Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 24th 2016

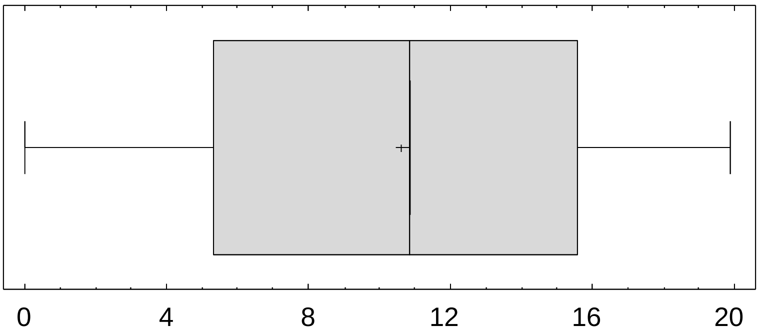
Surname, name:	
Group: **1E**	Signature:

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 the first partial (40%) and 3 about the second partial (60%). 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 time required to search a file in certain documentary database is a random variable with unknown type of distribution. This time was obtained for a set of files from a sample randomly taken, measured in milliseconds (ms). Based on the resulting data, the following plot was obtained:



a) Indicate the population under study. What is the random variable? What type is it? *(3 points)*

b) Calculate at least 5 descriptive parameters that can be obtained from this plot. Indicate the type of each parameter. *(3 points)*

c) Is it possible to deduce the number of files (sample size) from which this plot was obtained? *(1 point)*

d) Assuming that data follow a uniform model, calculate the probability to have a search time greater than 8 ms. *(3 points)*

2. (1st Partial) The factory ARTUDITU is interested in controlling the quality of transistors manufactured. We know that 1% of transistors manufactured in this process are defective.

a) A sample of 15 transistors is randomly taken in a routine control. What is the probability to find more than two defective transistors? *(3 points)*

b) If a sample of 15 transistors is taken at random and it turns out that at least one of them is defective, what is the probability to find exactly two defective transistors? *(3 points)*

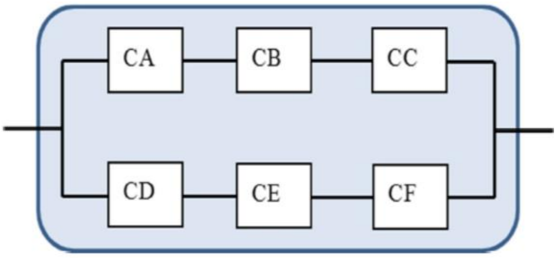
c) If a total of 100 random samples are taken in one month, each of size 15, what is the probability to find in one month more than 10 defective transistors? *(4 points)*

3. (1st Partial) Certain device consists of six identical components (CA, CB, CC, CD, CE, CF). The time of operation until failure of each component follows an exponential distribution. The reliability of these components after 1000 hours of operation is 0.9. The operation or failure is independent from each other. Answer the following questions by justifying the answers in detail, and by defining all the variables and events involved.

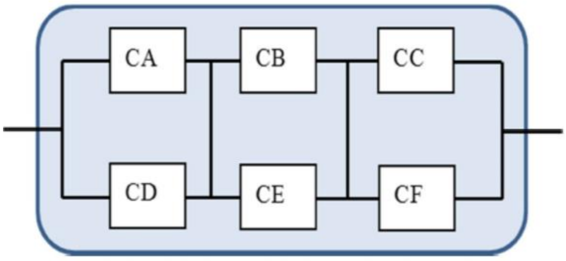
a) Calculate the half-life (average time of operation) of these components. *(3 points)*

b) Two students (STU1 and STU2) have a different opinion about the most appropriate type of assembly for the components in order to increase the reliability of the resulting device. The assembly proposed by each student for the six components is shown below.

Proposal STU1:



Proposal STU2:



Calculate the reliability of each proposal. Which of them is the most suitable taking into account the objective pursued? *(7 points)*

4. (2^{nd} Partial) Answer the following questions, justifying the reply.

a) From a normal population with standard deviation σ=5, a sample of size 16 is randomly extracted. What is the probability that the sample mean and the population mean differ in less than 2 units? *(3 points)*

b) A given quality parameter is routinely measured in a process. Certain operative changes are carried out to improve the quality. In order to study the effectiveness of such changes, a sample of 5 units was taken at random, resulting a sample mean of 20.5 and a sample variance of 1.2. Obtain a confidence interval for the population mean, considering α=0.05. Is it reasonable to admit that the population mean of the parameter is 22, after the changes introduced in the process? *(3.5 points)*

c) Is it reasonable to assume that the population variance of the parameter, after the operative changes carried out, is 9.0 units^2? *(3.5 points)*

5. (2^{nd} Partial) One engineer wants to study the effect of computer configuration (three possibilities: A, B and C) and cache memory (3 levels: low, medium and high) on the average performance of a computer system. Each treatment was tested three times.

a) Analyze what effects are statistically significant, after obtaining the summary table of ANOVA (consider α=5%), taking into account that: SS_{total}=11039.2; SS_{config}=604.47; SS_{cache}=8890.4; $SS_{residual}$=690.005. *(5 points)*

b) Assuming that the hypothesis of homoscedasticity is satisfied, calculate the estimated value of the variance of each population under study. *(2 points)*

c) In general, what additional information to that provided by the summary table of ANOVA can be obtained from the graphical representation of LSD intervals? *(3 points)*

6. (2nd Partial) An automatic system for the management of stocks is available in a warehouse. For the last 100 orders managed by the warehouse, the system records the number of units ordered and the time to process the order (minutes). A linear regression model was obtained to predict this time, and the results obtained with Statgraphics are the following:

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: TIME
Independent variable: UNITS
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept      -0,886722       0,212955        -4,1639         0,0001
Slope          0,0321159     0,00259249
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     39,9565
Residual                  25,5156
-----------------------------------------------------------------------------
Total (Corr.)             65,4721     99
```

a) Based on the information provided by Statgraphics, what is the regression model estimated? *(2 points)*

b) Is there a statistically significant linear effect of the number of units on the average process time? Consider a type I risk of 1%. *(2 points)*

c) Obtain the coefficient of determination. Indicate how this coefficient will be altered if the time is expressed in seconds instead of minutes. *(2 points)*

d) Knowing that the average process time expected for an order of 80 units is 1.68 minutes, what is the probability for an order of 80 units taken at random to be processed in more than 2 minutes? *(4 points)*

SOLUTION

1a) The population is formed by the set of files contained in the documentary database. Random variable: time (measured in milliseconds) required to search a file (chosen at random) in that database. This variable is one-dimensional, quantitative and continuous (because time is measured in a continuous scale).

1b) - Range = maximum - minimum ≈ 20 - 0 = 20.
- Interquartile range (IQR) = 3rd quartile - 1st quartile = 15.6 - 5.2 = 10.4.
- Median = 10.9 (vertical line inside the box)
- Average (mean) ≈ 10.7 (point inside the box)
- Third quartile = 15.6 (right edge of the box).

The range and IQR inform about the data dispersion, while the other three parameters inform about their position. Apart of these 5 parameters, others can be obtained from the graph:
- First quartile = 5.2 (left edge of the box), it is a parameter of position.
- Maximum ≈ 20 (parameter of position).
- Skewness coefficient ≈ 0 because the shape of the graph is nearly symmetrical. It is a parameter of shape.

1c) No, from a box-whisker plot like this one it is not possible to deduce the number data used to build the plot.

1d) X ≈ U (0; 20); $P(X \le x) = (x-a)/(b-a)$;
$P(X > 8) = 1 - P(X < 8) = 1 - (8-0)/(20-0) = 1 - (8/20) = \mathbf{0.6}$

2a) random var. X: number of defective transistors in a sample of 15. This variable follows a binomial distribution, with parameters: X ≈ Bi (n=15, p=0.01). By applying the probability function for this distribution:

$$P(X > 2) = 1 - P(X \le 2) = 1 - P(X=0) - P(X=1) - P(X=2) =$$
$$= 1 - \binom{15}{0} \cdot 0.01^0 \cdot 0.99^{15} - \binom{15}{1} \cdot 0.01^1 \cdot 0.99^{14} - \binom{15}{2} \cdot 0.01^2 \cdot 0.99^{13} =$$
$$= 1 - 0.8601 - 0.1303 - 0.0092 = \mathbf{0.000416}$$

2b) The following conditional probability is requested:
$$P[(X=2)/(X>0)] = \frac{P[(X=2)\cap(X>0)]}{P(X>0)} = \frac{P(X=2)}{P(X>0)} = \frac{0.00921}{1-0.8601} = 0.0658$$

2c) One hundred consecutive samples of 15 transistors are taken, and the total number of defective units found in total is counted. This is equivalent to taking only one sample of size 100 · 15. Hence, the random variable Y: number of defective transistors found in 100 samples of size 15 will follow a binomial distribution with parameters: Y ≈ Bi (n=1500, p=0.01).

As the value of *n* is very high and the probability is small, it can be approximated to a Poisson distribution: Y ≈ Ps (λ = n·p= 15). Drawing a vertical line for λ = 15 in the Poisson abacus, it crosses the curve 10 in a point that corresponds to a probability of 0.12 (reading on the vertical axis). Thus,
$P(Y > 10) = 1 - P(Y \le 10) = 1 - 0.12 = \mathbf{0.88}$

An alternative way to solve this problem is by considering Y as the sum of defective transistors found in the set of 100 samples: $Y = X_1 + \ldots + X_{100}$. As the number of addends is high, we can assume that Y follows a Normal distributions according to the central limit theorem:

$X_i \approx Bi(n=15, p=0.01)$; $E(X)=n\cdot p=0.15$; $\sigma_X^2 = n\cdot p\cdot(1-p) = 15\cdot 0.01\cdot 0.99 = 0.1485$

$$E(X_1+\ldots+X_{100}) = E(X_1)+\ldots+E(X_{100}) = 100\cdot E(X) = 100\cdot 0.15 = 15$$

$$\sigma^2(X_1+\ldots+X_{100}) = \sigma^2(X_1)+\ldots+\sigma^2(X_{100}) = 100\cdot\sigma^2(X) = 100\cdot 0.1485 = 14.85$$

$$P(Y>10)\approx P\left[N\left(15,\sqrt{14.85}\right)\geq 10.5\right] = P\left[N(0;\ 1)\geq\frac{10.5-15}{\sqrt{14.85}}\right] = P\left[N(0;\ 1)\geq -1.17\right] = \mathbf{0.88}$$

3a) Random variable T: time of operation (hours) until failure of the component. If the reliability of these components after 1000 h is 0.9, it implies that P(T>1000)=0,9 according to the definition of reliability.

$P(T>1000) = e^{-\alpha\cdot 1000} = 0.9$; $\alpha = -(\ln 0.9)/1000$.

Average value of an exponential var.: $E(X) = 1/\alpha = -1000/(\ln 0.9) = \mathbf{9491.2}$ h.

3b) Event A: component A is still operative after 1000 hours. P(A)=0,9. Events B, C, D, E and F are defined analogously.

Reliability of assembly proposed by STU-1:

The reliability of the assembly at 1000 h, which means the probability to be operative more than 1000 h, is the probability to work of the upper branch or (union of events) to work of the lower branch. For each branch, in order to be operative, all their components have to work (i.e. intersection of events). Thus,

Reliability = $P[(A\cap B\cap C)\cup(D\cap E\cap F)] =$

$= P[(A\cap B\cap C)] + P[(D\cap E\cap F)] - P[(A\cap B\cap C)\cap(D\cap E\cap F)] =$

Assuming independence of events, it turns out that:

$= P(A)\cdot P(B)\cdot P(C) + P(D)\cdot P(E)\cdot P(F) - P(A)\cdot P(B)\cdot P(C)\cdot P(D)\cdot P(E)\cdot P(F) =$

$= 0.9^3 + 0.9^3 - 0.9^6 = \mathbf{0.9266}$

Reliability of assembly proposed by STU-2:

The assembly is equivalent to 3 sub-circuits in series, each with 2 components in parallel. The reliability after 1000 h will be the probability to work the 3 sub-circuits (intersection of events). For each sub-circuit, it is required to be operative the component at the upper or lower branch (union of events). As the reliability is the same for all components and assuming independence:

Reliability = $P[(A\cup D)\cap(B\cup E)\cap(C\cup F)] = P(A\cup D)\cdot P(B\cup E)\cdot P(C\cup F) =$

$= [P(A\cup D)]^3 = [P(A)+P(D)-P(A)\cdot P(D)]^3 = [0.9+0.9-0.9^2]^3 = \mathbf{0.9703}$

The proposal of the 2nd student is better because it yields a higher reliability.

4a) Given that $\bar{x}\approx N(m\ ;\ \sigma/\sqrt{n})$ it turns out that: $\bar{x}-m\approx N(0\ ;\ \sigma/\sqrt{n})$.

$$P(|\bar{x}-m|<2) = P[(\bar{x}-m)\in[-2;\ 2]] = 1-2\cdot P[(\bar{x}-m)>2] = 1-2\cdot P[N(0;\sigma/\sqrt{n})>2] =$$

$$= 1-2\cdot P\left[N(0;1)>\frac{2-0}{5/\sqrt{16}}\right] = 1-2\cdot P[N(0;1)>1.6] = 1-2\cdot 0.0548 = \mathbf{0.8904}$$

4b) The confidence interval for the population mean (m) is calculated from the sample mean (20.5) and the sample variance (1.2):

$$\bar{x} \pm t_{n-1}^{\alpha/2} \cdot s\big/\sqrt{n} = 20.5 \pm t_4^{0,025} \cdot \sqrt{1.2}\big/\sqrt{5} = 20.5 \pm 2.776 \cdot 0.4899 = 20.5 \pm 1.36$$

$m \in [19.14;\ 21.86]$. It is not reasonable to admit a population mean of 22 for α=0.05 because 22 is out of the confidence interval obtained.

4c) We obtain firstly a confidence interval for the population variance: $\sigma^2 \in \left[(n-1)\cdot s^2/g_2;\ (n-1)\cdot s^2/g_1\right]$ being g_1 and g_2 the critical values of a chi-square distribution with 4 degrees of freedom comprising 95% of the values of that distribution. From the table of that distribution, we obtain that g_1=0.484, g_2=11.143. Thus, the confidence interval will be:

$$\sigma^2 \in \left[4\cdot 1.2/11.143;\ 4\cdot 1.2/0.484\right] = [0.43;\ 9.92]$$

It is reasonable to assume that the population variance of the parameter is σ^2=9 for α=0.05 because the value 9 is inside the interval obtained.

5a) There are 9 treatments (3 possible configurations combined with 3 levels of memory). Each treatment was tested 3 times which implies a set of 27 experimental data and, hence, the total degrees of freedom (d.f.) are 27-1=26. Each factor has 3 levels (2 d.f.), the interaction has 4 d.f. (the product 2·2). The mean square is obtained by dividing the sum of squares by d.f., and the F-ratio is obtained by dividing each mean square by the residual mean square.

```
----------------------------------------------------------------------
Source        Sum of Squares     df    Mean Square     F-Ratio     P-Value
----------------------------------------------------------------------
Config               604.47       2        302,235       7,88        <0,05
Cache               8890,4        2       4445,2       115,97        <0,05

Config*Cache         854,325      4        213,58       5,57         <0,05

RESIDUAL             690,005     18         38,33
----------------------------------------------------------------------
TOTAL              11039,2       26
----------------------------------------------------------------------
```

The critical value of a distribution $F_{2;18}$ is 3.55 for α=5%. As the F-ratio obtained for configuration is higher than the critical value, the effect of this factor is statistically significant. For cache memory, also F-ratio=115.97 > 3.55, which implies that this factor also affects significantly over the average performance.

The critical value of a distribution $F_{4;18}$ is 2.93 for α=5%. The effect of the interaction is also statistically significant because its F-ratio is 5.57 >2.93.

5b) Each treatment corresponds to a different population. Assuming that the hypothesis of homoscedasticity is satisfied, it implies that the variance of all populations will be the same, which is equivalent to the variance of residuals. In ANOVA, the residual variance coincides with the residual mean square, which is **38.33** according to the table obtained in section 5.a.

5c) LSD intervals are used to interpret the effect of qualitative factors with more than 2 variants. If the effect of a factor is not statistically significant according to the ANOVA table, then LSD intervals do not provide further useful information because all of them will be overlapped. But if the effect is significant, then at least 2 averages will be different at the population. In that case, LSD intervals are useful to identify which variants have a different mean.

If the factor is quantitative with more than two levels, the means plot showing LSD intervals provides an idea about the possible linear or quadratic effect of the factor over the response variable, but it is better to apply regression in order to study if that effect is statistically significant.

6a) Taking into account the estimated values of the slope and ordinate (intercept), the equation of the regression model will be:
Time = -0.8867 + 0.03212 · units

6b) The ratio between the slope (estimated value) and its standard error is: 0.03212 / 0.002592 =12.38. By considering true the null hypothesis that the slope of the line is zero at the population, this ratio will follow a Student-t distribution with N-1-I = 100-1-1 = 98 degrees of freedom. But the value 12.38 is rather infrequent for this distribution and, hence, the null hypothesis is rejected: there is enough evidence to affirm that the slope is different from zero at the population and, therefore, the linear effect is statistically significant.

6c) $R^2 = \frac{SC_{\text{modelo}}}{SC_{total}} = \frac{39.9565}{65.4721} = \mathbf{0.6103}$ = 61.03%

This coefficient is the same if variables are expressed in other units, because that change does not affect the degree of correlation between variables, which is quantified by means of this coefficient.

6d) The conditional distribution of Y when X=10 will be normal, with a mean of 1.68 (given in the statement) and with a variance equal to the residual variance (estimated from the residual mean square). For this purpose, we need to complete the table. The model has one degree of freedom because only one variable is included. Thus, 25.516 / 98 = 0.2604.

```
                              Analysis of Variance
-----------------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square     F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                         39,9565      1
Residual                      25,5156     98     0,2604
-----------------------------------------------------------------------------
Total (Corr.)                 65,4721     99
```

$$Y/X=10 \approx N\left(1.68;\ \sqrt{0.2604}\right); \qquad P(Y>2/X=10)=P\left[N\left(1.68;\sqrt{0.2604}\right)>2\right]=$$
$$=P\left[N(0;\ 1)>(2-1.68)/\sqrt{0.2604}\right]=P\left[N(0;\ 1)>0.627\right]=\mathbf{0.266}$$