Bachelor Degree in Computer Engineering

Statistics **group E (English)**

SECOND PARTIAL EXAM

June 6th 2018

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. The company **AIRFLY** Inc. has an aircraft model A-340 for transoceanic routes. When this aircraft was acquired 5 years ago, the manufacturer indicated that the average fuel consumption was 6900 kg of fuel per hour (kg/h) with a standard deviation of 180 kg/h. After analyzing a total of 50 routes carried out by this aircraft, **AIRFLY** has estimated an average consumption of 7200 kg/h with a standard deviation of 200 kg/h.

a) The company considers that the fuel consumption of this aircraft is not the same, on average, as when it was purchased. Justify if the company is right by proposing and solving a hypothesis test using a type-I risk of 5%. Note: do not use a confidence interval to solve this question. *(4 points)*

b) Without making additional calculations, justify if the reply of the previous question would change by considering a type-I risk of 1%. *(2 points)*

c) Calculate a 95% confidence interval for the current population standard deviation. Taking into account the interval obtained, do you consider that the variability of fuel consumption in transoceanic routes has changed? *(4 points)*

2. Certain manufacturer of batteries for mobile devices produces two different types of units (A and B). In order to analyze the effect of temperature on the battery duration, an experiment was carried out by measuring the lifetime of both battery under 3 different temperature conditions (5 ºC, 20 ºC and 35 ºC). For each one of the 6 combinations (battery type and temperature), the lifetime of three batteries was measured. The results of the statistical analysis are shown below. Taking into account the information obtained, answer the following questions:

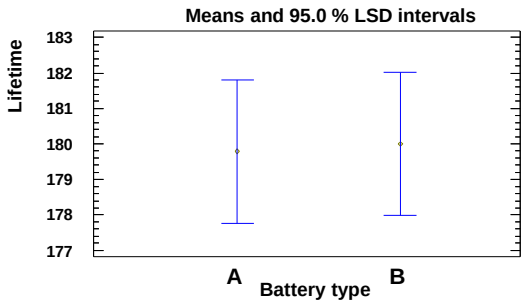
a) Fill in the Summary Table of ANOVA with those values that you may need to answer the remaining sections of this question. Justify your calculations.
(3 points)

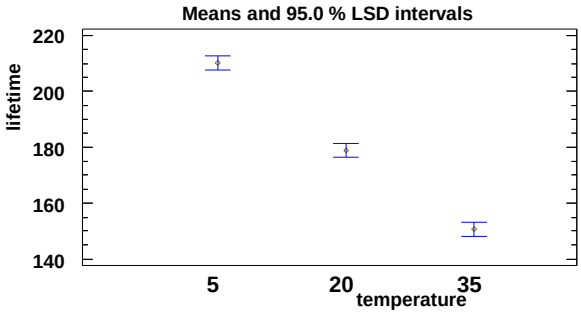
Analysis of Variance for Lifetime - Type III Sum of Squares

Source	*Sum of Squares*	*d.f.*	*Mean Square*	*F-ratio*	*P-value*
MAIN EFFECTS					
A: Battery type	0.222222				0.9067
B: Temperature	10630.8				0.0000
INTERACTIONS					
AB	1270.78				0.0000
RESIDUAL	186.0				
TOTAL (CORRECTED)	12087.8				

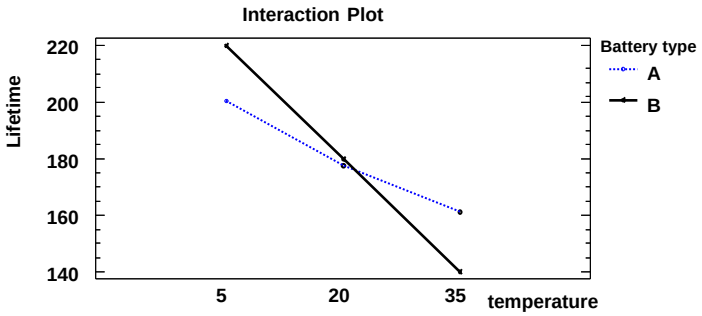
b) Study the statistical significance of the simple effects of both factors and their interaction (assume $\alpha = 5\%$ for significance). *(2 points)*

c) Interpret the plots of LSD intervals for the factors: battery type and temperature. Describe the nature of the effect of battery type and temperature. *(2 points)*





d) Interpret the interaction plot in consistency with the conclusions derived from the previous sections. Justify the reason why the interaction has been statistically significant or not in the ANOVA table. Taking into account this plot, what would you recommend to a possible customer to maximize the battery lifetime? *(3 points)*



3. The company **AIRFLY** Inc. has designed an experiment with 10 trials to measure the effect of a certain type of additive on the drying time of paint applied to their aircrafts. The following results were obtained:

	1	2	3	4	5	6	7	8	9	10
Concentration_Additive (%)	4.0	4.2	4.4	4.6	4.8	5.0	5.2	5.4	5.6	5.8
Drying_time (h)	8.7	8.8	8.3	8.7	8.1	8.0	8.1	7.7	7.5	7.2

Based on these data, the company has obtained the following results after applying certain statistical tools:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Drying_time
Independent variable: Concentration_Additive
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        12,1933       0,523576        23,2885         0,0000
Slope          -0,833333       0,106126       -7,85234         0,0000
------------------------------------------------------------------------------


                          Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                      2,29167      1      2,29167      61,66       0,0000
Residual                  0,297333      8    0,0371667
------------------------------------------------------------------------------
Total (Corr.)               2,589      9

Correlation Coefficient = -0,940827
R-squared = 88,5155 percent
Standard Error of Est. = 0,192787
```

a) The company **AIRFLY** Inc. considers that there is not any relationship between drying time and the concentration of additive, so that technicians have decided to use the maximum amount of additive to maximize the drying time. What is your opinion about it? Justify your answer by proposing and discussing the appropriate hypothesis tests. *(2.5 points)*

b) Calculate the residual that would result when using the estimated model for an additive concentration level of 4%. What does this value represent?
(3 points)

c) Assuming that the necessary hypotheses to apply the linear regression model are satisfied, how could we model the distribution of the random variable "drying time" for an additive concentration level of 4.8%? Estimate also the parameters of this distribution model. *(3 points)*

d) What are the hypotheses necessary to apply the linear regression model?
(1.5 points)

SOLUTION

1a) We want to test if the population mean of fuel consumption is 6900 (presumed mean value when the airplane was purchased) or if it has changed: H_0: m=6900; H_1: m≠6900.

$$\frac{\bar{x}-m}{s/\sqrt{n}} \approx t_{n-1};\quad \frac{7200-6900}{200/\sqrt{50}} = 10{,}61$$

The t-calculated follows a Student's t distribution with 49 degrees of freedom (n=50); the value obtained is very unlikely for this distribution, because 95% of their values are comprised between -2.01 and 2.01. Thus, the company is true: there is enough evidence to say that the airplane consumption is significantly higher than when it was purchased.

1b) Considering a type-I risk α=1%, it turns out that 99% of the values for a t_{49} distribution fluctuate between -2.68 and 2.68. The t-calculated value is outside this interval, which implies that the reply of section 1a) does not change.

1c) $\sigma^2 \in \left[\frac{(n-1)\cdot s^2}{g_2}; \frac{(n-1)\cdot s^2}{g_1}\right]$ being g_1 and g_2 the critical values of a χ^2 distribution with 49 degrees of freedom (d.f.) that comprise 95% of the values. It turns out that g_1 and g_2 cannot be directly obtained from the χ^2 table, so an approximate interval can be obtained by considering 50 d.f.:

$$\sigma^2 \in \left[\frac{49\cdot 200^2}{71{,}42}; \frac{49\cdot 200^2}{32{,}357}\right];\quad \sigma^2 \in [27443;\ 60574];\ \text{square root: } \sigma \in [165{,}7;\ 246]$$

The critical values for 49 d.f. can be obtained by interpolating between 45 and 50 d.f., resulting the following interval (more accurate than the previous one):

$$\sigma^2 \in \left[\frac{49\cdot 200^2}{70{,}22}; \frac{49\cdot 200^2}{31{,}55}\right];\quad \sigma^2 \in [27912;\ 62124];\ \sigma \in [167{,}1;\ 249{,}2]$$

The value 180 is comprised within this interval (standard deviation of fuel consumption 5 years ago). Thus, there is not enough evidence to affirm that the variability of the consumption has changed.

2a) The ANOVA table filled out is the following:

Source	*Sum of Squares*	*d.f.*	*Mean Square*	*F-ratio*	*P-value*
MAIN EFFECTS					
A: Battery type	0,222222	**1**	**0,2222**	**0,0143**	0,9067
B: Temperature	10630,8	**2**	**5315,4**	**342,93**	0,0000
INTERACCIONS					
AB	1270,78	**2**	**635,39**	**40,99**	0,0000
RESIDUALS	186,0	**12**	**15,5**		
TOTAL (CORRECTED)	12087,8	**17**			

Number of data = 18 = (6 combinations) x 3 batteries assayed per combination.
Total d.f. = N-1 = 17; D.f. of each factor = nº variants-1;
D.f. interaction = 2·1= 2; Mean Square = SS / d.f.; F-ratio = MS / $MS_{resid.}$

2b) The p-value associated to "battery type" is much higher than 0.05, which implies that he null hypothesis is accepted: the simple effect of this factor is not statistically significant. Thus, we have to accept the null hypothesis that the mean value of battery duration is the same for both models: $m_A = m_B$.

The p-value associated to factor "temperature" is much lower than 0.05 and therefore we have to reject the null hypothesis: the simple effect of this factor is statistically significant. Thus, we have to reject the null hypothesis that the population mean of duration is the same for the three temperatures assayed.

The p-value associated to the interaction is much lower than 0.05. Hence, we reject the null hypothesis: the effect of the interaction between both factors is statistically significant.

2c) Factor "battery type": the plot of LSD intervals is consistent with the conclusion derived from the ANOVA table: intervals are overlapped and, hence, we accept the null hypothesis: $m_A = m_B$. This factor has no effect on battery duration, on average, at the population level. Thus, changing from model A to B would not result in a change on battery duration, on average.

Factor "temperature": the plot shows that intervals are not overlapped, which reveals a statistically significant effect. This effect is approximately linear because a straight line can be fitted which passes through the intervals. It is a linear negative effect, because a higher temperature implies a lower duration (negative correlation). This effect can be modeled using linear regression, which would allow to study if there is additionally a quadratic effect.

2d) The interaction plot suggests a linear effect: the battery duration decreases as temperature increases, but this effect is different for each battery type (i.e., the slope is not the same). The fact that the lines corresponding to each type are not parallel indicates an interaction, which is consistent with the results of the ANOVA table, as the effect of the interaction is statistically significant because its p-value is very low. Thus, there is enough evidence to say that the interaction observed in the plot based on data from a sample also corresponds to an interaction at the population level.

Recommendation: if possible, it is recommended to reduce the temperature in order to maximize battery duration. If the temperature (T) is something fixed that cannot be chosen, the recommendation is the following. In case of T about 5ºC, type B is more convenient. In case of T about 35ºC, it would be recommended to work with type A. But as these batteries are intended for mobile devices, it seems quite unlikely to work normally at such a high temperature. If T is about 20ºC, it doesn't matter to use either type A or B, and the cheapest one would be recommended.

3a) Statement 1: the company considers that the drying time is not related with the concentration of additive. It is false, because the correlation between both variables is statistically significant. The hypothesis test in this case (H_0: $b=0$; H_1: $b\neq0$) is associated to the slope of the regression line: $Y = a+b\cdot x$. Given the low p-value associated with the slope (nearly zero), H_0 is rejected and we can affirm that the slope is significantly different from zero at the population level. As a result, the correlation will also be statistically significant.

Statement 2: technicians have decided to use the maximum amount of additive to maximize drying time. First of all, the target of maximizing the time seems striking, because the interest in industrial procedures is usually to minimize process times. Anyway, the company is not consistent: if they consider that there is no correlation, it is nonsense to think that increasing the amount of additive would result in an increase of drying time. Finally, this relationship is wrong due to the negative slope, so that a greater amount of additive would result in a lower drying time.

3b) The equation obtained is: Y = 12.1933 - 0.833 ·X. The p-values associated to the model coefficients are very low, which implies that both coefficients are statistically significant. For a concentration of 4%, the prediction according to the model is: $Y_{predicted}$ = 12.1933 - 0.833 · 4 = 8.8613.
The value of time observed is 8.7 for a concentration of 4% (first value in the table). By definition, **residual** = $Y_{observed}$ - $Y_{predicted}$ = 8.7 - 8.8613 = **-0.161**.

This residual represents the error of the model, i.e., the difference between the time observed and estimated. This error represents the effect on the response variable (drying time) of all remaining factors which were not included in the regression model. If we would know absolutely all variables implied in the prediction of drying time, we would be able to estimate this value with total precision and the residual would be null. It happens with many deterministic models which are common in physics. But in this case the regression equation is just fitted with one explicative variable, so that the rest of factors not considered produce a random variability in each observation, which is assumed to be normal. Residuals account for this variability. In this case the residual is negative, which means that the value of Y observed is less than the value estimated by the model.

3c) If the model hypotheses (see question 3d) are fulfilled, the drying time for an additive concentration X=4.8 can be modelled by means of a normal distribution with the following parameters:
- Mean: if X=4.8, the model predicts the following value, which will be the average time for this X: Y = 12.1933 - 0.833 · 4.8 = **8.1949**.
- Variance: estimated as the mean square of residuals = 0.03717.
- Standard deviation: square root of the previous value, which appears in the table as *standard error of est*. = **0.1928**.

3d) The hypotheses involved in a linear regression model are three, though the first two ones are more important:
- Hypothesis of normality: the variable under study follows a normal bivariate distribution, which implies that: (1) the marginal distributions of X and Y are normal, (2) the conditional distribution of Y for a particular value X follows a normal model, and (3) the residuals are normal, so that no outliers appear.

- Hypothesis of homoscedasticity: the variance of the conditional distribution of Y for a given value X is constant, it does not depend on X.

- Hypothesis of independence: individuals of the population have been chosen randomly, so that all of them have the same probability to belong to the sample. This hypothesis sometimes is not accomplished when observations are generated along the time, so that the value observed in a certain instant of time

can be partly dependent on those values obtained previously. If this hypothesis is not accomplished it would be recommended to use advanced regression models taking into account time series analysis.