeitos de ano, dentro das localidades. Após as restrições, existem apenas três parcelas correspondentes a este tipo de efeitos.

Hipóteses: $H_0 : \beta_{2(1)} = \beta_{2(2)} = \beta_{3(2)} = 0$ vs. $H_1 : (\beta_{2(1)} \neq 0) \vee (\beta_{2(2)} \neq 0) \vee (\beta_{3(2)} \neq 0)$.

Estatística do Teste: $F = \frac{QMB(A)}{QMRE} \cap F_{\left[\sum_{i=1}^{2}(b_i-1)\,,\, n-\sum_{i=1}^{2} b_i\right]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha[(b_1-1)+(b_2-1)\,,\, n-(b_1+b_2)]} = f_{0.05(3,15)} = 3.29$.

Conclusões: Como $F_{calc} = 16.570 > 3.29$, rejeita-se H_0, o que corresponde a admitir a existência de efeitos de anos.

No teste aos efeitos do factor `Local`, há uma única parcela (o efeito de `Braga`). Tem-se:

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

Estatística do Teste: $F = \frac{QMA}{QMRE} \cap F_{\left[a-1\,,\, n-\sum_{i=1}^{2} b_i\right]}$, sob H_0.

Nível de significância: $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{\alpha[a-1,n-(b_1+b_2)]} = f_{0.05(1,15)} = 4.54$.

Conclusões: Como $F_{calc} = 13.072 > 4.54$, rejeita-se H_0, o que corresponde a admitir a existência de efeitos de localidade.

Concluindo-se pela existência de efeitos de localidade, e uma vez que existem apenas dois locais, podemos afirmar que há diferenças nos pesos médios dos bolbos em Elvas e Braga, diferença essa representada pela parcela α_2 da equação do modelo.

(c) Pede-se para comparar as médias das células de Braga, isto é, as médias de célula μ_{21}, μ_{22} e μ_{23}. Sabemos que através das comparações múltiplas de Tukey, pode-se concluir pela diferença de qualquer par destas médias, caso a diferença entre as correspondentes médias amostrais exceda, em módulo, o termo de comparação:

$$q_{\alpha(b_1+b_2\,,\, n-(b_1+b_2))}\sqrt{\frac{QMRE}{n_c}} \quad = \quad q_{0.05(5,15)}\sqrt{\frac{12.189}{4}}\,.$$

Uma vez que pelas tabelas de Tukey $q_{0.05(5,15)} = 4.37$, o termo de comparação é 7.6284. Ora, a maior diferença de médias amostrais das células de Braga é $|\overline{y}_{22.} - \overline{y}_{23.}| = 19.9325 - 12.9425 = 6.99$, que é inferior ao termo de comparação. Assim, não se pode (ao nível de significância $\alpha = 0.05$) concluir pela diferença entre os pesos médios populacionais em Braga, nos três anos estudados. Esta conclusão, bem como a análise das duas médias anuais em Elvas, sugere que a conclusão muito clara do teste F aos efeitos de ano efectuado no ponto 2, se deve sobretudo à enorme diferença de pesos médios dos bolbos nos dois anos do estudo em Elvas.

(d) Tem-se agora uma ANOVA a um único factor (`Local`), com apenas $k = 2$ níveis. Este delineamento muito simples (que também poderia ser estudado através dos testes t de comparação de médias de duas populações com base em 2 amostras independentes, dado na disciplina de Estatística dos primeiros ciclos do ISA) fica um delineamento desequilibrado, uma vez que no nível `Elvas` ($i = 1$) há $n_1 = 8$ observações e no nível `Braga` ($i = 2$) há

$n_2 = 12$ observações. Esse facto não obsta a que se possa responder às perguntas feitas no enunciado.

i. Sabemos que, por definição, a Soma de Quadrados associada aos efeitos do factor subordinado, no modelo hierarquizado, é a diferença das Somas de Quadrados Residuais no modelo a um factor ajustado nesta alínea e no modelo hierarquizado, ou seja,

$$\begin{array}{rrcl} & SQB(A) & = & SQRE_A - SQRE_{A/B} \\ \Leftrightarrow & SQRE_A & = & SQB(A) + SQRE_{A/B} \quad = \quad 605.94 + 182.84 \quad = \quad 788.78 \end{array}$$

Os graus de liberdade residuais serão, como em qualquer modelo ANOVA a um factor, $n-k$, o que no nosso caso significa 18. Logo, $QMRE_A = \frac{SQRE_A}{n-k} = 43.8211$. Sabemos ainda que, por definição, a Soma de Quadrados associada ao factor dominante no modelo hierarquizado (SQA) é a Soma de Quadrados do factor (SQF) no modelo com apenas esse factor. Uma vez que os graus de liberdade também serão agora $k-1 = 1$, isso significa que SQF, os seus graus de liberdade e QMF são iguais aos indicados na tabela-resumo do modelo hierarquizado. No entanto, o valor da estatística F correspondente ao teste aos efeitos do factor `Local` será diferente, uma vez que mudou o Quadrado Médio Residual. Tem-se:

Variação	g.l.	SQ	QM	F
Factor	1	159.34	159.34	$F = \frac{QMF}{QMRE} = 3.636$
Residual	18	788.78	43.8211	–

ii. Há agora um único teste F a efectuar, semelhante ao teste aos efeitos do factor A no contexto do modelo hierarquizado, descrito na alínea 15b. Para optar entre as hipóteses em confronto, $H_0 : \alpha_2 = 0$ vs. $H_1 : \alpha_2 \neq 0$, a regra é rejeitar H_0 caso $F_{calc} > f_{\alpha(k-1,n-k)} = f_{0.05(1,18)} = 4.41$. Como $F_{calc} = 3.636$, não se rejeita H_0. A conclusão, com base neste modelo e ao nível $\alpha = 0.05$, é diferente da conclusão no modelo hierarquizado: não se pode rejeitar a igualdade de pesos médios dos bolbos nas duas localidades. Esta conclusão resulta do facto que, ao ignorar-se no modelo desta alínea a variabilidade entre anos, essa variabilidade foi juntar-se à variabilidade residual (isto é, não explicada pelo modelo). O aumento do QMRE nesta alínea resulta dessa maior variabilidade não explicada pelo modelo. Mas esse maior QMRE (que surge no denominador da estatística do teste) diminui o valor de F_{calc} e acabou por colocá-lo fora da região de rejeição ao nível 0.05. Este exemplo ilustra a importância de um delineamento e modelo contemplarem fontes de variabilidade importantes no estudo da variável resposta.

16. (a) Pede-se para mostrar que a soma dos n_i resíduos e_{ij}, correspondentes ao nível i do Factor ($i = 1, 2, ..., k$), 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,

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

uma vez que se trata duma soma de desvios dum conjunto de observações em relação à sua média (ou seja, do tipo $\sum_{i=1}^{n}(x_i - \overline{x})$, 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 $\overline{y}_{ij.}$ das observações da célula da referida observação. Assim, a soma dos resíduos das n_{ij} observações efectuadas na célula (i,j) é dada por:

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

17. 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, $\overline{Y}_{1.} = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{ij}$. Mas, como é sabido (ver apontamentos da UC de Estatística, dos primeiros ciclos do ISA), a média $\overline{X}$ duma amostra aleatória $\{X_i\}_{i=1}^{n}$ de n variáveis aleatórias com distribuição $\mathcal{N}(\mu, \sigma^2)$, tem distribuição $\overline{X} \cap \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Assim, e tendo em conta que $\alpha_1 = 0$, tem-se $Y_{1j} \cap \mathcal{N}(\mu_1 \,,\, \sigma^2)$ e $\hat{\mu}_1 = \overline{Y}_{1.} \cap \mathcal{N}(\mu_1 \,,\, \frac{\sigma^2}{n_1})$, como se quer mostrar.
- O estimador de $\alpha_i = \mu_i - \mu_1$, para $i > 1$, é dado pela correspondente diferença de médias amostrais, $\hat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{1.}$. 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 $\overline{Y}_{i.} \cap \mathcal{N}(\mu_1 + \alpha_i \,,\, \frac{\sigma^2}{n_i})$. As duas parcelas são independentes, uma vez que as parcelas que entram para o cálculo da média $\overline{Y}_{1.}$ são diferentes das que entram no cálculo da média $\overline{Y}_{i.}$. Logo, essa diferença de duas variáveis aleatórias Normais independentes tem distribuição Normal. Os parâmetros dessa distribuição são: $E[\hat{\alpha}_i] = E[\overline{Y}_{i.} - \overline{Y}_{1.}] = E[\overline{Y}_{i.}] - E[\overline{Y}_{1.}] = (\mu_1 + \alpha_i) - \mu_1 = \alpha_i$; e $V[\hat{\alpha}_i] = V[\overline{Y}_{i.} - \overline{Y}_{1.}] = V[\overline{Y}_{i.}] + V[\overline{Y}_{1.}] - 2\,\underbrace{Cov[\overline{Y}_{i.}, \overline{Y}_{1.}]}_{=0} = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_1}$ (a covariância é nula, tendo em conta a independência de duas médias de nível diferentes). Logo, $\hat{\alpha}_i \cap \mathcal{N}\left(\alpha_i \,,\, \sigma^2(\frac{1}{n_1} + \frac{1}{n_i})\right)$, 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: $W = \frac{SQRE}{\sigma^2} \cap \chi^2_{n-k}$;
- 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 raíz 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

$$P\left[-t_{\alpha/2\,(n-k)} \;<\; \frac{\hat{\mu_1}-\mu_1}{\sqrt{\frac{QMRE}{n_1}}} \;<\; t_{\alpha/2\,(n-k)}\right] \;=\; 1-\alpha$$

$$\Leftrightarrow \qquad P\left[-t_{\alpha/2\,(n-k)}\cdot\sqrt{\frac{QMRE}{n_1}} \;<\; \hat{\mu_1}-\mu_1 \;<\; t_{\alpha/2\,(n-k)}\cdot\sqrt{\frac{QMRE}{n_1}}\right] \;=\; 1-\alpha$$

$$\Leftrightarrow \qquad P\left[t_{\alpha/2\,(n-k)}\cdot\sqrt{\frac{QMRE}{n_1}} \;>\; \mu_1-\hat{\mu_1} \;>\; -t_{\alpha/2\,(n-k)}\cdot\sqrt{\frac{QMRE}{n_1}}\right] \;=\; 1-\alpha$$

$$\Leftrightarrow \qquad P\left[\hat{\mu_1}-t_{\alpha/2\,(n-k)}\cdot\sqrt{\frac{QMRE}{n_1}} \;<\; \mu_1 \;<\; \hat{\mu_1}+t_{\alpha/2\,(n-k)}\cdot\sqrt{\frac{QMRE}{n_1}}\right] \;=\; 1-\alpha$$

Calculando os extremos deste intervalo de probabilidade para a nossa amostra (e recordando que $\hat{\mu}_1 = \overline{Y}_{1.}$) 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 $W = \frac{SQRE}{\sigma^2} \cap \chi^2_{n-k}$ e repetindo o raciocínio anterior, obtém-se $\frac{Z}{\sqrt{W/(n-k)}} = \frac{\hat{\alpha_i}-\alpha_i}{\sqrt{QMRE\left(\frac{1}{n_1}+\frac{1}{n_i}\right)}} \cap t_{n-k}$.

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 $\sqrt{QMRE\left(\frac{1}{n_1}+\frac{1}{n_i}\right)}$.

18. Neste exercício, tem-se um delineamento a um factor, em que a variável resposta é o rendimento médio por árvore em cada parcela, e o factor é definido pelos genótipos, existindo $k = 35$ diferentes genótipos. No entanto, e tendo em conta que esses $k = 35$ níveis do factor representam uma amostra aleatória duma imensidão de genótipos existentes, é aconselhado tratar os efeitos do factor como *aleatórios* e não (como em todos os modelos ANOVA anteriores) como efeitos fixos. Na prática, isso significa que os efeitos de nível (até aqui representados como α_i) deixam de ser constantes, e passam a ser considerados *variáveis aleatórias* (cujos valores resultaram da experiência aleatória que consiste na selecção aleatória dos genótipos usados na experiência). Esta natureza nova das parcelas do modelo correspondentes aos efeitos do factor exige alguma adaptação do modelo e do estudo, como se verá nas alíneas seguintes.

    (a) Existem $n = 35\times 4 = 140$ observações, Y_{ij}, $n_i = 4$ das quais associadas ao genótipo i ($i = 1, ..., 35$). Tem-se:

        i. $Y_{ij} \;=\; \mu + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}\;, \quad \forall\, i = 1, ..., 35\;, \quad \forall\, j = 1, ..., 4.$
        ii. $\boldsymbol{\alpha}_i \cap \mathcal{N}(0, \sigma^2_\alpha), \forall\, i$
        iii. $\boldsymbol{\epsilon}_{ij} \;\cap\; \mathcal{N}(0\,,\,\sigma^2_\epsilon), \forall\, i, j$

iv. $\{\{\boldsymbol{\alpha}_i\}_i, \{\boldsymbol{\epsilon}_{ij}\}_{i,j}\}$ são $k+n = 35+140 = 175$ v.a.s independentes.

Como referido acima, a principal novidade deste modelo em relação ao modelo ANOVA a um factor de efeitos fixos, consiste no facto de as parcelas correspondentes aos efeitos do factor serem agora variáveis aleatórias $\boldsymbol{\alpha}_i$. Torna-se necessário especificar as distribuições de probabilidades correspondentes a essas novas v.a.s, e o modelo clássico com um factor de efeitos aleatórios admite que essa distribuição é Normal, de média zero e com variância σ^2_α. A fim de evitar confusões, a variância dos erros aleatórios (que continuam a ter distribuição Normal, de média zero e variância comum) passa agora a indicar-se por σ^2_ϵ. A hipótese final de independência passa a dizer respeito, não apenas aos erros aleatórios, mas também aos efeitos do factor. A parcela μ mantém-se constante, como no modelo de efeitos fixos, mas passa a ter uma interpretação diferente: é agora o valor esperado comum a **todas** as observações. De facto, e tendo em conta as hipóteses do modelo (em particular o valor esperado nulo dos efeitos de nível e dos erros), tem-se agora $E[Y_{ij}] = E[\mu + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}] = \mu$. Outra diferença em relação ao modelo de efeitos fixos, é que a variância das observações é agora $V[Y_{ij}] = V[\mu + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}] = \sigma^2_\alpha + \sigma^2_\epsilon$ (não existe covariância, dada a independência entre efeitos do factor e erros aleatórios). Estas duas parcelas designam-se as *componentes da variância*, nome que é também por vezes dado aos modelos de efeitos aleatórios.

(b) A nova hipótese nula a testar, que corresponde à inexistência de efeitos do factor, é $\sigma^2_\alpha = 0$. De facto, se esta hipótese fôr verdadeira, as variáveis aleatórias $\boldsymbol{\alpha}_i$ passam a apenas poder tomar o valor zero (não existindo variabilidade em torno do seu valor médio nulo), o que corresponde a dizer que não existe contribuição dos genótipos para o valor de Y, ou seja, inexistência de efeitos do factor. Esta variância σ^2_α é conhecida, no contexto de problemas genéticos deste tipo, como a *variabilidade genética* ou *genotípica*. Para testar as hipóteses

$$H_0 : \sigma^2_\alpha = 0 \qquad \text{vs.} \qquad H_1 : \sigma^2_\alpha > 0$$

no modelo a um factor de efeitos aleatórios usa-se uma estatística idêntica à usada no modelo de efeitos fixos a um factor, tendo-se a mesma distribuição sob H_0 e uma região crítica idêntica. Assim, é possível aproveitar a tabela-resumo do enunciado (construída para o correspondente modelo de efeitos fixos) e afirmar que a nova hipótese nula $\sigma^2_\alpha = 0$ é rejeitada de forma clara, uma vez o valor de prova associado a F_{calc} é muito baixo: $p = 4.61 \times 10^{-10}$. Assim, concluímos pela existência de variabilidade genética entre os genótipos.

(c) Em problemas semelhantes a este, associados à selecção genética, a variância observada na amostra é normalmente medida pela variância associada às médias amostrais de cada nível do factor (genótipo), que se mostra ser $V[\overline{Y}_{i.}] = \sigma^2_\alpha + \frac{\sigma^2_\epsilon}{n_c}$ (admitindo um delineamento equilibrado com n_c observações em cada nível). Esta variância é também designada por *variabilidade fenotípica*. O conceito de *heritabilidade* é a proporção desta variabilidade fenotípica (observada) que corresponde à variabilidade genotípica (logo, transmissível): $H^2 = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \frac{\sigma^2_\epsilon}{n_c}}$. A heritabilidade é estimada substituindo as variâncias σ^2_ϵ e σ^2_α pelos seus estimadores $\hat{\sigma}^2_\epsilon = QMRE$ (como de costume) e $\hat{\sigma}^2_\alpha = \frac{QMF - QMRE}{n_c}$. Logo, e tendo em conta que $F_{calc} = \frac{QMF}{QMRE}$ vem que a heritabilidade estimada é dada por $\hat{H}^2 = 1 - \frac{1}{F_{calc}}$. No nosso exemplo tem-se $\hat{H}^2 = 1 - \frac{1}{4.723} = 0.7883$. Assim, quase 79% da variabilidade fenotípica observada neste problema é de origem genética. Esta elevada percentagem sugere que a selecção dos melhores de entre os genótipos estudados pode originar ganhos importantes nos rendimentos.