Bachelor Degree in Computer Engineering

Statistics group E (English)

FIRST PARTIAL EXAM

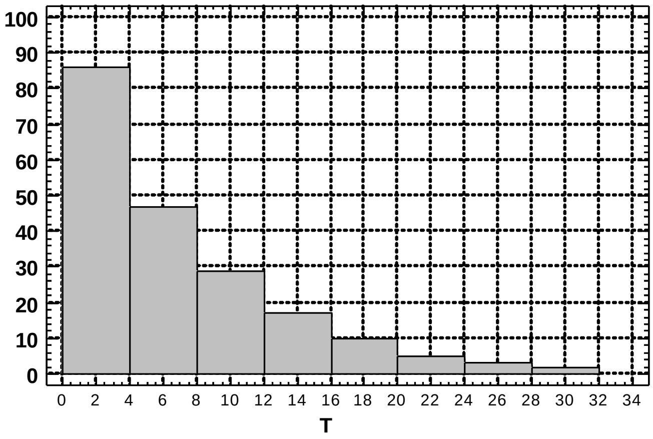
April 11th 2018

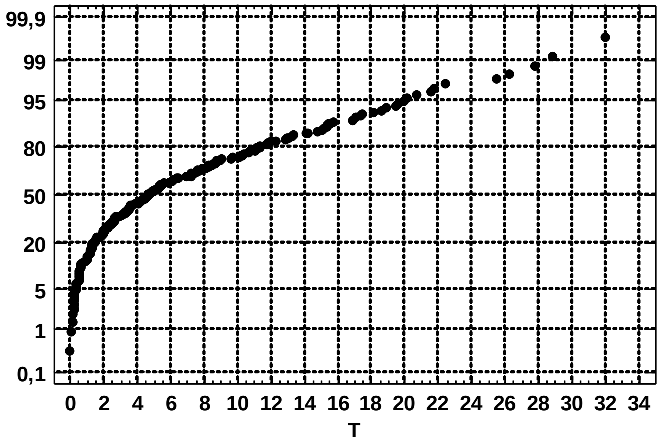
Surname, name	
Signature	

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 taken by an algorithm to process certain data matrices is a random variable. We have obtained 200 values of time (in milliseconds, ms) corresponding to a random sample of 200 matrices of very different size. The following graphs have been constructed using these values:





Answer the following questions by justifying conveniently the reply.

a) The graph on the right is a normal probability plot. What is the name of the graph on the left? For each graph, indicate its usefulness and what is represented in each of the axes. *(4 points)*

b) According to the plots shown above, indicate which theoretical model of probability distribution, among those studied, would properly describe the performance of the time taken by an algorithm to process a data matrix. *(3 points)*

c) The person responsible for quality control considers that the average is the most representative parameter to describe the position of these data, and the interquartile range is the most representative to describe their dispersion. Do you agree with this statement? *(3 points)*

2. Certain factory produces 10,000 containers per day. Machine A produces 3,000 of these containers, 1% of which are defective. The rest are produced by machine B, and it is known that 98% of them are of good quality.

a) Define the events involved in this problem. *(1 point)*

b) If one container is randomly selected, what is the probability of being defective? *(4 points)*

c) If one container is taken at random and it turns out to be of good quality, what is the probability of having been produced by machine A? Repeat the calculation for machine B. *(5 points)*

3. Certain universal USB chargers can be packaged in boxes of 5 or 10 units. It is known that 2% of the USB chargers are defective.

a) If a company buys one box with 5 universal USB chargers and another box with 10 units, calculate the probability to find, in total, less than two defective chargers. *(4 points)*

b) Define the random variable involved in the previous question. Calculate the mathematical expectation of that variable. *(2 points)*

c) The same company buys wires for the chargers, which present an average number of 0.5 esthetic defects per meter. If each charger needs one wire of two meters, what is the probability to find at least two esthetic defects in the wire of a charger randomly taken by the company? *(4 points)*

4. The time of operation until failure of certain type of electronic components, measured in hours, follows an exponential distribution with a median of 69 hours.

a) If a component is still in operation after 100 hours, what is the probability to be operative, in total, more than 250 hours? *(3 points)*

b) One device is formed by 10 components of this type. The device is assembled in such a way that the first component begins to work until it fails; at that moment, the next component becomes operative automatically, and so on until the 10th component fails. At that moment, the device stops working.

b.1) How is the duration (time of operation) of the device defined in this case? Calculate the mean and variance of this random variable. *(3 points)*

b.2) What is approximately the probability that the device is operative less than 1,700 hours? *(4 points)*

5. The shipping company GLOBAL LINE Inc. has a fleet of cargo ships to transport containers by different routes. The company is analyzing the fuel consumption according to the route. The consumption is determined by the speed and rout length based on the following expression:

$$\text{Consumption} = 0.2 \cdot \text{speed} + 0.8 \cdot \text{route\_length}$$

Moreover, the shipping company has estimated that the speed follows a normal model with average 100 units and standard deviation 10. The route length is modeled as a normal with average 5,000 and standard deviation 200. Assuming that the variables *speed* and *route length* are independents, calculate:

a) If one route is randomly chosen, what is the probability to have a consumption greater than 4,340 units? *(7 points)*

b) Calculate the second quartile of the random variable *Consumption*. *(3 points)*

SOLUTION

1a) The graph on the left is called frequency histogram.
- In a histogram, the vertical axis represents in this case the absolute frequency, i.e., the number of values comprised in each interval of the values of time.
- In a normal probability plot, the vertical axis represents the cumulative relative frequency (%): 100·P(X≤x), i.e., the percentage of values lower or equal to each value indicated in the horizontal axis.
- In both plots, the horizontal axis represents the values of the random variable T: time required to process a matrix.
- Usefulness of the histogram: it allows a graphical visualization of the pattern of data variability, so that we can get an approximate idea about the minimum, maximum, mode and median. It allows us to determine if the distribution is symmetric or skewed, and to detect abnormal patterns in the distribution, such as abnormal values, bimodal distributions, truncated data, etc.
- Usefulness of the normal probability plot: if data can be fitted approximately to a straight line, it can be concluded that the sample has been extracted from a normal population. The normal probability plot is more powerful than the histogram for this purpose as well as to detect outliers.

1b) This histogram indicates that the most frequent values are close to zero, and the frequency decreases progressively. This pattern corresponds to the exponential distribution. Thus, variable T can be modelled according to an exponential distribution with median = 5 (percentile 50, obtained from the normal probability plot).

1c) It is not correct to affirm that the average is the most representative parameter of position because it is influenced by extreme values, given the strong skewness of the distribution. It would be more convenient to use the median because it is not affected by extreme values. However, it is correct to say that the interquartile range is the most representative parameter of dispersion, because it indicates the range comprising 50% of the values, and it is not affected by the highest values.

2a) Event A: the container has been produced by machine A.
Event B: the container has been produced by machine B.
Event D: the container is defective.

2b) P(A) = 3,000/10,000 = 0.3; P(B) = 1 - 0.3 = 0.7
P(D/A) = 0.01; P(D/B) = 1 - 0.98 = 0.02
Aplicando el teorema de la probabilidad total:
P(D) = P(A)·P(D/A) + P(B)·P(D/B) = 0.3·0.01 + 0.7 · 0.02 = **0.017**

2c) The probability is obtained by applying Bayes' theorem:

$$P\left(A/\overline{D}\right)=\frac{P(A)\cdot P\left(\overline{D}/A\right)}{P\left(\overline{D}\right)}=\frac{P(A)\cdot[1-P(D/A)]}{1-P(D)}=\frac{0.3\cdot(1-0.01)}{1-0.017}=0.3021$$

$$P\left(B/\overline{D}\right)=\frac{P(B)\cdot P\left(\overline{D}/B\right)}{P\left(\overline{D}\right)}=\frac{P(B)\cdot[1-P(D/B)]}{1-P(D)}=\frac{0.7\cdot(1-0.02)}{1-0.017}=0.698$$

Alternative method: $P\left(B/\overline{D}\right)=1-P\left(A/\overline{D}\right)=1-0.3021=0.698$

3a) Taking a box with 5 chargers, another box with 10 and, next, counting the number of defective chargers in the set, it is equivalent to taking directly a sample of 15 chargers. The random variable X (number of defective chargers in the sample of 15) will follow a Binomial distribution with n=15 and p=0.02.

$$P(X<2)=P(X=0)+P(X=1)=\binom{15}{0}\cdot 0.02^0\cdot 0.98^{15}+\binom{15}{1}\cdot 0.02\cdot 0.98^{14}=$$

$$=0.98^{15}+15\cdot 0.02\cdot 0.98^{14}=0.7386+0.2261=0.965$$

Alternative method: it can be approximated to a Poisson distribution because a similar result is obtained.

$$X\approx Ps(\lambda=15\cdot 0.02=0.3)\ ;\quad P(X<2)=e^{-0.3}\cdot\frac{0.3^0}{0!}+e^{-0.3}\cdot\frac{0.3^1}{1!}=e^{-0.3}(1+0.3)=0.963$$

3b) Random variable X: number of universal USB chargers that are defective in the total sample of 15 chargers.
The mathematical expectation of a random variable is the average value, which is computed in a Binomial distribution as: E(X) = m = n·p = 15·0.02 = **0.3**.

3c) Variable Y: number of esthetic defects in a one-meter wire. The minimum value is zero and there is no maximum. Thus, Y will follow a Poisson model with a parameter λ coincident with the average: Ps(λ=0.5).
Variable Z: number of esthetic defects in a two-meter wire: $Z=Y_1+Y_2$ (i.e., the defects found in the first meter and in the second meter of wire). Z will follow a distribution Ps $(\lambda_1+\lambda_2)$, which is: Ps (λ=1).

$$P(Z\geq 2)=1-P(Z=0)-P(Z=1)=1-e^{-1}\cdot\frac{1^0}{0!}-e^{-1}\cdot\frac{1^1}{1!}=1-e^{-1}(1+1)=0.264$$

Alternative method: P(Z≥2) = 1-P(Z≤1), which is obtained from the Poisson abacus, with λ=1, reading on curve “1”.

4a) P(T>69)=0.5; $e^{-\alpha\cdot 69}=0.5$; -69α = ln(0.5); α=-(ln0.5)/69 = 0.01005
By using the lack-of-memory property of the exponential distribution:

$$P[(T>250)/(T>100)]=P(T>150)=e^{-0.01\cdot 150}=\mathbf{0.222}$$

4b1) Variable T_i: time of operation of component *i*.
Average: E(T)=1/α=1/0.01005 = 99.546. It is coincident with the std. deviation.
The total duration of the device is obtained by summing the time of operation of the 10 components: $T_{total}=T_1+T_2+...+T_{10}$ (they are independent).
Average = $E(T_1+...+T_{10})=E(T_1)+...+E(T_{10})=10\cdot E(T_i)=10\cdot 99.55=$ **995.5** hours

$$\sigma^2(T_{total})=\sigma^2(T_1+...+T_{10})=\sigma^2(T_1)+...+\sigma^2(T_{10})=10\cdot\sigma^2(T_i)=10\cdot 99.55^2=99094\ \text{h}^2$$

The standard deviation is the square root: 314.8 hours.

4b2) As T_{total} is obtained by summing 10 random variables, it will follow approximately a normal distribution according to the central limit theorem.

$$P(T_{total}<1700)=P[N(995.5;\ 314.8)<1700]=P[N(0;\ 1)<(1700-995.5)/314.8]=$$

$$=P[N(0;\ 1)<2.238]=1-0.0126=0.987$$

5a) Speed ~ Normal ($m=100$ $\sigma=10$); Route_length ~ N ($m=5000$ $\sigma=200$)
C (fuel consumption for any route) = 0.2 speed + 0.8 route_length;
C is linear combination of normal variables and, hence, it will also be normal.
$E(C) = 0.2\ m_{speed} + 0.8\ m_{route\_length} = 0.2\cdot100 + 0.8\cdot5000 = 4020$ units
Variance of C = $0.2^2 \cdot 10^2 + 0.8^2 \cdot 200^2 = 25604 \Rightarrow \sigma_{Consumption} = 160$
$P(C>4340) = P[N(4020; 160) > 4340] = P[N(0;1) > (4340-4020)/160] =$
$= P[N(0;1) > 2] =$ **0.0228**

5b) $P(C<Q_2)=0.50 \Rightarrow$ As C is a normal distribution, the median (second quartile) will be coincident with the average. Thus, $Q_2=$ **4020** units.

b) Calculate the second quartile of the random variable *Consumption*.

(3 points)