Bachelor Degree in Computer Engineering

Statistics **group E (English)**

FIRST PARTIAL EXAM

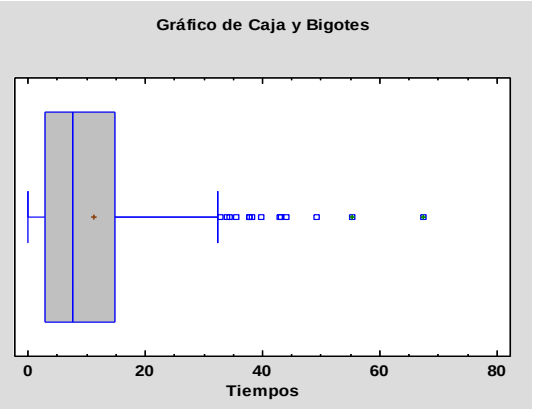
March 26th 2019

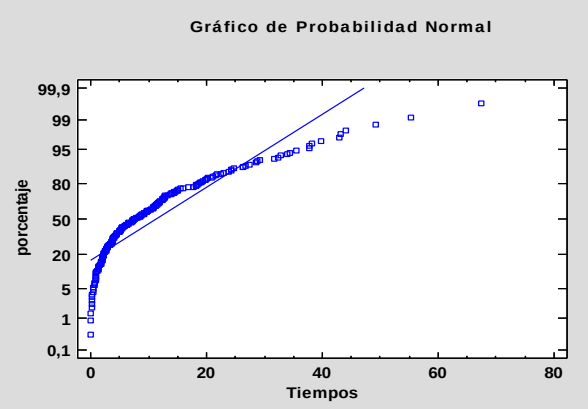
Surname, name	
Signature	

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 of operation until breaking (in months) of 200 components are represented in the two following graphs:





Answer the following questions, conveniently justifying the reply.

a) Indicate if the following statement is true or false (justify your answer): "*The random variable under study is defined in the population of the different months in which the time of operation until breaking is measured*". *(2.5 points)*

b) Generally speaking, what is the utility of each of the plots represented? *(3 points)*

c) If data are expressed in other units, does the value of position parameters and/or dispersion parameters change? *(2 points)*

d) Discuss the main conclusions derived from both plots regarding the pattern of variability observed in the data. *(2.5 points)*

2. Students enrolled in a certain degree at UPV have laptops of brands HPH, ADER and SANY (each student has only one laptop). It was checked that 30% are HPH and 15% are ADER. After making the appropriate inquiries, it has been obtained that 20% of HPH laptops, 30% of ADER and 25% of SANY laptops were purchased in an online store.

a) Define the events required to solve the calculations of the next sections. What events correspond to each of the percentages indicated in the statement? *(1 point)*

b) If a laptop is chosen at random, what is the probability to have been purchased in the online store? *(2.5 points)*

c) If a laptop is chosen at random and it turns out that it was purchased in the online store, what is the probability to be HPH? *(3 points)*

d) Taking into account the above information, if a class with 30 students is considered, what is the probability that 3 of them bought their laptop in the online store? Define the random variable under study, indicate the type of statistical distribution and the value of its parameters. *(3.5 points)*

3. The channel "Xes Music Official" of the YouTube platform receives 8 visits every hour on average. Assuming that the visits received at different hours of the day are independent of each other, answer the following questions:

a) Define the random variable under study, indicate the model of distribution and the value of its parameters. Over what population is defined this variable? *(2.5 points)*

b) Knowing that 8 is the average value of this variable, what is the probability to receive exactly 8 visits in one hour randomly chosen? *(2.5 points)*

c) Taking into account that the time of each visit follows an exponential distribution of median 3 minutes:

c.1) Calculate the percentage of visits with a time comprised between 3 and 5 minutes. *(3 points)*

c.2) If a visit randomly chosen has already taken two minutes, what is the probability to finish before two additional minutes? *(2 points)*

4. Some tourists who have come several days to Valencia's Fallas festival go out for a walk in the morning, afternoon and at night. In the morning, they walk an average of 4.6 km with a standard deviation (σ) of 0.8 km; in the afternoon they walk 7.8 km on average with σ = 0.6 km; and at night they walk 4.4 km on average with σ = 1 km. It is assumed that the three random variables are normally distributed and that distances walked are independent of each other.

a) The following random variable is considered: "*total distance traveled by walk in one day (morning, afternoon and night) by one tourist*". What is the distribution of this variable? Obtain the value of its parameters. *(3 points)*

b) If a tourist is randomly chosen, what is the probability to have travelled by walk more than 17 km in one day? *(3 points)*

c) From what length could be considered that a tourist, during one day, has performed an abnormally long walk (only exceeded in 3 cases per thousand)? *(4 points)*

SOLUTION

1a) The statement is false, because the individuals of this population are the components, not the months. The correct statement would be: *"The random variable under study is defined in the population of the different components in which the time of operation until breaking is measured".*

1b) Utility of both plots: they are graphical tools of descriptive statistics that reflect visually the pattern of data variability and provide information about the data dispersion, position parameters and degree of symmetry (skewness) of the distribution.

Utility of the box & whisker plot: in addition, it visualizes extreme values, some of which might be outliers (anomalous data) that should be discarded. It is possibly the most useful graphical tool to study the degree of skewness of a distribution from a reduced set of data. Furthermore, it provides an easy identification of quartiles and allows the comparison of two sets of data by means of a multiple box-whisker plot.

Usefulness of normal probability plot: its main advantage is to discuss the normality of a data set, i.e., if data can be considered as a sample randomly taken from a population with normal distribution. In that case, it is possibly the most powerful graphical tool to study the presence of anomalous values (outliers) that do not belong to the population and that might be discarded.

1c) If data are expressed in other units (which implies to be multiplied by a constant k), the value of all position parameters will change in a proportional way, being multiplied by k. The dispersion parameters will also change except the coefficient of variation: the standard deviation, range and interquartile range are multiplied by k, while the variance by k^2. Obviously, a parameter will not change if it is zero (for example, the minimum).

1d) The main conclusions derived from both plots are the following:
- There is a marked positive skewness in the data distribution. Moreover, as the variable in this case measures time, being zero the minimum value, an exponential distribution with median 7.5 might be appropriate for modeling the pattern of data variability.
- Regarding the position parameters: median = 7.5, mean = 12, first quartile = 3, third quartile = 15 (approximately).
- Regarding the dispersion: data are comprised between 0 and 67 (range = 67) being the interquartile range: 15 - 3 = 12 (approximately).
- The box-whisker plot shows about 12 extreme values that appear outside the right whisker. There is not enough evidence to claim that these values are outliers that should be discarded because they do not belong to the population.

2a) Event A: the laptop is of brand HPH; B: the laptop is of brand ADER.
C: the laptop is of brand SANY; T: the laptop was purchased in an online store. The percentages correspond to the following probabilities:
0.3 = P(A); 0.15 = P(B); 0.2 = P(T/A); 0.3 = P(T/B); 0.25 = P(T/C).

2b) P(C)=1-0.3-0.15= 0.55; By applying the total probability theorem:
$P(T) = P(A)\cdot P(T/A) + P(B)\cdot P(T/B) + P(C)\cdot P(T/C) = 0.3\cdot 0.2 + 0.15\cdot 0.3 + 0.55\cdot 0.25 =$ **0.2425**

2c) By applying the Bayes' theorem, it turns out that:

$$P(A/T) = \frac{P(A) \cdot P(T/A)}{P(T)} = \frac{0.3 \cdot 0.2}{0.2425} = 0.2474$$

2d) Random variable X: *number of students who have purchased their laptop in the online store out of a class of 30 students*. The maximum value of this discrete variable is 30, which implies that it follows a Binomial distribution, being the parameters n = 30 and p = 0.2425 (probability obtained in section 2b). The calculation requested is:

$$P(X = 3) = \binom{30}{3} \cdot 0.2425^3 \cdot (1 - 0.2425)^{27} = 4060 \cdot 0.0143 \cdot 0.7575^{27} = \mathbf{0.0321}$$

$$\binom{30}{3} = \frac{30!}{3! \cdot 27!} = \frac{30 \cdot 29 \cdot 28 \cdot 27!}{6 \cdot 27!} = 5 \cdot 29 \cdot 28 = 4060$$

3a) Random variable X: *number of visits received on the "Xes Music Official" channel in one hour*. In this discrete variable, the minimum value is zero and the maximum is not bounded; hence, it can be modeled according to a Poisson-type distribution of parameter $\lambda = 8$ (which is the average value). Since one value is available every hour, the individuals of this population are "hours": the population would be the whole set of hours in which the number of visits received on this channel is measured.

3b) By applying the probability function of the Poisson distribution, it turns out that:

$P(X=8) = e^{-8} \cdot 8^8 / 8! =$ **0.1396**

3c1) Considering T as the random variable, with a median = 3: $P(T>3) = 0.5 = e^{-\alpha \cdot 3}$; By solving this equation: $\alpha = -(\ln 0.5)/3 = 0.231$.

$P(3<T<5) = P(T>3) - P(T>5) = 0.5 - e^{-0.231 \cdot 5} = 0.5 - 0.315 = 0.185 =$ **18.5%**

3c2) If the visit has already taken 2 min, it implies that its duration will finally be greater than 2 (T>2). The requested calculation is about a conditional probability, which can be solved by applying the lack-of-memory property (*l.m.p.*) of the exponential distribution:

$P(T<4 / T>2) = 1 - P(T>4 / T>2) =$ (*lmp*) $= 1 - P(T>2) = 1 - e^{-0.231 \cdot 2} = 1 - 0.63 =$ **0.37**

4a) X_1, X_2 and X_3 are the distances travelled by walk (km) in the morning, afternoon and night, respectively, so that: $X_1 \approx N(4.6; 0.8)$; $X_2 \approx N(7.8; 0.6)$; $X_3 \approx N(4.4; 1)$.
The total distance D walked in one day will be: $D = X_1+X_2+X_3$, which will follow a normal distribution because it is obtained by summing three independent normal variables. The parameters of this distribution will be:
Average: $m_D = m_{X1} + m_{X2} + m_{X3} = 4.6 + 7.8 + 4.4 = 16.8$
As they are independent, $variance_D = var_{X1} + var_{X2} + var_{X3} = 0.8^2 + 0.6^2 + 1^2 = 2$
Standard deviation = square root of 2 = 1.414. Therefore, **$D \approx N(16.8; 1.414)$**;

4b) $P(D>17) = P[N(16.8; 1.414) >17] = P[N(0;1) > (17-16.8)/1.414] =$
$= P[N(0;1) > 0.14] =$ **0.444**

4c) The length *k* requested should satisfy: $P(D>k) = 0.003$;
$P[N(16.8; 1.414) >k] = 0.003$; $P[N(0;1) > (k-16.8)/1.414] = 0.003$.
Using the table of the N(0;1) distribution, it turns out that: $(k-16.8)/1.414 = 2.75$. Therefore, the value requested is: $k =$ **20.69 km**