Bachelor Degree in Computer Engineering

Statistics group E (English)

SECOND PARTIAL EXAM

June 7th 2016

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 time required to compile certain type of programs is a random variable which fluctuates uniformly between 5 and 30 milliseconds. Certain engineer needs to compile consecutively 50 programs.

a) Indicate the type of distribution of the following variable: total time required to compile a set of 50 programs. Justify your answer. Calculate the mean and variance of this random variable. *(4 points)*

b) Calculate approximately the probability to compile the 50 programs with a total time greater than one second. *(3 points)*

c) Calculate the limits that will comprise 95% of the values of total time to compile the 50 programs. *(3 points)*

2. The daily performance of a computing system (in MIPS) is a random variable following a normal distribution. Answer the following questions by justifying the reply.

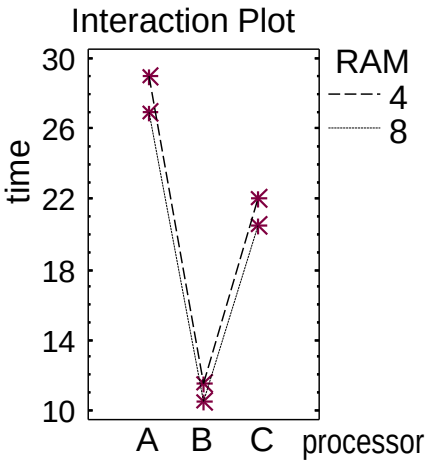
a) A set of 20 values of the performance are randomly taken from the system. It is assumed that the standard deviation of the population is $\sigma = 7$ MIPS. What is the probability to obtain a sample variance lower than 70 $MIPS^2$? *(4 points)*

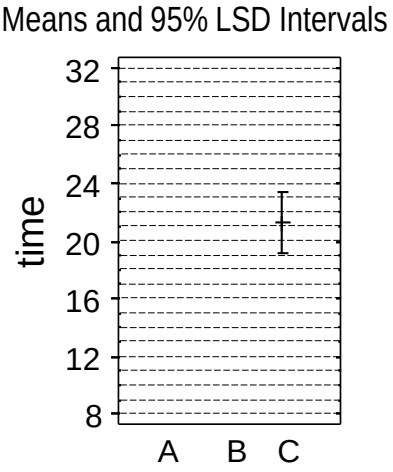
b) A sample of values about the performance of the computing system was obtained, corresponding to a set of 25 days randomly chosen. The sample mean was 85 MIPS and the sample standard deviation was 6.5 MIPS. By considering a type I risk of 5%, can we admit that the mean performance at the population is 95 MIPS? Justify conveniently your answer. *(4 points)*

c) What is the “type I risk” (also known as “type I error rate”) of a hypothesis test? *(2 points)*

3. Certain engineer intends to study how the type of processor and size of RAM memory affects the time required to execute certain algorithm. For this purpose, the algorithm was run on 12 computers with different characteristics, four of them with a processor of type A, four of type B and four of type C. Half of them have 4 GB of RAM memory, and the other half, 8 GB. The experimental values of time obtained (variable "t" in microseconds) are indicated below. The plots shown below were obtained using ANOVA, which was applied to analyze the values obtained. The following sums of squares were obtained: SS_{total} = 610.92; SS_{proc} = 586.17; SS_{RAM} = 6.75; $SS_{proc \cdot RAM}$ = 0.5.

PR.	RAM	t
A	4	31
A	4	27
A	8	28
A	8	26
B	4	12
B	4	11
B	8	10
B	8	11
C	4	21
C	4	23
C	8	22
C	8	19





a) Examine whether the simple effect of any of the two factors and the effect of their interaction is statistically significant (α=0.05). *(4 points)*

b) According to the interaction plot shown above, taking into account the results obtained in the previous section and considering α=0.05, how does the RAM memory affect the time of execution of the algorithm? *(2 points)*

c) The plot on the right shows the LSD interval corresponding to processor C. Draw the LSD intervals of the other two types of processors, by justifying conveniently the calculations. *(2 points)*

d) Processor of type A has a speed of 1 GHz; type B, 3 GHz and type C, 2 GHz. Taking into account all available information, can we describe as linear or quadratic the effect of processor's speed on the time of execution? *(2 points)*

4. The time (in microseconds) required by a computer program to perform certain operation with data arrays depends on two parameters: processor's speed (GHz) and array size. A set of 50 arrays were processed with this program. The matrix of variances-covariances that relates the three variables is shown below, as well as the result of two linear regression models: the first one relates time *vs*. speed, and the second model relates time *vs*. size.

	Time	Speed	Size
Time	1,46648	-1,16058	4,49981
Speed	-1,16058	1,07253	-1,85217
Size	4,49981	-1,85217	60,3879

Regression Analysis - Linear model: Y = a + b*X

Dependent variable: TIME
Independent variable: SPEED

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	7,31367	0,258654	28,2759	0,0000
Slope	-1,08209	0,0639626	[illegible]	[illegible]

Regression Analysis - Linear model: Y = a + b*X

Dependent variable: TIME
Independent variable: SIZE

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	2,02794	0,317818	6,3808	0,0000
Slope	0,074515	0,0197547	[illegible]	[illegible]

a) Can we affirm that the variable *Time* presents a stronger correlation with *size* rather than with *speed*? *(2 points)*

b) Based on the results obtained, indicate which are the two regression models estimated from the statistical analyses carried out. *(1 point)*

c) Is there enough evidence to affirm that the linear effect of *size* is statistically significant? And what about the linear effect of *speed*? Consider α=0.05. *(2 points)*

d) What is the mean time which is expected when performing the operation with an array, using a speed of 2 GHz? *(1 point)*

e) Calculate the residual variance corresponding to the first regression model. *(2 points)*

f) If the speed is 2 GHz, calculate the probability to perform the operation with a time greater than 5 microseconds. *(2 points)*

SOLUTION

1a) The random variable of time (X) follows a distribution: U(5; 30).
E(X)=(5+30)/2 = 17.5; $\sigma^2 = (b-a)^2/12 = (30-5)^2/12 = 52.08$
Total time: Y = X_1+ X_2 + ... + X_{50}. According to the central limit theorem, the sum of N random variables of any kind tends to follow a normal distribution if N is large enough and the variables are independent. Thus, the total time will follow a normal distribution, being the mean and variance:

$$E(X_1+...+X_{50}) = E(X_1)+...+E(X_{50}) = 50\cdot E(X) = 50\cdot 17.5 = \mathbf{875}$$
$$\sigma^2(X_1+...+X_{50}) = \sigma^2(X_1)+...+\sigma^2(X_{50}) = 50\cdot \sigma^2(X) = 50\cdot 52.08 = \mathbf{2604.2}$$

1b) As the units are milliseconds, P(Y > 1 second) = P (Y>1000)

$$P(Y>1000) = P\left[N\left(875,\sqrt{2604}\right)>1000\right] = P\left[N(0;1) > \frac{1000-875}{\sqrt{2604}}\right] = P[N(0;1)>2.45] = \mathbf{0.0071}$$

1c) In a normal distribution, the interval m ± 1.96·s comprises 95% of the values. Thus, the requested interval will be: $875 \pm 1.96\cdot\sqrt{2604} =$ **[775, 975]**
The solution is also correct using 2.0 instead of 1.96: $875 \pm 2\cdot\sqrt{2604} = [773, 977]$

2a) Taking into account the equation $(n-1)\cdot s^2/\sigma^2 \approx \chi^2_{n-1}$, being n=20 (sample size) and σ=7, the probability requested is calculated as following:

$$P(s^2<70) = P\left(s^2\cdot\frac{n-1}{\sigma^2} < 70\cdot\frac{n-1}{\sigma^2}\right) = P\left(\chi^2_{19} < 70\cdot\frac{19}{7^2}\right) = P(\chi^2_{19}<27.14) \approx \mathbf{0.9}$$

2b) The null hypothesis to test is H_0: m=95, being n=25, s=6.5.

$\left|\frac{\bar{x}-m_0}{s/\sqrt{n}}\right| = \left|\frac{85-95}{6.5/5}\right| = 7.7$ that has to be compared with $t_{n-1}^{\alpha/2} = t_{24}^{0.025} = 2.064$

The computed t-statistic is greater than the critical value of a Student-t distribution with 24 degrees of freedom, which implies that 7.7 is not a frequent value of this distribution. Thus, the null hypothesis has to be rejected: we **cannot** admit that the mean performance at the population is 95 MIPS, at α=0.05. Another procedure is to check that 95 is outside the confidence interval for the population mean, which turns out to be [82.32, 87.68].

2c) By definition, the "type I risk" (also known as "type I error rate" or "significance level") of a hypothesis test, is the probability of rejecting the null hypothesis when it is true.

3a) The ANOVA table is filled out based on the following calculations:
1) Total degrees of freedom = 11 (12 data minus one)
2) Degrees of freedom of factor supplier = 3 variants - 1 = 2.
3) Degrees of freedom of factor RAM = 2 levels - 1 = 1
4) Degrees of freedom of the interaction = 2·1 = 2
5) Residual degrees of freedom (obtained by difference) = 11 - 2 - 1 - 2 = 6
6) $SS_{resid} = SS_{total} - SS_{supplier} - SS_{RAM} - SS_{suppl\cdot RAM}$ = 586.17 - 6.75 - 0.5 = 17.5

7) $MS_{supplier} = SS_{suppl} / df_{suppl} = 586.17 / 2 = 293.08$
8) $MS_{RAM} = SS_{RAM} / df_{RAM} = 6.75 / 1 = 6.75$
9) $MS_{supp \cdot RAM} = SS_{supp \cdot RAM} / df_{suppl \cdot RAM} = 0.5 / 2 = 0.25$
10) $F_{supplier} = MS_{suppl}/MS_{res} = 293.08 / 2.917 = 100.49$
11) $F_{RAM} = MS_{RAM}/MS_{res} = 6.75 / 2.917 = 2.31$
12) $F_{supp \cdot RAM} = MS_{supp \cdot RAM}/MS_{res} = 0.25 / 2.917 = 0.09$

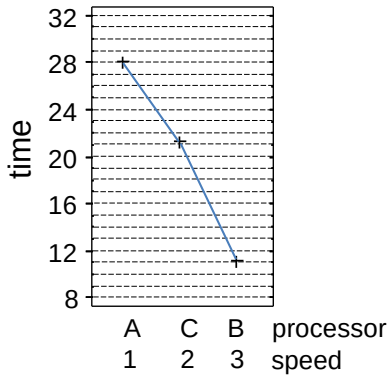
Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:SUPPLIER	586,17	2	293,085	100,49
B:RAM	6,75	1	6,75	2,31
INTERACTIONS				
AB	0,5	2	0,25	0,09
RESIDUAL	17,5	6	2,91667	
TOTAL (CORRECTED)	610,92	11		

The effect of supplier is statistically significant for α=0.05 because its F-ratio = 100.49 is greater than the critical value from the F table: $F_{2;6}^{0.05}$=5.14.
The effect of RAM is not statistically significant because its F-ratio= 2.31 is lower that the critical value: $F_{1;6}^{0.05}$=5.99.
The effect of the interaction is not statistically significant because its F-ratio = 0.09 is lower than the critical value: $F_{2;6}^{0.05}$=5.14

3b) The effect of the interaction is not statistically significant, which is consistent with the parallel lines observed in the interaction plot. This plot also shows that the mean times observed for RAM=8 are lower than the mean time for RAM=4. However, the effect of RAM memory is **not** statistically significant, which implies that there is not enough evidence to affirm that computers with RAM=8 will take in average less time than computers with RAM=4 to execute the algorithm.

3c) The width of LSD intervals depends on the number of values. Since there are 4 values for each supplier, the three intervals will have the same width as for supplier C shown in the plot, which ranges approximately from 19.1 to 23.3 (i.e., midpoint ±2.1). The midpoint is the sample average.
Average of values for A: (31+27+28+26)/4 = 28
Average of values for B: (12+11+10+11)/4 = 11
LSD intervals for A: 28 ±2.1 [25.9, 30.1]; supplier B: 11 ±2.1 [8.9, 13.1]

3d) The average values of time obtained for processors A, B, and C are: 28, 11, and 21.25, respectively. Taking into account the processor's speed, the plot shown on the right indicates the average time according to speed. The plot indicates that a straight line would fit quite accurately the points. Hence, it can be concluded that processor's speed exerts a linear effect on the time.



4a) By calling Y: time, X_1: speed, and X_2: size, the variances correspond to values in the main diagonal of the variance-covariance matrix:

$s_Y^2 = 1.4665;\quad s_{X_1}^2 = 1.0725;\quad s_{X_2}^2 = 60.388$

Moreover, from that matrix, it turns out: $\text{cov}(X_1, Y) = -1.1606;\quad \text{cov}(X_2, Y) = 4.4998$

$$R_{X_1,Y}^2 = r^2 = \frac{\text{cov}^2(X_1,Y)}{s_{X_1}^2 \cdot s_Y^2} = \frac{-1.1606^2}{1.0725 \cdot 1.4665} = 0.8564$$

$$R_{X_2,Y}^2 = r^2 = \frac{\text{cov}^2(X_2,Y)}{s_{X_2}^2 \cdot s_Y^2} = \frac{4.4998^2}{60.388 \cdot 1.4665} = 0.229$$

The answer is **no**, we cannot affirm that *time* presents a stronger correlation with *size* rather than with *speed*, because the coefficient of determination (R^2) between *time* and *size* (X_2) is lower than in the case of *speed*.

4b) Taking into account the estimated values indicated in the table of results, the equations are the following:
First model: Time = 7.314 - 1.082 · speed
Second model: Time = 2.028 + 0.0745 · size

4c) Regarding the first model, the linear effect of speed is statistically significant if the slope of the straight line is different from zero at the population level. Thus, being the regression model: $Y=\beta_0+\beta_1 \cdot X$, the null hypothesis to test is: H_0: $\beta_1=0$ versus the alternative hypothesis H_1: $\beta_1 \neq 0$. Given that $b_i / s_{b_i} \approx t_{N-1-I} \approx t_{48}$ being I=1 (model with one explicative variable), N=50 and $\alpha=0.05$, it turns out that: $b_i / s_{b_i} = -1.082/0.064 = -16.9$ which is not a frequent value for the t_{48} distribution. Thus, the null hypothesis is rejected: there is enough evidence to affirm that the linear effect of speed is statistically significant.

Regarding the second model: $b_i / s_{b_i} = 0.0745/0.0197 = 3.77$ which is not a frequent value for the t_{48} distribution (95% of values are comprised between -2.011 and 2.011). Thus, the linear effect of size is also statistically significant.

4d) The requested value is obtained by considering speed=2 in the first model:
E (time / speed=2) = 7.314 - 1.082 · 2 = **5.15**

4e) The residual variance can be computed as:
$s_{res}^2 = s_Y^2 \cdot (1 - r_{X_1,Y}^2) = 1.4665 \cdot (1 - 0.8564) =$ **0.2106**

4f) When the speed is 2, the conditional distribution of *time* will be a Normal distribution: the average is given by the regression model (5.15), and the variance will be the residual variance (0.199).

$$P(t > 5) = P\left[N\left(5.15; \sqrt{0.21}\right) > 5\right] = P\left[N(0;1) > \frac{5 - 5.15}{\sqrt{0.21}}\right] = P\left[N(0;1) > -0.327\right] = 1 - 0.37 = \mathbf{0.63}$$