**Bachelor Degree in Computer Engineering**

# Statistics

# SECOND PARTIAL

June 4th 2012

Surname, Name

Group:

Signature:

## Instructions

1. Write your name, group 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. All questions do NOT score the same (scores are indicated in each question).
7. At the end it is compulsory to sign in the list of the professor's table in order to justify that the exam has been handed in.
8. Time available: **2 hours**

**1. (10 points)** The website of certain company provides information about the new products that it offers. In order to study the interest of users for a new product launched by the company, the daily number of visits to this website has been measured during the 10 days after the launch of this product. Results obtained are the following:

```
Count = 10
Average = 145,0
Variance = 1072,22
Stnd. skewness = 1,14907
Stnd. kurtosis = 0,0464325
```

According to these results, answer the following questions:

**a)** Is it admissible that our sample comes from a normal distribution? (**2 points**)

**b)** Considering a type I risk α=0.05, is it admissible that the population average of the daily number of visits to the website is 120? **(4 points)**

**c)** Considering a type I risk α=0.05, is it admissible that the standard deviation of the daily number of visits is 25 in the population? **(4 points)**

**2. (10 points)** One experiment has been design to study the influence of configuration (two possibilities, A and B) on the average and dispersion of the performance of a computer system. Each experimental trial consists of running a test program under one of the configurations. Configuration A was assayed 8 times ($N_1$), and B was assayed 10 times ($N_2$). The performance was measured through the parameter *System Reaction Time* (i.e., time since pressing "enter" until the user begins to receive the requested service). After conducting the experiment and collecting the data, the following results were obtained:

| | | |
|---|---|---|
| Configuration A | average= 6.61 | variance= 4.45 |
| Configuration B | average= 12.43 | variance=14.47 |

Indicate if the following statements are true or false, justifying with detail the replies.

**a)** The difference of mean performance is not statistically significant considering a type I risk α=0.01. **(2.5 points)**

**b)** The difference of mean performance is statistically significant considering a confidence level of 95%. **(2.5 points)**

**c)** The p-value for the comparison of means is lower than 0.01. **(2.5 points)**

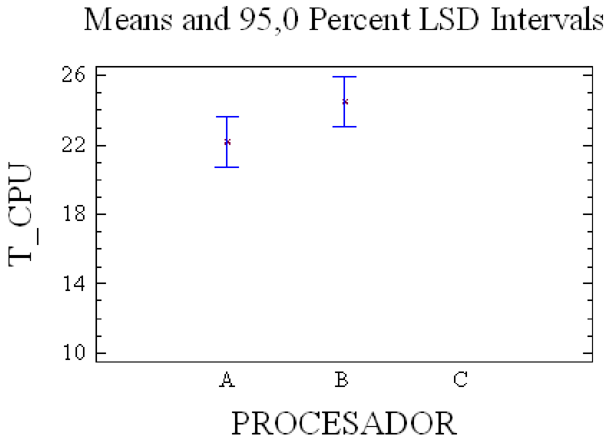
**d)** The standard deviations of each type of processor differ significantly, considering a type I risk α=5%. **(2.5 points)**

**3. (10 points**) In order to study the effect of processor and workload on the time of CPU use in the execution of certain type of procedures, tree processors (A, B and C) have been tested combined with three workloads (10, 20 and 30). Each treatment was repeated twice, and the resulting data were analyzed with ANOVA. The following table, which is incomplete, was obtained:

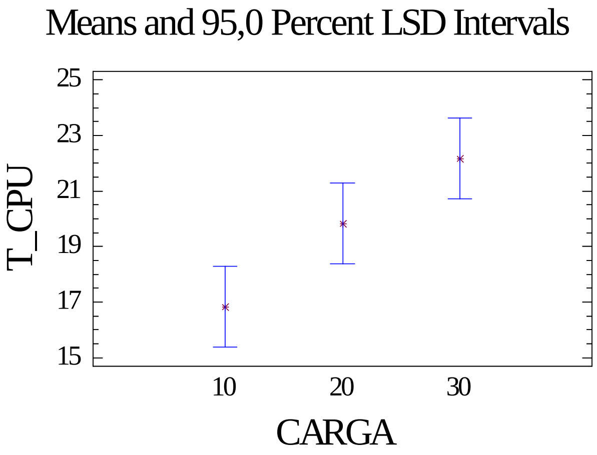
```
Analysis of Variance for T_CPU - Type III Sums of Squares
------------------------------------------------------------------------
Source       Sum of Squares     Df       Mean Square     F-Ratio     P-Value
------------------------------------------------------------------------
MAIN EFFECTS
 A:PROCESSOR      515,111
 B:WORKLOAD        85,7778

INTERACTIONS
 AB                96,8889

RESIDUAL           44,5
------------------------------------------------------------------------
TOTAL (CORRECTED)
------------------------------------------------------------------------
```

**a)** Complete the ANOVA table and indicate what effects are significant, considering a significance level of 5%. Justify the answer and all calculations. **(4 points)**

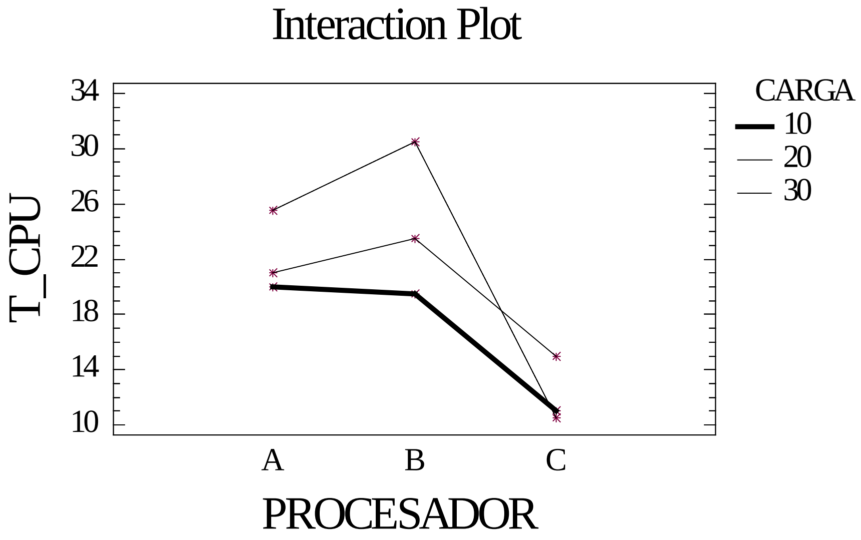
**b)** Given the following plot, draw **approximately** the LSD interval for processor C according to the results obtained in the ANOVA table and taking into account the plot shown in section **d)**. Justify your answer. **(2 points)**



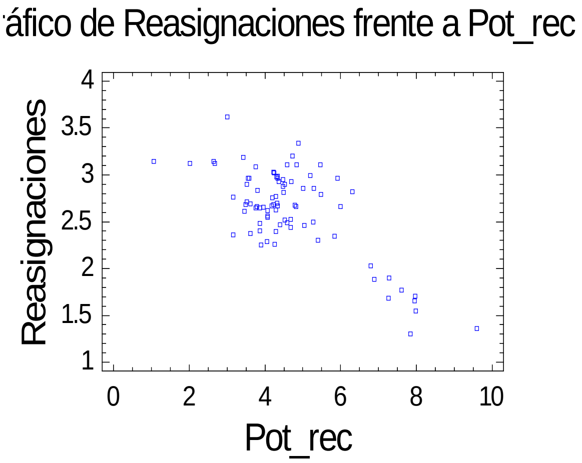
**c)** According to the following plot, indicate at descriptive level the nature of the effect of factor workload. **(2 points)**



**d)** According to the following plot and taking into account the results obtained in the ANOVA table, indicate at descriptive level the nature of the effect of factor workload. **(2 points)**



**4.** (**5 points**) A mobile phone company has been detecting problems in certain workstation. To try to solve them, the company has studied for a set of mobile terminals served by that workstation, the number of times that the transmission channel is reassigned once the call is in progress (variable *N_reassignments*), which is a quality indicator of the mobile call, and the power level (in mW) received by a mobile terminal every time that a call is initiated (variable *power*), which is an indicator of the signal coverage. In order to analyze the relationship between both variables, the following report has been obtained:



```
Summary Statistics
                Power   N_reassignm
------------------------------------
Count           82          82
Average         4.65893     2.63361
Median          4.332       2.67762
Stnd. deviation 1.43172     0.450234
Minimum         1.072       1.30222
Maximum         9.592       3.62142
Range           8.52        2.3192
First quartile  3.864       2.46312
Third quartile  5.032       2.96102
Interq. range   1.168       0.4979
-------------------------------------
Linear correlation coef. rxy = -0.7229
```

According to this information:

**a)** Describe the nature of the relationship between both random variables under study. **(1.5 points)**

**b)** Calculate the matrix of variances - covariances corresponding to both variables. **(3.5 points)**

**5.** (**10 points**) One computer engineer who works in a company dedicated to the assembly and sale of desktop computers needs to implement a regression model to predict the delivery time of new orders. This time is calculated as the number of days between the order for a new computer and the actual delivery of that computer (variable *Time*). The engineer considers that there might be a linear relationship between this time and the number of extra accessories requested with respect to the basic configuration of the computer (variable *N_extras*). A sample random of 16 orders is taken, and data of both variables (*Time* and *N_extras*) are recorded from each order. After analyzing the data by means of a linear regression model, results shown below were obtained:

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Time
Independent variable: N_extras
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic         P-Value
-----------------------------------------------------------------------------
Intercept       _______        1,59084         13,7823          0,0000
Slope           2,06871        0,116411        17,7707          ______
-----------------------------------------------------------------------------

Analysis of Variance
-----------------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                        2927,23       1      2927,23     315,80       0,0000
Residual                      129,77      14      9,26932
-----------------------------------------------------------------------------
Total (Corr.)                 3057,0      15

Correlation Coefficient = 0,978545
R-squared = ________ percent
Standard Error of Est. = _________
```

**a)** What is the equation that should be used to predict the delivery time as a function of the number of extra accessories? Indicate what estimated coefficients are statistically significant, considering α=5%. Justify your answer. **(3 points)**.

**b)** What conclusions can be derived from the global test for regression significance of the proposed model, considering a type I risk of 0.01? **(2 points)**

**c)** What percentage of the variability of delivery time is explained by the number of extra accessories? Indicate what parameter accounts for that percentage. **(2 points)**

**d)** If an order is received for a computer with 16 extra accessories, how many days will be required on average for the delivery according to the model? **(1 point)**

**e)** In what interval of values will be comprised approximately on average the delivery time for 95% of orders requesting 16 extra accessories? **(2 points)**

# SOLUTION

**1a)** As the standard skewness is comprised between -2 and 2, it can be assumed that the distribution is symmetric. As the standard kurtosis is also comprised between -2 and 2, it can be assumed that the sample comes from a Normal distribution.

**1b)** α=0.05; H$_0$: m=120; H$_1$: m≠120; H$_0$ is accepted if: $\left|\dfrac{\bar{x}-m_0}{s/\sqrt{n}}\right| < t_{n-1}^{\alpha/2}$

H$_0$ is accepted if: $\left|\dfrac{145-120}{\sqrt{1072.22}/\sqrt{10}}\right| < t_9^{0.025}$ ; H$_0$ accepted if: 2.4143 < 2.262.

This condition is not satisfied: it is **not** admissible that m$_0$ = 120.

**1c)** H$_0$: $\sigma^2=25^2$ ; H$_1$: $\sigma^2\neq 25^2$ ; Confidence interval for $\sigma^2$: $\left[(n-1)\cdot s^2/g_2;\ (n-1)\cdot s^2/g_1\right]$

From the chi-square table (α=0.05): g$_1$=2.7; g$_2$=19.023

$\sigma^2 \in \left[9\cdot 1072.22/19.023;\ 9\cdot 1072.22/2.7\right];\quad \sigma^2 \in \left[507.28;\ 3574\right]$

Given that $25^2 = 625 \in \left[507.28;\ 3574\right]$, there is not enough evidence to affirm that the standard deviation is different from 25. Therefore, it must be admitted that σ=25.

**2)** $s_{\left(\bar{x}_1-\bar{x}_2\right)} = \sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\cdot\sqrt{\dfrac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2}} = \sqrt{\dfrac{1}{8}+\dfrac{1}{10}}\cdot\sqrt{\dfrac{7\cdot 4.45+9\cdot 14.47}{8+10-2}} = 1.5065$

Confidence interval for the difference of means:

$m_1-m_2 \in \left[\left(\bar{x}_1-\bar{x}_2\right)\pm t_{n_1+n_2-2}^{\alpha/2}\cdot s_{\left(\bar{x}_1-\bar{x}_2\right)}\right];\ m_1-m_2 \in \left[(6.61-12.43)\pm t_{8+10-2}^{\alpha/2}\cdot 1.5065\right]$

If α=0.05, $t_{16}^{0.025}=2.12$: $m_1-m_2 \in \left[-5.82\pm 2.12\cdot 1.5065\right];\ m_1-m_2 \in \left[-2.63;\ -9.01\right]$

If α=0.01, $t_{16}^{0.005}=2.92$: $m_1-m_2 \in \left[-5.82\pm 2.92\cdot 1.5065\right]\ m_1-m_2 \in \left[-1.42;\ -10.22\right]$;

In both cases, $m_1-m_2<0 \Rightarrow m_1<m_2$. Thus, 2a) is false because the difference of means is statistically significant for α=0.01. Moreover, 2b) is true because the difference is statistically significant for (1-α)=0.95. Given that the null hypothesis m$_1$=m$_2$ is rejected for α=0.01, the p-value for the comparison of means is <0.01 and question 2c) is true.

**2d)** $\dfrac{\sigma_1^2}{\sigma_2^2} \in \left[\dfrac{s_1^2}{s_2^2\cdot f_2}, \dfrac{s_1^2}{s_2^2\cdot f_1}\right] = \left[\dfrac{4.45}{14.47\cdot 4.2}, \dfrac{4.45}{14.47\cdot 0.21}\right] = \left[0.073,\ 1.46\right]$

$1\in\left[0.073,\ 1.46\right]$ Thus, it can be assumed that $\sigma_1^2/\sigma_2^2=1$, which implies that the variances and standard deviations of both configurations do **not** differ significantly. As a result, 2d) is false.

**3a)** Total number of values = 3·3·2 = 18 (3 processors · 3 workloads · 2 replicates).
Total degrees of freedom (Df) = 18-1 = 17. Df (processor) = 3-1=2;
Df (workload) = 3-1 = 2. Df interaction = 2·2 = 4. Df residual = 17-2-2-4 = 9.
Mean square = Sum of squares / df; F-ratio = mean square / 84.94
Given that $52.09$ and $8.67 > (F_{2;9}^{0.05} = 4.26)$, then the effect of factors processor and workload is statistically significant.
Given that $4.90 > (F_{4;9}^{0.05} = 3.63)$, then the effect of the interaction is also significant.

```
Analysis of Variance for T_CPU - Type III Sums of Squares
------------------------------------------------------------------------
Source      Sum of Squares     Df      Mean Square    F-Ratio    P-Value
------------------------------------------------------------------------
MAIN EFFECTS
 A:PROCESSOR     515,111       2        257,555        52,09      <0,05
 B:WORKLOAD       85,7778      2         42,889         8,67      <0,05

INTERACTIONS
 AB               96,8889      4         24,222         4,90      <0,05

RESIDUAL          44,5         9         84,944
------------------------------------------------------------------------
TOTAL (CORRECTED) 742,277     17
------------------------------------------------------------------------
```

**3b)** The length of the thee intervals will be the same because the number of data available from each processor is the same. As the factor processor is significant, it implies that at least one of the intervals will not overlap. Thus, the interval for C will not overlap with the rest. Taking into account the plot in 3d), the mean T_CPU of processor C will be the mean of 10; 10.5 and 15: (10+10.5+15)/3 = **11.8.**

**3c)** If a straight line is drawn, it crosses rather accurately the center of the three intervals. Thus, the effect of factor workload is linear. Moreover, this effect is statistically significant because the first and third intervals do not overlap.

**3d)** The ANOVA table indicates that the interaction is statistically significant, which implies that the effect of workload will not be the same for the 3 processors. This issue is clearly reflected in the figure: the effect of workload on A and B is similar ($T\_CPU_{10} < T\_CPU_{20} < T\_CPU_{30}$), but it is different in the case of processor C: ($T\_CPU_{10} = T\_CPU_{30}$) < $T\_CPU_{20}$).
More precisely, the effect of workload on A and B seems to be quadratic because the increase of T_CPU between 10 and 20 is lower than the increase between 20 and 30. In the case of C there is also a quadratic effect but very different.

**4a)** The nature of the relationship is linear because a straight line fits the points and any indication of curvature (quadratic effect) is observed in the figure. The relationship can also be described as negative, because a higher power implies a lower number of reassignments. Moreover, the relationship is weak because the dispersion of points above and below the fitted line is considerable.

**4b)** $r = \mathrm{cov}/(s_x \cdot s_y) \Rightarrow$ cov=-0.7229·1.43172·0.450234 = -0.466

$$\begin{pmatrix} \mathrm{cov}_{x,x} & \mathrm{cov}_{x,y} \\ \mathrm{cov}_{y,x} & \mathrm{cov}_{y,y} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \mathrm{cov}_{x,y} \\ \mathrm{cov}_{x,y} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} 1.4317^2 & \mathrm{cov}_{x,y} \\ \mathrm{cov}_{x,y} & 0.4502^2 \end{pmatrix} = \begin{pmatrix} 2.050 & -0.466 \\ -0.466 & 0203 \end{pmatrix}$$

**5a)** $t_{statistic} = b_i / s_{b_i} \approx t_{N-1-I}$ Estimated intercept: $b_i = t_{statistic} \cdot s_{b_i} = 13.7823 \cdot 1.59084 = 21.925$

Slope: $t_{statistic} = 17.77 \approx t_{N-1-I} \approx t_{16-1-1} \approx t_{14}$. Considering α=0.05, $t_{14}^{0.025} = 2.14$

As 17.77>>2.14, the p-value of the slope is <0.05. Thus, the estimated slope can be regarded as different from zero at the population level, and the equation that should be used to predict the delivery time is: Time = 21.9254 + 2.0687 · N_extras

In this equation, the intercept is statistically significant because p-value=0,0000 which is below 0.05, and the slope is also significant as already mentioned.

**5b)** Global test for regression significance (p-value=0,000): $H_0 : \beta_1 = \beta_2 = ... = \beta_I = 0$.
In this case, there is only one variable in the model, and the coefficient of that variable is the slope. Thus, $H_0 : slope = 0$ ; $H_1 : slope \neq 0$.
The p-value of the model is below 0.01, which implies that the slope is significantly different from zero for α=0.01 and, hence, the correlation between both variables is statistically significant at the population level.

**5c)** The parameter that accounts for the percentage of variability of Y (delivery time) explained by X (number of extra accessories) is the coefficient of determination. The value of this parameter is: $R^2 = r^2 = 0.978545^2 = 0.9576$ =**95.76%**
(r is the correlation coefficient)

**5d)** If N_extras = 16, the estimated average time is:
Time = 21.9254 + 2.0687 · (N_extras=16) = **55.024**

**5e)** In a Normal distribution, the interval m±2σ comprises approximately 95% of the values. The mean square of residuals (9.2693) is the residual variance, that is, the variance of the conditional distribution of Y / X=x. Thus, when X=16, Time follows a Normal distribution N(m=55.024; $\sigma^2$=9.2693).
The interval comprising 95% of values is: $m \pm 2\sigma = 55.024 \pm 2\sqrt{9.2693} \Rightarrow$ **[50; 61]**