INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2015/16

19 de Janeiro de 2016 Segunda Chamada de Exame Duração: 3h30

I [3 valores]

Num estudo sobre enraizamento de estacas de videira, pretende-se utilizar um teste χ^2, baseado na estatística de Pearson, para decidir se é admissível considerar que o número de não enraizamentos em contentores de 10 plantas segue uma distribuição Binomial. Foram utilizados 164 contentores. Em cada contentor foi contado o número de plantas não enraizadas. Eis os resultados obtidos:

No. não enraizamentos	0	1	2	3	4	5	6	7	8	9	10
No. contentores	14	54	44	22	16	8	2	3	1	0	0

1. Determine, justificando, os valores mais adequados dos dois parâmetros na referida distribuição Binomial.

Caso não tenha obtido um valor estimado para o parâmetro p na alínea anterior, considere o valor $\hat{p} = 0.2$ nas alíneas seguintes.

2. Discuta a seguinte afirmação: *Caso se utilizassem as classes constantes da tabela no cálculo do valor da estatística de teste, não se poderia garantir a validade da distribuição assintótica correspondente. O agrupamento das classes com 5 ou mais não enraizamentos por contentor resolve este problema.*

3. Descreva o teste de hipóteses referido, sabendo que o valor calculado da estatística do teste (com o agrupamento de classes referido na alínea anterior) é $X^2_{calc} = 17.5129$. Comente as suas conclusões, para um nível de significância $\alpha = 0.05$.

4. Calcule o valor da parcela da estatística do teste correspondente a contentores com 3 plantas não enraizadas.

II [8 valores]

Num estudo sobre a casta Negra Mole foram analisadas várias características de qualidade do mosto: `brix`, em graus brix; `acidez`, em g/l de ácido tartárico; `pH`; teor de `antocianas` (g/l); peso do bago (variável `pesobago`, em g); e `rendimento` (em kg/planta). Eis alguns indicadores relativos a cada variável, e a respectiva matriz de correlações:

```
          rend     brix   acidez antocianas    pH     pesobago
Min.     1.660    19.53    3.090    69.50     3.760    1.920
Mean     2.471    21.53    3.381    98.52     3.865    2.297
Max.     3.450    23.63    3.670   149.92     4.010    2.700
St.Dev. 0.40270  0.94994  0.13761  20.02151  0.05223  0.19711
```

```
                rend   brix acidez antocianas    pH  pesobago
rend        1.000 -0.514 -0.168     -0.362 -0.113    0.146
brix       -0.514  1.000  0.187      0.561  0.099    0.207
acidez     -0.168  0.187  1.000     -0.303 -0.508    0.336
antocianas -0.362  0.561 -0.303      1.000  0.213   -0.277
pH         -0.113  0.099 -0.508      0.213  1.000   -0.241
pesobago    0.146  0.207  0.336     -0.277 -0.241    1.000
```

A medição do teor de antocianas no mosto é uma operação difícil, e seria desejável poder modelá-la em função das restantes variáveis. Foi ajustada uma regressão linear múltipla, tendo sido obtidos os seguintes resultados:

```
Coefficients:
            Estimate  Std.Error t value Pr(>|t|)
(Intercept)  274.796    243.281   1.130  0.26658
rend          -2.210      6.896  -0.320  0.75059
brix          14.605      2.917   5.008 1.68e-05
acidez       -59.258     20.360  -2.910  0.00633
pH           -54.928     51.261  -1.072  0.29147
pesobago     -31.644     12.916  -2.450  0.01959
---
Residual standard error: 13.83 on 34 degrees of freedom
Multiple R-squared:  0.5841, Adjusted R-squared:  0.523
F-statistic: 9.551 on 5 and 34 DF,  p-value: 9.242e-06
```

1. Discuta a qualidade de ajustamento do modelo. No teste de hipóteses adequado, utilize um nível de significância $\alpha = 0.05$.

2. Discuta o valor do coeficiente de determinação modificado. Justifique, no contexto deste problema, a diferença entre esse valor e o valor do coeficiente de determinação usual.

3. Utilize um intervalo a 95% de confiança para discutir a seguinte afirmação: *a magnitude elevada do valor* -54.928 *significa que a variável* `pH` *dá uma contribuição importante para a previsão do teor de antocianas (quando os restantes preditores não variam).*

4. Um algoritmo de exclusão sequencial produziu, no seu terceiro passo, o seguinte resultado:

   ```
   antocianas ~ brix + acidez + pesobago
              Df     AIC
   <none>          213.02
   - pesobago  1  218.05
   - acidez    1  218.84
   - brix      1  239.33
   ```

   (a) Diga, justificando, qual foi a variável excluída no primeiro passo, e qual a excluída no segundo passo, do algoritmo.

   (b) Calcule o valor do Quadrado Médio Residual do submodelo final do algoritmo. Compare com o valor estimado da variância dos erros aleatórios no modelo original e comente.

5. Considere agora uma regressão linear *simples* de `antocianas` sobre um único preditor.

   (a) Indique, justificando, a melhor variável preditora e o coeficiente de determinação do modelo escolhido. Comente.

(b) Calcule a recta de regressão ajustada no modelo que escolheu na alínea anterior.

(c) Calcule o erro padrão estimado do estimador do declive da recta de regressão escolhida.

III [5 valores]

No melhoramento de variedades tradicionais de tomate, uma característica importante é a resistência da película. Foram avaliadas 6 variedades tradicionais de tomate quanto a esta característica. De cada variedade foram colhidos aleatoriamente tomates em cada uma de 3 parcelas de tomateiros, sendo cada observação constituída pela resistência média dos frutos de cada parcela (medida num texturómetro, em grama força, *gf*). Eis os valores médios e variâncias obtidos, para cada variedade e para o conjunto das $n = 18$ observações:

Variedade	18	28	29	40C	Ace	Roma	Todas
Média	560.6	241.5	291.0	705.8	377.3	332.1	418.0361
Variância	14713.08	367.9434	5881.921	33132.64	5.414433	47.11163	34517.82

1. Descreva o delineamento experimental utilizado e o modelo ANOVA adequado para o estudar.
2. Construa a tabela resumo da ANOVA referida.
3. Diga, justificando, se é possível afirmar que existem diferenças significativas na resistência de diferentes variedades.
4. Quais as variedades cuja resistência difere significativamente da resistência da variedade 40C? Justifique com um teste adequado, ao nível $\alpha = 0.05$. (*Aviso:* Caso não tenha respondido à alínea 2, utilize o valore $QMRE = 9000$)
5. Um analista de dados efectuou um teste de Bartlett, que produziu os seguintes resultados:

```
> bartlett.test(resistencia ~ variedade , data=tomateresistencia)
data:  resistencia by variedade
Bartlett's K-squared = 24.2998, df = 5, p-value = 0.0001901
```

   Descreva sinteticamente o teste e comente as suas conclusões para o problema sob estudo. Indique uma razão pela qual a aplicação do teste de Bartlett neste problema é discutível.

IV [4 valores]

1. Num modelo de regressão linear simples, as n osbervações Y_i da variável resposta são independentes e têm distribuição $Y_i \cap \mathcal{N}\left(\beta_0 + \beta_1 x_i \, , \, \sigma^2\right)$. Considere o estimador do declive da recta de regressão, $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$, com $c_i = \frac{(x_i - \overline{x})}{(n-1)s_x^2}$. Mostre, justificando, que:

   (a) $\hat{\beta}_1$ tem distribuição Normal.

   (b) $E[\hat{\beta}_1] = \beta_1$ (Pode admitir que $\sum_{i=1}^{n} c_i x_i = 1$).

   (c) $V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)\, s_x^2}$.

2. Considere o modelo de regressão linear múltipla, o vector das n observações da variável resposta, $\vec{\mathbf{Y}}$ e o vector dos n valores ajustados de Y, ou seja o vector $\hat{\vec{\mathbf{Y}}} = \mathbf{H}\vec{\mathbf{Y}}$.

   (a) Mostre que a média dos n valores observados de Y é igual à média dos correspondentes valores ajustados.

   (b) Sabendo que o vector dos n valores observados tem distribuição de probabilidades $\vec{\mathbf{Y}} \cap \mathcal{N}(\mathbf{X}\vec{\boldsymbol{\beta}}\,,\, \sigma^2\mathbf{I}_n)$, determine a distribuição de probabilidades do vector $\hat{\vec{\mathbf{Y}}}$ dos n valores ajustados.