Bachelor Degree in Computer Engineering

Statistics **group E (English)**

SECOND PARTIAL EXAM

June 2nd 2014

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. In certain study to test a type of processor, 9 experimental trials were carried out to measure the execution time of certain type of operation. The average value of the data was 52.02 and the standard deviation was 0.82.

a) Calculate a confidence interval for the population average, considering a confidence level of 95%. Can it be admitted an average m=53? *(3.5 points)*

b) Assuming that the population average is 53, what is the probability to obtain a sample average higher than 52.02? *(3 points)*

c) Can it be admitted that the standard deviation is σ=1.6? Consider a type I risk α=1%. *(3.5 points)*

2. An experimental design is performed to study the effect of two factors (model of processor and RAM memory) in the time (in milliseconds) required to perform a search in a big database. Three models of processor and two memories (10 MB and 20 MB) are tested, and two replicates are carried out for each one of the possible combinations. The data obtained are indicated in the following table. It is assumed that data are normally distributed.

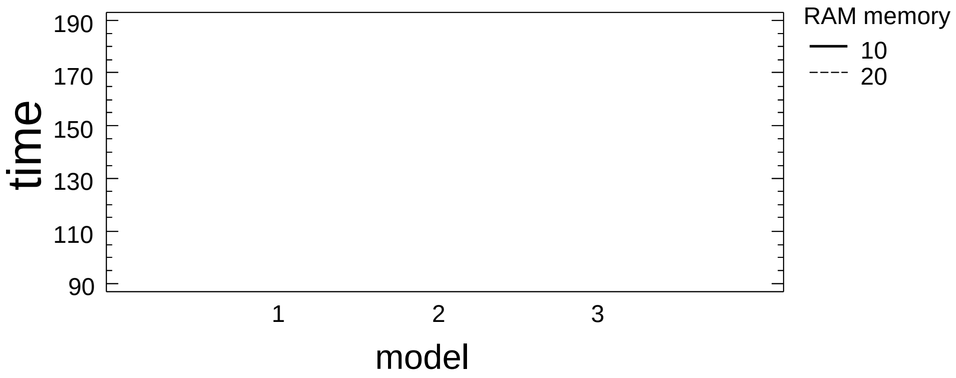
RAM memory	model 1	model 2	model 3
10 MB	102; 108	171; 176	119; 125
20 MB	92; 97	157; 164	117; 123

Results obtained with Statgraphics are the following:

```
Analysis of Variance for TIME - Type III Sums of Squares
------------------------------------------------------------------------
Source                    Sum of Squares     Df    Mean Square    F-Ratio
------------------------------------------------------------------------
MAIN EFFECTS
 A:RAM_memory                     216,75     [redacted]
 B:model                          9453,5     [redacted]

INTERACTIONS
 AB                           [redacted]     [redacted]

RESIDUAL                      [redacted]     [redacted]       17,25
------------------------------------------------------------------------
TOTAL (CORRECTED)                9840,25     [redacted]
------------------------------------------------------------------------
```

a) Determine if the simple effect of each factor or the interaction is statistically significant, considering α=5%. *(4 points)*

b) Draw the mean values in the interaction plot (justify your answer). Taking into account the results of ANOVA, what information can be deduced from this plot? *(3.5 points)*



c) Taking into account the results of ANOVA, determine the optimum operative conditions that lead to minimize the time of search in the database, considering α=1%. Calculate the average time expected under those conditions. *(2.5 points)*

3. Certain company of digital music recording is studying the possibility to predict the sales (thousands of compact disks/month) based on the investment in publicity (thousand euros/month). For this purpose, a set of historical data about sales and investment was analyzed by means of linear regression. The results obtained are the following (α=0.01):

Simple Regression - Sales vs. Investment in publicity
Dependent variable: Sales (thousand of disks/month)
Independent variable: Publicity (thousand euros/month)
Linear model: Y = a + b*X
Coefficients

Parameter	*Estimate*	*Standard Error*	*T Statistic*	*P-Value*
Intercept	134,14	7,53658		0,0000
Slope	0,0961245	0,00963236		

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	433688,	1	433688,	99,59	0,0000
Residual	862264,	198	4354,87		
Total (Corr.)	1,29595E6	199			

Correlation Coefficient = 0,578488

Based on the results shown above, answer the following questions:

a) Estimate the parameters of the regression model. Study the statistical significance of the parameters. What is the mathematical equation of the regression model? *(3 points)*

b) Calculate the average amount of monthly sales that is expected for an investment in publicity of 10000 euros in a certain month. *(2 points)*

c) Calculate the Coefficient of Determination of this model. What is the practical interpretation of this coefficient? *(2,5 points)*

d) What is the practical interpretation of the Residual Mean Square? Calculate the value of this parameter in this case. *(2,5 points)*

SOLUTION OF THE SECOND PARTIAL EXAM

1a) The confidence interval for the execution time, considering a confidence level of 95%, is calculated as:

$$[\bar{X} - t_8^{\alpha/2=0,025} \frac{s}{\sqrt{N}}, \bar{X} + t_8^{\alpha/2=0,025} \frac{s}{\sqrt{N}}] \qquad [52.02 - 2.306\frac{0.82}{\sqrt{9}}, 52.02 + 2.306\frac{0.82}{\sqrt{9}}]$$

The interval is: [**51.39, 52.65**] As the value m=53 is outside this interval, it cannot be admitted that m=53.

1b) $P(\bar{x} > 52.02) = P\left(\frac{\bar{x}-m}{s/\sqrt{n}} > \frac{52.02-m}{s/\sqrt{n}}\right) = P\left(t_8 > \frac{52.02-53}{0.82/\sqrt{9}}\right) = P(t_8 > -3.58) \approx$

$\approx$1-0.005 = **0.995** (exact value obtained with Statgraphics: 0.996)

1c) The interval for the standard deviation, considering α=1% is:

$$\left[\sqrt{(N-1)\frac{s^2}{g_2}}, \sqrt{(N-1)\frac{s^2}{g_1}}\right] \qquad \left[\sqrt{(9-1)\frac{0,82^2}{g_2}}, \sqrt{(9-1)\frac{0,82^2}{g_1}}\right]$$

Being g_1=**1.344** (from the Chi-square table, looking at the column 0.995, with 8 degrees of freedom) and g_2=**21.955** (from the Chi-square table, looking at the column 0.005).

The resulting interval is: [**0.49, 2**]. As the value σ=1.6 is comprised inside this interval, it can be admitted that the standard deviation of the population is 1.6.

2a) Total degrees of freedom (df) = 12 - 1 = 11
Degrees of freedom of factor RAM memory = 2 levels - 1 = 1
Degrees of freedom of factor model = 3 variants - 1 = 2
Degrees of freedom of the interaction: 1 · 2 = 2
Residual degrees of freedom, are obtained by difference: 11 - 1 - 2 - 2 = 6

$SS_{residual} = MS_{resid} \cdot df_{resid}$ = 17.25 · 6 = 103.5
$SS_{interac} = SS_{total} - SS_{res} - SS_{RAM} - SS_{model}$ = 9840.25-103.5-216.75-9453.5 =66.5
$F_{ratioRAM}$=(SS/df)/MS_{res} = (216.75/1) / 17.25 = 12.57
F_{ratio_model}=(SS/df)/MS_{res} = (9453.5/2) / 17.25 = 274.01

Considering α=0.05, the simple effect of factor RAM memory is statistically significant because the F-ratio (12.57) is higher than the critical value from the tables ($F_{1;6}$) which is 5.99.
The simple effect of factor model is statistically significant because the F-ratio (274.01) is higher than the critical values from tables ($F_{2;6}$) which is 5.14.
The effect of the interaction is NOT statistically significant because the F-ratio (1.93) is lower than the critical value from the tables ($F_{2;6}$) which is 5.14.

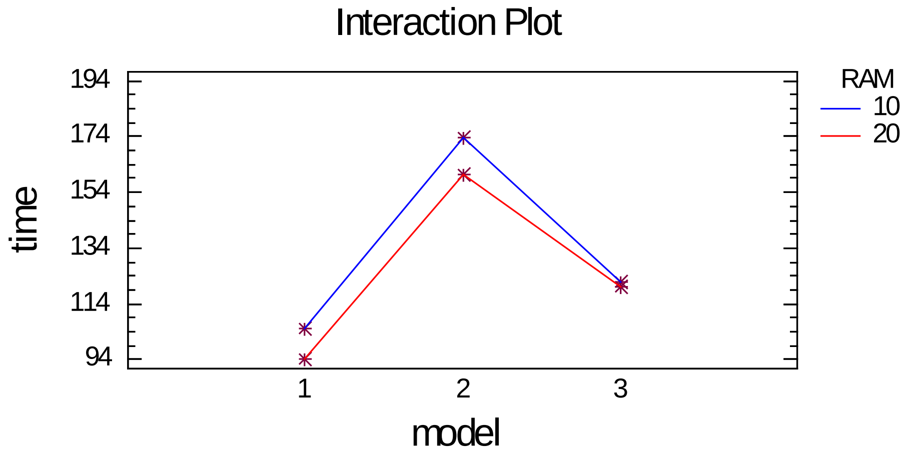
The complete summary table is shown next (the p-values are also indicated although these values can only be obtained with Statgraphics).

```
Analysis of Variance for TIME - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                    Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:RAM_memory                   216,75        1         216,75      12,57     0,0121
 B:model                        9453,5        2        4726,75     274,01     0,0000

INTERACTIONS
 AB                               66,5        2          33,25       1,93     0,2257

RESIDUAL                         103,5        6          17,25
--------------------------------------------------------------------------------
TOTAL (CORRECTED)              9840,25       11
--------------------------------------------------------------------------------
```

2b) Average values for each treatment:

(102+108)/2=105	(171+176)/2= 173,5	(119+125)/2=122
(92+97)/2= 94,5	(157+164)/2=160,5	(117+123)/2=120

These values are plotted in the graph, resulting:



The interaction is not statistically significant, which implies that there is not enough evidence to affirm that the effect of RAM memory is different for each one of the three models. Thus, at the population level, it can be admitted that the time for RAM=10 is significantly higher than for RAM=20, independently of the model type. With respect to the model, which is a qualitative factor, the time for model=2 is significantly higher than for model=1. In the case of model=3, the resulting time is intermediate, but from this plot it cannot be determined if the differences with respect to the other two models are statistically significant because LSD intervals are not available.

2c) Considering α=1%, factor model is statistically significant because the F-ratio is higher than the critical value for a $F_{2;6}$ distribution (274.01 >> 10.92). By contrast, factor RAM memory is <u>not statistically significant</u> because F-ratio is less than the critical value for a distribution $F_{1;6}$ (12,57 < 13,75). Thus, there is not enough evidence to affirm that RAM=10 takes more time, at the population level, than RAM=20. As a conclusion, only factor “model” is considered for the optimum operative conditions. Based on the interaction plot, a lower time will be obtained with model 1. The average time expected for model 1 will be: (102+108+92+97)/4 = **99.75**.

Solution problem 3:

3a) The regression model (Y=a+b·X) has two parameters: the intercept (a) and the slope (b). The estimated value of both parameters is obtained from the table of results: intercept = 134.14; slope = 0.09612.

The intercept is statistically significant (which means that it is different from zero at the population) because its p-value is less than 0.01. The p-value associated to the slope is not indicated in the table of coefficients, but it is the same as the p-value appearing in the table below "analysis of variance" (global significance test) which is less than 0.01, which implies that it is also statistically significant.
Thus, the model used to predict the sales will be:

Sales = 134.14 + 0.09612·Publicity

3b) For an investment in publicity of 10000 euros, as the units are thousand €, the variable "publicity" will take the value 10, and the expected average value of sales will be:
Sales = 134.14+0.09612·10 = **135.101** thousand disks/month = 135101 disks/month

3c) $R^2 = \frac{SS_{model}}{SS_{total}} \cdot 100 = \frac{433688}{1295950} \cdot 100 = 33.465\%$

In the particular case of simple linear regression, the coefficient of determination can also be obtained as the square of the correlation coefficient:
$\mathbf{R^2 = (r_{xy})^2 \cdot 100} = (0.5785)^2 \cdot 100 \cong \underline{\mathbf{33.5\%}}$
$\mathbf{R^2}$ is used to assess the goodness-of-fit for the model obtained. It expresses the percentage of variability (variance) of the dependent variable (in this case, monthly sales of disks) explained by the model (i.e., explained by the variability of the dependent variable, which is monthly investment in publicity).

3d) The residual mean square is an estimation of the residual variance (i.e., the variance of residuals). The residual of an observation is the difference between the observed value of the dependent variable and the value predicted by the regression model. The residual mean square accounts for the effect of all factors (random variables) not considered in the model.

Mean Square → $\underline{\mathbf{MS_{res} = 4354.87}}$ units^2

In the particular case of simple regression, as in this case, the residual variance could also be obtained as:

$\underline{\mathbf{S^2_{resid} = S^2_y (1-r^2_{xy})}} = 6512.32 \cdot (1-(0.578488)^2) \cong \underline{\mathbf{4332.98}}$ units^2