INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2019-20

January 27, 2020 **Final EXAM (Second date)** **Duration: 3h30m**

I [2.5 points]

Qualitative analyses suggest that, in arid climates, the distribution of grass vegetation surrounding isolated trees depends on the geographical orientation. We seek to determine whether this assumption is statistically significant in a given eco-system in the Sultanate of Oman. The number of individuals of three herbaceous species that germinated around isolated *Acacia tortilis* trees was counted, until a total of 2501 was reached. Each individual was classified by species and by the quadrant – North, East, South or West – in which it was located. The following results were obtained:

Species	North	East	South	West	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. What kind of test should one use to answer the question that is raised? Write down the hypotheses that are to be tested, as well as the test statistic, its asymptotic distribution and the critical (rejection) region.

2. Is the sample sufficiently large to carry out the test that you indicated? Justify your answer.

3. Compute the term in the test statistic that corresponds to the species *Zygophyllum simplex*, germinating towards the South. Knowing that the remaining terms in the test statistic add up to 229.6256, carry out the test and discuss your conclusions.

II [8.5 valores]

A study on grapevines of the Antão Vaz variety was carried out in 2019 in Pegões. For 109 different vines, data were collected relative to several characteristics: yield (variable `rend`, in kg/plant); mean weight of the grapes of a plant (variable `pesobago`, in g); soluble solids (variable `brix`, in degrees brix); tartaric acid (variable `acidez`, in g/l); and pH (variable `pH`). The goal is to model the variable `brix`. Here are some indicators for the resulting values, as well as the matrix of sample correlations between the variables:

	`rend`	`pesobago`	`brix`	`acidez`	`pH`
Minimum	4.037	1.321	15.23	4.5	3.51
1st quartile	7.847	1.898	17.5	5.05	3.64
Median	9.027	2.059	18.03	5.30	3.68
3rd quartile	10.046	2.271	18.67	5.55	3.73
Maximum	13.288	3.366	22.07	6.20	3.93
Mean	8.946073	2.084789	18.08339	5.299817	3.684495
Std. Deviation	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
```

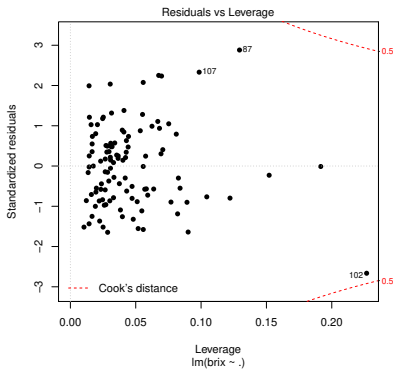
1. A multiple linear regression of `brix` on the remaining variables gave the following results.

```
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) Interpret the value of the model's coefficient of determination.

   (b) Based on an appropriate hypothesis test at a 0.01 significance level, comment the following statement: "*in the population, the mean values of brix decrease as the amount of tartaric acid increases, all other predictors being equal*".

   (c) Describe and discuss the following graph. To the right of the graph we see the observed values of observation 102, which appears in the bottom right of the graph.




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

2. Considere the simple linear regression model of `brix` over `pH`.

   (a) Test whether this simple linear regression has a goodness-of-fit that is significantly different from that of the model with all the predictors. Comment.

   (b) Compute the leverage of observation 102 in this regression model. Given that this observation's (internally) standardized residual is $R_{102} = -2.5833$, do you think that its Cook's distance will be very different in this model and in the model with all the predictors, that was fitted before?

3. The equation of the fitted linear regression line of `brix` over the predictor `acidez` is $y = 22.9924 - 0.9263\,x$. Its coefficient of determination is $R^2 = 0.1005$.

   (a) Calculate, showing all steps, the coefficient of linear *correlation* between `brix` and `acidez`.

(b) Carry out a goodness-of-fit test for this model. Comment your results, taking into account the value of the coefficient of determination.

III [5 points]

In a study of the Alvarinho grapevine variety in Monção, yields (variable `rend`, in kg/plant) were measured in 8 different environments, that were chosen due to their diversity. In each environment, 9 fields were chosen. It is known that the 72 fields have different characteristics, and it is not possible to associate fields in different environments. Each field was divided into 6 plots, so that 6 yields could be measured in each field. The sample mean and variance of all observed yields were, respectively, 2.949606 kg/plant and 6.05404 (kg/plant)^2. Below are shown the mean yields recorded in the experimental situations of environment 2.

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

1. What experimental design was used? Justify your response. Describe in detail the corresponding ANOVA model.
2. Build the ANOVA summary table for the model that you specified, given that the estimated variance of the random errors is 2.2347 and the Sum of Squares associated with the environment effects is 1666.2.
3. Which kind of effects are significant? Describe in detail one of your tests and more briefly the remaining test(s). Discuss your conclusions, based on the available information.
4. Use Tukey's tests to determine whether the smallest and the largest sample mean yields observed in environment 2 can be considered significantly different for a $\alpha = 0.05$ significance level. Comment your results, also taking into account your conclusions from the previous question. **Note:** the quantile for the appropriate Tukey distribution is 5.939.
5. In what way would you change your answer to *question 1* if your fields had been previously classified as belonging to one of nine different groups and, for every environment, the nine fields had been chosen so as to belong to each of those groups.

IV [4 points]

1. Consider the logistic relation, with equation $y = \frac{1}{1+\mathrm{e}^{-(c+d\,x)}}$.
   (a) Show that this relation can be linearized by taking the *logit* of the response variable y, that is, $\ln\left(\frac{y}{1-y}\right)$.
   (b) Considering y as a function of x, show that the relative rate of change of y is $d[1-y(x)]$.
2. Consider a multiple linear regression with p predictor variables, fitted using n observations. Let $\mathbf{X}$ be the model matrix and $\mathbf{H}$ the matrix of orthogonal projections on the column-space of $\mathbf{X}$.
   (a) Show that the Residual Sum of Squares, $SQRE$, is given by the squared norm of the vector $(\mathbf{I}_n - \mathbf{H})\vec{\mathbf{Y}}$, where $\vec{\mathbf{Y}}$ is the vector of observations of the response variable and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

(b) Show that $\mathbf{H}\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n$, where $\vec{\mathbf{1}}_n$ is the vector with n ones. Explain why this equation means that the elements in each row of matrix $\mathbf{H}$ add up to 1.

(c) Show that the mean of the observed values of the response variable is equal to the mean of the corresponding fitted values.

(d) Justify the following statement: "*each fitted value $\hat{Y}_j$ is a weighted average of all the observations Y_i, with the weight of the same observation Y_j given by its leverage*". What is the consequence of this fact for observations with high leverages?