

Exercícios - Estatística e Delineamento - 2015/16

4 Análise de Variância

AVISO: Os conjuntos de dados de vários Exercícios desta secção encontram-se disponíveis numa área de trabalho (só acessível na rede local do ISA). Para disponibilizar estes conjuntos de dados deve-se:

- Descarregar o ficheiro `exerANOVA.RData` para a directoria onde tem a sua sessão de trabalho (de preferência uma pasta chamada `AulasED` num *pen*).
- Executar, a partir duma sessão do R nessa directoria, o comando `load("exerANOVA.RData")`.

Na sessão do R deverão estar agora disponíveis (confirme com o comando `ls()`) os objectos `toxicos` (Exercício 1), `CO2` (Exercício 3), `zinco` (Exercício 4), `terrenos` (Exercício 5), `decomposicao` (Exercício 6), `absorcao` (Exercício 8), `leitoeis` (Exercício 12) e `TabRegua` (Exercício 14).

EXERCÍCIOS

1. Um estudo sobre a presença de substâncias tóxicas no meio ambiente, associadas à utilização de um pesticida envolve análises efectuadas por quatro laboratórios diferentes. Suspeita-se que os resultados destas análises diferem devido à utilização de diferentes técnicas laboratoriais o que, a ser verdade, exigiria maior controlo das técnicas laboratoriais usadas.

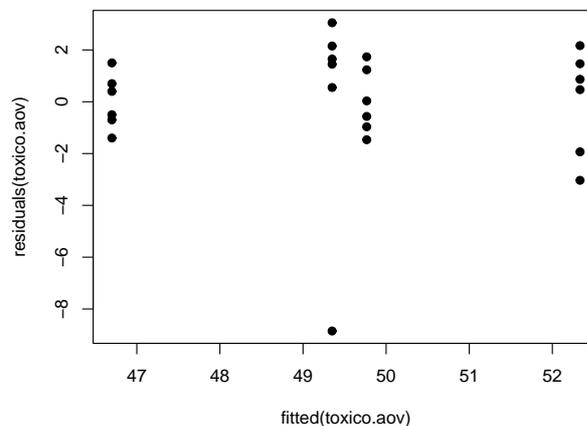
A fim de avaliar a questão, foram entregues a cada laboratório 6 pequenos contentores com solo recolhido aleatoriamente num terreno que antes fora tratado com o referido pesticida. Os resultados laboratoriais das análises químicas medem a concentração dum composto químico nocivo (em ppm). Os valores observados são indicados na tabela (e que estão disponíveis na *data frame* `toxicos`, que pode ser carregada para o R, seguindo as instruções dadas à cabeça destes Exercícios).

Laboratórios				
	1	2	3	4
	53.2	51.0	47.4	51.0
	54.5	40.5	46.2	51.5
	52.8	50.8	46.0	48.8
	49.3	51.5	45.3	49.2
	50.4	52.4	48.2	48.3
	53.8	49.9	47.1	49.8
Média ($\bar{y}_{i.}$)	52.3333	49.3500	46.7000	49.7667
variância amostral (s_i^2)	4.1507	19.4750	1.1200	1.5867

- A média global das observações é $\bar{y}_{..} = 49.5375$;
- a variância amostral da totalidade das observações é $s_y^2 = 9.8868$.

- (a) Indique qual o tipo de delineamento experimental em causa. Explícite o modelo correspondente e todas as hipóteses adicionais que sejam necessárias à consideração do problema em estudo.
- (b) Construa a tabela-resumo da análise de variância correspondente a este caso,
 - i. Utilizando apenas uma máquina de calcular e a informação disponível neste enunciado.
 - ii. Utilizando, no R, o comando `aov(concentracao ~ laboratorio, data=toxicos)`.

- (c) Formalize e efectue o teste F adequado ao problema acima referido, com um nível de significância de 5%. Parece-lhe necessário exigir aos laboratórios a adopção de uma técnica padronizada?
- (d) Repita agora o teste de hipóteses da alínea anterior, mas utilizando o nível de significância de 1%. Compare os resultados e comente.
- (e) Utilize o comando `model.matrix` do R para inspecionar a natureza da matriz do modelo, \mathbf{X} , neste contexto.
- (f) Utilize o comando `fitted` do R para identificar os valores ajustados da variável resposta, nesta Análise de Variância.
- (g) O gráfico dos resíduos (usuais) das observações, contra os valores ajustados pelo modelo de análise de variância, é apresentado a seguir. Comente o gráfico e as suas possíveis implicações. Identifique a observação cujo resíduo é, em módulo, mais elevado.



2. Num estudo sobre o café Robusta em Angola¹ analisa-se a frequência de defeitos no grão de café para lotes provenientes de 6 diferentes regiões de Angola. Admita que, em cada região, foram escolhidos aleatoriamente 11 lotes². Na tabela seguinte são apresentadas, para cada região, as médias e desvios-padrão da percentagem do peso total de grãos sem defeito, assim como as variâncias e log-variâncias associadas.

Região	Média	Variância (s_i^2)	$\ln(s_i^2)$
Cabinda	44.19	48.1636	3.8746
Ambriz	58.87	360.2404	5.8868
Encoje	51.80	193.7664	5.2667
Cazongo	60.61	176.6241	5.1740
Libolo	61.96	599.7601	6.3965
Amboim	42.11	454.1161	6.1184

- (a) Descreva o delineamento experimental e o modelo ANOVA que lhe está associado.
- (b) Construa o quadro-resumo associado a esta ANOVA.

¹ Contribuição para o Estudo das Características dos Cafés em Angola, A. Baião Esteves e J. Santos Oliveira.

² Esta hipótese não corresponde ao estudo original, antes citado.

- (c) Determine a variância da totalidade das $n = 66$ observações.
- (d) Teste a hipótese de a percentagem de grão sem defeito depender da região de proveniência do grão. Utilize os níveis de significância 0.05 e 0.01 e comente. Determine o valor de prova (p -value) associado ao valor calculado da estatística.
- (e) Independentemente do resultado da alínea anterior, verifique quais as regiões cujas médias de grãos com defeito devem ser consideradas diferentes, e quais aquelas em que se pode admitir a igualdade, utilizando um teste de Tukey, com um nível de significância 0.05.
- (f) Pretende-se testar a validade da hipótese de homogeneidade de variâncias exigida no modelo que indicou na primeira alínea. Verifique que o valor calculado da estatística do teste de Bartlett é $K_{calc}^2 = 15.53$. Quais as suas conclusões?
3. Sabe-se que o dióxido de carbono tem um efeito crítico no crescimento de populações microbianas; pequenas quantidades de CO_2 podem estimular o crescimento de algumas espécies enquanto que, pelo contrário, grandes concentrações têm de forma geral uma acção inibitória. Este último efeito é usado comercialmente para preservar alimentos armazenados.

Realizou-se um estudo para investigar a acção de diferentes concentrações de CO_2 na taxa de crescimento de *Pseudomonas fragi*; os diferentes níveis (tratamentos) foram pré-fixados e a variável resposta medida foi a percentagem de variação na massa das culturas após uma hora de crescimento nas respectivas condições, originando os dados da seguinte tabela.

Concentração de CO_2					
	0.0	.083	.29	.50	.86
	62.6	50.9	45.5	29.5	24.9
	59.6	44.3	41.1	22.8	17.2
	64.5	47.5	29.8	19.2	7.8
	59.3	49.5	38.3	20.6	10.5
	58.6	48.5	40.2	29.2	17.8
	64.6	50.4	38.5	24.1	22.1
	50.9	35.2	30.2	22.6	22.6
	56.2	49.9	27.0	32.7	16.8
	52.3	42.6	40.0	24.4	15.9
	62.8	41.6	33.9	29.6	8.8

Estes dados estão disponíveis na *data frame* C02, sendo as concentrações de CO_2 repetidas em duas colunas: numa sob a forma de factor e noutra sob a forma de variável numérica.

- (a) Pretende-se testar a hipótese nula $H_0 : \mu_1 = \mu_2 = \dots = \mu_5$, onde μ_i indica a taxa de crescimento esperada para a i -ésima concentração de CO_2 . É sugerida a utilização de uma Análise de Variância. Enuncie os pressupostos necessários para poder efectuar o teste referido.
- (b) Haverá evidência suficiente para rejeitar H_0 com uma significância de $\alpha = .05$?
- (c) Caso tenha respondido afirmativamente na alínea anterior, indique os pares de concentrações de CO_2 que têm efeitos diferentes sobre o crescimento do *Pseudomonas fragi*, recorrendo aos intervalos de confiança de Tukey.
- (d) Estude a validade dos pressupostos do modelo ANOVA.
- (e) Dada a natureza da variável preditora, também poderia ser considerada uma regressão linear das taxas de crescimento sobre as concentrações de dióxido de carbono, encaradas como uma variável numérica. Utilizando a coluna de C02 com as concentrações dadas como variáveis numéricas (isto é, a coluna C02.numerico), responda às seguintes questões.
- i. Construa a nuvem de pontos da variação de massa sobre concentração de CO2.

- ii. Ajuste a regressão linear simples referida, traçando a recta de regressão sobre a nuvem de pontos. Comente.
 - iii. Compare os resultados do teste F de ajustamento global obtidos usando os comandos `lm` e `aov`. Comente.
4. Pretende-se avaliar se uma técnica de processamento de feijão verde altera a concentração de zinco neste alimento. Há especial interesse em avaliar a referida concentração em quatro fases do processamento: no início; imediatamente antes do chamado branqueamento; logo após o branqueamento; e no final do processamento.

Uma vez que se sabe que diferentes lotes de feijão diferem nas concentrações iniciais de zinco, decidiu-se utilizar nove diferentes lotes, medindo as respectivas concentrações nas quatro fases do processamento acima indicadas. As medições obtidas estão indicadas na Tabela (e na *data.frame* zinco).

Lote	Fases do processamento			
	1	2	3	4
1	2.23	3.71	2.53	5.46
2	2.20	4.67	2.87	5.19
3	2.44	3.45	2.83	5.51
4	2.11	2.73	2.33	4.82
5	2.30	2.58	2.19	6.63
6	1.72	1.85	1.80	2.39
7	1.78	1.81	1.75	2.09
8	2.36	2.32	1.83	2.27
9	2.91	2.50	1.97	2.39

- (a) Identifique o delineamento experimental utilizado e descreva o respectivo modelo ANOVA (de efeitos fixos).
 - (b) Construa a tabela de Análise de Variância para o caso em estudo.
 - (c) Efectue o teste F para saber se diferentes fases do processamento afectam as concentrações médias de zinco ($\alpha = 0.05$). Comente.
 - (d) Por engano, um analista considerou que existia apenas um factor no estudo, correspondendo às diferentes fases do processamento. Compare as tabelas-resumo da ANOVA obtidas em cada caso e comente.
5. Pretende-se comparar o rendimento obtido com quatro variedades de trigo. Identificaram-se 13 terrenos com características de solos diferentes, que correspondem aos tipos de terrenos nos quais se pretende fazer as culturas. Os 13 terrenos são então divididos em quatro parcelas de igual dimensão. Em cada terreno associa-se, de forma aleatória, uma parcela a cada uma das quatro variedades. Após a colheita registam-se os rendimentos obtidos (em t/ha) na tabela (e disponíveis na *data.frame* terrenos).
- (a) As médias amostrais de cada variedade sugerem que há variedades com desempenho superior. Mas serão essas diferenças significativas? A fim de responder, efectue uma Análise da Variância adequada, construindo a tabela-resumo correspondente. Comente as suas conclusões.
 - (b) Teste se, entre terrenos, existem diferenças significativas, como seria de supôr. Comente.
 - (c) Calcule o rendimento ajustado para a variedade A no terreno I. Compare com o rendimento observado e comente.

Terreno	Variedade			
	A	B	C	D
I	1.800	2.457	0.722	0.789
II	1.709	1.839	1.546	1.304
III	1.277	1.293	1.515	1.273
IV	1.675	1.745	0.800	0.846
V	1.814	1.833	1.678	1.732
VI	1.896	1.203	1.192	1.580
VII	1.078	1.689	1.583	1.168
VIII	1.740	1.518	1.050	1.305
IX	1.200	1.133	0.778	1.033
X	1.500	0.722	0.636	0.925
XI	1.932	1.700	1.203	0.850
XII	1.169	1.209	1.112	0.986
XIII	1.438	1.577	1.355	1.525
Médias	1.556	1.532	1.167	1.178
Variâncias	0.0879	0.1855	0.1266	0.0934

6. Deseja-se estudar a decomposição de folhas em quatro meios-ambiente diferentes e para três diferentes durações da exposição ao meio-ambiente. Mede-se a perda de peso de pacotes de folhas, em gramas. O estudo deu os seguintes resultados (disponíveis na *data.frame* `decomposicao`):

AMBIENTE	TEMPO		
	1 mês	2 meses	3 meses
1	1.09	1.35	1.60
	1.06	1.53	1.40
2	1.16	1.38	2.18
	1.03	1.35	1.77
3	1.01	1.63	1.66
	1.04	1.51	1.98
4	0.90	1.60	1.73
	1.03	1.72	1.76

- Identifique o delineamento experimental utilizado e descreva o modelo ANOVA adequado.
 - Construa a tabela-resumo respectiva.
 - Efectue um teste à existência de interacção entre tempo de decomposição e meio-ambiente. Utilize o nível de significância $\alpha = 0.01$.
 - Teste a existência de efeitos principais significativos de meios ambiente, ao nível de 5%.
 - Teste a existência de efeitos principais significativos de tempos de decomposição ($\alpha = 0.01$).
 - Construa os gráficos de interacção e comente-os.
 - Qual a combinação de meio-ambiente e duração que apresenta o maior grau de decomposição das folhas na amostra? Pode-se concluir que essa é a combinação a que, na população, está associada a maior decomposição das folhas?
7. Uma experiência visa estudar o rendimento duma variedade de trigo em função de diferentes formas de aplicar dois adubos, um com fósforo (um adubo fosfatado), e outro com potássio. Consideram-se três dosagens de aplicação do adubo fosfatado, designadas por Baixa, Média e Elevada. Igualmente, consideram-se três dosagens de aplicação do fertilizante com potássio, igualmente designadas por

Baixa, Média e Elevada. A experiência realiza-se num terreno com 27 parcelas de igual dimensão. Repartem-se, de forma totalmente casualizada, três parcelas por cada combinação de dosagem de um e outro fertilizante. Os resultados obtidos (em t/ha) foram os seguintes:

		Potássio (K)									Média	Variância
		Baixa			Média			Elevada				
Fósforo (P)	Baixa	4.6	4.9	4.3	6.3	6.1	6.4	6.6	6.7	6.9	5.8667	0.9775
	Média	5.4	5.6	5.2	6.8	5.7	6.7	7.5	8.0	7.3	6.4667	1.045
	Elevada	5.3	5.7	5.1	7.5	7.0	7.2	7.1	7.4	6.1	6.4889	0.88861
Média		5.1222			6.6333			7.0667				
Variância		0.20944			0.3200			0.3175				

As médias observadas para cada combinação de dosagens de cada tipo de fertilizante foram as seguintes:

		Potássio		
		Baixa	Média	Elevada
Fósforo	Baixa	4.600	6.267	6.733
	Média	5.400	6.400	7.600
	Elevada	5.367	7.233	6.867

A Tabela-Resumo associada a esta experiência é a seguinte:

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
fosforo	?	2.24296	1.121481	?	0.00366530
potassio	2	18.75630	?	?	0.00000001
fosforo:potassio	?	1.93926	0.484815	3.36504	0.03187154
Residuals	18	2.59333	?		

Responda às seguintes questões, utilizando a informação disponível no enunciado.

- Complete a Tabela-Resumo, indicando como obtem cada um dos valores omissos.
- Que *tipo* de efeitos do modelo associado a este caso devem ser considerados significativos? Justifique, explicitando as hipóteses dos testes que efectuou, as estatísticas dos testes e os níveis de significância utilizados, bem como a natureza das regiões críticas, os valores obtidos e as conclusões.
- Diga se considera que os rendimentos justificam uma opção entre as duas hipóteses seguintes: (i) Elevada dosagem de potássio com Baixa dosagem de fósforo; e (ii) Elevada dosagem de potássio com Média dosagem de fósforo? Justifique.
- Ajuste agora um modelo a dois factores, mas que não preveja os efeitos de interacção. Construa a tabela-resumo correspondente. Identifique as diferenças entre esta tabela e a que se indicou acima (associada ao modelo que prevê efeitos de interacção). Comente as diferenças e identifique as consequências de não prever a existência de efeitos de interacção quando na realidade esses efeitos parecem existir.

8. Uma experiência pretende estudar o efeito de tempo de exposição e temperatura numa dada solução sobre a dissolução dum produto químico presente num material que é submerso na solução. O estudo apenas se debruça sobre três níveis de temperatura e três tempos de exposição. Os dados recolhidos são as absorções registadas do produto químico pela solução (disponíveis na *data.frame* `absorcao`):

Quantidade (mg) absorvida do químico			
TEMPO DE EXPOSIÇÃO	TEMPERATURA (°C)		
	T_1	T_2	T_3
E_1	35.5	91.2	70.1
	29.7	100.7	64.1
	31.5	82.4	70.1
E_2	52.5	71.0	79.4
	53.3	77.0	77.7
	55.0	75.6	75.1
E_3	85.9	87.0	83.0
	85.2	86.1	87.0
	80.2	88.1	78.5

Médias de células

```

exposicao:temperatura
  temperatura
exposicao T1  T2  T3
E1 32.23 91.43 68.10
E2 53.60 74.53 77.40
E3 83.77 87.07 82.83

```

- (a) Identifique o delineamento experimental utilizado e descreva a equação do modelo ANOVA correspondente.
- (b) Indique as hipóteses nulas e alternativas nos testes à existência de cada tipo de efeito previsto na equação do modelo.
- (c) Efectue os testes de hipóteses referidos na alínea anterior e indique as suas conclusões ao nível de significância de 5%.
- (d) Construa os gráficos de interacção entre tempos de exposição e temperaturas e comente.
- (e) Explícite estimativas da quantidade média de químico absorvido para as combinações de tratamentos que dão origem à maior e à menor absorção.
- (f) É legítimo concluir que, na população, o tratamento resultante de associar o tempo de exposição E_1 com a temperatura T_2 tem uma maior absorção média do que qualquer outro tratamento? Utilize testes de Tukey para responder.
- (g) Reformule o problema considerando que as nove combinações de tratamentos constituem uma Análise de Variância a um factor com nove níveis. Determine se essas nove combinações diferem ao nível de 5%. Comente.
9. Considere de novo os dados do Exercício 8. Utilize o R, e a *data.frame* `absorcao`, para responder às alíneas seguintes.

- (a) Compare as tabelas-resumo da ANOVA resultantes de trocar a ordem dos factores na fórmula indicando o modelo factorial a dois factores, isto é, considere primeiro a fórmula

$$\text{abs} \sim \text{temperatura} * \text{exposicao}$$

e depois a fórmula

$$\text{abs} \sim \text{exposicao} * \text{temperatura}$$

Comente.

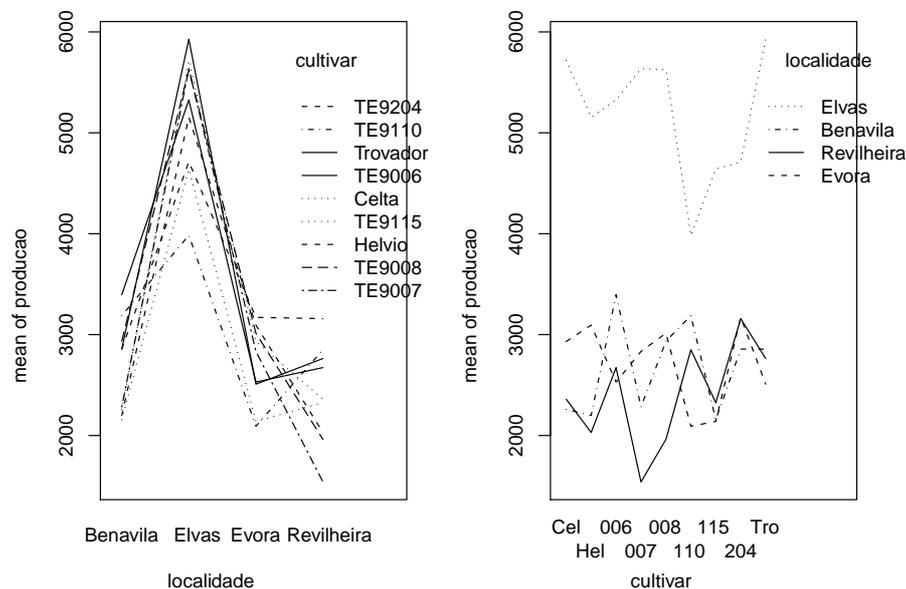
- (b) Considere agora o subconjunto de dados resultante de omitir a primeira observação da primeira célula (de valor 35.5) e as duas últimas observações da última célula (de valores 87.0 e 78.5). Neste caso, estar-se-ia a trabalhar com um *delineamento desequilibrado*. Repita a alínea anterior, mas para este novo conjunto de dados. Comente.

10. Uma engenheira agrônoma pretende seleccionar cultivares de trigo para as quatro explorações agrícolas pelas quais é responsável, que se localizam em Elvas, Évora, Benavila e Revilheira. Em cada uma destas explorações, definem-se 36 parcelas de terra, associando aleatoriamente quatro parcelas a cada uma de nove cultivares: Celta, Helvio, TE9006, TE9007, TE9008, TE9110, TE9115, TE9204 e Trovador. Em cada parcela foi medido o rendimento, em kg/ha. A variância da totalidade dos rendimentos observados é $s^2 = 1\,714\,242$.

- (a) Especifique o delineamento experimental utilizado, e descreva em pormenor o modelo ANOVA adequado a esta experiência.
- (b) Foi ajustado um modelo ANOVA, com o programa R. Resultados parciais desse ajustamento são dados de seguida.

	Df	Sum Sq	Mean Sq	F value
localidade	???	183759916	???	234.9531
cultivar	???	???	964060	???
localidade:cultivar	???	???	???	4.0768
Residuals	???	28156076	260704	

- i. Complete a tabela, indicando como obtém cada um dos valores omissos.
- ii. Teste formalmente (a um nível de significância $\alpha = 0.01$) quais os tipos de efeitos do modelo que devem ser considerados significativos. Descreva um teste em pormenor e discuta os restantes de forma sintética.
- iii. Os gráficos de interacção associados a esta experiência são os seguintes. Comente-os, relacionando as suas conclusões das alíneas anteriores com os gráficos.



- iv. Discuta o efeito de mudar as unidades de medida da variável resposta de kg/ha para toneladas por hectare. Quais os valores da tabela que se alteram, e quais os que ficam iguais? Quais os efeitos da mudança de unidades nas conclusões dos testes F ?
- (c) Os rendimentos médios das quatro parcelas correspondentes a cada combinação de localidade e cultivar foram os seguintes.

localidade	cultivar								
	Celta	Helvio	TE9006	TE9007	TE9008	TE9110	TE9115	TE9204	Trovador
Benavila	2260	2198	3395	2283	2935	3188	2157	2857	2857
Elvas	5719	5154	5326	5636	5624	3984	4640	4713	5927
Evora	2932	3094	2530	2833	3013	2092	2140	3171	2509
Revilheira	2360	2030	2673	1540	1962	2850	2325	3158	2763

Utilizando o teste de Tukey, indique quais as cultivares que, em Elvas, não diferem significativamente do melhor rendimento observado nessa localidade ($\alpha = 0.10$). Comente.

11. Com o objectivo de analisar as alterações no conteúdo em taninos da polpa de sapotis (frutos do sapotizeiro, *Manilkara achras*) provocadas pela temperatura de conservação (alta/baixa) e pelo tempo de armazenamento (0, 3, 6 ou 9 dias) foi efectuado um estudo que forneceu os seguintes dados:

Temperatura	Tempo							
	0 dias		3 dias		6 dias		9 dias	
alta	20.8	19.7	26.5	27.5	26.5	26.4	26.5	26.9
	18.0	19.5	27.0	26.4	27.0	24.0	25.9	26.3
baixa	32.3	34.1	20.8	20.5	16.4	15.7	10.3	9.7
	30.7	31.8	21.0	20.9	15.9	16.0	7.8	9.8

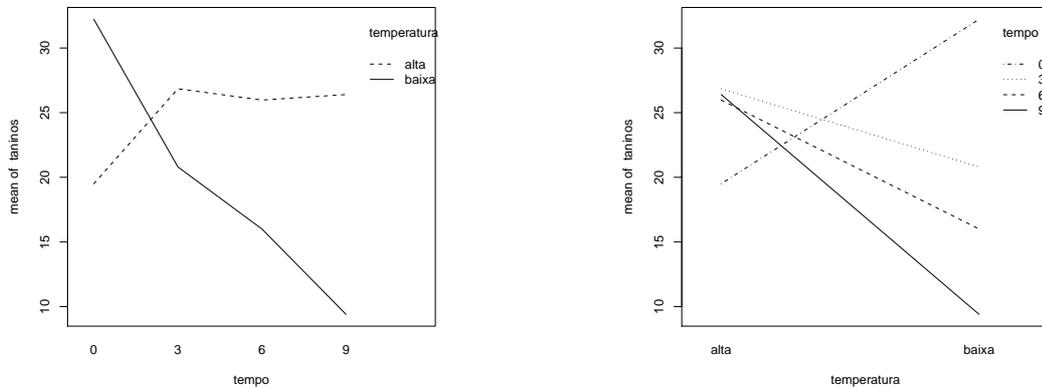
A média e a variância do conjunto das 32 observações são 22.14375 e 47.83222, respectivamente. As médias associadas a cada tempo de armazenamento, cada temperatura e cada combinação de tempo e temperatura, são:

Tables of means

tempo				tempo:temperatura				
0	3	6	9	temperatura				
				tempo	alta	baixa		
25.862	23.825	20.987	17.900	0	19.50	32.23		
				3	26.85	20.80		
				6	25.97	16.00		
				9	26.40	9.40		



- Identifique o delineamento experimental utilizado no estudo e descreva de forma pormenorizada o melhor modelo ANOVA que lhe está associado.
- Sabendo que a Soma dos Quadrados dos Resíduos é 20.72 e que o Quadrado Médio associado aos diferentes tempos de armazenamento é 96.01, construa o Quadro-Resumo da Análise de Variância associado a esta experiência.
- Podem considerar-se que os diferentes tempos de armazenamento influenciam o teor de taninos na polpa destes frutos? Responda a esta questão utilizando testes de hipóteses.
- Interprete os seguintes gráficos de interação e compare as conclusões obtidas com os resultados do Quadro-Resumo desta ANOVA.



- (e) Um investigador afirma que se a temperatura de conservação é alta, o tempo de armazenamento não afecta o teor médio de taninos na polpa dos sapotis. Realize uma análise estatística que lhe permita afirmar se concorda, ou não, com esta afirmação.
12. Um estudo efectuado pela Secção de Produção Animal do ISA visa estudar se existem efeitos na digestibilidade em leitões, associados à natureza da fibra (factor *Fibra*, com 2 níveis) e à adição, ou não de enzimas digestivas (factor *Enzimas*) no alimento dos leitões. Foram consideradas como variáveis resposta os Coeficientes de Utilização Digestiva (CUDs) para matéria seca (*MS*), proteína bruta (*PB*), fibra neutro-detergente (*NDF*), fibra ácido-detergente (*ADF*), hemi-celuloses (*HC*) e celulose (*CEL*). A experiência utilizou doze repetições por célula. Os dados obtidos estão disponíveis na *data.frame* `leitoes` (caso tenham sido efectuadas as instruções no Aviso que antecede estes exercícios).
- Identifique o delineamento experimental utilizado no estudo da variável resposta *CEL* e descreva o modelo ANOVA que lhe está associado.
 - É admissível considerar que os valores médios da variável *CEL* são afectados por interacção entre os factores *Fibra* e *Enzima*? E existem efeitos principais do factor *Fibra*? E do factor *Enzima*? (Utilize um nível de significância $\alpha = 0.05$).
 - Independentemente da sua resposta na alínea anterior, utilize a distribuição de Tukey para dizer se existe alguma combinação de *Fibra*/*Enzima* que esteja associada a um *CEL* médio significativamente mais elevado, a um nível de significância $\alpha = 0.05$. A sua resposta é coerente com os testes *F* correspondentes?
 - Utilize o teste de Bartlett para validar a hipótese de variâncias de célula homogêneas.
13. Considere de novo os dados descritos no Exercício 12.
- Considere o modelo a 2 factores, com interacção, para a variável resposta *CEL*.
 - Indique a natureza da matriz \mathbf{X} associada a este modelo.
 - Construa, com o auxílio do R, a matriz \mathbf{H} de projecção ortogonal sobre o subespaço gerado pelas colunas da matriz \mathbf{X} referida na subalínea anterior.
 - Obtenha o vector $\hat{\mathbf{Y}}$ dos valores ajustados, multiplicando a matriz \mathbf{H} agora obtida pelo vector das observações de *CEL*.
 - Utilize o vector ajustado $\hat{\mathbf{Y}}$ obtido na alínea anterior para calcular a Soma de Quadrados Residual do modelo em causa.

- (b) Repita a alínea 13a), mas considerando agora o modelo a 2 factores *sem* interacção.
- (c) Repita a alínea 13a), mas considerando agora o modelo a 1 factor, o factor Enzima. Além de determinar a Soma de Quadrados Residual $SQRE_B$, determine também a Soma de Quadrados associada aos efeitos do Factor nesse modelo, SQF_B .
- (d) Repita a alínea 13a), mas considerando agora o modelo a 1 factor, o factor Fibra. Além de determinar a Soma de Quadrados Residual $SQRE_A$, determine também a Soma de Quadrados associada aos efeitos do Factor nesse modelo, SQF_A .
- (e) Relacione as Somas de Quadrados obtidas nas quatro alíneas anteriores com a tabela-resumo da ANOVA do modelo a 2 Factores com interacção. Comente.
14. Num estudo sobre um clone de uma variedade de videira (casta Tinta Francisca), estudou-se o rendimento (em kg/planta) em dois diferentes locais (Tabuaço e Régua), e em diferentes anos. Os dados disponíveis para o Tabuaço referem-se aos anos 1999 e 2003, e para a Régua aos anos 1999, 2000 e 2002. Em cada local e ano foram observadas 8 unidades experimentais (parcelas), tendo sido obtidos os seguintes resultados:

Tabuaço		Régua		
1999	2003	1999	2000	2002
1.292	1.188	0.181	0.992	1.594
1.025	0.458	0.917	0.881	2.315
0.958	0.258	0.435	0.835	2.125
0.625	0.892	0.195	0.758	0.835
0.625	0.225	0.142	0.488	1.675
0.125	0.244	0.235	0.925	0.995
0.425	0.444	0.029	0.488	0.679
0.092	0.092	0.019	0.125	0.594

Pretende-se saber se existem efeitos de local e/ou efeitos de ano no rendimento do referido clone.

- (a) Indique o tipo de delineamento experimental utilizado, e um modelo ANOVA adequado para esta experiência.
- (b) Efectue os testes necessários para indicar se considera que existem efeitos de local e/ou efeitos de ano.
- (c) Utilize os testes de Tukey para estudar pares de combinações de local e ano que tenham rendimentos médios significativamente diferentes.
- (d) Efectue um teste de Bartlett para avaliar se é adequado admitir homogeneidade de variâncias nas cinco situações experimentais.
15. Após um ensaio para selecção de variedades de alho, um agricultor de Braga foi aconselhado a utilizar um genótipo comercializado com a designação V7088. Sabendo que o estudo que levava ao aconselhamento fora feito em Elvas nos anos de 2000 e 2004, o referido agricultor decidiu repetir o ensaio nos seus terrenos, em condições semelhantes às do ensaio inicial. Durante três anos (de 2007 a 2009) seleccionou aleatoriamente quatro parcelas, tendo utilizado o referido genótipo. Tal como no estudo inicial, o agricultor optou por medir a variável, comercialmente importante, do peso dos bolbos de alho. Os pesos médios (em g) obtidos nas 4 parcelas de cada situação referida foram os seguintes.

1	Elvas	Ano2000	13.9200
2	Elvas	Ano2004	29.8200

```

3 Braga Ano2007 15.4500
4 Braga Ano2008 19.9325
5 Braga Ano2009 12.9425

```

Tendo pedido apoio estatístico, o agricultor foi confrontado com o seguinte modelo ANOVA ajustado:

```

> summary(aov(peso.bolbo ~ Local/Ano, data=exame300113))
              Df Sum Sq Mean Sq F value    Pr(>F)
Local          1 159.34  159.345    13.072 0.002543
Local:Ano      3 605.94  201.980    16.570 5.005e-05
Residuals     15 182.84   12.189

```

- Indique que tipo de delineamento experimental descreve a experiência realizada, e descreva em pormenor o modelo ANOVA que foi ajustado, explicando o significado de todas as parcelas da equação do modelo, bem como os pressupostos adicionais admitidos.
 - Efectue os testes F que considerar apropriados. Comente os resultados e, com base nesses testes, diga se é possível afirmar que os pesos médios dos bolbos do referido genótipo são diferentes em Braga e Elvas.
 - Há diferenças significativas nos pesos médios anuais em Braga? Justifique mediante um teste de Tukey ($\alpha = 0.05$).
 - Confuso com os resultados, o agricultor, que havia frequentado um curso de introdução à Estatística onde apenas fora leccionada a Análise de Variância a um factor, decidiu ajustar um modelo desse tipo, utilizando todos os dados disponíveis, mas ignorando os anos a que os dados se referem.
 - Construa, justificando, a tabela resumo da ANOVA que o agricultor obteve.
 - Descreva sinteticamente os resultados obtidos através da ANOVA do agricultor ($\alpha = 0.05$). Compare esses resultados com os obtidos através do estudo do modelo anterior e comente.
16. Mostre que é nula a soma dos resíduos das observações em:
- cada nível do Factor, numa ANOVA a 1 Factor (de efeitos fixos);
 - cada célula, numa ANOVA a 2 Factores, com interacção.
17. Considere um Modelo ANOVA a 1 Factor. Tendo em conta as expressões dos estimadores dos parâmetros,
- Mostre que os estimadores têm a seguinte distribuição:
 - $\hat{\mu}_1 \cap \mathcal{N}(\mu_1, \sigma^2/n_1)$.
 - $\hat{\alpha}_i \cap \mathcal{N}\left(\alpha_i, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_i}\right)\right) \quad (i > 1)$.
 - Sabendo que a Soma de Quadrados Residual, $SQRE$, é independente dos estimadores de qualquer dos parâmetros, mostre que:
 - um intervalo de confiança a $(1 - \alpha) \times 100\%$ de confiança para μ_1 é dado por

$$\left[\bar{y}_1 - t_{\alpha/2(n-k)} \cdot \sqrt{QMRE/n_1} \quad , \quad \bar{y}_1 + t_{\alpha/2(n-k)} \cdot \sqrt{QMRE/n_1} \right]$$

- um intervalo de confiança a $(1 - \alpha) \times 100\%$ de confiança para α_i ($i > 1$) é dado por

$$\left[(\bar{y}_i - \bar{y}_{1.}) - t_{\alpha/2(n-k)} \cdot \sqrt{QMRE \cdot \left(\frac{1}{n_1} + \frac{1}{n_i}\right)}, \right. \\ \left. (\bar{y}_i - \bar{y}_{1.}) + t_{\alpha/2(n-k)} \cdot \sqrt{QMRE \cdot \left(\frac{1}{n_1} + \frac{1}{n_i}\right)} \right]$$

18. Num estudo inicial do ISA e da DRA Centro para melhoramento das maçãs Bravo Esmolfe, pretendeu-se avaliar se diferentes genótipos estavam associados a diferentes rendimentos médios para, em caso afirmativo, submeter os genótipos que neste estudo inicial revelaram melhor desempenho a um estudo com mais repetições, visando seleccionar genótipos merecedores de utilização intensiva.

Neste estudo inicial foram seleccionados $k = 35$ genótipos na zona da Beira Interior. Foi registado o rendimento médio (em kg/árvore) em 4 parcelas.

Ajustou-se um modelo ANOVA (de efeitos fixos), tendo-se obtido a seguinte tabela-resumo:

```
> summary(aov(rendimento ~ genotipo , data=bravoeshmolfe))
              Df Sum Sq Mean Sq F value    Pr(>F)
genotipo      34  169.3    4.981    4.723 4.61e-10 ***
Residuals    105  110.7    1.055
```

Mas o facto de os genótipos escolhidos para este estudo inicial serem apenas uma amostra dum número enorme de possíveis genótipos, sugere a utilização dum modelo tipo ANOVA, mas com *efeitos aleatórios* do factor.

- Descreva pormenorizadamente um modelo tipo ANOVA com *efeitos aleatórios* do factor. Em particular, indique em que se distingue do modelo de *efeitos fixos* estudados anteriormente.
- Teste, para este estudo, a existência de variabilidade genética. Explícite as hipóteses, estatística do teste, região crítica e interprete as suas conclusões.
- Qual o valor do índice de heritabilidade associado a este exemplo? Comente.