Bachelor Degree in Computer Engineering

Statistics **group E (English)**

SECOND PARTIAL EXAM

May 29th 2019

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. One industry manufactures certain model of capacitors used in electronic circuits of computer equipment. The capacitance value is a quality parameter that is controlled periodically. The quality control department assumes that this parameter follows a Normal distribution of mean 20 nF and standard deviation 0.2 nF. It is assumed that the process performs correctly when the capacitance of one capacitor randomly selected is between 19.5 and 20.5 nF. Otherwise, it is considered that the process is out of statistical control and needs to be reviewed.

a) If a hypothesis test is applied to this case, what would be the interpretation of the type-I risk? Justify your answer. *(2 points)*

b) What is the most common range of values of the following parameters?

b.1) Type-I risk (justify your answer) *(1 point)*

b.2) p-value (justify your answer) *(1 point)*

c) The company decides to implement a change in the manufacturing process, which is suspected to affect the parameters of the statistical distribution model of capacitance. In order to study this issue, a sample of 15 capacitors is randomly taken, resulting an average capacitance value of 20.15 nF and a sample variance of 0.09 nF^2.

c.1) Based on this information and considering $\alpha = 0.05$, is there enough evidence to affirm that the change implemented in the process has altered the <u>average</u> capacitance at the population level? To answer this question, indicate firstly the hypothesis test that is considered in this case, solve this test and interpret the results in this context, justifying your answer. *(3 points)*

c.2) Can we assume a standard deviation at the population of 0.2? Answer this question by using the confidence interval method, considering a confidence level of 95%. *(3 points)*

2. A public library has six computers for users' queries, three located in room A and another three in room B. Computers differ, among other things, in the processor speed, which can be 2, 3 or 4 GHz. A random sample is taken of the consultation time of 24 users: 8 of them use a 2 GHz computer, another 8 one of 3 GHz and another 8 one of 4 GHz. Half of the users perform the query in room A, and the other half in room B. The values of time obtained (measured in minutes) are indicated below, as well as the summary table of the ANOVA.

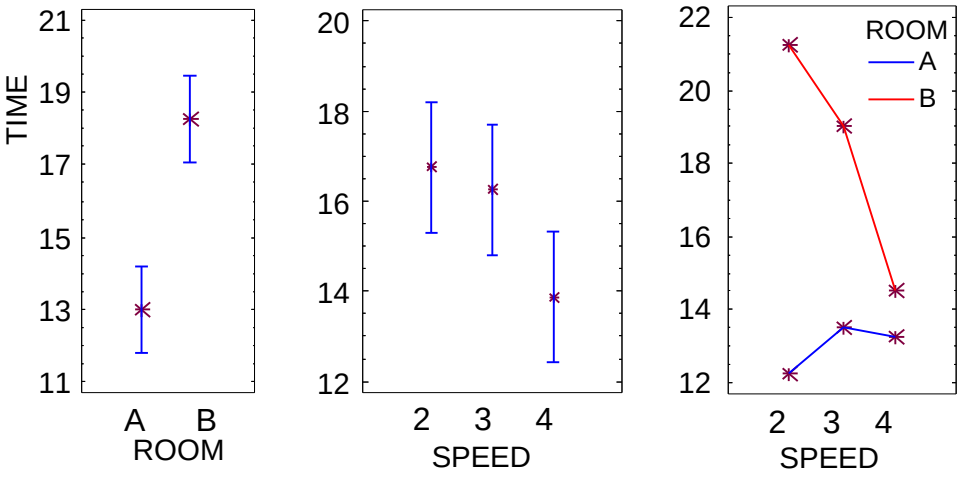
Room A	
speed	time
2 GHz	9; 12; 13; 15
3 GHz	11; 13; 16; 14
4 GHz	14; 16; 12; 11

Room B	
speed	time
2 GHz	23; 25; 20; 17
3 GHz	23; 15; 21; 17
4 GHz	14; 17; 15; 12

Analysis of Variance for TIME - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:SPEED	37,75	2	18,875	2,49	0,1107
B:ROOM	165,375	1	165,375	21,85	0,0002
INTERACTIONS					
AB	60,25	2	30,125	3,98	0,0370
RESIDUAL	136,25	18	7,56944		
TOTAL (CORRECTED)	399,625	23			

The means plot with 95% LSD intervals for the two factors are shown below as well as the interaction plot.

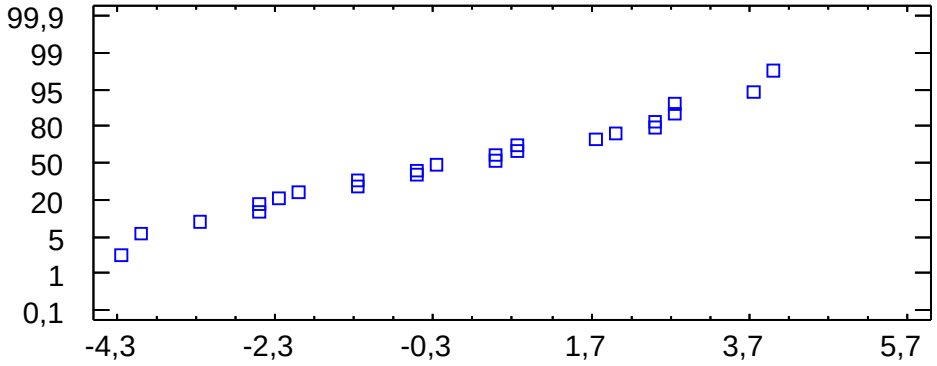


a) With the computer equipment available, the library considers that, at the population level, the consultation time does not depend on the processor speed. Do you agree? Justify conveniently your answer with the appropriate statistical method(s) and considering $\alpha = 5\%$. *(2 points)*

b) Taking into account what is deduced from the two means plots with LSD intervals, what additional information is provided in this case by the interaction plot to describe how factors "room" and "processor speed" affect the consultation time, at the population level? Consider α = 1%. *(2.5 points)*

c) If a new user wants to choose the room and computer where the consultation times are expected to be the lowest, what would you recommend, considering α = 1%? What would be the estimated average time under those conditions? *(2.5 points)*

d) The following plot has been obtained with the residuals of the model:

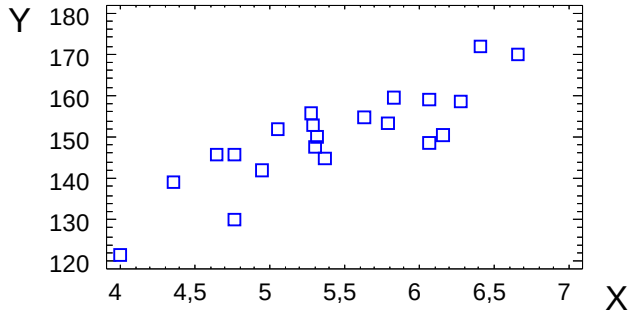


d.1) What hypothesis of ANOVA is usually checked using this plot? Do you consider that this hypothesis is reasonably satisfied in this case? Justify your answer. *(1 point)*

d.2) What would you recommend doing in case this hypothesis is not satisfied? *(1 point)*

d.3) In addition to this hypothesis, what other assumptions must the data fulfill so that the conclusions derived from ANOVA can be regarded as reliable? *(1 point)*

3. Certain manufacturer of breakfast cereals wants to establish the relationship that allows predicting sales based on expenses on child advertising on television (both variables, measured in thousands of euros). For this purpose, one study has collected the monthly data corresponding to the last 21 months. The following plot has been obtained based on these data:



a) What would be the independent variable X and the dependent variable Y in this context? What is deducted from the plot at the sample level? Justify conveniently your answer. *(2 points)*

b) Using *Statgraphics*, the following table of results was obtained corresponding to a linear regression model fitted with the data:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        75,0224       10,7008        7,01093         0,0000
Slope            13,8411        1,9562         7,0755         0,0000
------------------------------------------------------------------------------

                          Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio       P-Value
------------------------------------------------------------------------------
Model                      1919,29      1      1919,29      50,06        0,0000
Residual                   728,415     19      38,3376
------------------------------------------------------------------------------
Total (Corr.)               2647,7     20
```

What percentage of the variability in monthly sales is explained by other variables apart from the monthly expenses on advertising, for example: price, level of sales and price of competitor products, etc.? *(2.5 points)*

c) Write the mathematical equation of the proposed regression model, estimating the value of its parameters and their statistical significance (α = 5%). What is the practical interpretation of the values obtained for the estimated parameters, within the context of the problem? *(2.5 points)*

d) Assuming that all hypotheses of a simple linear regression model are satisfied, what is the probability of obtaining monthly sales greater than 140,000 euros if the monthly expenses on child TV advertising are 4,000 euros? What is the statistical model of conditional distribution that should be considered to calculate this probability? *(3 points)*

SOLUTION

1a) The type I risk (α) is defined as the probability to reject the null hypothesis when it is true. In this case, the null hypothesis is that the process operates correctly, being the capacitance (C) a random variable with a Normal distribution, being mean = 20 and σ = 0.2 → H_0: C ≈ N(20; 0.2). This hypothesis is rejected when C<19.5 or if C>20.5. Therefore, in this case:

α = P[N(20; 0.2) < 19.5] + P[N(20; 0.2) > 20.5] = 2 · P[N(20; 0.2) >20.5] =

= 2 · P[N(0; 1) > 0.5/0.2] = 2 · 0.0062 = **0.0124**

The resulting value is small, which makes sense (generally, $\alpha \leq 5\%$).

1b1) The type I risk (α) is a value that has to be decided when having to interpret the statistical results. As it is the probability to make an error, this probability is established as a small value. The most frequent values are 5% and 1%, and occasionally (in case of having few values) it could be admitted up to 10%, but never greater. On the contrary, if the amount of data is very big, we could consider lower values, like 0.1%. In summary, the most frequent range is: **$0.001 \leq \alpha \leq 5\%$**.

1b2) The *p*-value, also called "observed significance level", is a probability and, hence, it ranges from 0 to 1. It is defined as the smallest significance level (α) that can be considered, so that the null hypothesis would be rejected, according to the observed values. If *p*-value < α, then H_0 is rejected, while it is accepted if *p*-value > α. In practice, we can find any *p*-value between 0 and 1, it is not possible to established *a priori* a range of frequent values, because it is quite often to find differences which are not statistically significant. However, when working with very big samples (thousands of values), the differences tend to be statistically significant and, hence, in that case it is more likely to find *p*-values < 0.05.

1c1) Hypothesis test: H_0: m = 20; H_1: m ≠ 20.

Computed t-statistic: $t_{calc} = \frac{\bar{x}-m_0}{s/\sqrt{n}} = \frac{20.15-20}{\sqrt{0.09}/\sqrt{15}} = 1.936$

This parameter follows a Student's t distribution with 14 degrees of freedom. 95% of values from this distribution are comprised between -2.145 and 2.145. As t_{calc} is inside this interval, we can accept H_0; i.e., there is not enough evidence, according to the sample obtained, to affirm that the change in the process has altered the average capacitance at the population level.

1c2) Confidence interval for the population standard deviation (α=5%):

$$\sigma \in \left[\sqrt{\frac{(n-1)\cdot s^2}{g_2}}; \sqrt{\frac{(n-1)\cdot s^2}{g_1}}\right] = \left[\sqrt{\frac{14\cdot 0.09}{26.119}}; \sqrt{\frac{14\cdot 0.09}{5.629}}\right] = [0.219; 0.473]$$

Being g_1 (5.629) and g_2 (26.119) the interval of a chi-squared distribution with 14 degrees of freedom that comprises 95% of the data.

H_0: σ = 0.2 ; H_1: $\sigma \neq 0.2$; As the value 0.2 is outside the confidence interval obtained, the null hypothesis is rejected, **it cannot be admitted that σ = 0.2**.

2a) As the *p*-value of the interaction is 0.037, less than α =0.05, it can be concluded that the effect of the interaction is not statistically significant. It implies that the effect of processor speed on the time is statistically different for each room. The interaction plot shows that the differences on mean time for the three speeds assayed are very small in room A, while a linear decreasing effect is observed for foom B. Therefore, the library is not right: it cannot be admitted that, at the population level, the time does not depend on processor speed. Actually, time depends on speed in room B, because a lower average time is expected when the speed increases. However, this effect is not observed in room A.

2b) From the first plot it is deduced that the consultation time in room B is significantly greater than in room A. The ANOVA table indicates that the simple effect of factor speed is not statistically significant (p=0.11 > 0.01), which is consistent with the second plot because the three LSD intervals appear overlapped. Considering α=1%, the effect of the interaction is not statistically significant because p=0.037 > 0.01. Therefore, although certain differences have been observed with the samples corresponding to the two rooms for the different speeds, there is not enough evidence, based on the ANOVA results, to affirm that these differences also correspond to the population level. Hence, the effect of processor speed on the time should be considered equal in both rooms (i.e., as if the lines were parallel in the interaction plot at the population level).

The ANOVA reveals that only the simple effect of factor room is statistically significant; this conclusion is consistent with the two plots of means with LSD intervals, so that the interaction plot does not provide any additional information at the population level.

However, if data were analyzed using multiple linear regression, it is possible that the effect of the interaction might appear as statistically significant considering α = 1%, so that the interpretation of results would be different.

2c) As this question also considers α=1%, we start from the reasoning of the previous question. We expect lower times at the population level in room A. Since the interaction is not statistically significant nor the simple effect of processor speed, there is not enough evidence to affirm that a higher speed will lead to lower times on average at the population level. Thus, the recommendation is to use any computer in room A. In those conditions, the estimated average time will be the mean of all times measured in room A, which is 13 minutes according to the means plot.

2d1) This figure shows a normal probability plot of the model residuals. This kind of plot is commonly used to verify the hypothesis of normality, because ANOVA assumes that all the populations involved in the study follow a normal model of statistical distribution. In this case, the points fit "reasonably well" to a straight line, so that it can be admitted a normal distribution of residuals: there is not enough evidence to reject the hypothesis of normality. On the other hand, no outliers are detected that should be discarded.

2d2) It might turn out that the hypothesis of normality is not accomplished by different reasons: (**a**) If residuals fit reasonably well to a straight line in the normal probability plot but some few points are clearly separated from the line, such points are considered as outliers (abnormal values) that should be removed, or corrected if it is possible to identify the reason for abnormality.
(**b**) If the pattern of residuals suggests a skewed distribution, it is recommended to transform the original values. In case of a positively skewed distribution, the most common transformations are: logarithm or square root. In case of a negatively skewed distribution, it is necessary to transform firstly in order to achieve a positive skew (for example, changing the sign and adding an appropriate constant).

(**c**) The hypothesis of normality cannot always be achieved, for example when residuals reveal a mixture of distributions, or the presence of truncated data, etc.

2d3) In addition to the hypothesis of normality, ANOVA assumes the hypothesis of homoscedasticity (i.e., the different populations involved in the study have the same variance), and the hypothesis of independence (i.e., the different experimental observations are independent of each other, that is, they have been obtained randomly so that each observation has the same probability of appearing in the sample).

3a) Taking into account that monthly sales depend on expenses on child TV advertising, it turns out that the dependent variable Y will be "monthly sales of the company" while the independent variable X will be "expenses on child TV advertising".
Based on the sample, it can be observed a direct relationship between both variables (positive correlation): when the expenses on child TV advertising are higher, it is expected to obtain a higher amount of sales on average, which makes sense. The degree of correlation could be described as "moderate". It can be observed a linear effect that, quite probably, would turn out to be statistically significant if data were analyzed by means of simple linear regression.

3b) The percentage of variability of Y explained by X is computed as SS_{model} / SS_{total}. Therefore, el percentage of variability of Y **not** explained by X will be: $SS_{residual}$ / SS_{total} = 728.415 / 2647.7 = 0.2751 = **27.51%**.

3c) The mathematical equation of the regression model proposed is:

$$\text{Monthly sales} = 75.022 + 13.8411 \cdot \text{expenses}_{\text{advertising}}$$

The estimated value of both parameters (slope and ordinate) appears on the table of results from Statgraphics. For both parameters, their *p*-value is nearly zero (less than α) which implies that both are statistically significant; i.e., both can be regarded as different from zero at the population level.

- Practical interpretation of the ordinate: it is the estimated value of monthly sales that would be expected on average if the company does not spend on children's TV advertising. However, the value X = 0 is quite far away from the range of X values tested (from 4 to 6.5), so that the model prediction for X = 0 is not reliable.

- Practical interpretation of the slope: for each thousand euros that the company decides to increase the monthly expenses on child TV advertising, it is expected an average monthly sales increase of 13841.1 euros.

3d) Assuming that all hypotheses of the theoretical model are fulfilled, the distribution of sales (Y) conditioned to an expense of 4 thousand euros (i.e., X=4 since the units are in thousand euros), will be a normal distribution with average obtained from the regression model and with a variance equal to the residual variance:

$E(Y/X=4) = 75.022 + 13.8411 \cdot 4 = 130.3868$
$\sigma^2(Y/X=4) = s^2_{residual}$ = Mean Square of residuals = 38.3376 (from the table). By applying the square root: $s_{resid} = 6.1917$.
Therefore: $\mathbf{Y/X=4 \approx N\ [m=130.39,\ \sigma = 6.19]}$.
In such conditions, the probability to obtain monthly sales greater than 140 thousand euros:
$P[(Y>140)/(X=4)] = P[N(130.39; 6.19) > 140] =$
$= P[N(0;1) > (140-130.39)/6.19] = P[N(0; 1)> 1.5526] =$ **0.0603**