Bachelor Degree in Computer Engineering

Statistics **group E (English)**

FIRST PARTIAL EXAM

March 31st 2014

Surname, name	
Signature	

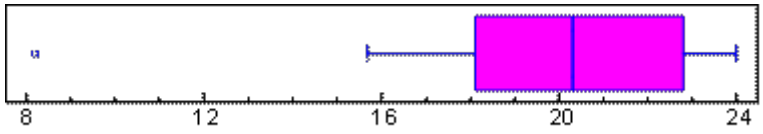
Instructions

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. Certain computer program performs searches in a database of considerable size. In order to improve the efficiency of this program, certain modification is made, and the program is tested 13 times. The following values of time (in milliseconds) are obtained:

{ 8,2 ; 15,7 ; 17,3 ; 18,1 ; 18,5 ; 19,2 ; 20,3 ; 21,1 ; 21,6 ; 22,8 ; 23,2 ; 24 ; 24 }

In order to explore the sample, the following plot is obtained with the data:



a) What is the population under study? What are the individuals of this population? *(1.5 points)*

b) Calculate the first quartile and the range. *(2 points)*

c) Indicate the name of this type of plot. *(1 point)*

d) Based on the plot shown above and taking into account that the standardized coefficient of skewness is -2.276, can we affirm that the values of time of search in the database follow a negatively skewed distribution? *(2 points)*

e) Obtain, approximately, the value of one parameter of position that is representative of the values of the sample. *(1.5 points)*

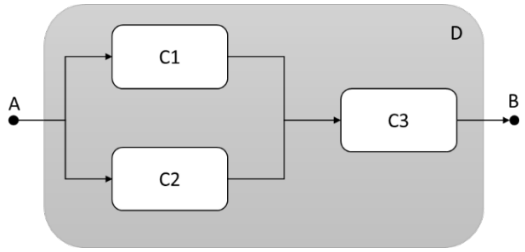
f) What method would you use to study if these data follow a Normal distribution? What are the advantages of that method with respect to the plot shown above? *(2 points)*

2. One company manufactures electronic components (C) with an average life of 4 years. Assuming that the duration of this type of components can be modelled by means of an exponential distribution,

a) Calculate the percentage of manufactured components that will be operative for more than 5 years (define the random variable under study, its distribution and parameters). *(2.5 points)*

b) If one component has been operative for one year, what is the probability to work correctly for 5 additional years (that is, for more than 6 years in total)? *(2.5 points)*

c) Calculate the reliability of device **D** (shown below) after 5 years of operation considering that C1, C2 and C3 are electronic components of type C (described above) with an independent performance. Define and describe the random variables under study, and the concept of reliability in this case. *(5 points)*



3. One company manufactures motherboards and sells them in big sets of several thousands of units. The company wants to ensure that the proportion of defective motherboards manufactured is less than 1%. For this purpose, a sampling plan is established consisting of selecting at random N motherboards and accepting the set if there are less than 2 defective ones.

What is the minimum value of N so that the probability of rejecting a set that does not satisfy the desired requirement is greater than 99%? *(10 points)*

4. For a certain computer program, the time of compilation fluctuates uniformly between 20 and 30 seconds. Indicate the random variable and answer the following questions:

a) What is the probability of the program to compile in less than 25 seconds? *(2 points)*

b) If 10 programs are compiled sequentially, what is the probability to get a total running time lower than 4,5 minutes? *(4 points)*

c) If the total running time is expressed in minutes, what is the variance? What would be the units of this variance? *(4 points)*

SOLUTION:

1a) The population is the set of searches that can be carried out with the modified program. One individual is each one of the possible search that can be performed on the database with the program.

1b) The first quartile is 18.1 (left edge of the box), that corresponds to the fourth lowest value.
By definition, range = maximum - minimum = 24 - 8.2 = 15.8

1c) The name is box-whisker plot.

1d) We cannot affirm that the values of time follow a negatively skewed distribution in the population. The right whisker is shorter than the left one, which suggests a negatively skewed distribution if the median is positioned to the right. But this is not the case (median = 20.3), which implies that there is not enough evidence to affirm that the distribution is skewed. Assuming that the distribution is Normal, the value 8.2 is very separated from the end of the left whisker, which suggests that it is an outlier that should be discarded. The value of the skewness coefficient is strongly affected by this outlier, and therefore this coefficient is not useful in this case to discuss the pattern of data distribution.

1e) Given the presence of an outlier as discussed in the previous section, the median (20.3) is more representative than the average (19.54) as parameter of position in this case. If the outlier is discarded, the average is 20.48, which is also an appropriate parameter of position.

1f) Data can be represented on a normal probability plot, which is more useful than the box-whisker plot to study the data normality in the case of few values and in order to discuss the presence of abnormal values.

2a) variable **T** = {Duration (years) of one component C} ~ Exp(α)
$m=1/\alpha \rightarrow \alpha = 1/m = ¼$; $P(T>5) = e^{-5/4} = 0{,}287 \rightarrow$ **28.7%**

2b) By applying the lack-of-memory property of the exponential distribution:
P(T>6/T>1) = P(T > 6-1) = P(T > 5) = $e^{-5/4}$ =**0.287**.

2c) random variables:
T1={Duration (years) of component C1}
T2={Duration (years) of component C2}
T3={Duration (years) of component C3}
TD={Duration (years) of component D}
T1, T2 and T3 follow and exponential distribution of parameter $\alpha = ¼$

The **reliability at 5 years** is defined as: **P(TD>5)**

- **P(TD>5) = P[((T1>5)∪(T2>5))∩(T3>5)]** = P[(T1>5)∪(T2>5)] P(T3>5) =

= [P(T1>5) + P(T2>5) – P((T1>5)∩(T2>5))] P(T3>5) =
= [P(T1>5) + P(T2>5) – P(T1>5) P(T2>5)] P(T3>5)

- P(T1>5) = P(T2>5) = P(T3>5) = 0.287 (section 2a)

- **P(TD>5)** = [P(T1>5) + P(T2>5) – P(T1>5) P(T2>5)] P(T3>5) = (0.287 + 0.287 - $0{,}287^2$) 0.287 = **0.141**

3a) X: number of defective motherboards in the sample of size N.
X follows a Binomial distribution with parameters N and p≥0.01.
It can be approximated by means of a Poisson distribution with $\lambda=N\cdot 0.01$
P(X>1)>0.99 so that: P(X≤1)≤0.01
Using the Poisson abacus considering a probability of <0.01, it turns out that the parameter λ has to be greater than 6.5. Thus, N≥6.5/0.01 = **650**

4a) X: time of program compilation in seconds, X ~ U (20, 30)
P(X < 25) = 0.5 because 25 is the median (average value of 20 and 30).

4b) E(X) = 25, Var(X) = $(b-a)^2/12 = 10^2/12 = 8.33$
Y= time of compilation of 10 sequential programs, in seconds:
$Y = X_1 + ... + X_{10}$; Y ~ N ($m=10\cdot 25$, $\sigma^2=10\cdot 8.33$) ; Y ~ N (250, $\sqrt{83.3}$)
4.5 minutes = 270 seconds
P(Y< 270) = P(Z < (270-250)/$\sqrt{83.3}$) = P(Z < 2.19) = 1 – P(Z > 2.19) =
= 1 – 0.0143 = **0.9857**

4c) Z= time of compilation of 10 sequential programs, in minutes
Z = Y / 60 = (1/60)·Y
Var (Z) = $(1/60)^2\cdot$var(Y) = $83.3/60^2$ = **0.02315**
The variance has squared units, and in this case the units are: $minutes^2$.