Bachelor Degree in Computer Engineering

Statistics **group E (English)**

SECOND PARTIAL EXAM

June 1st 2015

Surname, name	
Signature	

Instructions

1. Write your name and sign in this page.
2. Answer each question in the corresponding page.
3. All answers must be justified.
4. Personal notes in the formula tables will not be allowed.
5. Mobile phones are not permitted over the table. It is only permitted to have the DNI (identification document), calculator, pen, and the formula tables. Mobile phones cannot be used as calculators.
6. Do not unstaple any page of the exam (do not remove the staple).
7. All questions score the same (over 10).
8. At the end, it is compulsory to sign in the list on the professor's table in order to justify that the exam has been handed in.
9. Time available: **2 hours**.

1. A study about the performance of a computer system has obtained a sample of 14 values of time of access to a certain database (in seconds). The sample mean has been 5.4 s and sample standard deviation is 1.9 s. It is assumed that data are normally distributed.

a) Calculate a confidence interval for the mean time of access with a confidence level of 95%. What is the practical interpretation of the 0.95 probability associated with this confidence interval? *(3 points)*

b) Can we admit that the mean time of access is 4 s in the population considering a type I risk of 0.05? Justify your answer, indicating what is the hypothesis test considered. *(1.5 points)*

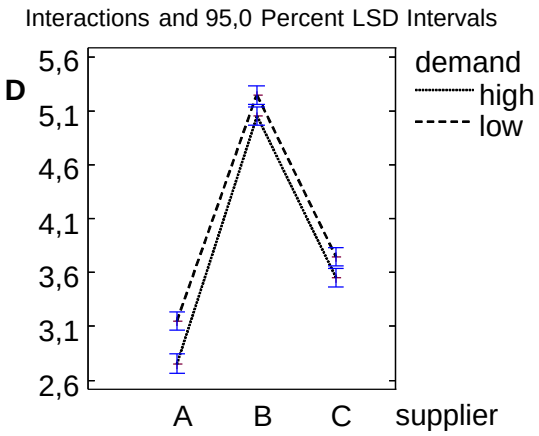
c) Based on the calculations performed in section a), can we admit that the mean time of access is 4 s in the population considering a type I risk of 0.1? Justify your reply. *(1 point)*

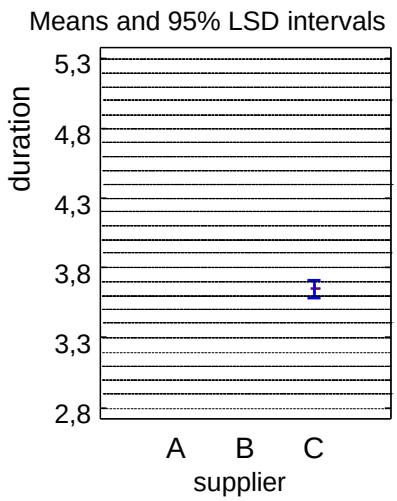
d) Can we admit that the sample was extracted from a population with a standard deviation (σ) equal to 2 s? (α=5%). *(3 points)*

e) What kind of wrong decisions can be taken when studying a hypothesis test? What are the names of this kind of errors? *(1.5 points)*

2. In order to study if the duration (D) of batteries (in years) is different according to the supplier (A, B or C), a sample of 4 batteries is taken from each supplier, and they are tested in similar conditions, two of them under a high energy demand and the other two under a low energy demand. The resulting data are shown below.

Sup.	energy	D
A	high	2.7
A	high	2.8
A	low	3.2
A	low	3.1
B	high	5.0
B	high	5.1
B	low	5.2
B	low	5.3
C	high	3.5
C	high	3.6
C	low	3.7
C	low	3.8





Analysis of Variance for DURACION - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:SUPPLIER				
B:DEMAND	0,213333			
INTERACTIONS				
AB				2,67
RESIDUAL			0,005	
TOTAL (CORRECTED)	10,3767			

a) Fill in the table of ANOVA, indicating the calculations carried out. What effects are statistically significant? (consider α=0.05) *(3.5 points)*

b) Are the results obtained in section a) consistent with the interaction plot shown? Justify your reply. *(1.5 points)*

c) Taking into account the results obtained, how does the energy demand affect the duration of batteries? *(1.5 points)*

d) The plot of LSD intervals only shows the interval corresponding to supplier C. Draw the other two intervals, justifying the calculations. Is it correct to describe the effect of any of the factors as linear or quadratic? *(2.5 points)*

e) Describe the procedure to study in this case the presence of outliers. *(1 point)*

3. A regression analysis was carried out to evaluate the time employed by a new algorithm of vector management. All vectors in the study contain between 500 and 800 items. Two variables were considered: different size of the problem (variable called VECTOR_SIZE, which is the number of items in the vector) and, also, the mean time required for the vector management (variable called TIME) expressed in microseconds. The fitted model obtained by means of Statgraphics from the simple linear regression analysis is the following:

```
Regression Analysis - Linear model: Y = a + b*X
-------------------------------------------------------------------------
Dependent variable: TIME
Independent variable: VECTOR_SIZE
-------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
-------------------------------------------------------------------------
Intercept       -323,203       216,543                        0,1476
Slope            1,83038       0,319413
-------------------------------------------------------------------------
                           Analysis of Variance
-------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square   F-Ratio    P-Value
-------------------------------------------------------------------------
Model                     1,1833E6
Residual                  936897,0
-------------------------------------------------------------------------
Total (Corr.)           2,12019E6     27
```

Based on the results obtained, answer the following questions:

a) What is the estimated regression equation that should be used to predict the time of vector management as a function of the vector size? Do you think that the estimated values of the model parameters are reasonable taking into account their interpretation? Justify conveniently your reply. *(2 points)*

b) Justify if the linear effect of the vector size is statistically significant (α = 0.05). *(2 points)*

c) Between what limits will fluctuate, in 95% of cases, the time of management for a vector containing 800 items if this algorithm is used? *(2.5 points)*

d) Calculate the value of the correlation coefficient between the time of management of a vector and its size. *(1,5 points)*

e) Apart from the time of management, two other continuous variables measured for the same population of vectors are available (Y_2 and Y_3). The following sample covariances were obtained from the three variables (Y_1, Y_2 and Y_3): $cov(Y_1, Y_2) = 24.16$; $cov(Y_1, Y_3) = 18.39$; $cov(Y_2, Y_3) = -57.12$. Can we affirm that the strongest sample correlation appears between variables Y_2 and Y_3? Justify your answer. *(2 points)*

SOLUTION

1a) Confidence interval for the population mean (m) with a confidence level of 95% (α=0.05):

$[\overline{X} - t_{13}^{0,025}\frac{s}{\sqrt{N}}, \overline{X} + t_{13}^{0,025}\frac{s}{\sqrt{N}}]$; $[5.4 - 2.160\frac{1.9}{\sqrt{14}};\ 5.4 + 2.160\frac{1.9}{\sqrt{14}}]$; **[4.303; 6.497]**

The probability 0.95 associated with the confidence interval obtained implies that, if 100 samples are extracted from the population and the confidence interval is obtained from each sample, 95 of these intervals (in average) will contain the real value of the population mean, while 5 of them will not contain it. Therefore, there is a probability of 0.95 to find the real unknown value of the population mean within the computed interval.

1b) The null hypothesis to test is that the mean access time in the population is 4 seconds, versus the alternative hypothesis that the mean time is different from 4: H_0: m=4 ; H_1: m≠4. Given that the value 4 is not included in the interval computed in the previous section (obtained with 1-α=0.95), it cannot be admitted that m=4 (considering α=0.05).

1c) The value 4 is outside the confidence interval obtained in section a) with α=0.05. Considering α=0.10, the interval becomes narrower (smaller), so surely the value 4 will also be outside the interval. Thus, it cannot be admitted m=4. The interval will be narrower because the critical value of Student's t with α=10% is $t_{13}^{0.05}$ = 1.771 instead of $t_{13}^{0.025}$ = 2.160.

1d) Confidence interval for σ: $\left[\sqrt{(n-1)\cdot s^2/g_2};\sqrt{(n-1)\cdot s^2/g_1}\right]$ being g_1 the critical value that satisfies P($\chi^2_{13} > g_1$)=0.975 and being g_2 the critical value so that P($\chi^2_{13} > g_2$)=0.025. We obtain from the tables: g_1 =5.009 ; g_2 =24.736. Being n-1=13: $\left[\sqrt{13\cdot 1.9^2/24.736};\sqrt{13\cdot 1.9^2/5.009}\right] = [1.377;\ 3.061]$

As the value σ=2 is within the obtained interval, it can be admitted that the sample was extracted from a population with σ=2.

1e) Two types of wrong decisions can be adopted in a hypothesis test:
- To reject the null hypothesis when it is actually true: it is called type I error, or α error.
- To accept the null hypothesis when it is actually false: it is called type II error or β error.

2a) The ANOVA table is filled out based on the following calculations:
1) Total degrees of freedom = 11 (12 data minus one)
2) Degrees of freedom of factor supplier = 3 variants - 1 = 2.
3) Degrees of freedom of factor demand = 2 levels - 1 = 1
4) Degrees of freedom of the interaction = 2·1 = 2

5) Residual degrees of freedom (obtained by difference) = 11 - 2 - 1 - 2 = 6
6) $SS_{resid} = MS_{resid} \cdot df_{res} = 0.005 \cdot 6 = 0.03$
7) $MS_{A\cdot B} / MS_{res} = \text{F-ratio}_{A\cdot B}$; $MS_{A\cdot B} = 2.67 \cdot 0.005 = 0.013333$
8) $SS_{A\cdot B} = MS_{A\cdot B} \cdot df_{A\cdot B} = 0.01333 \cdot 2 = 0.02667$
9) $SS_{suppl} = SS_{total} - SS_{res} - SS_{AB} - SS_{dem} = 10.377 - 0.03 - 0.0267 - 0.213 = 10.1067$
10) $MS_{supplier} = SS_{suppl} / df_{suppl} = 10.1067 / 2 = 5.0533$
11) $MS_{dem} = SS_{dem} / df_{dem} = 0.2133 / 1 = 0.2133$
12) $F_{supplier} = MS_{suppl}/MS_{res} = 5.0533 / 0.005 = 1010.67$
13) $F_{dem} = MS_{dem}/MS_{res} = 0.2133 / 0.005 = 42.67$

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:PROVEEDOR	10,1067	2	5,05333	1010,67
B:DEMANDA	0,213333	1	0,213333	42,67
INTERACTIONS				
AB	0,02667	2	0,0133333	2,67
RESIDUAL	0,03	6	0,005	
TOTAL (CORRECTED)	10,3767	11		

The effect of supplier is statistically significant for α=0.05 because its F-ratio = 1010.67 is greater than the critical value from the F table: $F_{2;6}^{0.05}$=5.14.
The effect of demand is also significant because its F-ratio=42.67 is greater that the critical value: $F_{1;6}^{0.05}$=5.99.
However, the effect of the interaction is not statistically significant because its F-ratio=2.67 is lower than the critical value: $F_{2;6}^{0.05}$=5.14

2b) Results obtained in the previous section are the following:
- The highest sum of squares corresponds to factor supplier, which implies that this factor is the main responsible of the data variability. This fact is consistent with the interaction plot, which reveals substantial differences particularly between suppliers A and B.
- The effect of factor demand is statistically significant, but its sum of squares is much lower than in the case of supplier. This is consistent with the plot, because the differences between low and high demand are clearly lower than those observed between suppliers. Moreover, the fact that not all LSD intervals overlap between high and low demand also suggests that this factor is statistically significant.
- The interaction is not statistically significant, which is consistent with the plot because the lines corresponding to high or low demand are nearly parallel.

2c) The duration of batteries is significantly lower for a high energy demand. This effect of demand is independent of the type of supplier because the interaction is not statistically significant. Taking into account that the average duration with high energy demand is 3.78 years, and the mean for a low demand is 4.05 years (difference = 0.27), it can be concluded that, when the energy

demand changes from high to low, the duration increases in average 0.27 years for any one of the three suppliers.

2d) The width of the LSD intervals depends on the number of values and the confidence level. Since there are 4 values for each supplier and the confidence level is the same, the three intervals will have the same width as for supplier C shown in the plot (approximately ±0.05).
Average of values for A: (2.7+2.8+3.2+3.1)/4 = 2.95
Media de los datos de B: (5.0+5.1+5.2+5.3)/4 = 5.15
LSD intervals for supplier A: 2.95 ±0.05; supplier B: 5.15 ±0.05

- Supplier: this is a qualitative factor and, hence, it is **not** possible to discuss about the linear or quadratic effect. Actually, the shape of the plot changes totally just changing the order of suppliers (e.g. A-C-B instead of A-B-C).
- Demand: this factor is quantitative, but it is not possible to know if the effect is linear or quadratic because it was only tested at two levels. For that purpose, it would be necessary to have information at more than two levels.

2e) In order to study the existence of outliers in ANOVA, the most powerful method is to save the residuals (corresponding to the model including all effects that are statistically significant) and, next, to plot those residuals on a normal probability plot. All values that clearly deviate from the straight line, either for being abnormally high or low, can be considered as outliers.

3a) Taking into account the estimated values indicated in the table of results, the equation of the regression model is the following:
Time = -323.203 + 1.83038 · vector_size
Although the p-value of the constant is not statistically significant, it is not correct to remove it and use the model: Time = 1.83038·vector_size.

- The estimated value of the slope (1.83) is reasonable because more time will be required for the vector management as its size increases.
- The estimated value of the constant (-323.203) might seem unreasonable because the time would become negative when size tends to zero, which is impossible. But this situation never occurs because the model was fitted within the range of size between 500 and 800 items. When size is 500, the estimated time is 592 μs, which guarantees that the predicted times will always be positive within the range of size under study.

3b) The linear effect of vector size is statistically significant if the slope of the straight line is different from zero at the population level. Thus, being the regression model: $Y=\beta_0+\beta_1 \cdot X$, the null hypothesis to test is: H_0: $\beta_1=0$ versus the alternative hypothesis H_1: $\beta_1 \neq 0$. Given that $b_i / s_{b_i} \approx t_{N-1-I} \approx t_{26}$ being I=1 (model with one explicative variable), N=28 (total number of observations = total d.f. + 1) and $\alpha=0.05$, it turns out that: $b_i / s_{b_i} = 1.83/0.319 = 5.73$ which is not a frequent value for the t_{26} distribution. Thus, the null hypothesis is rejected: there is enough evidence to affirm that the linear effect is statistically significant.

OTHER ALTERNATIVE METHOD: In simple linear regression, the p-value associated with the slope is always exactly the same as the p-value of the global significance test (ANOVA). By completing that table:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1,1833E6	1	1,1833E6	32,8380	
Residual	936897,0	26	36034,5		
Total (Corr.)	2,12019E6	27			

The F-ratio is greater than the critical value from the F table: $F_{1;26}^{0.05}$=4.23. Therefore, the null hypothesis is rejected and it is concluded that the linear effect is statistically significant.

3c) When the size is 800, the conditional distribution of time will be a Normal distribution: the average is given by the regression model, and the variance will be the residual variance, which coincides with the residual mean squares:

$MS_{res} = SS_{res} / df_{res} = 936897 / 26 = 36034.5$

Expected average time (for size=800) = -323.203 + 1.83038 · 800 = 1141.1

In a Normal distribution, m±1.96·s comprises 95% of the values. Therefore, the requested interval is: $1141.1 \pm 1.96 \cdot \sqrt{36034.5}$ = **[769 ; 1513]**

Nonetheless, if the value 1.96 is rounded up to 2, the result can also be considered as correct.

3d) $r = \sqrt{R^2} = \sqrt{\frac{SS_{\text{mod}}}{SS_{total}}} = \sqrt{\frac{1.1833 \cdot 10^6}{2.1202 \cdot 10^6}} = \mathbf{0.747}$

The sign of the correlation coefficient is positive because the slope is positive.

3e) In order to determine if the sample correlation is stronger of weaker, it is necessary to calculate the coefficient of correlation from the covariance: $r = \text{cov} / (s_X \cdot s_Y)$. But this calculation is not possible in this case because the standard deviations of Y_2 and Y_3 are unknown. Thus, we cannot affirm that the strongest sample correlation in this case appears between variables Y_2 and Y_3.