Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 15th 2012

Surname, name:

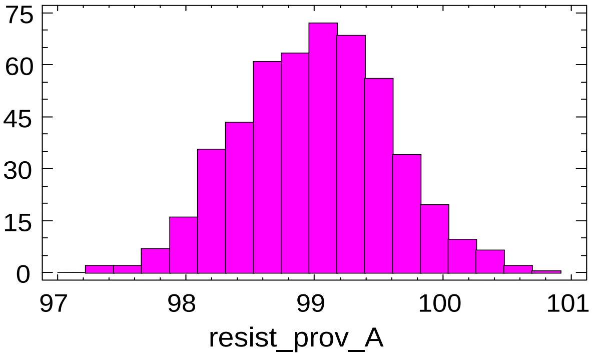
Group: **1E**	Signature:

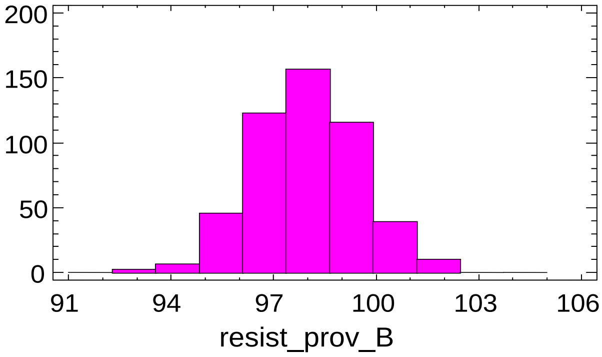
Indicate with a tick mark the partials examined 1st □ 2nd □

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 each partial. 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 company that manufactures computer equipment uses certain electronic component with a resistance of 100 ohms, which can be purchased from two different suppliers (A or B). In order to study the differences in the resistance of components sold by each supplier, a sample of 500 components is taken from supplier A and another sample of 500 components from B. After measuring the resistance of these components, the following histograms were obtained:





According to these histograms, answer the following questions justifying conveniently the replies.

a) What does the vertical scale indicate? Why is it so different in the two cases? (**2.5 points**)

b) In which of the two suppliers is there a higher dispersion in the values of resistance? Why? (**2.5 points**)

c) Do you think that a histogram is an adequate technique to detect outliers? What other techniques would you use? (**2.5 points**)

d) Indicate what parameters are the most appropriate in this case to quantify the data position of resistance. Calculate approximately such parameters for each supplier. (**2.5 points**)

2. (1^{st} Partial) One cybernaut has created a blog and has found that it receives 5 visits on average every day.

a) What is the random variable under study? What is the type of distribution of this variable? (**2 points**)

b) Calculate the probability to receive daily at least 2 visits. (**3 points**)

c) Calculate the probability to receive in one week at least 40 visits (use the approximation to the Normal distribution. (**5 points**)

3. (1^{st} Partial) One device is comprised by three identical components that operate independently. The time of operation until failure of these components follows an exponential distribution, and their reliability after 120 hours of operation is 80%.

a) Calculate the reliability of one component after 400 hours. **(4 points)**

b) The device operates correctly if at least two of the three components are operative. Calculate the reliability of the device after 400 hours. **(6 points)**

4. (2nd Partial) Hash tables are data structures with associated keys that allow the search of elements with low computational costs without the need of ordering the elements previously when being introduced. A hash function is available to achieve an efficient search, whose mission is to indicate the position of the searched element through a key.
One programmer proposes a hash function and pretends to evaluate it. For this purpose, the programmer performs a random search of 30 elements in the hash table and measures the time (in ms) taken to find each item. Some statistical results of the study are the following:

```
Average = 23,17
Standard deviation = 5,72
Stnd. Skewness = 1,01
Stnd. kurtosis = 0,13
```

a) Is it acceptable a population mean of 37 ms for the search time, considering a risk α=0.05? Answer the question using the *t-test*. (**3 points**)

b) Can we accept that the standard deviation of the search time at the population level is 3 ms, with a confidence level of 95%? (**3 points**)

A second programmer proposes another hash function. In order to evaluate it, a random search of 12 elements in the hash table is performed, measuring on the same machine as the previous programmer the time in ms taken to perform the search for each one of the 12 elements. The following results are obtained:

```
Average = 19,32
95% confidence interval for the difference of means: [-0,1516;7,8516]
```

c) Can we affirm with a significance level α=0.05 that the differences of mean search times between both hash functions are statistically significant? Are the differences statistically significant considering a confidence level of 99%? (**2 points**)

d) What is the practical interpretation of this Confidence Interval? (**2 points**)

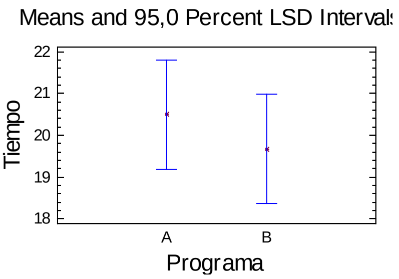
5. (2^{nd} Partial) One company of mathematical software wants to find out which of two programs for numerical integration works faster. For this purpose, an experiment is carried out with 12 functions to integrate: 4 ones of type 1, four of type 2 and four functions of type 3. For each function, the company measures the time (in ms) required by the program to run the integration procedure, using program A in 6 functions and B in the remaining 6 ones.

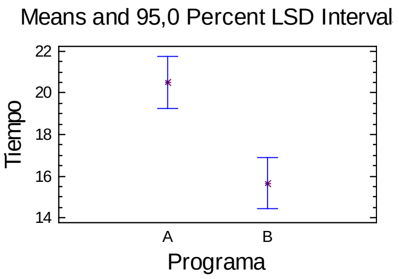
a) Complete the summary table of ANOVA and indicate which effects are statistically significant (α=0.05). Justify all calculations. (**3 points**)

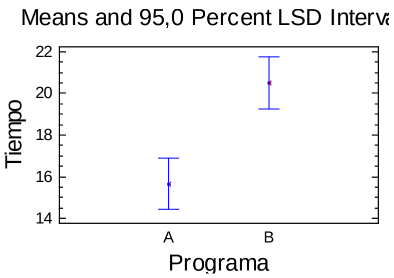
```
Analysis of Variance for Time - Type III Sums of Squares
----------------------------------------------------------------------
Source          Sum of Squares   Df   Mean Square    F-Ratio    P-Value
----------------------------------------------------------------------
MAIN EFFECTS
 A:Function_type     555,167            277,5830      90,03
 B:Program                               70,0833      22,73     0,0031

INTERACTIONS
 AB                   41,1667            20,5833

RESIDUAL                                  3,0833
----------------------------------------------------------------------
TOTAL (CORRECTED) 684,917
----------------------------------------------------------------------
```

b) Taking into account the information provided by the summary table of ANOVA, what kind of additional information would be obtained from the graphical representation of LSD intervals? (**2 points**)

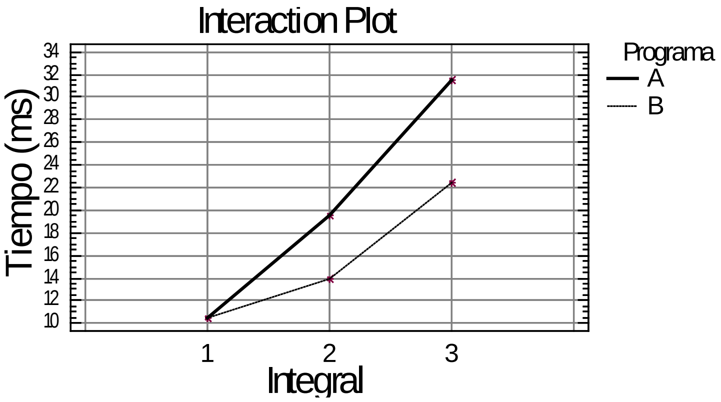
c) Indicate which of the following plots is the correct one in this case, taking into account the plot shown in section **d)**. Justify your answer. (**2.5 points**)



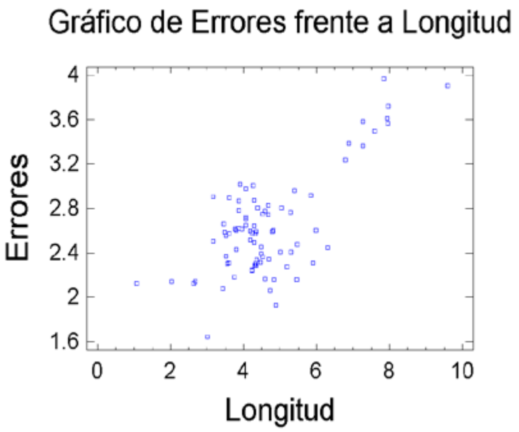




d) Given the following plot, describe the effect of both factors on the average time taken to run the integration procedure. **(2.5 points)**



6. (2^{nd} Partial) Two computer systems are connected through a high performance port that undergoes interferences and noise problems due to different reasons. In order to investigate this problem, a study is performed to assess the relationship between the average length of bit frames between the two systems (variable *length*) and the number of errors detected during the transmission of certain test files (variable *errors*). Data of both variables corresponding to the transmission of 82 large files between the two systems were compiled, and the following results were obtained:



```
Summary statistics
                        Length     Errors
----------------------------------------
Count                   82         82
Mean                    4.65893    2.63361
Median                  4.332      2.5896
Variance                2.04982    0.202711
Standard deviation      1.43172    0.450234
Minimum                 1.072      1.6458
Maximum                 9.592      3.965
Range                   8.52       2.3192
First quartile          3.864      2.3062
Third quartile          5.032      2.8041
Interq. Range           1.168      0.4979
Stnd. skewness          3.87271    3.65235
Stnd. kurtosis          3.38387    2.05077
-----------------------------------------
Correlation coefficient = 0.722933
```

a) According to the plot shown above, describe the nature of the relationship between the two variables under study. (**1 point**)

b) Calculate the parameters of the regression line corresponding to the relationship between the two variables under study, and obtain the mathematical equation of the simple regression model. (**3 points**)

c) Calculate the expected number of errors for a length of bit frame = 6. (**1 pt.**)

Attempting to further investigate the results, a regression analysis was performed and the following results were obtained with Statgraphics:

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Errors
Independent variable: Length
-----------------------------------------------------------------------------
                                   Standard          T
Parameter      Estimate              Error        Statistic         P-value
-----------------------------------------------------------------------------
Intercept                          0.118335        13.3049
Slope                              0.0242919       9.35872          0.0000
-----------------------------------------------------------------------------
Analysis of variance
-----------------------------------------------------------------------------
Source            Sum of Squares     Df  Mean Square    F-ratio     P-value
-----------------------------------------------------------------------------
Model                    8.58140      1      8.58140      87.59
Residual                 7.83818     80      0.09797
-----------------------------------------------------------------------------
Total (Corr.)            16.4196     81
```

d) Is the proposed model significant according to the global test? Are the parameters of this model statistically significant? Use a significance level α=0.01. **(3 points)**

e) What is the interpretation of the Residual Mean Square in the ANOVA test of the regression model? What is this parameter used for? **(2 points)**

SOLUTION OF THE FINAL EXAM 15/06/2012

Question 1:

a) The vertical scale is absolute frequency, i.e., number of data contained in each interval of the hystogram. This scale is much bigger in the histogram of supplier B because it contains mess intervals (i.e., a lower number of bars). Taking into account that both histograms were built with 500 data, when dividing the range of variation of resistance by a lower number of intervals, it turns out a higher number of data in each interval, and as a result the absolute frequency becomes higher.

b) Range of A ≈ 101 – 97 = 4 ohms

Range of B ≈ 102.5 – 92.5 = 10 ohms

Given that the ranges are so different and taking into account that in both cases data follow approximately a Normal distribution (as the hystogram resembles a Gauss function), supplier B presents a bigger variability than supplier A (i.e., bigger standard deviation, variance and interquartile range).

c) Generally speaking, a histogram is not a convenient tool to detect outliers because a single extreme value will result in a bar with a small height that can easily be unnoticed except if the total number of values is rather small. In order to detect outliers it would be more convenient to use a box-whisker plot or a normal probability plot.

d) Both histograms are rather symmetric and resemble a Gauss function, which suggests a Normal distribution of the data. In such cases, the axis of symmetry of the histogram corresponds to the average and the median, which are the most convenient parameters of position. Thus, it can be deduced from the plot that the average of supplier A is approximately 99 ohms, and 98 ohms in the case of B.

Question 2:

a) The random variable under study is X: number of visits to the blog in one day. This is a discrete variable. The minimum value is zero and there is not a maximum value, which suggests a Poisson distribution: $X \sim Ps(\lambda)$. In a Poisson distribution, λ is the average, which is 5 visits/day in this case. Thus, $X \sim Ps\ (\lambda=5)$.

b) $P(X\geq2)=1 - P(X=0) - P(X=1) = 1 - e^{-5}\cdot(1+5) =$ **0.9596**

Calculation with the abacus: $P(X\geq2)=1-P(X\leq1)= 1 - 0.04= 0.96$

c) Y={number of visits to the blog during one week (7 days)}. $Y=X_1+X_2+...+X_7$

As Y is the sum of 7 Poisson variables (one corresponding to each day), it turns out that Y will follow a Poisson distribution with $\lambda=7\cdot5=35$. As $\lambda>9$, Y can be approximated to a Normal distribution with average =35 and $\sigma=(35)^{½}=5.92$. Thus, applying the correction of continuity:

$P(Y\geq40)=P[N(35; 5.92)>39.5]=P[N(0;1)>\frac{39,5-35}{5,92}]=P[N(0;1)>0.76] =$ **0.224**

Question 3:

a) Random variable X_i: {time of operation until failure of *i*-th component (in hours)}

$X_i \sim \text{Exp}(\alpha)$ The reliability of *i*-th component after 120 hours is 80% →

→ $P(X_i \geq 120) = 0.8 = e^{-120\alpha}$ → $\alpha = 0.00186$

Reliability of this component after 400 hours: $P(X_i \geq 400) = e^{-400\cdot 0.00186}$ = **0.475**

b) Random variable Y: number of components operating correctly in the device after 400 hours. Y ~ Bi (n=3, p=0.475).

P (the device operates after 400 h) = $P(Y\geq 2) = P(Y=2)+P(Y=3) =$

$$= \binom{3}{2}\cdot 0.475^2\cdot(1-0.475)+\binom{3}{3}\cdot 0.475^3\cdot(1-0.475)^0 = 3\cdot 0.487^2\cdot(1\text{-}0.487)+0.487^3=\mathbf{0.463}$$

Question 4:

a) H_0: m=37; H_1: m≠37. Accept H_0 if $t_{calc} > t_{29}^{0.025}$

$$t_{calc} = \left|\frac{23.17-37}{5.72/\sqrt{30}}\right| = 13.24 \ > 2.045 \ \longrightarrow$$ H_0 should be rejected: it is not acceptable m=37.

b) H_0: σ = 3; H_1: σ ≠ 3 $\sigma^2 \in \left[(n-1)\cdot s^2/g_2\ ;(n-1)\cdot s^2/g_1\right]$

From the χ^2_{29} table, considering α=0.05 we obtain that g_1=16.047 and g_2=45.722.

$\sigma^2 \in \left[29\cdot 5.72^2/45.722;\ 29\cdot 5.72^2/45.722\right]$; $\sigma^2 \in \left[20.75;\ 59.12\right]$; $\sigma \in \left[4.55;\ 7.69\right]$

Given that the value 3 is not comprised within this interval, H_0 should be rejected. Thus, it is not acceptable that σ = 3.

c) H_0: $m_1 = m_2$; H_1: $m_1 \neq m_2$ $m_1 - m_2 \in \left[-0.1516;\ 7.85\right]$ considering α=0.05.

$0 \in \left[-0.1516;\ 7.85\right] \Rightarrow m_1 - m_2 = 0 \Rightarrow m_1 = m_2$

Thus, with α=0.05, we accept H_0 : the differences of mean search times between both hash functions are **not** statistically significant.

$IC_{m_1-m_2} \Rightarrow \left(\bar{x}_1-\bar{x}_2\right)\pm t^{\alpha/2}_{n_1+n_2-2}\cdot S_{\left(\bar{x}_1-\bar{x}_2\right)}$ For α=0.05 $t^{0.025}_{40} = 2.021$; For α=0.01 $t^{0.005}_{40} = 2.704$;

For α=0.01 the interval becomes wider and will also comprise the value zero. Thus, H_0 is also accepted: the differences are **not** statistically significant.

d) Being m_1 and m_2 the search time means at the population level of the hash function proposed by the first and second programmers, respectively, if an infinite number of samples are taken from both populations and a confidence level for (m_1-m_2) is calculated from the sample parameters, it turns out that the real value of (m_1-m_2) will be comprised within the estimated interval in 95% of the cases. Thus, if this interval contains the value zero it can be deduced with a type I risk of 5% that 0=m_1-m_2 and, hence, the null hypothesis m_1=m_2 should be accepted.

Question 5:

a) The complete summary table of the ANOVA is the following:

```
Analysis of Variance for Time - Type III Sums of Squares
----------------------------------------------------------------------
Source          Sum of Squares   Df   Mean Square    F-Ratio    P-Value
----------------------------------------------------------------------
MAIN EFFECTS
 A:Function_type     555,167      2     277,5830      90,03      <0,05
 B:Program            70,0833     1      70,0833      22,73       0,0031

INTERACTIONS
 AB                   41,1667     2      20,5833       6,676     <0,05

RESIDUAL              18,4998     6       3,0833
----------------------------------------------------------------------
TOTAL (CORRECTED) 684,917        11
----------------------------------------------------------------------
```

$\mathrm{Df}_{\mathrm{funct}}$= (3 variants)-1 = 2; $\mathrm{Df}_{\mathrm{prog}}$= (2 vari.)-1 = 1; $\mathrm{Df}_{\mathrm{AB}} = \mathrm{Df}_{\mathrm{A}} \cdot \mathrm{Df}_{\mathrm{B}} = 2\cdot1=2$.

$\mathrm{Df}_{\mathrm{total}}$ = 12 values - 1 = 11; $\mathrm{Df}_{\mathrm{residual}} = 11\text{-}2\text{-}1\text{-}2 = 6$.

$\mathrm{SS}_{\mathrm{program}} / 1 = 70.0833 \rightarrow \mathrm{SS}_{\mathrm{program}} = 70.0833$.

$\mathrm{SS}_{\mathrm{residual}} / 6 = 3.0833 \rightarrow \mathrm{SS}_{\mathrm{residual}} = 3.0833\cdot6= 18.4998$

$\text{F-ratio}_{\mathrm{AB}} = \mathrm{MS}_{\mathrm{AB}} / \mathrm{MS}_{\mathrm{resid}} = 20.5833/3.0833 = 6.676$

Factor program is statistically significant because p-value=0.0031 < 0.05

Factor function is also significant because F-ratio = 90.03 >> $\left(F_{2;6}^{0.05}=5.14\right)$

The interaction is statistically significant because F-ratio = 6.68 > $\left(F_{2;6}^{0.05}=5.14\right)$

Thus, the three effects are statistically significant considering α=0.05.

b) The summary table of ANOVA indicates if the simple effect of one factor is statistically significant. However, if that factor has more than two levels or variants, it is not possible to know which variant is significantly different from the rest. In such case, LSD intervals are used to study the differences among levels. This information is useful for the interpretation of results.

In this case, factor program is statistically significant and it has two variants. Consequently, the LSD intervals will not overlap and the only information provided by such intervals is the average value of each variant. By contrast, factor function has three variants and the LSD intervals will show the mean values and will reflect if the differences between variants are statistically significant.

c) As factor program is statistically significant, the LSD intervals will not overlap. Consequently, the left figure is discarded. According to the interaction plot in d), the average time with A is higher than with B, which implies that the central figure is the correct one. As a further check, the average time for each program according to d) are:

Average time A = (10+19.5+31.5)/3= 20.3

Average time B = (10+14+22.5)/3= 15.5

These average values correspond to the central figure.

d) The interaction is statistically significant according to the ANOVA table, which implies that the effect of factor program is not the same for all variants of the other factor. Actually, according to the plot, the average time is the same for both programs with function type "1", but not for the other types. Regarding types "2" and "3", program A takes more time than B. On average, the average time for function type "1" is lower than for "2" and "3" (i.e., $\text{time}_1 < \text{time}_2 < \text{time}_3$).

[Note: it is not correct to say that the effect of function is linear or quadratic, because this factor is qualitative.]

<u>Question 6:</u>

a) Both variables are positively correlated because the number of errors tends to increase for greater values of length (r>0). A straight line fits properly the data, which implies that the relationship is linear. No quadratic trend is observed in the graph. The correlation is not strong because $R^2=0.5$.

b) $b = r_{xy} \cdot S_y/S_x = 0.722933 \cdot 0.450234/1.43172 = 0.2273$

$a = \bar{y} - b \cdot \bar{x} = 2.63361 - 0.227341 \cdot 4.65893 = 1.5744$

The equation of the regression model would be: Errors = 1.5744 + 0.2273·length

c) $E(errors/length = 6) = 1.5744 + 0.2273 \cdot 6 =$ **2.938**

d) According to the global test: $MS_{\mathrm{mod}el}/MS_{residual} \approx F_{I;(N-1-I)} \approx F_{1;80}$ In this case: $8.58/0.098 = 87.59 >> \left(F_{1;80}^{0.01} = 6.96\right)$. Thus, the model is significant at $\alpha=0.01$.

The model has 2 parameters: *a* and *b*. The slope is statistically significant because the p-value associated to this parameter is < 0.01. The intercept is also significant because $(t_{statistic} = b_i/s_{b_i}) > t_{N-1-I}^{\alpha/2} \rightarrow 13.305 > (t_{80}^{0.005} = 2.6)$.

e) The Residual Mean Square is the variance of the residuals, also called residual variance. Residual is the difference of observed minus predicted Y values. In regression, the residual mean square is used to obtain a confidence interval for the prediction. It is also used in the global significance test given that $MS_{\mathrm{mod}el}/MS_{residual} \approx F_{I;(N-1-I)}$.