Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 19th 2014

Surname, name:	
Group: **1E**	Signature:
Indicate with a tick mark the partials examined	1st □ 2nd □

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. Over the table it is only permitted to have the DNI (identification document), calculator, pen, and the formula tables.
5. **Do not unstaple any page of the exam** (do not remove the staple).
6. The exam consists of 6 questions, 3 ones corresponding to the first partial (50%) and 3 about the second partial (50%). The lecturer will correct those partial exams indicated by the student with a tick mark in this page. **All questions of each partial exam score the same** (over 10).
7. 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.
8. Time available: **3 hours**

1. (1st Partial) One shopping center sells three models of laptop (A, B and C). The following frequency table displays the sales of each model during the year 2013, as a function of the customer's age: young customer (age < 30), adult (from 30 to 50) and senior customer (age > 50). It is assumed that each customer indicated in the table only purchased one laptop.

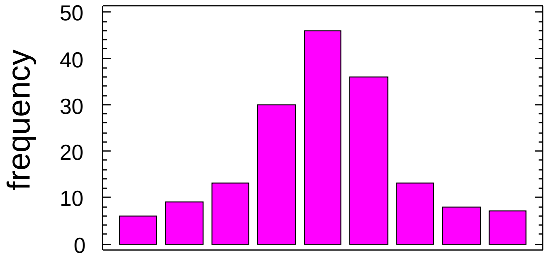
	Model A	model B	model C	Row Total
young	30 26,79%	46 41,07%	36 32,14%	112 66,67%
adult	9 30,00%	13 43,33%	8 26,67%	30 17,86%
senior	6 23,08%	13 50,00%	7 26,92%	26 15,48%
Column Total	45 26,79%	72 42,86%	51 30,36%	168 100,00%

Answer the following questions justifying properly the reply.

a) If a customer is randomly chosen, what is the probability to be young and to have purchased a laptop of model A or B? *(3 points)*

b) Based on the information contained in the table, are the events "*to be a young customer*" and "*to have purchased a laptop of model A*" independent? *(4 points)*

c) The following figure shows a bar chart obtained from the data reflected in the frequency table (the information in the horizontal axis has been omitted). What conclusion can be deduced from this chart taking into account the information provided in the frequency table? *(3 points)*



2. (1st Partial) The time taken by a computer server to run a request (access) follows an exponential distribution, being the median of 3 milliseconds.

a) What is the probability to run a request in less than 2 milliseconds?

(3 points)

b) Calculate the percentile 95 of the random variable under study. *(2 points)*

c) If 10 consecutive accesses to the server are randomly selected, what is the probability to get a total time greater than 50 milliseconds? *(5 points)*

3. (1st Partial) One computer program records daily the number of failures that occur in the machines of certain industry. The average is 3 failures per day.

a) Calculate the probability of occurring more than 20 failures in 6 days. *(5 points)*

b) What is the probability that in one week (6 working days) occur exactly 15 failures? *(5 points)*

4. (2nd Partial) As part of a study on the academic performance of students in the first degree course of Computer Engineering, a set of 111 scores obtained by students in the second partial of Statistics has been analyzed. It was obtained that the sample average was 5.54 and the confidence interval at 95% for the population average score was [5.2 ; 5.9].

Assuming that the 111 students evaluated can be regarded as a representative sample of all students in the first degree course, and that the score obtained by students follows a Normal distribution, answer the following questions:

a) What is the interpretation of the confidence interval obtained? *(3.5 points)*

b) Can it be accepted that the average score of students in the first degree course is 6.2 considering a confidence level of 95%? What would be the answer considering a confidence level of 90%? Justify all your replies. *(3 points)*

c) Taking into account that the standard deviation obtained in the scores of the 111 students was 1.88, calculate an interval (with a confidence level of 95%) for the population standard deviation σ. *(3.5 points)*

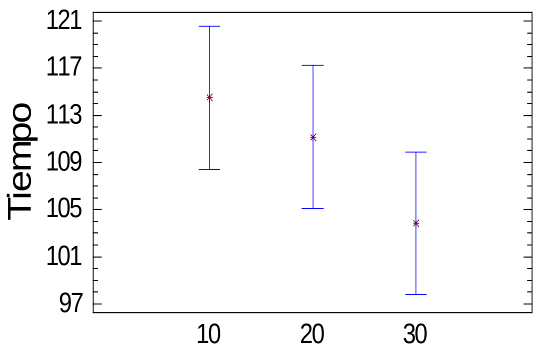
5. (2nd Partial) One experiment has been carried out to assay the effect of two factors (model of processor and RAM memory) in the time (in milliseconds) taken to perform a search in a big database. Three models of processors (10, 20 and 30) are assayed and three memories (10 MB, 20 MB and 30 MB). Two replicates were performed for all possible combinations. It is assumed that data are normally distributed.

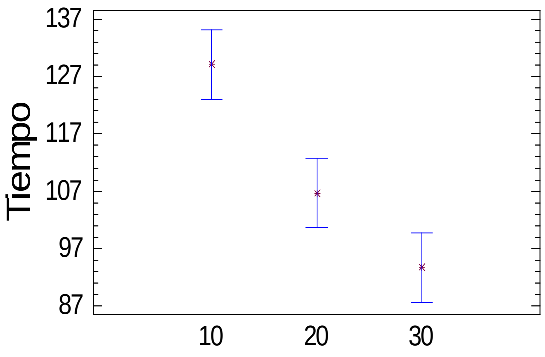
Results obtained with Statgraphics are the following:

```
Analysis of Variance for TIME - Type III  Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df      Mean Square       F- Ratio
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:Memory                 3871,0
 B:Model                 357,333

INTERACTIONS
 AB

RESIDUAL                                               86,3889
--------------------------------------------------------------------------------
TOTAL (CORRECTED)         5028,5
--------------------------------------------------------------------------------
```

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

b) The plots below show the LSD intervals for the two factors analyzed, considering a confidence level of 95%. Indicate which factor (memory or model) corresponds to each one of the plots, justifying the answer. *(3 points)*





c) What would be the optimum operative conditions to minimize the search time in the database? (consider α=10%). *(3 points)*

6. (2nd Partial) The average response time (in seconds) of a computer system, as well as the mean load (in queries per minute), has been measured during 13 days. The resulting data were analyzed by means of linear regression.

```
Multiple Regression Analysis
-----------------------------------------------------------------
Dependent variable: TIME
-----------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic
-----------------------------------------------------------------
CONSTANT               0,0747486       0,190814
LOAD                    0,789228       0,084857
-----------------------------------------------------------------

Analysis of Variance
--------------------------------------------------------------
Source           Sum of Squares     Df  Mean Square    F-Ratio
--------------------------------------------------------------
Model                   17,1757
Residual
--------------------------------------------------------------
Total                   17,5169
```

a) Write the mathematical equation of the estimated model. Which parameters of this model are statistically significant? (α=5%). *(4 points)*

b) What is the interval that comprises approximately 95% of the cases for the response time when the load is equal to 4? *(6 points)*

SOLUTION OF THE EXAM

EXERCISE - 1

1a) (30+46)/168 = 0.4524 = 45.24%

1b) Event A: the customer has purchased a laptop of model A.
Event B: the customer is young.

P(A/B) = 30/112 = 0.2679 ; P(A) = 45/168 = 0.2679

P(B/A) = 30/45 = 0.6667 ; P(B) = 112/168 = 0.6667

Considering that P(A/B) = P(A) and, moreover, P(B/A) = P(B), then it can be concluded that the events A and B are independent.

1c) The bar-chart shows the values of the 9 absolute frequencies indicated in the table. The main conclusion is that three frequencies are quite higher than the rest, which are: 46 (young customer - model B), 36 (young customer - model C) and 30 (young customer - model A). Thus, highest values correspond to young customers which correspond to the highest proportion (66.67%).

EXERCISE - 2

2a) $P(X<x)=1-e^{-\alpha\cdot x}$; $P(X<3)=0.5=1-e^{-\alpha\cdot 3}$; $\alpha=-(\ln 0.5)/3=0.231$

$P(X<2)=1-e^{-0.231\cdot 2}=\mathbf{0.37}$

2b) Percentile 95 (Z_{95}) is the value so that: $P(X<Z_{95})=0.95$

$P(X<Z_{95})=1-e^{-0.231\cdot Z_{95}}=0.95$; $Z_{95}=-[\ln(1-0.95)]/0.231=\mathbf{12.96\ ms}$

2c) Random variable X: time (ms) taken by the server to run a request

X~exp(α) ; $E(X)=1/\alpha=4.328$; $\sigma^2(X)=1/\alpha^2=18.732$

Random variable Y: total time (ms) of 10 consecutive requests

$Y=X_1+X_2+...+X_{10}$; $E(Y) = E(X_1+...+X_{10}) = E(X_1)+...+E(X_{10})=10\cdot E(X)=43.28$

$\sigma^2(Y)=\sigma^2(X_1+...+X_{10})=\sigma^2(X_1)+...+\sigma^2(X_{10})=10\cdot\sigma^2(X)=10\cdot 18.732=187.32$

Assuming that Y is a Normal distribution according to the central limit theorem

$$P(Y>50)=P\left[N\left(m=43.28;\sigma=\sqrt{187.32}\right)>50\right]=P\left[N(0;1)>\frac{50-43.28}{\sqrt{187.32}}\right]=P[N(0;1)>0.49]=\mathbf{0.312}$$

EXERCISE - 3

3a) Variable X: number of failures in one day. X~Ps(λ); E(X)= λ =3

Variable Y: number of failures in 6 days. $Y=X_1+...+X_6$; $Y\approx Ps(\lambda=\lambda_{X_1}+...+\lambda_{X_6})$;

$Y\approx Ps(\lambda=3\cdot 6=18)$; $P(Y>20)=1-P(Y\le 20)=$(abacus) = 1 - 0.73=**0.27**

3b) $P(Y=15)=e^{-18}\cdot\frac{18^{15}}{15!}=\mathbf{0.0786}$

EXERCISE - 4

4a) If 100 samples of equal size are taken from a Normal population and the confidence interval (CI) is calculated for the population mean score (m), we would obtain a different sample mean and sample variance, which would lead to different confidence intervals. However, 95 of these 100 CI calculated would contain the true value of the population mean (i.e., the mean score of Statistics for all students of first degree course).

Therefore, the confidence interval obtained comprises the set of null hypotheses for m that are consistent with the data of a given sample, for a certain significance level. The population mean score (for all students of first degree course) is comprised between 5.2 and 5.9 with a probability of 95%.

4b) Given that 6.2 is outside the confidence interval at 95%, we cannot accept (with a confidence level of 95%) the hypothesis that the mean score is 6.2 for students of first year course.

The confidence interval at 90% is narrower than the previous one and, hence, 6.2 will also be outside this interval. Consequently, it cannot be accepted that the mean score for students of first year is 6.2 considering a confidence level of 90%.

4c)

$$IC_{\sigma}^{95\%} = \left[\sqrt{(n-1)\frac{s^2}{g_2}}, \sqrt{(n-1)\frac{s^2}{g_1}}\right] = \left[\sqrt{110\frac{1.88^2}{140.9}}, \sqrt{110\frac{1.88^2}{82.9}}\right] = [1.66;\ 2.16]$$

$$Tabla\ g_1 / P\left(\chi^2_{n-1} \le g_1\right) = P\left(\chi^2_{111-1} \le g_1\right) = \frac{\alpha}{2} = \frac{0.05}{2} = 0.025 \rightarrow g_1 = 82.86$$

$$Tabla\ g_2 / P\left(\chi^2_{n-1} \ge g_2\right) = P\left(\chi^2_{111-1} \ge g_2\right) = \frac{\alpha}{2} = \frac{0.05}{2} = 0.025 \rightarrow g_2 = 140.92$$

EXERCISE - 5

5a) Total number of values = 9 treatments x 2 repetitions = 18
Total degrees of freedom = 18 - 1 = 17
Degrees of freedom of factor RAM memory = 3 levels - 1 = 2
Degrees of freedom of factor model = 3 variants - 1 = 2
Degrees of freedom of the interaction: 2 · 2 = 4
Residual degrees of freedom are obtained by difference: 17 - 2 - 2 - 4 = 9

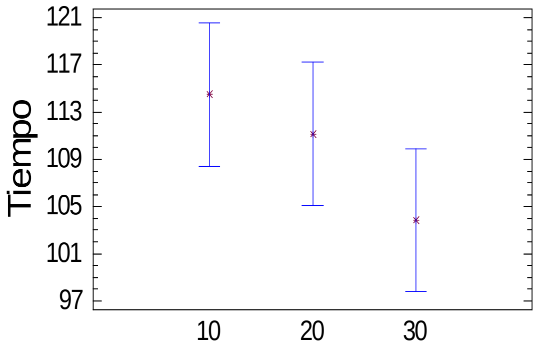
$SS_{residual}$ = MS_{resid} · df_{resid} = 86.3889 · 9 = 777.5
$SS_{interac}$=SS_{total} - SS_{res} - SS_{RAM} - SS_{model} = 5028.5-777.5-3871-357.33 =22.67
$F_{ratioRAM}$=(SS/df)/MS_{res} = (3871/2) / 86.39 = 22.4
F_{ratio_modelo}=(SS/df)/MS_{res} = (357.33/2) / 86.39 = 2.07
$F_{ratio\_interac}$=(SS/df)/MS_{res} = (22.67/4) / 86.39 = 0.07

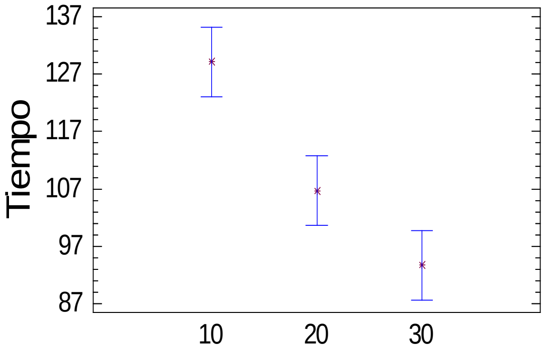
Considering α=0.05, the simple effect of factor *RAM memory* is statistically significant because its F-ratio (22.4) is higher than the critical values from tables ($F_{2;9}$ = 4.26).
The simple effect of factor model is NOT statistically significant because the F-ratio (2.07) is less than the critical values from tables ($F_{2;9}$ = 4.26).
The effect of the interaction is NOT statistically significant because the F-ratio (0.07) is lower than the critical values from tables ($F_{4;9}$ = 3.63).

The summary table is shown below (the p-value is also indicated though it can only be calculated 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:Memory                       3871,0      2          1935,5       22,40      0,0003
 B:Model                       357,333      2         178,667        2,07      0,1824

INTERACTIONS
 AB                            22,6667      4         5,66667        0,07      0,9907

RESIDUAL                         777,5      9         86,3889
--------------------------------------------------------------------------------
TOTAL (CORRECTED)               5028,5     17
--------------------------------------------------------------------------------
```

5b) The plot on the right corresponds to factor memory because the first LSD interval does not overlap with the others, which indicates that the effect is statistically significant. Given that this is the only significant factor, according to the previous question, necessarily the plot should correspond to factor memory.
By contrast, in the plot on the left, all LSD intervals are overlapped, which suggests that the simple effect of this factor is not statistically significant. Hence, the plot on the left should correspond to factor model which is not significant considering α=5%.





5c) The means plot shown above corresponds to a confidence level of 95% (α=5%). As factor model is not statistically significant, as well as the interaction, it doesn't matter to choose any of the three models because they will provide in average the same average time. However, the mean time for RAM=30 is significantly lower than the others, and consequently this will be the optimum operative condition to minimize the time, considering α=5%.

If the significance level is α=10%, as indicated in the statement, the LSD intervals will be narrower, and the conclusion will be the same because:
- In the case of factor memory (right plot), none of the three LSD intervals will overlap as they become narrower (lower amplitude).
- In the case of factor model (left figure), though the intervals become somewhat narrower, it is likely that the LSD for 20 will still overlap with the LSD for 30.

EXERCISE - 6

6a) Based on the estimated values for the parameters shown in the table, the estimated model will be: E(T_resp/load) = 0.0747+0.789· load

The values of t-calc (t-statistic) are obtained as: estimate/standard_error:

For the ordinate (constant): t-calc=0.07475/0.1908 = 0.39

For the slope: t-calc= 0.7892/0.08486 = 9.3

Critical values from the t table: $t_{11}^{0,025}$=2.201. The ordinate is not statistically significant because |0.39| < 2.201. The slope is significant because |9.3|>2.201.

The statistical significance for the slope can also be studied from the ANOVA:

```
Analysis of Variance
-----------------------------------------------------------------------------
Source                  Sum of Squares      Df  Mean Square     F-Ratio
-----------------------------------------------------------------------------
Model                          17,1757       1    17,1757       554,05
Residual                         0,341      11      0,031
-----------------------------------------------------------------------------
Total                          17,5169      12
```

Critical value from the F table: $F_{1;11}^{0.05}$=4.84. As F-ratio>F-table, it can be concluded that the linear effect of load on the average response time is statistically significant.

6b) (T_resp/load=4) = 0.0747+0.789·4 = 3.232 ; $S_{residual}=(0.031)^{1/2}$= 0.176

T_resp/load=4 ~ N(m=3.232; σ=0.176)

Considering that the interval m±2σ comprises approximately 95% of the Normal distribution, the requested interval for the distribution in this case that comprises 95% of the values will be: 3.232 ± 2·0.176 = [**2.88 ; 3.82**]