
INSTITUTO SUPERIOR DE AGRONOMIA
ESTATÍSTICA E DELINEAMENTO EXPERIMENTAL– 2024-25
Resoluções dos Exercícios de Delineamento Experimental e de Análise de
Variância de Efeitos Fixos

1. A variável resposta nesta experiência é o tamanho das raízes de estacas de oliveira e o factor que se pretende estudar é a concentração de cálcio no solo com 4 níveis (C1, C2, C3, C4).

(a) As unidades experimentais do estudo são os vasos. Cada vaso vai dar origem a uma observação da variável resposta.

(b) Há cinco repetições. Cada tratamento é aplicado aleatoriamente a 5 vasos.

(c) O delineamento experimental é a um factor (concentração de cálcio com 4 níveis) totalmente casualizado (cuja sigla em inglês é CRD).

(d) Vamos utilizar o *package agricolae* do R. Em geral, a primeira vez que utilizamos um *package* é necessário instalá-lo. Para tal devemos correr o comando

```
install.packages("nome_do_package")
```

(neste caso, `install.packages("agricolae")`). Em alternativa e caso se esteja a utilizar o ambiente RStudio, podemos ir ao separador *Packages*, na janela inferior do lado esquerdo, clicar no botão *Install* e escrever o nome do *package* pretendido (ou ir ao menu *Tools* → *Install Packages ...*).

A seguir é necessário carregar o *package* em memória através do comando

```
library("nome_do_package")
```

(neste caso, `library("agricolae")`). Em próximas sessões de trabalho no R, em que se pretenda utilizar este *package*, já não é necessário instalá-lo, basta carregá-lo em memória com o comando `library`.

De acordo com o delineamento experimental efectuado, podemos gerar uma casualização possível para este ensaio com os seguintes comandos:

```

trt<-c("C1","C2","C3","C4")
EX1.CRD <- design.crd(trt,r=5,seed=123,serie = 0)
EX1.CRD
EX1.CRD$book
$book
  plots r trt
1      1 1 C2
2      2 1 C1
3      3 2 C1
4      4 2 C2
5      5 3 C1
6      6 4 C1
7      7 1 C4
8      8 1 C3
9      9 3 C2
10     10 2 C3
11     11 3 C3
12     12 4 C2
13     13 4 C3
14     14 2 C4
15     15 3 C4
16     16 5 C3
17     17 4 C4
18     18 5 C2
19     19 5 C1
20     20 5 C4

```

- (e) O controlo desta fonte de variação na experiência (diferente temperatura ao longo da bancada) pode ser feita utilizando um outro factor **blocos**. Neste caso, o novo delineamento passa por considerar cinco diferentes posições (linhas) na bancada (os blocos), a primeira mais perto da porta da estufa e as restantes cada vez mais afastadas dessa porta. Em cada posição, ir-se-à distribuir de forma aleatória quatro vasos, cada um sujeito a uma das quatro concentrações de cálcio disponíveis. Este delineamento experimental designa-se por delineamento em blocos completos casualizados (na sigla inglesa RCBD).

Uma casualização possível para este ensaio pode ser gerada com os comandos:

```

EX1.RCBD <- design.rcbd(trt,r=5,seed=257)
EX1.RCBD
EX1.RCBD$book
  plots block trt
1    101     1  C1
2    102     1  C2
3    103     1  C3
4    104     1  C4
5    201     2  C3
6    202     2  C2
7    203     2  C1
8    204     2  C4
9    301     3  C2
10   302     3  C4
11   303     3  C3
12   304     3  C1
13   401     4  C3
14   402     4  C2
15   403     4  C1
16   404     4  C4
17   501     5  C4
18   502     5  C1
19   503     5  C2
20   504     5  C3

```

2. A variável resposta nesse estudo é o rendimento da batata e o factor que se pretende analisar é o tipo de fungicida com 5 níveis, incluindo o tratamento de controlo, sem fungicida (C, F1, F2, F3, F4). Para controlar a heterogeneidade das condições experimentais, foi considerado um outro factor (bloco, com 6 níveis)(as seis parcelas de terreno homogéneas). Cada parcela de terreno (bloco) deverá ser subdividida em cinco talhões e, a cada um deles, será atribuído aleatoriamente um dos cinco tratamentos de fungicida em estudo. Estamos perante um clássico delineamento experimental em blocos completos casualizados, cuja sigla em inglês é RCBD.

Uma casualização possível para este estudo pode ser gerada no R com os comandos:

```

trt<-c("C","F1","F2","F3","F4")
EX2.RCBD <- design.rcbd(trt,r=6,seed=-123)
EX2.RCBD
EX2.RCBD$book
  plots block trt
1    101     1  F1
2    102     1  F4
3    103     1  F2
(...)
28   603     6  F3
29   604     6   C
30   605     6  F2

```

EX2.RCBD\$sketch

```
      [,1] [,2] [,3] [,4] [,5]
[1,] "F1" "F4" "F2" "C"  "F3"
[2,] "C"  "F2" "F4" "F3" "F1"
[3,] "F3" "F2" "C"  "F4" "F1"
[4,] "C"  "F2" "F1" "F3" "F4"
[5,] "F1" "F3" "F4" "F2" "C"
[6,] "F4" "F1" "F3" "C"  "F2"
```

3. A variável resposta nesse estudo é o crescimento do tremoço e pretende-se estudar os efeitos de dois factores: tipo de solo com 2 níveis (G e R) e disponibilidade de água, também com 2 níveis (B e A). Para a implementação da experiência, foram aplicados 4 tratamentos, resultantes da combinação `solo_agua` (GB, GA, RB, RA).

- As unidades experimentais do estudo são os vasos. Cada vaso vai dar origem a uma observação da variável resposta.
- Como cada vaso tem 2 plantas, cada uma delas representa uma pseudo-repetição. Pseudo-repetições ocorrem quando há a possibilidade de se obter mais de uma observação por unidade experimental. Deve-se substituir cada grupo de pseudo-repetições por uma única observação média. Deste modo pode diminuir-se a variabilidade entre as observações (independentes) das várias unidades experimentais. As pseudo-repetições podem, no entanto, causar problemas quando são incorretamente identificadas como unidades experimentais numa análise estatística.
- Como o delineamento é equilibrado, com igual número de repetições por nível do factor em estudo (tratamento), e existem 16 vasos disponíveis para a experiência, cada tratamento pode ser repetido 4 vezes.
- De acordo com a informação disponível, o delineamento experimental irá ser um delineamento totalmente casualizado.
- Uma possível aleatorização pode ser obtida no R com os seguintes comandos:

```
trt<-c("GB","GA","RB","RA")
EX3.rcd<-design.crd(trt, r=4, seed=909)
EX3.rcd
$parameters
$parameters$design
[1] "crd"
$parameters$trt
[1] "GB" "GA" "RB" "RA"
$parameters$r
[1] 4 4 4 4
(...)
$book
plots r trt
1    101 1  GB
2    102 2  GB
3    103 1  RA
(...)
14   114 4  RA
15   115 4  GA
16   116 4  GB
```

- (f) Dado existirem gradientes de temperatura e luminosidade na bancada onde se vão colocar os vasos da experiência, o delineamento considerado deve ter agora em atenção esses dois factores adicionais, temperatura e luz.
- i. Como há agora duas novas fontes de heterogeneidade associadas a esta experiência (temperatura e luz), que podem ter o mesmo número de tratamentos em estudo (combinações solo_agua), o delineamento experimental pode ser feito em **Quadrado Latino** (LSD na sigla inglesa): subdivide-se a bancada em 4 linhas e 4 colunas onde se colocam 16 vasos. Os tratamentos são atribuídos aleatoriamente neste quadrado de forma a que cada um só ocorra uma única vez em cada linha e em cada coluna.
 - ii. A criação de um esquema em quadrado latino é mais complexo que os dos exercícios anteriores. A utilização do *package agricolae* facilita essa tarefa.

```
trt<-c("GB","GA","RB","RA")
EX3.LSD <- design.lsd(trt,seed=9876)
EX3.LSD
EX3.LSD$book
  plots row col trt
1    101  1  1  GA
2    102  1  2  RA
3    103  1  3  GB
4    104  1  4  RB
5    201  2  1  GB
6    202  2  2  RB
7    203  2  3  RA
8    204  2  4  GA
9    301  3  1  RA
10   302  3  2  GA
11   303  3  3  RB
12   304  3  4  GB
13   401  4  1  RB
14   402  4  2  GB
15   403  4  3  GA
16   404  4  4  RA
EX3.LSD$sketch
  [,1] [,2] [,3] [,4]
[1,] "GA" "RA" "GB" "RB"
[2,] "GB" "RB" "RA" "GA"
[3,] "RA" "GA" "RB" "GB"
[4,] "RB" "GB" "GA" "RA"
```

Nota: Num delineamento em quadrado latino, o número de repetições é sempre igual ao número de tratamentos do factor de interesse, neste caso quatro, e por isso não é necessário fornecer essa informação quando se utiliza a função `design.lsd`.

4. Pretende-se estudar os efeitos do factor variedade, com 5 níveis, na produção de frutos em macieiras.
 - (a) Neste caso há uma disposição sistemática das variedades de macieira (uma em cada fila). As unidades experimentais serão as diferentes filas e cada uma das 50 plantas será uma pseudo-repetição. Deste modo, cada grupo de 50 plantas da mesma variedade, irá dar origem a uma única observação por variedade. Logo, não há observações em número suficiente (não há repetições) para se poder estudar o efeito da variedade na produção de fruto das macieiras.
 - (b) Caso se estivesse a plantar um novo pomar com as características descritas no enunciado (5 linhas com 50 plantas/linha) e não havendo heterogeneidade relevante na zona do pomar, podia-se dividir cada linha em 5 talhões de 10 plantas (as unidades experimentais) e distri-

buir de forma totalmente aleatória as 5 variedades pelos 25 talhões disponíveis, de modo a ter 5 repetições por variedade.

Em alternativa e caso este pomar já existisse, podia tentar encontrar-se outros pomares parecidos também plantados com as mesmas 5 variedades. Cada pomar funcionaria como um bloco, dando origem a 5 novas observações, uma de cada variedade, independentes das provenientes de outros pomares (blocos). Neste caso, o delineamento seria em blocos completos casualizados, assumindo-se condições homogêneas dentro de cada pomar e distribuição aleatória das 5 variedades em cada um deles.

Estas são duas possíveis soluções de delinear uma experiência de modo a conseguir responder à questão da existência de efeitos do factor variedade na produção de frutos em macieiras. De acordo com possíveis condições de heterogeneidade dentro de cada pomar, outros delineamentos podiam ser propostos.

5. Neste ensaio há dois factores que interessa estudar: a espécie infestante, com quatro níveis (A, B, C, D), e o stress hídrico/rega, com dois níveis (I, NI). Há, no entanto, diferentes restrições na aplicação destes dois factores: o stress hídrico/rega (factor principal com 2 níveis) só pode ser atribuído a grandes unidades experimentais, ao passo que as espécies de infestantes (factor subplot com 4 níveis) podem ser semeadas em pequenas unidades experimentais. Existem também condições de heterogeneidade, ao nível da fertilidade do solo, que se pretendem controlar: a fertilidade do solo será um novo factor (com quatro níveis) que dá origem à existência de quatro blocos.

Deve assim adoptar-se um delineamento em parcelas divididas (*split-plot design* na terminologia inglesa), em blocos completos casualizados. Cada mancha de solo (bloco), associada a uma diferente fertilidade, é subdividida em duas grandes parcelas (*whole plots* em inglês) onde se atribui aleatoriamente rega (I) e não rega (NI). Finalmente, cada grande parcela é subdividida em quatro pequenas parcelas (*subplots* em inglês) onde se atribui aleatoriamente cada uma das quatro espécies infestantes (A, B, C, D).

Um possível esquema para este delineamento pode ser obtido no R com os seguintes comandos:

```

whole.plot <- c("I","NI")
sub.plot <- LETTERS[1:4]
EX5.SPLIT <- design.split(whole.plot,sub.plot,r=4,design="rcbd",seed=354)
EX5.SPLIT$book

```

	plots	splots	block	whole.plot	sub.plot
1	101	1	1	I	D
2	101	2	1	I	C
3	101	3	1	I	A
4	101	4	1	I	B
5	102	1	1	NI	B
6	102	2	1	NI	C
7	102	3	1	NI	D
8	102	4	1	NI	A
9	103	1	2	NI	D
10	103	2	2	NI	A
11	103	3	2	NI	B
12	103	4	2	NI	C
13	104	1	2	I	A
14	104	2	2	I	D
15	104	3	2	I	B
16	104	4	2	I	C
17	105	1	3	I	B
18	105	2	3	I	D
19	105	3	3	I	A
20	105	4	3	I	C
21	106	1	3	NI	B
22	106	2	3	NI	A
23	106	3	3	NI	C
24	106	4	3	NI	D
25	107	1	4	NI	B
26	107	2	4	NI	A
27	107	3	4	NI	C
28	107	4	4	NI	D
29	108	1	4	I	B
30	108	2	4	I	C
31	108	3	4	I	D
32	108	4	4	I	A

6. O factor em estudo nesta experiência é o clone de eucalipto, com 4 níveis (E1, E2, E3, E4).

- (a) É o técnico B que propõe o procedimento correcto. Os tabuleiros serão as unidades experimentais e as estacas presentes em cada um são pseudo-repetições. No caso do procedimento proposto pelo técnico A, só existiriam 4 tabuleiros e uma única observação por clone de eucalipto, não havendo a informação necessária para permitir estudar o sucesso de enraizamento das estacas para os diferentes clones de eucalipto.

Pelo contrário, o procedimento proposto pelo técnico B, utiliza um total de 32 tabuleiros, 8 por cada clone, com 10 estacas (pseudo-repetições) em cada um. Assim, os 8 tabuleiros (repetições) irão dar origem a oito observações independentes que permitirão estudar a variabilidade da variável resposta dentro do mesmo clone, estimar a variância dos erros aleatórios e comparar a variabilidade entre clones.

(b) Há agora um novo factor (luz) que deve ser controlado para garantir condições adequadas à realização da experiência.

- i. As estacas de cada um dos tabuleiros são pseudo-repetições. Logo há 10 pseudo-repetições por tabuleiro.
- ii. As unidades experimentais são os tabuleiros.
- iii. Há oito repetições por clone nesta experiência.
- iv. O delineamento adequado será em blocos completos casualizados: em cada prateleira (bloco) será colocado aleatoriamente um tabuleiro de cada um dos 4 clones. No total deverão existir 8 prateleiras, cada uma com diferentes condições de luminosidade. Um possível esquema para este delineamento pode ser obtido no R com os seguintes comandos:

```
trt <- c("E1","E2","E3","E4")
EX6.RBCD <- design.rcbd(trt, r=8,seed=579)
EX6.RBCD$book
```

	plots	block	trt						
	1	101	1	E1	[1,]	"E1"	"E3"	"E2"	"E4"
	2	102	1	E3	[2,]	"E2"	"E4"	"E3"	"E1"
	3	103	1	E2	[3,]	"E3"	"E2"	"E1"	"E4"
	(...)				[4,]	"E4"	"E1"	"E3"	"E2"
	30	802	8	E1	[5,]	"E3"	"E4"	"E1"	"E2"
	31	803	8	E3	[6,]	"E2"	"E3"	"E1"	"E4"
	32	804	8	E4	[7,]	"E3"	"E1"	"E2"	"E4"
					[8,]	"E2"	"E1"	"E3"	"E4"

```
EX6.RBCD$sketch
```

7. Pretende-se estudar a acidez total do mosto (variável resposta) em três variedades de videira (factor videira com 3 níveis, V1, V2, V3). Como se sabe que os resultados são afetados pelo técnico de laboratório que efectua a análise e com o objectivo de controlar esta variabilidade adicional na experiência, vai considerar-se os diferentes técnicos como níveis de um novo factor (A, B, C, D, E, F). O delineamento será assim feito em blocos completos casualizados, atribuindo aleatoriamente a cada técnico (bloco) um saco de bagos de cada uma das variedades de videira. Uma possível aleatorização desta experiência pode ser obtida no R através dos seguintes comandos:

```
trt <- c("V1","V2","V3")
EX7.RBCD <- design.rcbd(trt, r=6,seed=102938)
EX7.RBCD
EX7.RBCD$book
```

	plots	block	trt	
	1	101	1	V1
	2	102	1	V2
	3	103	1	V3
	4	201	2	V2
	(...)			
	15	503	5	V1
	16	601	6	V1
	17	602	6	V3
	18	603	6	V2

```
row.names(EX7.RBCD$sketch) <- LETTERS[1:6]
```



```
EX7.RBCD$sketch
  [,1] [,2] [,3]
A "V1" "V2" "V3"
B "V2" "V3" "V1"
C "V2" "V1" "V3"
D "V3" "V1" "V2"
E "V2" "V3" "V1"
F "V1" "V3" "V2"
```

8. O factor em estudo é a marca comercial do sumo de maçã com 8 níveis (S1, S2, S3, S4, S5, S6, S7, S8). De modo a controlar a subjectividade (heterogeneidade) das respostas dadas por cada provador, o delineamento mais adequado será em blocos completos casualizados. Cada provador é considerado um bloco e a ordem pela qual vai provar cada um dos 8 sumos será atribuída de forma aleatória.

Uma possível aleatorização desta experiência pode ser obtida no R através dos seguintes comandos:

```
#criar o vector (S1, S2, ..., S8)
trt <- paste("S",1:8,sep="")
EX8.RBCD <- design.rcbd(trt,r=10,
                        seed=675489)
EX8.RBCD$book
  plots block trt
1    101     1  S7
2    102     1  S8
3    103     1  S3
4    104     1  S4
(...)
77  1005    10  S2
78  1006    10  S7
79  1007    10  S5
80  1008    10  S1

#criar o vector (P1, P2, ..., P10)
provador <- paste("P",1:10,sep="")
row.names(EX8.RBCD$sketch) <- provador
EX8.RBCD$sketch
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
P1 "S7" "S8" "S3" "S4" "S6" "S2" "S1" "S5"
P2 "S5" "S2" "S1" "S7" "S8" "S4" "S3" "S6"
P3 "S1" "S8" "S7" "S3" "S5" "S6" "S2" "S4"
P4 "S1" "S7" "S8" "S3" "S2" "S5" "S4" "S6"
P5 "S6" "S3" "S5" "S4" "S1" "S7" "S8" "S2"
P6 "S3" "S4" "S6" "S8" "S5" "S1" "S7" "S2"
P7 "S1" "S4" "S6" "S8" "S3" "S5" "S7" "S2"
P8 "S4" "S8" "S2" "S1" "S5" "S6" "S7" "S3"
P9 "S6" "S1" "S7" "S5" "S2" "S3" "S8" "S4"
P10 "S3" "S4" "S8" "S6" "S2" "S7" "S5" "S1"
```

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

O modelo ANOVA a um factor corresponde a:

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

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

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

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

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

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

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

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

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

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

```
> tomate.aov <- aov(res.pel ~ variedade , data=tomate)
> summary(tomate.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
variedade	5	478507	95701	10.6	0.000448
Residuals	12	108296	9025		

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

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

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

Nível de significância: $\alpha = 0.05$.

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

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

- (d) O valor de prova (p -value) associado ao valor calculado da estatística de teste é $p = 0.000448$. Pela própria definição de p -value, esta é a área à direita de $F_{calc} = 10.6044$, numa distribuição $F_{[5,12]}$. Logo, seria preciso fazer um teste de hipóteses com nível de significância $\alpha = 0.000448$ (ou inferior) para que F_{calc} não pertencesse à Região Crítica e a conclusão do teste pudesse ser a de não rejeitar H_0 .
- (e) Tal como nas regressões lineares, a primeira coluna da matriz \mathbf{X} é uma coluna de uns. No contexto duma ANOVA a um factor, as restantes colunas são variáveis indicatrizes de pertença de cada observação a um dos níveis do factor, ou seja, colunas com apenas dois valores: “1” associado a observações que pertencem ao nível do factor em causa, e “0” associado a observações associadas a outros níveis do factor. A restrição imposta no modelo ($\alpha_1 = 0$) implica que não há indicatriz do primeiro nível do factor, neste caso, o nível “18”. Assim, neste caso teremos uma primeira coluna de $n = 18$ uns e cinco colunas indicatrizes dos segundo, terceiro, quarto, quinto e sexto níveis do factor ($\mathcal{I}_2, \mathcal{I}_3, \mathcal{I}_4, \mathcal{I}_5$ e \mathcal{I}_6), como se pode confirmar através do comando referido no enunciado:

```

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

```

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

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

```

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

```

Estas são as médias de variedade dadas no enunciado.

- (g) Os parâmetros do modelo são: $\mu_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ e α_6 . As estimativas destes parâmetros são:

$\hat{\mu}_1 = 560.6433$ *gf*, a resistência média observada nos frutos da variedade 18,

$\hat{\alpha}_2 = 241.4833 - 560.6433 = -319.1600$ *gf*, a diferença da resistência média observada nos frutos da variedade 28 face à resistência média observada nos frutos da variedade 18;

$\hat{\alpha}_3 = 290.9500 - 560.6433 = -269.6933$ *gf*, a diferença da resistência média observada nos frutos da variedade 29 face à resistência média observada nos frutos da variedade 18;

$\hat{\alpha}_4 = 705.7800 - 560.6433 = 145.1367$ *gf*, a diferença da resistência média observada nos frutos da variedade 40C face à resistência média observada nos frutos da variedade 18;

$\hat{\alpha}_5 = 377.2533 - 560.6433 = -183.3900$ *gf*, a diferença da resistência média observada nos frutos da variedade Ace face à resistência média observada nos frutos da variedade 18;

$\hat{\alpha}_6 = 332.1067 - 560.6433 = -228.5367$ *gf*, a diferença da resistência média observada nos frutos da variedade Roma face à resistência média observada nos frutos da variedade 18.

No R estas estimativas obtêm-se com o comando:

```

> coef(tomate.aov)

```

```

(Intercept)  variedade28  variedade29  variedade40C
      560.6433    -319.1600    -269.6933     145.1367
variedadeAce variedadeRoma
      -183.3900    -228.5367

```

Ou, uma vez que estamos perante um modelo análogo a um modelo de Regressão Linear Múltipla, diferindo apenas na natureza das variáveis preditoras, que são aqui variáveis indicatrizes dos níveis 2 a 6 do factor, as estimativas dos parâmetros obtêm-se com o comando:

```

> tomate.lm<-(lm(res.pel ~ variedade , data=tomate))
> summary(tomate.lm)
Call:
lm(formula = res.pel ~ variedade, data = tomate)

```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	560.64	54.85	10.222	2.82e-07
variedade28	-319.16	77.57	-4.115	0.00143
variedade29	-269.69	77.57	-3.477	0.00457
variedade40C	145.14	77.57	1.871	0.08590
variedadeAce	-183.39	77.57	-2.364	0.03577
variedadeRoma	-228.54	77.57	-2.946	0.01223

```

---
Residual standard error: 95 on 12 degrees of freedom
Multiple R-squared:  0.8154, Adjusted R-squared:  0.7385
F-statistic: 10.6 on 5 and 12 DF,  p-value: 0.0004476

```

- (h) Para comparar médias de variedade iremos utilizar o teste de Tukey. Sabemos que podemos considerar diferentes duas médias populacionais de nível, μ_i e $\mu_{i'}$, caso as respectivas médias amostrais de nível difiram mais do que o termo de comparação do teste de Tukey, ou seja, se $|\bar{y}_i - \bar{y}_{i'}| > q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}}$, onde $q_{\alpha(k,n-k)}$ indica o valor que deixa à sua direita uma região de probabilidade α , na distribuição de Tukey com parâmetros k e $n-k$. No nosso caso, $k=6$ e $n-k=12$, sendo, pelas tabelas da distribuição Tukey, $q_{0.05(6,12)}=4.75$. Como $\sqrt{\frac{QMRE}{n_c}} = \sqrt{\frac{9024.685}{3}} = 54.84732$, podemos decretar a diferença significativa entre a média amostral das resistências da variedade 40C, $\bar{y}_4 = 705.8$ (que é a maior de todas), e a de qualquer outra variedade cuja média difira desta em mais de $4.75 \times 54.84732 = 260.5248$ gf. Ora, $705.8 - 260.5248 = 445.2752$, e apenas a variedade 18 não tem média amostral inferior a esse valor. Logo, podemos concluir que as resistências médias das variedades 28, 29, Ace e Roma são diferentes (inferiores) à resistência média da variedade 40C. O *package agricolae* tem a função `HSD.test` que permite realizar as comparações múltiplas de médias com base no resultado de Tukey. No final do *output* desta função, todas as médias amostrais referentes ao modelo ajustado são apresentadas por ordem decrescente e têm associadas letras que permitem identificar quais as que (não) são significativamente diferentes."

```

> library(agricolae)
> HSD.test (tomate.aov, "variedade", console=TRUE)

```

```
Study: tomate.aov ~ "variedade"
```

```
HSD Test for res.pel
```

```
Mean Square Error: 9024.685
```

```
variedade, means
```

	res.pel	std r	Min	Max
18	560.6433	121.297482	3 420.59	632.04
28	241.4833	19.181852	3 219.34	253.00

29	290.9500	76.693685	3	223.71	374.48
40C	705.7800	182.023736	3	503.51	856.39
Ace	377.2533	2.326893	3	375.18	379.77
Roma	332.1067	6.863791	3	324.82	338.45

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

Minimum Significant Difference: 260.5375

Treatments with the same letter are not significantly different.

	res.pel	groups
40C	705.7800	a
18	560.6433	ab
Ace	377.2533	bc
Roma	332.1067	bc
29	290.9500	c
28	241.4833	c

- (i) O facto dos resíduos se encontrarem ‘empilhados’ em seis colunas é o reflexo natural do facto, referido na alínea anterior, de apenas haver seis diferentes valores ajustados nesta ANOVA: as seis médias amostrais de cada variedade, $\hat{y}_{ij} = \bar{y}_i$. ($j = 1, 2, 3$). Este facto ajuda a identificar as observações associadas aos resíduos de maior magnitude. Assim, por exemplo, o maior resíduo (em módulo) corresponde ao ponto no canto inferior direito. Por estar associado a uma média \bar{y}_i de aproximadamente 700, tem de corresponder à variedade 40C. Por ser um resíduo negativo, tem de corresponder a uma observação com valor inferior à média dessa variedade, o que apenas acontece com a primeira das três observações desse nível. Assim, a observação a que corresponde o referido resíduo é a observação $y_{4,1} = 503.51$. Embora o número de repetições em cada nível ($n_c = 3$) seja muito baixo, e portanto susceptível de gerar impressões enganadoras, o gráfico sugere alguma heterogeneidade nas variâncias de Y_{ij} em cada nível. Os valores das variâncias amostrais de nível indicam que há variedades com muito pouca variabilidade nas resistências observadas (como a *Ace*, com $s_5^2 = 5.414433$) e outras com uma variabilidade muito maior (como a *29*, com $s_3^2 = 5881.921$, mais de mil vezes maior).

10. Neste exercício sobre os estomas das folhas de café, não estão disponíveis os dados originais. Apenas se conhece a tabela dos valores médios e variâncias amostrais de cada variedade.

- (a) A variável resposta Y é o comprimento médio dos estomas das folhas duma planta. Para explicar a variabilidade dos valores desta variável, apenas se dispõe de um factor: o factor variedade, com $k = 3$ níveis (as três variedades indicadas no enunciado). É um delineamento totalmente casualizado, a 1 factor, equilibrado, pois existem $n_i = 12$ observações para qualquer variedade ($i = 1, 2, 3$), perfazendo um total de $n = 3 \times 12 = 36$ observações Y_{ij} . Eis o modelo:

i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall i = 1, 2, 3$, $j = 1, 2, \dots, 12$, com $\alpha_1 = 0$, onde

- Y_{ij} indica o comprimento médio dos estomas das folhas da planta j da variedade i ;
- μ_1 indica o comprimento médio populacional dos estomas das folhas de plantas da primeira variedade ($i = 1$) que é, por ordem alfabética, a variedade **CA**;
- α_i indica o efeito (acréscimo em relação à média da variedade **CA**) da variedade i ; e
- ϵ_{ij} indica o erro aleatório associado à observação Y_{ij} .

- ii. $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\forall i, j$.
 iii. $\{\epsilon_{ij}\}_{i,j}$ constitui um conjunto de variáveis aleatórias independentes.
- (b) Começemos pelo cálculo das Somas de Quadrados. Uma vez que o delineamento é equilibrado (igual número de observações em cada nível), a média global da totalidade das 36 observações ($\bar{y}_{..}$) é a média simples das três médias de nível dadas na tabela: $\bar{y}_{..} = (22.85833 + 19.49333 + 25.31583)/3 = 22.55583$. Tendo em conta as fórmulas vistas nas aulas teóricas e os valores dados no enunciado, temos:

$$SQRE = (n_c - 1) \sum_{i=1}^3 s_i^2 = 11 \times (13.69303 + 2.725424 + 9.388936) = 284.1983 ;$$

$$SQF = n_c \sum_{i=1}^3 (\bar{y}_i - \bar{y}_{..})^2$$

$$= 12 \times ((22.85833 - 22.55583)^2 + (19.49333 - 22.55583)^2 + (25.31583 - 22.55583)^2)$$

$$= 205.0561,$$

Logo, tem-se a seguinte tabela-resumo:

Fonte	g.l.	SQ	QM	F_{calc}
Factor	$k - 1 = 2$	$SQF = 205.0561$	$QMF = \frac{SQF}{k-1} = 102.5281$	$\frac{QMF}{QMRE} = 11.90516$
Resíduos	$n - k = 33$	$SQRE = 284.1983$	$QMRE = \frac{SQRE}{n-k} = 8.61207$	

- (c) Neste caso, e uma vez que não são conhecidas as observações individuais, apenas é possível calcular a variância da totalidade das $n = 36$ observações recorrendo à decomposição da Soma de Quadrados Total correspondente a esta ANOVA:

$$s_y^2 = \frac{SQT}{n-1} = \frac{SQF + SQRE}{n-1} = \frac{205.0561 + 284.1983}{35} = \frac{489.2544}{35} = 13.9787 .$$

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

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

Nível de significância: $\alpha = 0.05$.

Região Crítica (Unilateral Direita): Rejeitar H_0 se $F_{calc} > f_{0.05(2,33)} \approx 3.30$ (entre os valores tabelados 3.23 e 3.32).

Conclusões: O valor da estatística do teste foi calculado na alínea anterior: $F_{calc} = 11.90516$. 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 o comprimento médio dos estomas das folhas não é igual em todas as variedades.

NOTA: o valor de prova associado à estatística calculada é (tendo em conta a natureza unilateral direita do teste) $P[F_{(2,33)} > F_{calc}] = P[F_{(2,33)} > 11.90516]$. Não é possível obter este valor nas tabelas, mas pode calcular-se essa probabilidade com o auxílio do **R**:

```
> 1-pf(11.90516, 2,33)
[1] 0.000128065
```

Assim, tem-se $p = 0.000128065$.

- (e) Sabemos que duas médias de nível μ_i e $\mu_{i'}$ devem ser consideradas diferentes caso as respectivas médias amostrais difiram (em módulo) 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 $k = 3$ e $n = 36$. 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 $q_{0.05(3,33)} = 3.47$. Um valor mais preciso pode ser obtido através do comando `qtukey` do **R**:

```
> qtukey(0.95, 3, 33)
[1] 3.470189
```

Sabemos pela alínea (b) que $QMRE = 8.61207$ e também que $n_c = 12$. Logo, o termo de comparação é dado por $q_{\alpha(k,n-k)} \sqrt{\frac{QMRE}{n_c}} = 3.470189 \times \sqrt{\frac{8.61207}{12}} = 2.93979$. Calculando as diferenças, em módulo, entre as médias amostrais de cada variedade, obtém-se a seguinte tabela:

$ \bar{y}_i - \bar{y}_{i'} $	CA ($i'=1$)	CL ($i'=2$)	PR ($i'=3$)
CA ($i=1$)	–	3.3650	2.4575
CL ($i=2$)	3.3650	–	5.8225
PR ($i=3$)	2.4575	5.8225	–

Assim, ao nível de significância $\alpha = 0.05$, o comprimento médio dos estomas de folhas da variedade **CL** é diferente, quer do comprimento médio da variedade **CA**, quer do comprimento médio da variedade **PR**. No entanto, não se pode considerar (por pouco) significativamente diferentes os comprimentos médios dos estomas de folhas das variedades **CA** e **PR**.

Existem duas formas frequentes de representar esta conclusão, sendo usual em ambas ordenar os níveis do factor por ordem crescente das respectivas médias, e:

- i. sublinhando-se com traços os grupos de níveis cujas médias não diferem significativamente o que, nesta alínea (ao nível $\alpha = 0.05$) produz o seguinte resultado:

CL	CA	PR
19.49333	<u>22.85833</u>	<u>25.31583</u>

- ii. ou colocando uma mesma letra ao lado das variedades cujas médias não se consideram significativamente diferentes, por exemplo:

CL	CA	PR
19.49333 ^a	22.85833 ^b	25.31583 ^b

Assim, a média de **CL** é significativamente diferente das médias, quer de **CA**, quer de **PR** (com quem não partilha letras em comum), mas já a média da variedade **CA** não difere significativamente da média de **PR** (uma vez que partilham a mesma letra).

11. A variável resposta Y é, neste caso, a variação de massa (coluna `variacao.massa` no ficheiro `C02.txt`). Existem ao todo $n = 50$ observações.

- (a) A experiência foi estabelecida de acordo com um delineamento totalmente casualizado. 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 totalmente casualizado a 1 factor, equilibrado, é o seguinte:

- i. $Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}$, $\forall i = 1, 2, 3, 4, 5$, $j = 1, 2, \dots, 10$, 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 a 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. $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\forall i, j$.
- 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 no ficheiro `C02.txt`, com os valores da variável resposta na coluna `variacao.massa` e os diferentes níveis de CO_2 no factor `C02.factor` (note-se que, por omissão, o R vai ler estes dados como valores numéricos; para esta coluna ser considerada um factor deve-se utilizar a função `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 $n - k = 45$.

```
> C02$C02.factor <- factor(C02$C02.factor)
> summary(aov(variacao.massa ~ C02.factor, data=C02))
              Df Sum Sq Mean Sq F value Pr(>F)
C02.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: $F = \frac{QMF}{QMRE} \sim F_{(k-1, n-k)}$, 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) Pedem-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:

$$\left[(\bar{y}_i - \bar{y}_j) - q_{\alpha(k, n-k)} \sqrt{\frac{QMRE}{n_c}}, (\bar{y}_i - \bar{y}_j) + q_{\alpha(k, n-k)} \sqrt{\frac{QMRE}{n_c}} \right].$$

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: $\bar{y}_1 = 59.14$. De forma análoga, a média amostral no segundo nível é: $\bar{y}_2 = 46.04$. 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.

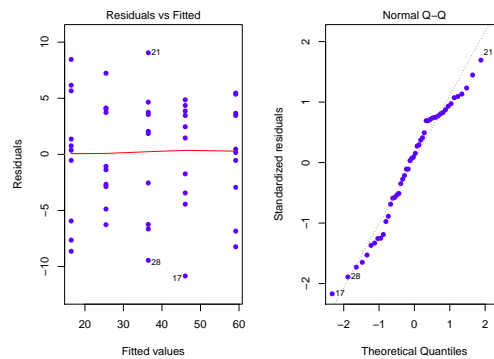
- (d) Como em qualquer modelo linear, o resíduo é a diferença entre cada valor observado da variável resposta e o correspondente valor ajustado pelo modelo, ou seja, e usando a notação da ANOVA a 1 Factor, $e_{ij} = y_{ij} - \hat{y}_{ij}$. Sabe-se que, num modelo ANOVA a um factor, o valor ajustado duma dada observação corresponde à média amostral das observações no mesmo nível do factor: $\hat{y}_{ij} = \bar{y}_i$. Assim, todas as observações do primeiro grupo têm valor

ajustado igual a $\hat{y}_{1j} = \bar{y}_1 = 59.14$. 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 $\hat{y}_{2j} = \bar{y}_2 = 46.04$. O resíduo da terceira observação do segundo grupo é assim $e_{23} = y_{23} - \bar{y}_2 = 47.5 - 46.04 = 1.46$. Para calcular a totalidade dos resíduos podemos recorrer ao R (arredondando a três casas decimais):

```
> round(residuals(C02.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(C02.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 $\hat{y}_{ij} = \bar{y}_i$, 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 dos erros aleatórios, uma vez que a variabilidade dos resíduos não parece diferir muito nos cinco níveis do factor.

O *qq-plot* (gráfico à direita) não indicia problemas graves com a Normalidade dos erros aleatórios, 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 \bar{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.

- (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=C02, pch=16)
```

- ii. A regressão linear pedida é dada por:

```
> C02.lm <- lm(variacao.massa ~ CO2.numerico, data=C02)
```

```
> summary(C02.lm)
```

Coefficients:

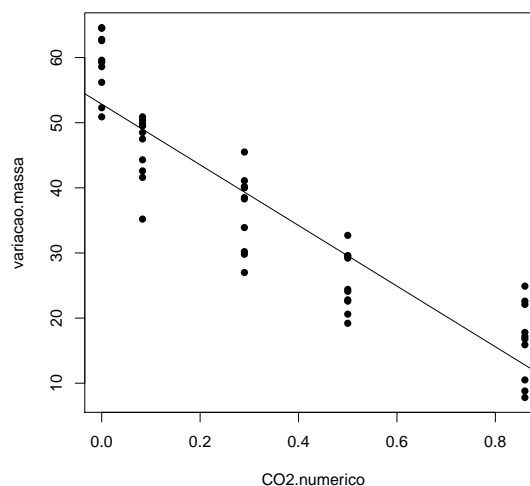
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	52.849	1.408	37.52	<2e-16 ***
CO2.numerico	-46.569	3.030	-15.37	<2e-16 ***

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(C02.lm)`) é:



Apesar de alguma tendência para uma relação curvilínea, 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 ($R^2 = 0.8312$). Esta regressão explica 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 `C02.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 `C02.factor`:

```
> summary(lm(variacao.massa ~ C02.factor, data=C02))
(...)
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).

12. (a) A descrição da experiência corresponde a um delineamento factorial a dois factores sem interação, sendo o primeiro factor constituído pelas fases do processamento e o segundo factor constituído 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 (estamos, portanto, perante um típico delineamento experimental em blocos completos casualizados)¹.

Após importar para o R o ficheiro de dados `zinco.txt` e converter a segunda e terceira colunas em factores deverá ter-se a *data frame* `zinco` com três colunas: a variável resposta

¹Seria 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.

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

- i. $Y_{ijk} = \mu_{11} + \alpha_i + \beta_j + \epsilon_{ijk}$, $\forall i = 1, 2, 3, 4$, $j = 1, 2, \dots, 9$, $k = 1$ (o índice k é dispensável porque não há repetições nas células), com $\alpha_1 = 0$ e $\beta_1 = 0$, e onde
 - Y_{ijk} indica a 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. $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$.
- 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:

```
> zinco.aov <- aov(concentracao ~ fase + lote, data=zinco)
> summary(zinco.aov)
```

	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* interação, 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(zinco.aov, 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: $SQA = b n_c \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 = 9 \times ((2.228 - 2.847778)^2 + (2.847 - 2.847778)^2 + (2.233 - 2.847778)^2 + (4.083 - 2.847778)^2) = 20.59066$, e $SQB = a n_c \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2 = 4 \times ((3.483 - 2.847778)^2 + (3.733 - 2.847778)^2 + \dots + (2.443 - 2.847778)^2) = 17.76391$. 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: $F = \frac{QMA}{QMRE} \sim F_{(a-1, n-(a+b-1))}$, 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) É pedido o valor ajustado para a (única) observação na célula (1,1), ou seja, é pedido o valor de \hat{y}_{111} . Com o auxílio do comando `fitted` do R, verifica-se que esse valor é $\hat{y}_{111} = 2.862500$:

```
> fitted(zinco.aov)
      1      2      3      4      5      6      7      8      9
2.862500 3.112500 2.937500 2.377500 2.805000 1.320000 1.237500 1.575000 1.822500
      10     11     12     13     14     15     16     17     18
3.481389 3.731389 3.556389 2.996389 3.423889 1.938889 1.856389 2.193889 2.441389
      19     20     21     22     23     24     25     26     27
2.868056 3.118056 2.943056 2.383056 2.810556 1.325556 1.243056 1.580556 1.828056
      28     29     30     31     32     33     34     35     36
4.718056 4.968056 4.793056 4.233056 4.660556 3.175556 3.093056 3.430556 3.678056
```

Importa salientar que, ao contrário do que sucede numa ANOVA a um factor, este valor ajustado *não é* a média das observações nessa célula (o que neste caso corresponderia a dizer a única observação na célula (1,1), ou seja, $y_{111} = 2.23$). Tratando-se da célula de referência (a célula para a qual os efeitos de ambos os factores foram igualados a zero, $\alpha_1 = \beta_1 = 0$), sabemos que o valor médio nessa célula é o parâmetro $\mu_{11} = E[Y_{111}]$. Como se viu nas aulas teóricas, esse valor esperado é estimado por $\hat{Y}_{111} = \bar{Y}_{1..} + \bar{Y}_{.1.} - \bar{Y}_{...}$, ou seja, pela soma das médias das observações na respectiva linha e respectiva coluna, menos a média global de todas as observações. Essas três médias já foram calculadas na alínea 12b, mas para minorar os erros de arredondamento serão de novo calculadas:

```
> mean(zinco[zinco$fase=="1",1])
[1] 2.227778
> mean(zinco[zinco$lote=="1",1])
[1] 3.4825
> mean(zinco$conc)
[1] 2.847778
> 2.227778 + 3.482500 - 2.847778
[1] 2.8625
```

- (e) 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**. Assim, tenderá a aumentar-se o denominador ($QMRE$) na estatística $F = \frac{QMF}{QMRE}$ aos efeitos do factor A.
- 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.

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

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

- (b) Sabemos que os graus de liberdade associados aos efeitos de factor correspondem ao número de níveis do factor, menos um. Assim, no nosso caso, tem-se que os g.l. de factor genótipo são $a - 1 = 5$, e os do factor terreno são $b - 1 = 5$. Os graus de liberdade residuais podem ser calculados como o que falta para que a soma dê $n - 1 = 35$, ou seja, $n - (a + b - 1) = 25$, e assim se completa a primeira coluna da tabela. Tendo em conta que o Quadrado Médio Residual é, por definição, $QMRE = \frac{SQRE}{n - (a + b - 1)}$, tem-se $SQRE = QMRE \times (n - (a + b - 1)) = 0.3660 \times 25 = 9.15$, e assim se completa a última linha da tabela. Os dois Quadrados Médios em falta (QMA e QMB) podem ser ambos calculados através do conhecimento dos valores calculados das duas estatística F , disponíveis na tabela. De facto, por definição, $F_A = \frac{QMA}{QMRE}$, pelo que $QMA = F_A \times QMRE = 4.204 \times 0.3660 = 1.538664$. Por um raciocínio análogo, tem-se $QMB = F_B \times QMRE = 2.691 \times 0.3660 = 0.984906$, e assim se completa a penúltima coluna da tabela. Faltam apenas os valores das Somas de Quadrados associadas aos dois factores: SQA e SQB. Mas, por definição, tem-se $QMA = \frac{SQA}{g.l.(SQA)}$, pelo que $SQA = QMA \times (a - 1) = 1.538664 \times 5 = 7.69332$. De forma inteiramente análoga, obtém-se o valor de SQB: $SQB = QMB \times (b - 1) = 0.984906 \times 5 = 4.92453$. Resumindo, tem-se:

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

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

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

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

Nível de significância: $\alpha = 0.01$.

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

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

Agora o teste a efeitos de terreno:

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

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

Nível de significância: $\alpha = 0.01$.

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

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

- (d) O enunciado pede para considerar o que aconteceria se, aos mesmos dados, fosse ajustado um modelo ANOVA com um único factor, o factor **genótipo**. O pedido corresponde a ignorar a existência do factor terreno (embora ele tenha sido considerado no delineamento experimental que foi efectivamente usado), tratando-se as seis observações de cada genótipo como meras repetições. Nesse caso, e como se viu nas aulas teóricas, a tabela ANOVA terá apenas duas linhas: uma correspondente ao único factor agora considerado (genótipo) e outra residual. A linha da tabela correspondente ao factor genótipo permanece inalterada quanto a graus de liberdade (na notação dos modelos a um factor tem-se $k = a = 6$, logo continua a ter-se $a - 1 = 5$ g.l. associados aos genótipos); Soma de Quadrados ($SQA = SQF = n_c \sum_{i=1}^6 (\bar{y}_i - \bar{y}_{..})^2$); e (por conseguinte) Quadrado Médio ($QMA = \frac{SQA}{a-1}$). Já quanto à nova Soma de Quadrados Residual, tem de corresponder à soma das antigas parcelas SQB e $SQRE_{2f}$ no modelo a dois factores, sem interacção, ajustado inicialmente. De facto, e como se viu nas aulas teóricas, a Soma de Quadrados Total não depende do modelo ajustado, mas apenas dos valores de Y observados ($SQT = (n - 1) s_y^2$). No modelo a dois factores, sem interacção, essa Soma de Quadrados foi decomposta como $SQT = SQA + SQB + SQRE_{2f}$. A mesma Soma de quadrados é agora decomposta como $SQT = SQA + SQRE_{1f}$. Sendo igual o total (SQT) e a primeira parcela em cada decomposição (SQA), necessariamente se tem $SQRE_{1f} = SQB + SQRE_{2f}$. Logo, $SQRE_{1f} = 4.92453 + 9.15 = 14.07453$. Assim, o novo Quadrado Médio Residual é $QMRE_{1f} = \frac{SQRE_{1f}}{n-a} = \frac{14.07453}{30} = 0.469151$. O valor da (única) estatística F existente no modelo a um factor será agora: $F = \frac{QMA}{QMRE_{1f}} = \frac{1.538664}{0.469151} = 3.279678$. Assim, a tabela do modelo a um único factor será:

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

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

- (e) A hipótese cujo estudo se pede é a hipótese de existirem *efeitos de interacção* entre genótipos e terrenos. Trata-se efectivamente duma hipótese possível (que seria um caso particular duma interacção genótipo \times ambiente). Mas não é possível ajustar um modelo que preveja

essa possibilidade (o modelo a dois factores *com* interacção) pois, como já se referiu, não existem repetições nas células.

14. (a) Trata-se de um clássico delineamento em blocos completos casualizados, portanto, com dois factores (**variedade**, com 4 níveis e **terreno** ("bloco"), com 13 níveis). Em cada terreno ou "bloco", apenas há uma parcela com cada variedade. Logo, só é possível ajustar um modelo a dois factores factorial sem interacção, tal como nos exercícios 12 e 13.

```
> terrenos.aov <- aov(rend ~ variedade + terreno, data=terrenos)
```

- (b) A tabela-resumo correspondente é:

```
> summary(terrenos.aov)
              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
```

Efectuando, em pormenor, o teste aos 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: $F = \frac{QMA}{QMRE} \sim F_{(a-1, n-(a+b-1))}$, 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: $F_{calc} = 6.145$. 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.

Em alternativa pode-se utilizar o p-value associado ao valor calculado da estatística, concluindo-se que, ao nível $\alpha = 0.05$, deve considerar-se a existência de efeitos do factor variedade ($p - value < 0.05$).

- (c) 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: $H_0 : \beta_j = 0, \forall j = 2, \dots, 13$ vs. $H_1 : \exists j = 2, \dots, 13$ tal que $\beta_j \neq 0$.

Estatística do teste: $F = \frac{QMB}{QMRE} \sim F_{(b-1, n-(a+b-1))}$, 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 depende 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 $QMRE$.

- (d) 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 $\hat{y}_{ijk} = \bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...}$, 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, $\bar{y}_{1..} = 1.556$, admitindo que o factor A é o factor variedade. A média das quatro observações associadas ao terreno I é $\bar{y}_{.1.} = (1.800 + 2.457 + 0.722 + 0.789)/4 = 1.4420$. 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 $\bar{y}_{...} = 1.358308$. 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(terrenos.aov, 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(terrenos.aov)
      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
[...]
```

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

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

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

- Admite-se que os erros aleatórios são todos Normais, de média zero e variâncias homogêneas: $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k .
- Admite-se que os erros aleatórios ϵ_{ijk} são independentes.

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

Sabemos que, neste tipo de modelo, os graus de liberdade associados a $QMRE$ são dados por $n-ab$, onde $n=60$ é o número total de observações e $ab=10$ é o número de parâmetros existentes no modelo. Assim, $g.l.(SQRE)=50$. Sabemos ainda que, para os vários tipos de efeitos, os graus de liberdade são dados pelo número de parcelas de cada tipo de efeito, após a introdução das restrições, ou seja, associado a SQA há $a-1=4$ g.l., associado a SQB há $b-1=1$ g.l., e associado a $SQAB$ há $(a-1)(b-1)=4$ graus de liberdade.

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

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

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

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

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

```
(c) pmanso<-read.table("pmanso.txt",header=T )
pmanso.aov<-aov(alt2~prov*local, data=pmanso)
summary(pmanso.aov)
> summary(pmanso.aov)
              Df Sum Sq Mean Sq F value    Pr(>F)
prov           4   280.6     70.2    4.228  0.00503 **
local          1   785.6    785.6   47.340  9.23e-09 ***
prov:local     4   139.4     34.8    2.100  0.09471 .
Residuals     50   829.7     16.6
```

- (d) Vai-se efectuar em pormenor o teste aos efeitos de interacção entre proveniência e local.

Hipóteses: $H_0 : (\alpha\beta)_{ij} = 0$, para todo o i e j vs. $H_1 : \exists i, j$ tal que $(\alpha\beta)_{ij} \neq 0$.

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

Nível de significância: $\alpha = 0.05$.

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

Conclusões: Como $F_{calc} = \frac{QMA}{QMRE} = 2.093 < 2.57$, não se rejeita H_0 (para $\alpha = 0.05$), e conclui-se pela inexistência de efeitos significativos de interacção.

Em alternativa pode-se utilizar o p -value associado ao valor calculado da estatística, concluindo-se que, ao nível $\alpha = 0.05$, não se rejeita H_0 (p -value > 0.05), concluindo-se, portanto, que não existem efeitos de interacção.

O teste aos efeitos principais do factor proveniência, as hipóteses do teste podem ser escritas apenas como $H_0 : \alpha_i = 0, \forall i$ vs. $H_1 : \exists i$ tal que $\alpha_i \neq 0$. O valor calculado da estatística de teste é $F_{calc} = \frac{QMA}{QMRE} = 4.229$. Nas tabelas o limiar da região crítica unilateral direita: $f_{0.05(4,50)} \approx 2.57$ (entre os valores tabelados 2.53 e 2.61). Como $F_{calc} = \frac{QMA}{QMRE} = 4.229 > 2.57$, rejeita-se H_0 , sendo possível concluir pela existência de efeitos principais de proveniência (ao nível $\alpha = 0.05$) (ou, alternativamente, p -value $= 0.00503 < 0.05$, logo, rejeita-se H_0).

No teste aos efeitos principais do factor local do estudo, as hipóteses do teste podem ser escritas apenas como $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, uma vez que após a imposição da restrição $\beta_1 = 0$, apenas sobra um efeito deste tipo, o efeito β_2 associado a Tavira. O valor calculado da estatística de teste é muito grande ($F_{calc} = 47.394$) deixando antever a rejeição de H_0 , facto que é confirmado determinando nas tabelas o limiar da região crítica unilateral direita: $f_{0.05(1,50)} \approx 4.04$ (entre os valores tabelados 4.00 e 4.08) (ou, alternativamente, p -value $= 9.23e - 09 < 0.05$, logo, rejeita-se H_0). Assim, conclui-se claramente pela existência de efeitos principais de localidade, o que neste caso significa que existe um efeito associado à passagem do local de plantação de Sines para Tavira. Uma rápida inspecção das médias de local sugere que se trata dum maior crescimento dos pinheiros em Tavira, pelo que se deduz que β_2 terá um valor positivo.

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

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

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

16. Trata-se dum delineamento totalmente casualizado, 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 factorial, 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) Começamos pelo teste aos efeitos de interacção entre os dois tipos de adubação, associados às parcelas $(\alpha\beta)_{ij}$. Em concreto:

Teste à interacção. As hipóteses são:

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

Se efetuarmos seguidamente os testes aos efeitos principais dos factores:

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

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

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

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

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 $\bar{y}_{13} = 6.733$ e $\bar{y}_{23} = 7.6$. 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 calcular 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 ($QMRE$).

17. (a) A experiência foi conduzida em 2 locais, em cada local o ensaio com as diferentes cultivares foi estabelecido de acordo com um delineamento totalmente casualizado. 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 factorial, com interacção, dado por:

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

(b) i. Os nove valores em falta na tabela são dados por:

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

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

iii. Em primeiro lugar, 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 $j = 2, 3, \dots, 9$ [não há interacção]
vs. $H_1 : \exists i = 2, 3, 4$, $j = 2, 3, \dots, 9$ tais que $(\alpha\beta)_{ij} \neq 0$ [há interacção].

Estatística do teste: $F = \frac{QMAB}{QMRE} \sim F_{[(a-1)(b-1), n-ab]}$, 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 $H_0 : \beta_j = 0, \forall j = 2, 3, \dots, 9$ vs. $H_1 : \exists j = 2, 3, \dots, 9$, tal que $\beta_j \neq 0$. A Região Crítica é agora dada pela rejeição de H_0 caso $F_{calc} > f_{0.01(8,108)} \approx 2.68$. O valor da estatística calculada $F_{calc} = 3.698$ pertence à Região Crítica, levando à rejeição de H_0 , concluindo-se também pela existência de efeitos de cultivar sobre os rendimentos.

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

- iv. Os dois gráficos de interacção reflectem a mesma informação, embora de formas diferentes. No gráfico da esquerda, as quatro localidades definem posições no eixo horizontal. Por cima de cada localidade encontram-se nove pontos, associados às nove cultivares. A ordenada de cada um desses nove pontos é dada pelo rendimento médio das parcelas correspondentes a essa combinação de localidade e cultivar. Os segmentos de recta unem os pontos correspondentes a cada cultivar (segundo a legenda indicada no gráfico). Embora haja algum paralelismo nas nove curvas seccionalmente lineares, para as três primeiras localidades, os rendimentos na Revilheira sugerem a existência de efeitos de interacção. Por exemplo, a cultivar *TE9110* que, quer em Évora, quer em Elvas, regista o rendimento mais baixo (facto que se pode confirmar na tabela de médias dada na alínea c) tem o segundo mais elevado rendimento na Revilheira. Também a cultivar *Celta*, cujo rendimento em Benavila é o terceiro mais baixo, regista o segundo maior rendimento em Elvas. Assim, há cultivares que manifestam “preferências” ou “aversões” por diferentes localidades, reflectindo efeitos de interacção. O teste à interacção efectuado na alínea anterior confirma que esses efeitos são significativos, ao nível $\alpha = 0.01$. O gráfico da direita dá, como se disse, uma perspectiva diferente sobre a mesma informação. Agora, são as cultivares que definem nove posições no eixo horizontal. Por cima de cada uma dessas posições (cultivares) há quatro pontos, com ordenadas dadas pelos rendimentos médios da referida cultivar, nas quatro localidades consideradas no ensaio. Segmentos de recta unem os pontos correspondentes a uma mesma localidade. Neste gráfico torna-se evidente que os rendimentos são sempre bastante superiores em Elvas (no gráfico da esquerda, esse facto reflectia-se no “pico” por cima de Elvas). Essa será a principal razão pela clara rejeição da hipótese nula no teste à existência de efeitos principais de localidade. Por outro lado, os efeitos de interacção reflectem-se na mais visível ausência de paralelismo, nomeadamente nos traços correspondentes a Évora e Revilheira, que para várias cultivares parecem ter comportamentos quase antagónicos.
- v. Pedem-se para discutir o efeito sobre a tabela resultante de dividir a variável resposta por mil (passando o rendimento a ser expresso em t/ha). Os graus de liberdade não são, naturalmente, afectados. O mesmo não se passa com as Somas de Quadrados. À nova variável $Y^* = Y/1000$ corresponderão novas médias de nível, de célula e global, que também resultam de dividir por mil (para ficarem em t/ha). Tendo em conta que no modelo em questão, as médias de célula definem os valores ajustados, tem-se $\hat{Y}_{ijk}^* = \hat{Y}_{ijk}/1000$. Assim, as novas Somas de Quadrados resultam de dividir as suas congéneres

originais por 1000^2 , ou seja, por um milhão. De facto, $SQT^* = \sum_i \sum_j \sum_k (Y_{ijk}^* - \bar{Y}_{...}^*)^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \bar{Y}_{...}/1000)^2 = SQT/(1000^2)$. Também $SQRE^* = \sum_i \sum_j \sum_k (Y_{ijk}^* - \hat{Y}_{ijk}^*)^2 = \sum_i \sum_j \sum_k (Y_{ijk}/1000 - \hat{Y}_{ijk}/1000)^2 = SQRE/(1000^2)$. De forma análoga, e utilizando as fórmulas para delineamentos equilibrados,

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

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

Por diferença, tem igualmente de verificar-se $SQAB^* = SQAB/(1000^2)$. Assim, toda a coluna de Somas de Quadrados na tabela será dividida por um milhão. Essa mesma transformação aplica-se à coluna de Quadrados Médios (que resulta de dividir Somas de Quadrados por graus de liberdade). Mas na coluna final, correspondente aos valores calculados das estatísticas F , o quociente de Quadrados Médios mantém-se inalterado (a transformação multiplicativa de numerador e denominador é igual). Logo, as conclusões de todos os testes (incluindo os respectivos p -values) mantêm-se inalterados.

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

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

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

18. (a) Admite-se um delineamento totalmente casualizado, 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 (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) do conteúdo de taninos na célula definida pelo nível i do Factor A (*Temperatura*) e o nível j do Factor B (*Tempo de armazenamento*);
 - μ_{11} indica a média (populacional) das observações do conteúdo de taninos na célula (1, 1), ou seja, com temperatura alta e 0 dias de armazenamento;

- α_i indica o efeito principal do nível i do Factor A (*Temperatura*);
- β_j indica o efeito principal 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. $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, $\forall i, j, k$.

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 é 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 $SQA = bn_c \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 = 16 [(24.681 - 22.14375)^2 + (19.606 - 22.14375)^2] = 16 \times 12.87781 = 206.045$. A última Soma de Quadrados em falta ($SQAB$) pode ser calculada a partir das restantes quatro: $SQAB = SQT - (SQA + SQB + SQRE) = 1482.799 - (206.045 + 288.03 + 20.72) = 968.004$. Assim,

Variacção	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: $F = \frac{QMAB}{QMRE} \sim F_{[(a-1)(b-1), n-ab]}$, 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: $F = \frac{Q_{MB}}{Q_{MRE}} \sim F_{(b-1, n-ab)}$, 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

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

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 $\alpha = 0.05$.

19. Para modelar a variável resposta rendimento, existem dois factores: o local e ano. Em cada local o delineamento foi totalmente casualizado (em cada local foram aleatoriamente atribuídas as unidades experimentais onde se efectuaram as avaliações nos diferentes anos) mas não se trata dum delineamento factorial: os anos observados em cada local não são os mesmos.

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) :

$Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, $\forall i=1,2$, $j=1,2,3=b_1$ (se $i=1$) ou $j=1,2=b_2$ (se $i=2$), $k=1,2,\dots,8$, 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,

• Y_{ijk} indica o rendimento (em kg/planta) na parcela k , do local i , no ano j ;

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

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.

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 aov 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 “/”):

```
(a) > TabRegua <- read.delim("TabRegua.txt", stringsAsFactors=TRUE)
> TabRegua$ano <- factor(TabRegua$ano)
> 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: $F = \frac{QMB(A)}{QMRE} \sim F_{[(b_1-1)+(b_2-1), n-(b_1+b_2)]}$, 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).

Em alternativa pode-se utilizar o p-value associado ao valor calculado da estatística do teste.

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: $F = \frac{QMA}{QMRE} \sim F_{[a-1, n-(b_1+b_2)]}$, 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.

Em alternativa pode-se utilizar o p-value associado ao valor calculado da estatística do teste.

- (b) 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 \bar{y}_{13} (Régua em 2002) é significativamente diferente de todas as outras:

$\bar{y}_{11.}$	$\bar{y}_{22.}$	$\bar{y}_{21.}$	$\bar{y}_{12.}$	$\bar{y}_{13.}$
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:

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

Este mesmo resultado pode ser obtido no R utilizando o *package agricolae*.

```
> library(agricolae)
> HSD.test(TabRegua.aov, trt=c("local", "ano"), console=TRUE)
Study: TabRegua.aov ~ c("local", "ano")
HSD Test for rend
Mean Square Error: 0.188483
local:ano, means
      rend      std r      se  Min  Max   Q25   Q50   Q75
Regua:1999 0.269125 0.2922975 8 0.1534939 0.019 0.917 0.11375 0.1880 0.28500
Regua:2000 0.686500 0.2950288 8 0.1534939 0.125 0.992 0.48800 0.7965 0.89200
Regua:2002 1.351500 0.6665498 8 0.1534939 0.594 2.315 0.79600 1.2945 1.78750
Tabuaco:1999 0.645875 0.4286825 8 0.1534939 0.092 1.292 0.35000 0.6250 0.97475
Tabuaco:2003 0.475125 0.3766671 8 0.1534939 0.092 1.188 0.23925 0.3510 0.56650
```

```
Alpha: 0.05 ; DF Error: 35
Critical Value of Studentized Range: 4.065949
```

```
Minimum Significant Difference: 0.6240985
```

```
Treatments with the same letter are not significantly different.
```

```
      rend groups
Regua:2002 1.351500 a
Regua:2000 0.686500 b
Tabuaco:1999 0.645875 b
Tabuaco:2003 0.475125 b
Regua:1999 0.269125 b
```

20. (a) Trata-se dum delineamento totalmente casualizado, com dois factores em estudo, o factor *casta* (factor A), e o factor *genótipo* (factor B). As oito repetições dos diferentes genótipos de cada casta foram aleatoriamente dispostas no terreno. O objectivo do estudo é avaliar os eventuais efeitos destes factores sobre a variável resposta (rendimento). Pela própria natureza dos factores em questão, o delineamento deve ser considerado *hierarquizado*, com genótipos subordinados a castas. Não faria sentido considerar o delineamento factorial: não há cruzamentos entre cada um dos oito genótipos e cada uma das duas castas, já que um genótipo apenas faz sentido quando referido à sua casta.

Assim, temos $a = 2$ castas (níveis do factor A) e, para o factor subordinado genótipos, há $b_1 = 4$ genótipos para a casta 1 (Antão Vaz) e $b_2 = 4$ genótipos para a casta 2 (Malvasia Fina). Ao todo há $b_1 + b_2 = 8$ situações experimentais, e $n_c = 8$ repetições em cada uma das situações experimentais, num total de $n = 64$ observações. O modelo mais adequado será o modelo hierarquizado:

- $Y_{ijk} = \mu_{11} + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$, $\forall i, j, k$, onde Y_{ijk} indica o rendimento da repetição k ($k = 1, 2, \dots, 8$) do genótipo j ($j = 1, 2, 3, 4$) da casta i ($i = 1, 2$). Impõem-se as

restrições $\alpha_1 = 0$, $\beta_{1(i)} = 0$ para $i = 1, 2$. Com estas restrições, o parâmetro μ_{11} é o rendimento médio populacional do primeiro genótipo da casta 1, isto é, do genótipo AN105 da casta Antão Vaz; α_2 é o efeito da casta Malvasia Fina; $\beta_{j(i)}$ ($j = 2, 3, 4$) é o efeito do genótipo j na casta $i = 1, 2$, e ϵ_{ijk} é o erro aleatório associado à observação Y_{ijk} , que corresponde à variabilidade não explicada pelos efeitos previstos no modelo.

- $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, para qualquer i, j, k .
- Os erros aleatórios ϵ_{ijk} são independentes.

- (b) Sabemos que os graus de liberdade na tabela-resumo da ANOVA são dados por: $a - 1 = 1$ para o efeitos de castas; $(b_1 - 1) + (b_2 - 1) = 6$ para os efeitos do factor subordinado, genótipos; e $n - (b_1 + b_2) = 64 - 8 = 56$ para o residual. Por outro lado, conhecemos a partir do enunciado a Soma de Quadrados do Factor A (castas), $SQA = 79.73597$ e o Quadrado Médio Residual, $QMRE = \frac{SQRE}{n - (b_1 + b_2)} = 2.873782$, de onde é possível obter a Soma de Quadrados Residual $SQRE = 2.873782 \times 56 = 160.9318$. A Soma de Quadrados associada ao factor subordinado (genótipos) pode ser obtida pela diferença da soma das outras SQs já calculadas em relação à Soma de Quadrados Total, que sai do conhecimento da variância amostral da totalidade das 64 observações. Assim, $SQT = (n - 1)s_y^2 = 63 \times 5.389415 = 339.5331$, logo $SQB(A) = SQT - (SQA + SQRE) = 339.5331 - (79.73597 + 160.9318) = 98.86533$. Os Quadrados Médios restantes obtêm-se dividindo Somas de Quadrados pelos respectivos graus de liberdade e os valores das duas estatísticas F resultam de dividir o correspondente quadrado médio pelo $QMRE$. Os valores resultantes são sintetizados na tabela em baixo.

Variacão	g.l.	SQs	QMs	F
Casta (A)	1	79.73597	79.73597	$F_A = \frac{79.73597}{2.873782} = 27.74601$
Genótipo [B(A)]	6	98.86533	16.47755	$F_{B(A)} = \frac{16.47755}{2.873782} = 5.733751$
Residual	56	160.9318	2.873782	–
Total	63	339.5331	5.389415	–
	$(n - 1)$	(SQT)	(s_y^2)	–

- (c) Para responder será necessário efectuar um teste F aos efeitos do factor subordinado (genótipos), cuja hipótese nula corresponde à inexistência desse tipo de efeitos.

Hipóteses: $H_0 : \beta_{j(i)} = 0, \forall i, j$ vs. $H_1 : \exists i, j$ tal que $\beta_{j(i)} \neq 0$.

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

Nível de significância: O enunciado pede o nível $\alpha = 0.05$.

Região Crítica: (Unilateral direita) Rejeitar H_0 se $F_{calc} > f_{0.05(6,56)}$ que, pelas tabelas é um valor entre os valores tabelados 2.25 e 2.34.

Conclusões: Como $F_{calc} = 5.733751 > 2.34$, rejeita-se H_0 , o que corresponde a admitir a existência de efeitos de genótipos.

Assim, foi importante prever este tipo de efeitos. Ignorar a existência de efeitos de genótipos iria inflacionar a Soma de Quadrados Residual, o que poderia mascarar a existência de efeitos do outro factor (casta), mesmo que eles existam.

- (d) Um teste análogo, mas aos efeitos do factor dominante (casta) terá como hipóteses $H_0 : \alpha_2 = 0$ (uma vez que apenas existem duas castas e impôs-se a restrição $\alpha_1 = 0$) vs. $H_1 : \alpha_2 \neq 0$. A região crítica deste teste (igualmente unilateral direita) é $f_{0.05(1,56)}$, um valor entre os valores tabelados 4.00 e 4.08. Como $F_{calc} = 27.746 > 4.08$, rejeita-se a hipótese nula. Assim, conclui-se (ao nível de significância $\alpha = 0.05$) que o efeito $\alpha_2 \neq 0$, ou seja que, para além de existirem efeitos de genótipos, há um efeito significativo de casta, e havendo apenas duas castas, pode-se afirmar que os rendimentos da casta Malvasia Fina são significativamente diferentes dos da casta Antão Vaz.

(e) O genótipo MF201 referido no enunciado tem o maior rendimento médio amostral $\bar{y}_{2,4} = 7.678$ (ordenando os genótipos como o R). Pretende-se saber que outras médias amostrais \bar{y}_{ij} diferem significativamente de $\bar{y}_{2,4}$. Utilizaremos as comparações múltiplas de Tukey ao nível global $\alpha = 0.05$. O termo de comparação correspondente é $q_{\alpha(b_1+b_2, n-(b_1+b_2))} \sqrt{\frac{QMRE}{n_c}} = q_{0.05(8,56)} \sqrt{\frac{2.873782}{8}} \approx 4.45 \times 0.5993519 = 2.667$. Qualquer média amostral de rendimento de genótipo inferior a $7.678 - 2.667 = 5.011$ deverá assim ser considerada significativamente diferente da média do genótipo MF201. Há apenas dois genótipos que não têm rendimentos significativamente diferentes, ambos da casta Malvasia Fina: MF1420 e MF1426. Assim, não se rejeitam as hipóteses $\mu_{MF201} = \mu_{MF1420}$ e $\mu_{MF201} = \mu_{MF1426}$. Os três genótipos em questão são da casta Malvasia Fina, o que é coerente com a conclusão da alínea anterior: para além de efeitos de genótipo, é possível falar de efeitos de casta, sendo os rendimentos da casta Malvasia Fina globalmente superiores.

21. (a) Trata-se dum delineamento a dois factores, totalmente casualizado por local (em cada local foram aleatoriamente atribuídas as unidades experimentais onde se efectuaram as avaliações nos diferentes anos) – 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 $n = n_c \left(\sum_{i=1}^2 b_i \right) = 4 \times 5 = 20$ 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.

- $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, 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. Começamos 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: $F = \frac{QMB(A)}{QMRE} \sim F \left[\sum_{i=1}^2 (b_i - 1), n - \sum_{i=1}^2 b_i \right]$, 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: $F = \frac{QMA}{QMRE} \sim F_{\left[a-1, n-\sum_{i=1}^2 b_i \right]}$, 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:

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

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 $|\bar{y}_{22} - \bar{y}_{23}| = 19.9325 - 12.9425 = 6.99$, 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,

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

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 21b. 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.