INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2019-20

27 de Janeiro de 2020 Segunda Chamada de EXAME Duração: 3h30

I [2,5 valores]

Análises qualitativas sugerem que a distribuição da vegetação herbácea em torno de árvores isoladas, em climas áridos, depende da orientação geográfica. Pretende-se validar estatisticamente esta suspeita, para um dado ecossistema no Sultanato de Omã. Contabilizou-se (até totalizar 2501) o número de indivíduos de três espécies herbáceas que germinaram em torno de árvores isoladas de *Acacia tortilis*, em cada um dos quadrantes com orientação Norte, Leste, Sul e Oeste. Obtiveram-se os seguintes resultados:

Espécie	Norte	Leste	Sul	Oeste	Total
Sisymbrium irio	309	91	18	60	478
Spergularia fallax	690	249	223	395	1557
Zygophyllum simplex	150	26	243	47	466
Total	1149	366	484	502	2501

1. Identifique o tipo de teste que lhe permite responder à questão colocada. Explicite as hipóteses, a estatística do teste, a respectiva distribuição assintótica e a região crítica.

2. Diga, justificando, se considera a dimensão da amostra suficiente para efectuar o teste indicado na alínea anterior.

3. Calcule a parcela da estatística do teste que corresponde à espécie *Zygophyllum simplex*, germinando com orientação a Sul. Sabendo que as restantes parcelas da estatística do teste totalizam 229.6256, complete o teste e comente as suas conclusões.

II [8,5 valores]

Um estudo vinícola com a casta Antão Vaz, realizado em 2019 em Pegões, avaliou, para 109 diferentes videiras, dados relativos a várias características: rendimento (variável `rend`, em kg/planta); peso médio dos bagos duma planta (variável `pesobago`, em g); sólidos solúveis (variável `brix`, em graus brix); ácido tartárico (variável `acidez`, em g/l); e pH (variável `pH`). Pretende-se modelar a variável `brix`. Eis alguns indicadores dos valores obtidos, bem como a matriz de correlações amostrais entre as variáveis:

	`rend`	`pesobago`	`brix`	`acidez`	`pH`
Mínimo	4.037	1.321	15.23	4.5	3.51
1°quartil	7.847	1.898	17.5	5.05	3.64
Mediana	9.027	2.059	18.03	5.30	3.68
3°quartil	10.046	2.271	18.67	5.55	3.73
Máximo	13.288	3.366	22.07	6.20	3.93
Média	8.946073	2.084789	18.08339	5.299817	3.684495
Desv.Padrão	1.826938	0.303144	1.072828	0.367239	0.075136

```
           rend pesobago    brix  acidez      pH
rend      1.0000  -0.2462 -0.4822 -0.0255 -0.4822
pesobago -0.2462   1.0000  0.4649  0.2668  0.4717
brix     -0.4822   0.4649  1.0000    ???   0.8305
acidez   -0.0255   0.2668    ???   1.0000 -0.2655
pH       -0.4822   0.4717  0.8305 -0.2655  1.0000
```

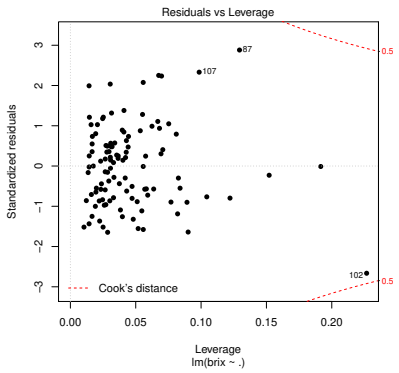
1. Uma regressão linear múltipla de `brix` sobre as restantes variáveis deu os seguintes resultados.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.56710    4.23937  -2.728  0.00747
rend         -0.08447    0.03441  -2.455  0.01575
pesobago      0.69669    0.22833   3.051  0.00289
acidez       -0.61539    0.17525  -3.512  0.00066
pH            8.74345    1.03514   8.447 1.93e-13
---
Residual standard error: 0.5615 on 104 degrees of freedom
Multiple R-squared: 0.7363,  Adjusted R-squared: 0.7261
F-statistic: 72.58 on 4 and 104 DF,  p-value: < 2.2e-16
```

   (a) Interprete o valor do coeficiente de determinação do modelo.

   (b) Discuta, com base num teste de hipóteses adequado, ao nível de significância 0.01, a seguinte afirmação: "*na população, os teores médios de brix diminuem com o aumento do teor de ácido tartárico, para valores fixos dos restantes preditores*".

   (c) Descreva e discuta o seguinte gráfico. À direita do gráfico surgem os valores observados das variáveis na observação 102, que surge no canto inferior direito do gráfico.




```
> egav2019[102,]
     rend pesobago  brix acidez   pH
102 4.037    1.666 19.53    4.5 3.93
```

2. Considere o modelo de regressão linear simples de `brix` sobre `pH`.

   (a) Teste se esta regressão linear simples tem um ajustamento significativamente diferente do modelo com todos os preditores. Comente.

   (b) Calcule o efeito alavanca da observação 102 nesta regressão. Sabendo que o resíduo (internamente) estandardizado dessa observação é $R_{102} = -2.5833$, parece-lhe que a distância de Cook desta observação será muito diferente neste modelo e no modelo com todos os preditores acima ajustado?

3. A recta de regressão linear ajustada de `brix` sobre o preditor `acidez` tem equação $y = 22.9924 - 0.9263\,x$ e coeficiente de determinação $R^2 = 0.1005$.

   (a) Calcule, justificando, o coeficiente de *correlação* linear entre `brix` e `acidez`.

(b) Efectue o teste de ajustamento global deste modelo. Comente os seus resultados, tendo em atenção o valor do coeficiente de determinação.

III [5 valores]

Num estudo sobre o desempenho da casta Alvarinho em Monção foi medido o rendimento (variável `rend`, em kg/planta) em 8 diferentes ambientes, seleccionados pela sua diversidade. Em cada ambiente foram demarcados 9 terrenos. Sabe-se que as características dos 72 terrenos são diferentes, nada permitindo associar terrenos de ambientes diferentes. Em cada terreno demarcaram-se 6 parcelas, de forma a haver 6 observações de rendimento para cada terreno. A média e a variância amostrais da totalidade dos rendimentos observados foram, respectivamente, 2.949606 kg/planta e 6.05404 $(\text{kg/planta})^2$. Em baixo indicam-se os rendimentos médios obtidos nas situações experimentais do ambiente 2.

terreno	t1	t2	t3	t4	t5	t6	t7	t8	t9
rendimento	4.873	7.314	7.202	5.840	6.885	8.617	7.247	5.898	6.007

1. Diga, justificando, qual o delineamento experimental utilizado e descreva pormenorizadamente o modelo ANOVA correspondente.
2. Construa a tabela de síntese da ANOVA que indicou, sabendo que a estimativa da variância dos erros aleatórios é 2.2347 e que a Soma de Quadrados associada a efeitos de ambiente é 1666.2.
3. Que tipos de efeitos podem ser considerados significativos? Descreva em pormenor um dos testes e de forma mais sucinta o(s) restante(s). Discuta as suas conclusões à luz da informação disponível.
4. Utilize o teste de Tukey para determinar se é possível considerar significativamente diferentes, ao nível $\alpha = 0.05$, o menor e o maior rendimento médio amostral obtidos no ambiente 2. Comente, à luz das suas conclusões na alínea anterior. **Nota:** o valor do quantil da distribuição de Tukey adequado é 5.939.
5. Diga de que forma alteraria a sua resposta *no ponto 1* caso tivessem sido previamente definidos nove diferentes tipos de terrenos e, em todos os ambientes, os nove terrenos tivessem sido seleccionados de forma a corresponderem a cada um desses tipos de terrenos.

IV [4 valores]

1. Considere a relação logística, de equação $y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}}$.
   (a) Mostre que a relação pode ser linearizada tomando o *logit* da variável resposta y, ou seja, $\ln\left(\frac{y}{1-y}\right)$.
   (b) Considerando y como função de x, mostre que a taxa de variação relativa de y é igual a $d[1 - y(x)]$.
2. Considere uma regressão linear múltipla com p variáveis preditoras e ajustada com base em n observações. Seja $\mathbf{X}$ a matriz do modelo e $\mathbf{H}$ a matriz de projecção ortogonal sobre o espaço das colunas de $\mathbf{X}$.

(a) Mostre que a Soma de Quadrados Residual, *SQRE*, é dada pela norma ao quadrado do vector $(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$, onde $\vec{\mathbf{Y}}$ é o vector das observações da variável resposta e $\mathbf{I}_n$ a matriz identidade $n \times n$.

(b) Justifique que $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, sendo $\vec{\mathbf{1}}_n$ o vector de n uns. Justifique que esta equação significa que os elementos de cada linha da matriz $\mathbf{H}$ somam 1.

(c) Mostre que a média dos valores observados da variável resposta é igual à média dos correspondentes valores ajustados pelo modelo.

(d) Justifique a seguinte afirmação: "*cada valor ajustado $\hat{Y}_j$ é uma média ponderada de todas as observações Y_i, sendo a ponderação da própria observação Y_j equivalente ao respectivo efeito alavanca*". Qual a consequência deste facto para observações com elevado efeito alavanca?