Bachelor Degree in Computer Engineering

Statistics **group E (English)**

FIRST PARTIAL EXAM

May 13th 2013

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. 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.
5. Do not unstaple any page of the exam (do not remove the staple).
6. All questions 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: 2 hours.

1. Justify if the following statements are true or false:

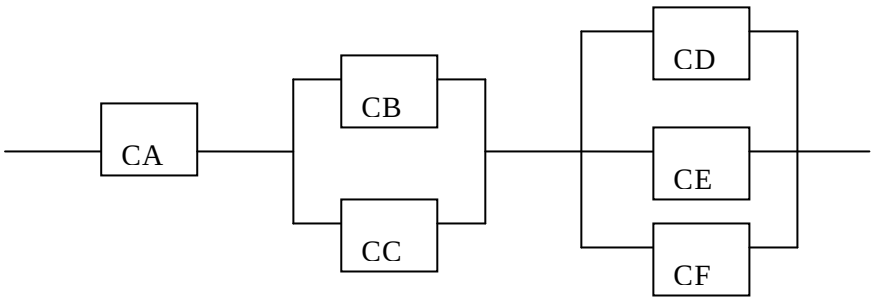
a) The standard deviation is not a useful parameter when the objective of the analysis is to study if a set of data follow a Normal distribution. **(2.5 points)**

b) The discrete variable with two values ("0" and "1") reflecting the situation of computer technicians as unemployed (value "1") or not unemployed (value "0") is defined on the set of all unemployed workers in that sector. **(2.5 points)**

c) Measurements of breaking strength obtained for computer mice manufactured by two computer companies (A and B) constitute a two-dimensional random variable. **(2.5 points)**

d) The third quartile of a data sample provides information about the data dispersion of the sample. **(2.5 points)**

2. One device is formed by six identical components (CA, CB, CC, CD, CE, CF) assembled as indicated in the following diagram:



For each component, the time of operation until failure follows an exponential distribution with a median of 650 hours.

a) If one component has been working for 300 hours, what is the probability to be operative correctly more than 400 hours in total? **(5 points)**

b) Assuming that the 6 components of the device work independently, what is the reliability of this device after 800 hours? Define all variables and events involved in this problem. **(5 points)**

3. The company ELECTONYS uses AA network cards (marketed by Chisco) for the assembly of computer equipment. This company wants to ensure that the percentage of defective cards in each set of purchased cards is less than 10%. To verify this requirement, a sample of N cards is randomly selected and tested for each purchased set. The set is accepted only if the number of defective cards is less than 3. What is the minimum value N so that the probability of accepting a set with 10% or more defective cards does not exceed 5%?

4. In certain library, the time spent by users to conduct book searches follows a Normal distribution with an average of 200 seconds and standard deviation of 30 seconds.

a) If a user performs a book search, what is the probability to spend a time comprised between 170 and 200 seconds? **(5 points)**

b) If the time of book search for 3 users is added ($T_1+T_2+T_3$), what is the probability to exceed 590 seconds? **(5 points)**

5. In certain research about the use of video games according to gender in Spanish Universities, values from 112 students were collected, and the following results were obtained:

	Time spent on video games (hours/week)				
Sample parameters	**N**	$\overline{X}$	**S**	**standard skewness**	**standard kurtosis**
WOMEN	51	3.28	0.78	-0.90	-0.73
MEN	61	5.16	0.84	-0.53	-0.75

According to these results, answer the following questions:

a) Can we admit that the average time of video game usage is the same in men and women (α = 5%)? Use the appropriate TEST. **(3 points)**

b) Can we admit the hypothesis of equal variances? (α=5%) **(3 points)**

c) Justify if the following conclusion can be derived from the research study: "Spanish male university students spend in average 5 hours per week playing video games (confidence level 90%)" **(3 points)**

d) What basic hypotheses must be assumed when applying these inference statistical methods? Justify if such hypotheses are satisfied in this case.
(1 point)

SOLUTION

1a) True: the standard deviation provides information about the data dispersion, not about the shape of the distribution. When the objective is to study if a set of data follow a Normal distribution, the skewness and kurtosis coefficients are useful parameters.

1b) False: this discrete variable is defined on the set of all workers in that sector, not only on the unemployed workers. The population is determined by all workers.

1c) False: the measurements constitute two one-dimensional random variables (not a two-dimensional variable) because we have two populations: computer mice manufactured by company A, and computer mice from company B. The elements of these populations are different.

1d) False: the third quartile (Q3) as well as the first quartile (Q1) are parameters of position, and do not provide information about the data dispersion. The interquartile range (Q3-Q1) is a parameter of dispersion.

2a) Random variable T: time of operation until failure of one component; T~exp(α)
Median = 650 ; P(T>650)=0.5; $e^{-650\alpha} = 0.5$; $\alpha = -\ln(0.5)/650 = 0.001066$
Applying the property of lack of memory (exponential distribution):
$P(T > 400/T > 300) = P[T > (400-300)] = P(T > 100) = e^{-100\cdot 0.001066} = e^{-0.1066} =$ **0.899**

2b) Event C_i: component *i* is operative correctly at 800 h, being i=A, B, ..., F
Event DEV: the device is operative correctly at 800h of operation.
$DEV = CA \cap (CB \cup CC) \cap (CD \cup CE \cup CF)$
Variable T_i: time of operation until failure (hours) of component *i*. T~ exp (α=0.001066)
Variable T_{DEV}: time of operation until failure (hours) of the device.
As all components are of the same type: P(CA) = P(CB) =... = P(CF) = P(C_i)
$P(C_i) = P(T_i > 800) = e^{-800\cdot 0.001066} = 0.426$; $P(\overline{C_i}) = 1 - P(C_i) = 0.574$
As the components work independently:
$P(CB \cup CC) = 1 - P(\overline{CB \cup CC}) = 1 - P(\overline{CB} \cap \overline{CC}) = \xrightarrow{independent} = 1 - P(\overline{CB}) \cdot P(\overline{CC})$
$P(CD \cup CE \cup CF) = 1 - P(\overline{CD \cup CE \cup CF}) = 1 - P(\overline{CD} \cap \overline{CE} \cap \overline{CF}) = 1 - P(\overline{CD}) \cdot P(\overline{CE}) \cdot P(\overline{CF})$

Reliability of the device after 800h = P(T_{DEV}>800) = P(DEV)=
$= P[CA \cap (CB \cup CC) \cap (CD \cup CE \cup CF)] = \xrightarrow{indep.} == P(CA) \cdot P(CB \cup CC) \cdot P(CD \cup CE \cup CF) =$
$= P(CA) \cdot \left\{1 - \left[P(\overline{C_i})\right]^2\right\} \cdot \left\{1 - \left[P(\overline{C_i})\right]^3\right\} = 0.426 \cdot (1 - 0.574^2) \cdot (1 - 0.574^3) =$ **0.232**

3) Random variable X: number of defective network cards in a set of N cards
X~Bi(N, p); The set is accepted if X<3.
P(accept a set with p≥0.1) < 0.05 ; P(X<3 if p≥0.1) < 0.05 ; P(X≤2 if p≥0.1) < 0.05
Assuming the most unfavorable case (p=0.1) and that the binomial distribution can be approximated by a Poisson distribution, with λ= N·p =N·0.1
In the Poisson abacus, reading <0.05 in the vertical axis, this horizontal line crosses the curve "≤2" in the value λ>6.25 (horizontal axis).
λ= N·0.1 > 6.25; N > 62.5; The minimum value that satisfies this condition is **N=63**

Is it correct to assume that the binomial distribution can be approximated by a Poisson? P(X=0) + P(X=1) + P(X=2) < 0.05 ;
$0.9^n + n \cdot 0.1 \cdot 0.9^{n-1} + 0.5 \cdot n \cdot (n-1) \cdot 0.1^2 \cdot 0.9^{n-2} < 0.05$
Solving the equation with Excel, this condition is satisfied for N≥61 which is nearly the same solution. Thus, the approximation was correct in this case.

4a) Random variable X: time spent by one user to conduct a book search.
X~N (m=200; σ=30)

$$P(170 < X < 200) = P(X < 200) - P\left[N(200;30) < 170\right] = 0.5 - P\left[N(0;1) < (170-200)/30\right] =$$
$$= 0.5 - P\left[N(0;1) < -1\right] = \text{(table)} = 0.5 - 0.1587 = \mathbf{0.3413}$$

4a) Variable Y: time of book search for 3 users; $Y = X_1 + X_2 + X_3$; $Y \approx N(m\ ;\ \sigma)$

$$E(Y) = E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = 3 \cdot 200 = 600$$

Assuming that the times of the three users are independent variables:

$$\sigma^2(Y) = \sigma^2(X_1 + X_2 + X_3) = \sigma^2(X_1) + \sigma^2(X_2) + \sigma^2(X_3) = 3 \cdot 30^2 = 2700$$
$$P(Y > 590) = P\left[N(600; \sigma^2 = 2700) > 590\right] = P\left[N(0;1) > (590-600)/\sqrt{2700}\right] =$$
$$= P\left[N(0;1) > -0.1925\right] = 1 - P\left[N(0;1) > 0.1925\right] = (table) = 1 - 0.424 = \mathbf{0.576}$$

5a) The inference test is $H_0 : m_1 = m_2$; $H_1 : m_1 \neq m_2$ being m_1 and m_2 the average time of video game usage in the population of women and men, respectively.

$$S = \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{(n_1-1)+(n_2-1)}} = \sqrt{\frac{0.6084 \cdot 50 + 0.7056 \cdot 60}{50+60}} = 0.81$$
$$S_{\overline{x}_1 - \overline{x}_2} = S\sqrt{(1/n_1) + (1/n_2)} = 0.81 \cdot \sqrt{(1/51) + (1/61)} = 0.15$$
$$t^{\alpha/2}_{(n_1-1)(n_2-1)} = t^{0.025}_{110} = (table) = 1.98$$
$$t_{computed} = (\overline{x}_1 - \overline{x}_2) \Big/ S_{\overline{x}_1 - \overline{x}_2} = (3.28 - 5.16)/0.15 = -12.53$$

The null hypothesis is rejected because $\left|t_{coomp}\right| = 12.53 > 1.98$. Thus, we cannot admit that the average time is the same in men and women, considering α=0.05.

5b) Now the inference test is $H_0 : \sigma_1^2 = \sigma_2^2$; $H_1 : \sigma_1^2 \neq \sigma_2^2$

$P(F_{n_1-1;n_2-1} < f_1) = \alpha/2$; $P(F_{50;60} < f_1) = 0.025$; from the table: $f_1 = 0.58$

$P(F_{n_1-1;n_2-1} > f_2) = \alpha/2$; $P(F_{50;60} > f_2) = 0.025$; from the table: $f_2 = 1.7$

$$IC_{\sigma_1^2/\sigma_2^2} = \left[\frac{s_1^2}{s_2^2 \cdot f_2}; \frac{s_1^2}{s_2^2 \cdot f_1}\right] = \left[\frac{0.6048}{0.7056 \cdot 1.7}; \frac{0.6048}{0.7056 \cdot 0.58}\right] = \left[0.5;\ 1.47\right]$$

As the confidence interval for the ratio of variances contains the value 1, it can be accepted that $\sigma_1^2 / \sigma_2^2 = 1$, which implies that we can admit the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$

5c) Now the inference test is $H_0 : m = 5$; $H_1 : m \neq 5$ The null hypothesis is accepted if:

$\left|\frac{\overline{x} - m}{s/\sqrt{n}}\right| < t^{\alpha/2}_{n-1}$ being $t^{0.1/2}_{61-1} = t^{0.05}_{60} = (table) = 1.671$; $\left|\frac{\overline{x} - m}{s/\sqrt{n}}\right| = \left|\frac{5.16 - 5}{0.84/\sqrt{61}}\right| = 1.488$ which is lower than 1.67. Thus, H_0 is accepted: there is not enough evidence to reject that Spanish male university students spend in average 5 hours per week playing video games, considering α=0.1.

5d) The inference statistical method applied in question 5a) assumes that (i) the variances of both populations are equal, (ii) data are normally distributed and (iii) the two samples are simple independent random samples. The hypothesis of normality is reasonable in this case because the standard skewness and kurtosis belong to the interval [-2, 2].

The inference statistical method applied in question 5b) assumes that (i) data are normally distributed and (iii) the two samples are simple independent random samples. The inference statistical method applied in question 5b) assumes that data are normally distributed.