Bachelor Degree in Computer Engineering

Statistics group E (English)

FIRST PARTIAL EXAM

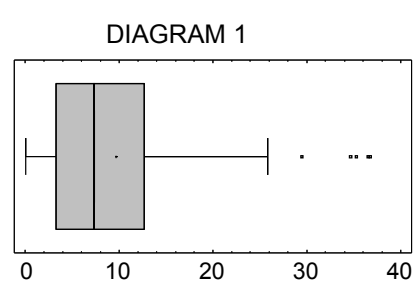
April 7st 2017

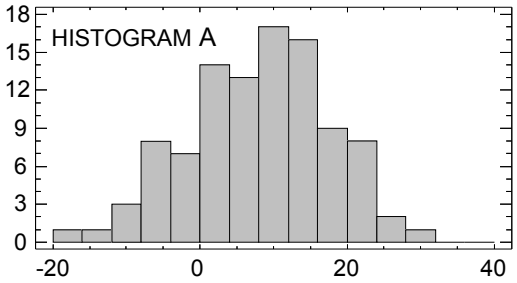
Surname, name	
Signature	

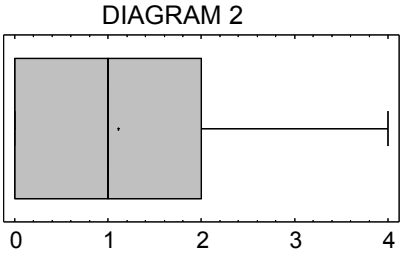
Instructions

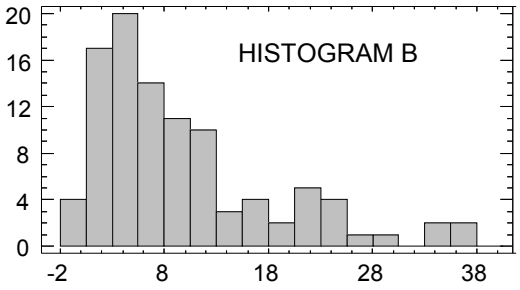
1. Write your name and sign in this page.
2. Answer each question in the corresponding page.
3. All answers must be justified.
4. Personal notes in the formula tables will not be allowed.
5. Mobile phones are not permitted over the table. It is only permitted to have the DNI (identification document), 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**.

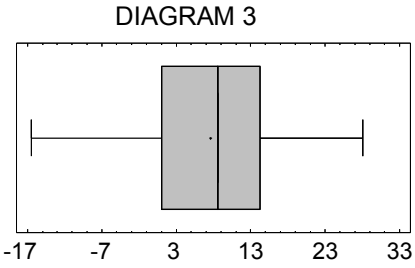
1. A descriptive study has been carried out with three data samples of **size 100** corresponding to different types of random variables. Two plots were obtained from each dataset: a frequency histogram and a box-whisker plot, which are shown below in a random position.

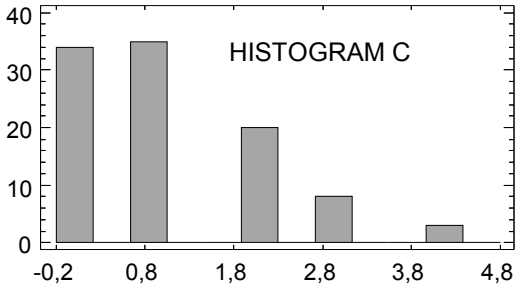












Apart from these plots, the standardized skewness coefficient (SSC) and the standardized kurtosis coefficient (SKC) for each datasets are also available, which are shown in the table below.

a) For each row in the table, select one of the three options of diagram and histogram. Indicate also the correspondence between the plots above by means of arrows. Justify conveniently your answer. *(6 points)*

Std. skewness and kurtosis coef.	DIAGRAM	HISTOGRAM
SSC = 3.29556, SKC = 0.13803	1 - 2 - 3	A - B - C
SSC = -1.08673, SKC = -0.83587	1 - 2 - 3	A - B - C
SSC = 5.51562, SKC = 2.81815	1 - 2 - 3	A - B - C

b) Indicate what would be the most representative parameters to describe the position and dispersion of the dataset represented by the Box & Whisker plot coded as "3". Calculate approximately, if possible, their value. Justify all your answers. *(4 points)*

2. In a hospital section, a clinical test is carried out to detect certain congenital pathology by means of a DNA chip whose data are analyzed using bioinformatics methods. Based on historical data, it is known that this disease is suffered by 4% of patients admitted to this hospital section. If a patient really suffers from this disease, the test gives a positive result in 95% of cases. But if the patient does not suffer the disease, the test is positive (i.e., indicates the presence of the pathology when it is not really true) in 2% of cases.

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

b) Calculate the percentage of cases in which the test gives a positive result. *(3 points)*

c) Calculate the percentage of cases in which the test gives an erroneous diagnosis. *(3 points)*

d) If the test is carried out with a patient and it gives a positive result, what is the probability that the patient actually suffers the pathology? *(3 points)*

3. Certain company that develops software applications has stated that its programmers commit on average 2 errors per page of code.

a) Calculate the probability to commit more than 6 errors in one page of code. *(4 points)*

b) If certain application contains 50 pages of code, what is the probability to find at least one page with more than 6 errors? *(6 points)*

4. Certain company manufactures a given model of diagnostic equipment based on NMR (nuclear magnetic resonance) which is used at hospitals. The sale price is negotiated with each client, and can be modelled as a random variable following a normal distribution, with an average of 500,000 € and a standard deviation of 80,000 €. The manufacturing cost of each equipment is comprised by a fixed term of 200,000 € and a variable cost that fluctuates as a normal model, with a mean of 100,000 € and a standard deviation of 20,000 €. It is assumed that the two variables are independent.

a) Indicate the type of distribution of the random variable "total manufacturing cost" and calculate the value of its parameters. *(3 points)*

b) Indicate the type of distribution of the random variable "profit" (calculated as sale price minus total manufacturing cost), indicating the value of its parameters. *(4 points)*

c) Calculate the probability to obtain a profit in a sale of an NMR equipment below 230,000 €. *(3 points)*

5. The time required by a user to perform a query on a library server fluctuates exponentially. It is known that 25% of the queries take less than 5 minutes to be performed.

a) What is the probability that a user requires exactly 5 minutes to perform a query? *(1 point)*

b) Calculate the average time required to attend a query. *(3 points)*

c) It turns out that a new user arrives and finds that the server is occupied by someone else who has been using it for 7 minutes. Calculate the probability to have to wait more than five additional minutes until the server becomes available. *(2 points)*

d) During one morning, 20 consecutive queries are carried out on the server. Calculate the probability that their total duration exceeds 4 hours. *(4 points)*

SOLUTION

1a) The correspondence between coefficients, diagrams and histograms is indicated below:

Std. skewness and kurtosis coef.:	DIAGRAM	HISTOGRAM
SSC = 3.29556, SKC = 0.13803	2	C
SSC = -1.08673, SKC = -0.83587	3	A
SSC = 5.51562, SKC = 2.81815	1	B

CAE = -1.09; SKC = -0.84: As both are comprised between -2 and 2, they can be considered null at the population level, which suggests a sample taken from a normal distribution. It corresponds to the most symmetric plots: **diagram 3 and histogram A**.

Diagram 1 is linked with **histogram B** because the range of values is coincident. The former shows the highest skewness among all plots (SSC = 5.52) because the mean is "quite close" to the 3rd quartile, and some isolated points appear, which justifies the highest value of SKC (2.81).

The dataset corresponding to **diagram 2 and histogram C** is comprised by discrete values from 0 to 4 and also reflects a positive skew (SSC = 3.29 > 2), though not so pronounced as in diagram 1, as the mean is similar to the median. This reason and the lack of extreme values explain why SKC is close to zero.

1b) As the dataset represented by the box-whisker plot number 3 can be considered as a sample taken from a normal population, the most representative parameters to describe the **position** are the mean (m = 7.5 which is the point appearing inside the box), which is nearly coincident with the median (8.5: vertical line inside the box).

In a normal distribution, the most representative parameters of **dispersion** are the standard deviation (s) and the variance (s^2).

a) Calculation of the standard deviation: There are 2 values below -12 and 3 values are higher than 24. As histogram A has been obtained with 100 values, this interval [-12; 24] comprises 95% of the values and corresponds to $m \pm 2 \cdot s$ assuming a normal distribution. This interval spans 4 times the standard deviation (2s below m, and 2s above m) and, hence, **s ≈** (24+12)/4 = **9**.

b) Calculation of the variance: $s^2 = 9^2 =$ **81** approximately.

Comment: In a normal distribution, the interquartile range is less representative than s or s^2. The value is: IQR = Q_3 - Q_1 = 14,5-1 = 13,5.

2a) Event A: the patient admitted to this hospital section suffers from the congenital pathology. Event $\overline{A}$: the patient does not suffer from the pathology. Event B: the test gives a positive result. The probabilities mentioned in the statement are: P(A) = 0.04; P(B/A) = 0.95; P(B/$\overline{A}$) = 0.02.

2b) By applying the total probability theorem, it turns out that:

$$P(B) = P(A)\cdot P(B/A) + P(\overline{A})\cdot P(B/\overline{A}) = 0.04\cdot 0.95 + 0.96\cdot 0.02 = 0.0572 \rightarrow \mathbf{5.72\%}$$

2c) The diagnosis can be erroneous because the patient does not suffer from the pathology **and** the test gives a positive result, **or** because the patient suffers from the pathology **and** the test gives a negative result. Therefore:

P (erroneous diag.) = $P\left[(\overline{A}\cap B)\cup(A\cap\overline{B})\right] = P(\overline{A}\cap B) + P(A\cap\overline{B}) =$

$= P(\overline{A})P(B/\overline{A}) + P(A)\cdot P(\overline{B}/A) = 0.96\cdot 0.02 + 0.04\cdot 0.05 = 0.0212 \rightarrow \mathbf{2.12\%}$

2d) Bayes' theorem: $P(A/B) = \dfrac{P(A\cap B)}{P(B)} = \dfrac{P(A)\cdot P(B/A)}{P(B)} = \dfrac{0.04\cdot 0.95}{0.0572} = 0.664$

3a) Random variable X: number of errors committed per page of code. This variable follows a Poisson distribution, X ≈ Ps (λ=2), because it ranges from 0 to infinite, and the parameter λ is equal to the average.

$P(X > 6) = 1 - P(X \le 6) = 1 - 0.9955 = \mathbf{0.0045}$

The value 0.9955 is obtained from the Poisson abacus, by using a vertical line at λ=2, cutting the curve "6" and reading the probability in the vertical axis.

3b) Random variable Y: number of code pages that satisfy the condition "to have more than 6 errors" in one software application with 50 code pages. The probability of this condition is 0.0045 (previous section). Thus, Y follows a distribution: Binomial (n=50, p=0.0045) because the minimum value is 0 and the maximum is 50. By applying the probability function, it turns out:

$$P(Y \ge 1) = 1 - P(Y = 0) = 1 - \binom{50}{0}\cdot 0.0045^0 \cdot 0.9955^{50} = 1 - 0.9955^{50} = \mathbf{0.203}$$

4a) TMC (total manufacturing cost) = FC (fixed cost) + VC (variable cost)

FC = 200,000; VC ≈ N (m=100,000; σ=20,000).

The fixed cost increases the average but does not affect the dispersion. Thus:

TMC ≈ N (m=300,000; σ=20,000) → The total manufacturing cost follows a normal distribution with average 300,000 € and standard deviation 20,000 €.

4b) SP (sale price) ≈ N (m=500,000; σ=80,000).

Profit = SP - TMC. It will follow a normal distribution because it is calculated as the difference of two normal variables. Assuming that SP and TMC are independent variables, the parameters of the distribution will be:

E(profit) = E(SP) - E(TMC) = 500,000 - 300,000 = 200,000

$\sigma^2(profit) = \sigma^2(SP - TMC) = \sigma^2(SP) + \sigma^2(TMC) = 80{,}000^2 + 20{,}000^2 = 6.8\cdot 10^9$

$\sigma(profit) = \sqrt{6.8\cdot 10^9} = 82{,}462 \rightarrow$ **Profit ≈ N (m=200,000, σ = 82,462)**

4c) $P\left[N(200{,}000;\ 82{,}462) < 230{,}000\right] = P\left[N(0;1) < \dfrac{230{,}000 - 200{,}000}{82{,}462}\right] =$

$= P[N(0;1) < 0.364] = 1 - 0.358 = \mathbf{0.642}$

5a) In a continuous random variable (i.e., with many decimals), the probability to find an "exact" value tends to zero.

5b) By applying the distribution function, it turns out:

$P(X<5)=0.25=1-e^{-\alpha\cdot 5};\ e^{-\alpha\cdot 5}=0.75;\ \alpha=-(\ln 0.75)/5=0.0575$

The average value of an exponential distribution is the inverse of the parameter:

$m=1/\alpha=1/0.0575=$ **17.38** minutes.

5c) To calculate the requested conditional probability, we can apply in this case the lack of memory property (l.m.p.) of the exponential distribution. Taking into account that 5 is the value of the first quartile, it turns out that:

$$P\left[(X>12)/(X>7)\right]\xrightarrow{l.m.p.}P(X>5)=\mathbf{0.75}$$

5d) In the exponential distribution, the standard deviation is equal to the mean. By applying the central limit theorem, the sum (Y) of independent exponential variables tends to be normal, so that the average of Y is the sum of the means, and the variance of Y is the sum of the variances:

m=17.38; $\sigma^2=17.38^2=302.1$; $Y=X_1+...+X_{20}$

$$E(Y)=E(X_1+...+X_{20})=E(X_1)+...+E(X_{20})=20\cdot 17.38=347.6$$

$$\sigma^2(Y)=\sigma^2(X_1+...+X_{20})=\sigma^2(X_1)+...+\sigma^2(X_{20})=20\cdot 17.38^2=6041.5$$

$$P(Y>4\cdot 60)=P\left[N\left(347.6;\sqrt{6041.5}\right)>240\right]=P\left[N(0;1)>\frac{240-347.6}{\sqrt{6041.5}}\right]=$$

$$=P[N(0;1)>-1.384]=1-0.0831=\mathbf{0.917}$$