**Bachelor Degree in Computer Engineering**

## Statistics

# THIRD PARTIAL EXAM

May 30th 2011

| | |
|---|---|
| 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 note be allowed. Over the table it is only permitted to have the DNI (identification document), calculator, pen, and the formula tables.
5. Do not unstable any page of the exam (do not remove the staple).
6. All questions 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: 2 hours

**1.** One company wants to analyze the use of network in two departments A and B in order to optimize the Internet connection in this company. For this purpose, the number of computers connected daily to Internet is sampled in each department. Data are collected during 6 days in department A and during 8 days in department B, obtaining the following results:

| | | n | $\bar{x}$ | S |
|---|---|---|---|---|
| Department | A | 6 | 20 | 1.67 |
| | B | 8 | 20.75 | 3.34 |

Considering a confidence level of 95%, indicate if the following sentences are true or false, justifying conveniently the reply:

**a)** A different number of computers is connected, in average, in departments A and B.

**b)** It can be admitted that the number of computers connected, in average, is the same in the two departments.

**c)** As the number of days sampled is different, it is not possible to compare the number of computers connected in the two departments.

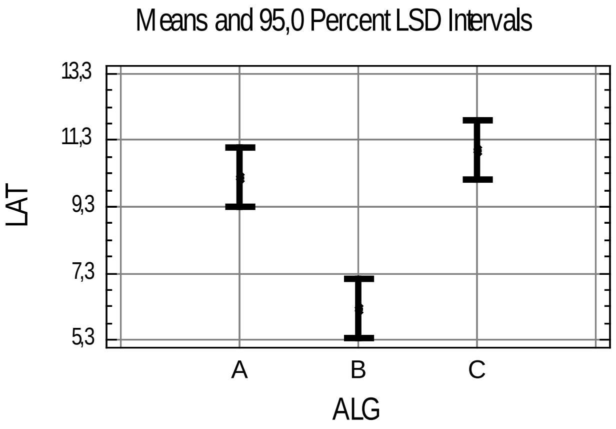
**2.** One study has been carried out with different algorithms of message routing in direct networks. The experiment has been designed assaying three different algorithms: A, B and C (factor ALG) that were combined with three multiplexing levels of physical channels: 4, 5 and 6 virtual channels (factor NCV). Each one of the 9 treatments was assayed two times, measuring in each experimental trial the latency of sent messages, in nanoseconds (variable LAT). After conducting the experiment and collecting the data, one Analysis of Variance was applied, resulting the following Summary Table:

```
Analysis of Variance for LAT - Type III Sums of Squares
------------------------------------------------------------------------
Source               Sum of Squares   Df   Mean Square  F-Ratio  P-Value
------------------------------------------------------------------------
MAIN EFFECTS
 A:ALG                 77,7733        ___    ________    _____
 B:NCV                 _______        ___     41,4867    _____
INTERACTIONS
 AB                    _______        ___    ________    _____
RESIDUAL               16,56          ___    ________
------------------------------------------------------------------------
TOTAL (CORRECTED)      250,52         ___
------------------------------------------------------------------------
```

**a)** Complete the Summary Table of ANOVA, indicating what effects are statistically significant (α=0.05). Justify the reply and the calculations.

(3 points)

**b)** According to the following plot, indicate if the average message latency is significantly different between the algorithms (A versus B, A versus C and B versus C). Justify the conclusion obtained and indicate if it is consistent with conclusions obtained in the previous section. (2 points)



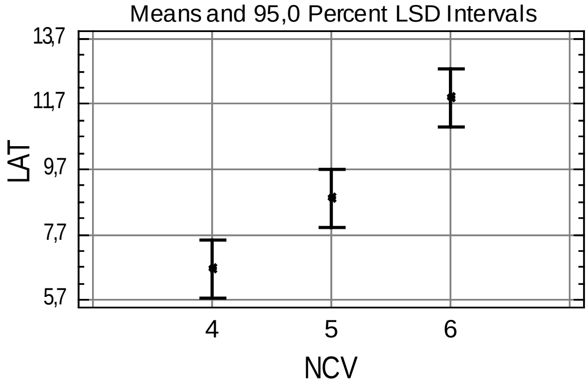
**c)** According to the following plot, study at descriptive level the effect nature of the number of virtual channels used in the physical channel multiplexation by answering the following questions: (2 points)

- Is the correlation coefficient between both variables positive, negative or close to zero? Why?

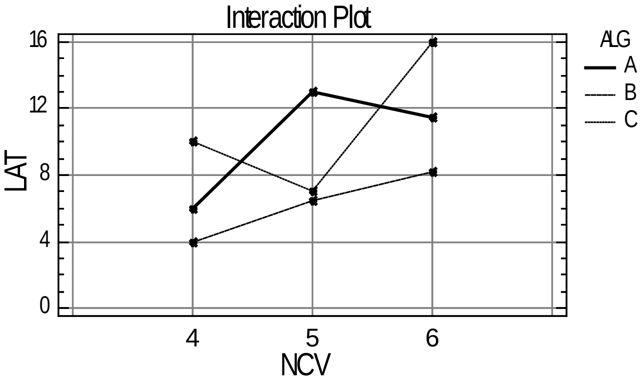
- Is the correlation between both variables statistically significant? Why?

- In order to predict LAT as a function of NCV, would you recommend a linear, or a quadratic model? Why?

- Would you recommend applying simple linear regression in order to study with more detail the linear and quadratic effect of NCV? Why?



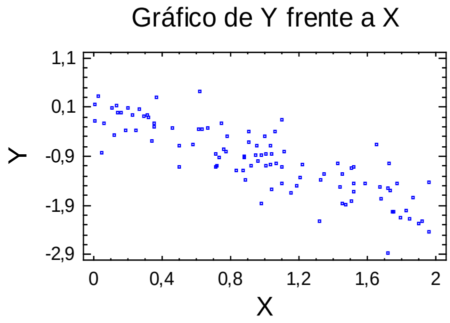
**d)** What additional useful information is provided by the following plot? Interpret its content by detailing the conclusions obtained. (2 points)



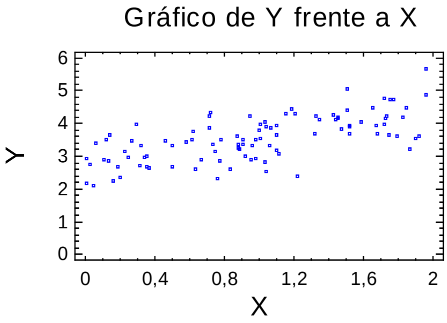
**e)** What would be the optimum treatment in order to minimize the average latency of messages routed through the network? (1 point)

**3.** Four scatterplots are shown below (1, 2, 3 and 4) as well as two Variance-Covariance matrices (A and B). Each one of these matrices corresponds to scatterplot 1 or scatterplot 2.

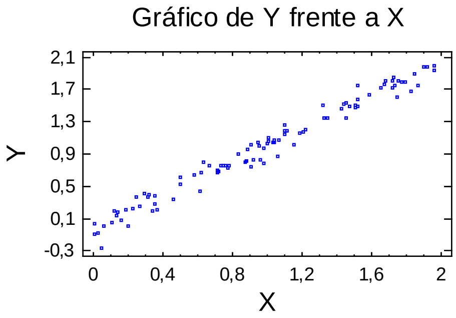
1



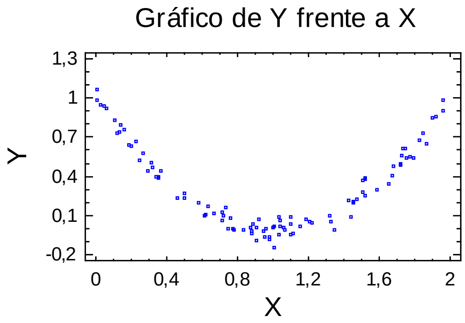
2



3



4



A

$$\begin{pmatrix} S^2_{XX} & S^2_{XY} \\ S^2_{YX} & S^2_{YY} \end{pmatrix} = \begin{pmatrix} 0{,}316815 & 0{,}262704 \\ 0{,}262704 & 0{,}482217 \end{pmatrix}$$

B

$$\begin{pmatrix} S^2_{XX} & S^2_{XY} \\ S^2_{YX} & S^2_{YY} \end{pmatrix} = \begin{pmatrix} 0{,}316815 & -0{,}335979 \\ -0{,}335979 & 0{,}503934 \end{pmatrix}$$

**a)** Answer and justify conveniently the following questions:

**a.1.** The exact value of the linear correlation coefficient in scatterplot 1 is: ________________________________________________________________

**a.2.** The exact value of the coefficient of determination in scatterplot 2 is: ________________________________________________________________

**b)** Focusing only on scatterplots 3 and 4, answer the following questions:

**b.1.** The approximate value of the linear correlation coefficient in scatterplot 3 is: ______________________________________________________________

**b.2.** The approximate value of the linear correlation coefficient in scatterplot 4 is: ______________________________________________________________

**4.** The following results have been obtained in one study about the relationship between the load of a computer system and the response time:

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------
Dependent variable: RES_TIME
Independent variable: LOAD
-----------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------
Intercept       0,888424       0,192557        4,61382         0,0007
Slope           0,362455       0,030144       12,0241          0,0000
-----------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------
Source         Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------
Model                 16,2784      1      16,2784     144,58       0,0000
Residual               1,2385     11      0,112591
-----------------------------------------------------------------------
Total (Corr.)         17,5169     12

Correlation Coefficient = 0,964
R-squared = 92,9297 percent
Standard Error of Est. = 0,335546
```

**a)** What is the equation of the estimated regression line? Are the parameters of this equation statistically significant? Consider $\alpha = 5\%$.

**b)** What information provides the ANOVA of this model?

**c)** Obtain the value of the coefficient of determination. What is the interpretation of this coefficient?

**d)** Calculate the probability to obtain a response time higher than 2 when the system load is equal to 4.