Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 20th 2018

Surname, name:

Group: **1E** Signature:

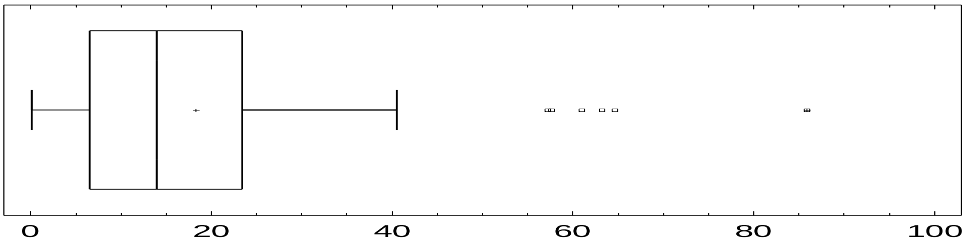
Indicate with a tick mark 1st 2nd

the partials examined □ □

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 the first partial (50%) and 3 about the second partial (50%). 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) The execution speed of 100 computer programs of a certain type has been registered. The following plot has been obtained from these values:



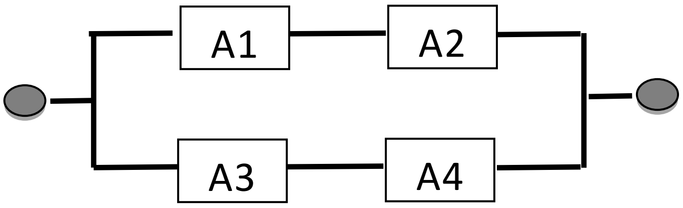
a) According to the plot, what can you say about the expected value of the *skewness coefficient* and the *standardized skewness coefficient*? Note: it is not necessary to calculate the exact value. *(2.5 points)*

b) What would be the most representative parameter of dispersion for this data set? Calculate the value and justify the answer. *(2.5 points)*

c) What would be the most representative parameter of position for this data set? Calculate the value and justify the answer. *(2.5 points)*

d) Calculate approximately the probability that the speed is greater than 50. *(2.5 points)*

2. (1st Partial) The company Robot Inc. sells electronic circuits like the ones shown below:



a) The company has modeled the life (i.e., time of operation until failure) of each component as an exponential model. If 80% of the components work more than 2000 days, calculate the mean lifetime of a component. *(2 points)*

b) What is the probability of the circuit shown above to operate more than 2500 days? *(3 points)*

c) The circuits are sold in packages of 20 units. If a customer detects at least two defective circuits, the entire package is returned to the company.

c.1) Identify the random variable involved in this problem. What is the model of statistical distribution for this variable? *(2 points)*

c.2) If 2% of the circuits sold are defective, what is the probability that a random customer will return a package of circuits to the company? *(3 points)*

3. (1st Partial) One elevator is used to carry packages whose weight fluctuates normally with a mean of 200 kg and standard deviation 19.6 kg.

a) If one package is taken at random, calculate the probability that the weight falls within the range 195 ± 10 kg. *(2.5 points)*

b) Calculate the percentile 5 of this distribution. Interpret the result obtained. *(2.5 points)*

c) If 5 packages are loaded on the elevator, what would be the statistical model for the random variable: "weight of 5 packages"? Estimate the value of the model parameters. *(3 points)*

d) If 5 packages are loaded on the elevator, calculate the probability that the total weight does not exceed 1100 kg. *(2 points)*

4. (2nd Partial) Certain study has sampled the annual income (in millions of euros) of the private clinics in a certain autonomous community of Spain during the year 2004. Another similar random sample is also available for the year 2003. The following information is available for both random samples:

```
Summary Statistics:   Income 2003        Income 2004
Count                          19                 19
Average                    0,1785             0,2549
Median                     0,1663             0,2285
Standard deviation         0,0739             0,1509
Minimum                    0,0606             0,0860
Maximum                    0,3408             0,6537
Range                      0,2802             0,5676
Stnd. skewness             1,2539             1,9805
Stnd. kurtosis             0,3863             1,3216
```

a) Obtain a confidence interval at 95% for the average annual income in private clinics of that community during the year 2004. Interpret the results. *(3 points)*

b) A health councilor indicates that the private clinics of that autonomous community had an average annual income of 200,000 euros during 2004. Study the hypothesis test for this statement with a significance of 5%. *(3 points)*

c) The same health councilor considers that the variability of the income of private clinics in 2004 was 0.1 (million euros)^2. Study the hypothesis test for this statement with a confidence level of 95%. *(4 points)*

5. (2nd Partial). One company has carried out an experiment to study the effect of three types of processors, as well as the power of the fan incorporated, on the processor speed. Both factors were tested at three levels, and each treatment was tested three times. These are the average results obtained in each treatment:

		Fan power		
		0.25	**0.5**	**0.75**
Processor type	A	52.5	56.4	58.0
	B	59.7	52.9	55.0
	C	53.0	57.8	57.4

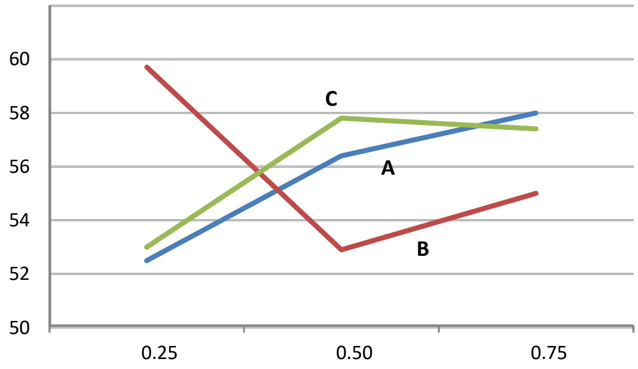
The following table was obtained by analyzing the resulting data with ANOVA:

```
Analysis of Variance for Speed
-----------------------------------------------------------------------------
Source                   Sum of Squares     Df     Mean Square     F-Ratio     P-Value
-----------------------------------------------------------------------------
MAIN EFFECTS
 A:Processor type             0.846667                                          0,4120
 B:fan power                   13.8467                                         0,00013

INTERACTIONS
 AB                            149.473                                          0,0000

RESIDUAL                          8.18
-----------------------------------------------------------------------------
TOTAL                          172.347
-----------------------------------------------------------------------------
```

a) Fill out the ANOVA table, justifying the calculations in detail. Considering a type-I risk of 5%, what conclusions can be drawn? *(4 points)*

b) What speed would be expected on average for a fan power of 0.25? Justify your answer. *(3 points)*

c) What can be deduced from the interaction plot, in accordance with the conclusions derived from the ANOVA table? *(3 points)*



6. (2nd Partial) In certain survey, a set of students were asked about their weight (in kg) and height (cm). Data were entered in STATGRAPHICS, and a regression line was obtained to relate the weight of women as a function of their height. The following results were obtained:

```
Regression Analysis - Linear model: Y = a + b*X
----------------------------------------------------------------------
Dependent variable: WEIGHT
Independent variable: HIGHT-150
Selection variable: SEX="women"
----------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
----------------------------------------------------------------------
Intercept         45,002        1,92181       23,4165          0,0000
Slope           0,767583       0,132065       5,81218          0,0000
----------------------------------------------------------------------

                           Analysis of Variance
----------------------------------------------------------------------
Source         Sum of Squares     Df  Mean Square    F-Ratio    P-Value
---------------------------------------------------------------------
Model               777,89         1      777,89       33,78      0,0000
Residual            921,086       40       23,0271
----------------------------------------------------------------------
Total (Corr.)     1698,98         41
```

a) What is the estimated model? Analyze the statistical significance of the model parameters based on the resulting table. *(2.5 points)*

b) In the context of this study, what would be the practical interpretation of the parameters of the estimated model? *(2.5 points)*

c) Assuming that all hypotheses of the regression model are fulfilled, what type of statistical distribution would be adequate to model the variability of the weight for those women with height equal to 167 cm? Additionally, calculate the second quartile of such distribution. *(2.5 points)*

d) Calculate the coefficient of determination. What is the practical usefulness of this parameter in the proposed model? *(2.5 points)*

SOLUTION

1a) The plot shows that data follow a positively skewed distribution, because the right whisker is longer than the left one, the median is not centered but displaced to the left, and a few isolated points are observed on the right side. Thus, the expected value of the skewness coefficient is positive. The standardized skewness coefficient will be greater than two.

1b) As it is a skewed distribution, the most representative parameter of dispersion is the interquartile range because it is not affected by the presence of extreme values or outliers, as it happens with the variance. Calculation: third quartile (right end of the box: 23.5) minus first quartile (left end: 6.5) = **17**.

1c) Due to the skewness, the most representative parameter of position is the median, because the average is affected by extreme values. It is the line inside box, which corresponds to a value **14**.

1d) Among the 100 values used to represent the box-whisker plot, 6 isolated points are observed, all of whom are greater than 50. Thus, the probability is approximately 6/100 = **6%**.

2a) Random variable X: time of operation until failure of one component.

X ≈ exp (α) ; P(X>2000) = 0.8 ; $e^{-\alpha\cdot 2000}=0.8$;

$\alpha=-(\ln 0.8)/2000$ = -0.0001116 ; Mean = E(X) = -2000/(ln 0.8) = **8962.84** days

2b) P(X>2500) = $e^{-0,0001116\cdot 2500}$ = 0.7566

Event A_i: component A_i works more than 2500 days.

Event $A_{1\text{-}2}$: components A_1 and A_2, assembled in series, work >2500 days.

Assuming that events A_i are independent among them:

$P(A_{1\text{-}2})$ = $P(A_1\cap A_2)$ = $P(A_1)\cdot P(A_2)$ = 0.7566^2 = 0.5724.

Analogously: $P(A_{3\text{-}4})=P(A_3\cap A_4)$ = 0.5724.

Prob. that the upper branch, in paralel with the lower branch, works >2500 =

$P[A_{1-2}\cup A_{3-4}]=P(A_{1-2})+P(A_{3-4})-P(A_{1-2}\cap A_{3-4})$=0.5724 +0.5724 -$0.5724^2$ = **0.817**

2c1) Random variable X: number of defective circuits in a package of 20 units. The minimum value is zero and the maximum is 20. Thus, this discrete random variable follows a Binomial distribution with n=20, being *p* the probability of a circuit to be defective, which is 2% (from section c2). Thus, **X ≈ B (20, 0.02)**.

2c2) P(return the package) = P(X≥2) $=1-P(X=0)-P(X=1)=$

$$=1-\binom{20}{0}\cdot 0.02^{0}\cdot 0.98^{20}-\binom{20}{1}\cdot 0.02^{1}\cdot 0.98^{19}=1 -0.6676 -0.2725 = \mathbf{0.0599}.$$

3a) Variable X: weight of a package; X ≈ N (m=200, σ=19.6).

$$P(X \in [185;\ 205]) = P(X > 185) - P(X > 205) = P[N(200;19.6) > 185] - P[N(200;19.6) > 205] =$$

$$= P[N(0;\ 1) > [(185-200)/19.6]] - P[N(0;\ 1) > [(205-200)/19.6]] =$$

=P[N(0; 1)> -0.7653] - P[N(0; 1)> 0.2551] = 1 - 0.2221 - 0.3993 = **0.379**.

3b) The value k will be percentile 5 if it satisfies that: P(X<k) = 0.05.
P[N(200; 19.6) < k] = 0.05; P[N(0;1) < ((k-200)/19.6)] = 0.05.
Searching the value 0.05 inside the table N(0;1) it turns out that:
(k-200)/19.6 = -1.645. Then, the result is: **167.76 kg**.
Interpretation: 5% of the values of weight are lower than 167.76 kg.

3c) Random variable X_i: weight of one package; Variable Y: weight of 5 packages.
$Y=X_1+X_2+X_3+X_4+X_5$. The sum of normal variables is also a normal distribution, with two parameters (mean and variance) computed as following:
$E(Y)=E(X_1+...+X_5) = E(X_1)+...+E(X_5) = 5\cdot E(X) = 5\cdot 200 = 1000$.

$$\sigma^2(Y) = \sigma^2(X_1 + ... + X_5) = \sigma^2(X_1) + ... + \sigma^2(X_5) = 5\cdot\sigma^2(X) = 5\cdot 19.6^2 = 1920.8.$$

This calculation assumes that variables X_i are independent among them.
Solution: $Y \approx N(m = 1000;\ \sigma^2 = 1920.8) \approx$ **N(m=1000; σ=43.83)**.

3d) P(Y<1100) = P[N(1000; 43.83) <1100] = P [N(0;1) < ((1100-1000)/43.83)] = P[N(0; 1)<2.282] = 1- 0,0113 = **0.9887**.

4a) Sample parameters for 20004: $\overline{x_i} = 0.2549$; s = 0.1509, n=19.

$$m \in \left(\overline{x_i} \pm t_{n-1}^{\alpha/2} \cdot s/\sqrt{n}\right)$$; critical value of Student's t: $t_{n-1}^{\alpha/2} = t_{18}^{0.025} = 2.101$

$$m \in \left[0.2549 \pm 2.101 \cdot 0.1509/\sqrt{19}\right];\quad m \in (0.2549 \pm 0.0727);\quad m \in \mathbf{[0.1822;\ 0.3276]}$$

Interpretation: if infinite samples are taken from a population and, from each sample, a confidence interval for the population mean (*m*) is obtained, it turns out that 95% of these intervals will contain the true value of *m*. Hence, we can say that there is a probability of 95% that the real value of *m* is within the interval obtained.

4b) As the units of the variable are million euros, the hypothesis test to check is the following: H_0: m=0,2; H_1: m≠0,2. As this value is contained in the interval obtained in the previous question, we can accept the null hypothesis (the significance level of 5% is coincident with the one of the previous question). Thus, there is not enough evidence to say that the health councilor is wrong.

4c) We want to study if the "variability" of income was 0.1 or not. As a parameter of variability, this value can correspond to variance or standard deviation. In this case, it is variance because the units indicated are quadratic: 0.1 (million euros)^2. Thus, the hypothesis test to check is: H_0: $\sigma^2 = 0.1$ versus the alternative: H_1: $\sigma^2 \neq 0.1$. In order to solve this test, it is required firstly to obtain a confidence interval for σ^2:

$\sigma^2 \in \left[\frac{(n-1)\cdot s^2}{g_2}; \frac{(n-1)\cdot s^2}{g_1} \right]$ being g_1 and g_2 the range of values for a χ^2 distribution with 18 degrees of freedom that comprises 95% of the values. Based on the table it turns out that: $g_1 = 8.231$; $g_2 = 31.526$.

$$\sigma^2 \in \left[\frac{18\cdot 0.1509^2}{31.526}; \frac{18\cdot 0.1509^2}{8.231} \right]; \quad \sigma^2 \in [0.0130;\ 0.0498]$$

As the value 0.1 is **not** contained inside the interval, the null hypothesis is rejected: the councilor is wrong; the variance of income of clinics in 2004 is much lower than 0.1.

5a) The ANOVA table filled out is the following:

```
Analysis of Variance for Speed
--------------------------------------------------------------------------------
Source                      Sum of Squares      Df     Mean Square     F-Ratio      P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:Type of Processor              0.846667       2          0.4233       0.932       0.4120
 B:Fan power                       13.8467       2          6.9234      15.235      0.00013

INTERACTIONS
 AB                                149.473       4          37.368      82.228       0.0000

RESIDUAL                              8.18      18          0.4544
--------------------------------------------------------------------------------
TOTAL                              172.347      26
--------------------------------------------------------------------------------
```

Nº total of data = 9 combinations (treatments) x 3 values obtained per combination = 27. Total degrees of freedom = N-1 = 26; D.f. for each factor = nº variants-1; D.f. interaction = 2·2= 4; Mean Square = SS / d.f.;
F-ratio = MS / $MS_{residual}$.

Conclusions: Factor "type of processor": as the *p*-value is >0.05, the null hypothesis is accepted: $m_A = m_B = m_C$. The simple effect of this factor is **not** statistically significant, which means that there is not enough evidence to affirm that the mean value of speed is different according to the type of processor.

Factor "fan power": as p-value < 0.05, we reject the null hypothesis H_0: $m_{0.25} = m_{0.5} = m_{0.75}$. The simple effect of this factor is statistically significant, which implies that we have to reject the null hypothesis that the mean speed at the population level is the same for the three fan powers.

Interaction: as *p*-valor is much lower than 0.05, the null hypothesis is rejected: the effect of the interaction between both factors is statistically significant.

5b) As the simple effect of factor "fan power" is statistically significant, at least one of the means ($m_{0.25}$, $m_{0.5}$ or $m_{0.75}$) will be different than the rest at the population level. As the type of processor is not specified, the expected value of speed for a fan power of 0.25 will be the mean value of experimental results obtained under such conditions: (52.5+59.7+53)/3 = **55.067**.

5c) The effect of the interaction is statistically significant, which is consistent with the fact that the three processors do not present a parallel performance: type A and C show a similar pattern (generally speaking, an increasing trend), but type B reflects a totally opposite behavior: it decreases suddenly from 0.25 to 0.5.

A similar value is obtained if we average the data obtained for each fan power, which justifies that the simple effect of *processor type* is not statistically significant. However, if we average the values for the three processor types, an increasing trend is obtained.

6a) The estimated value of the model coefficients is indicated at the column "*estimate*", so that the resulting predictive equation is the following:
Weight = 45.002 + 0.7676 · (height - 150).

- The parameter 45.002 (*intercept*) is statistically significant because its *p*-value is much lower than the values of α considered most frequently. Thus, the null hypothesis is rejected: there is enough evidence to say that this value is different from zero at the population level.
- The parameter 0.7676 (*slope*) is also statist. significant for the same reason.

6b) Practical interpretation of the value 45.002: *mean value expected for women with a height of 150 cm*. Actually, from the equation, if height = 150, it turns out that weight = 45.002.

Practical interpretation of 0.7676: in the subset of women, *if their height increases 1 cm, we expect an increase of their weight, on average, of 0.768 kg*.

6c) If the hypotheses of the regression model are fulfilled, the conditional distribution of weight, in case of height=167, will be a Normal model. The second quartile of this distribution, i.e. the median (which is coincident with the mean) is computed from the estimated model:
Weight = 45.002 + 0.7676 · (167 - 150) = **58.05 kg**.

6d) The coefficient of determination is obtained as: the sum of squares explained by the model divided by the total sum of squares:

$$\mathbf{R^2} = 777.89 / 1698.98 = 0.4579 = \mathbf{45.79\%}$$

This coefficient indicates that 45.79% of the variance of Y (weight) is explained by the variable *height*. As this value is not high (the maximum would be 100%), it can be concluded that the degree of correlation between both variables is moderate. Thus, the model has certain predictive ability, but with a considerable degree of error.