Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 10th 2013

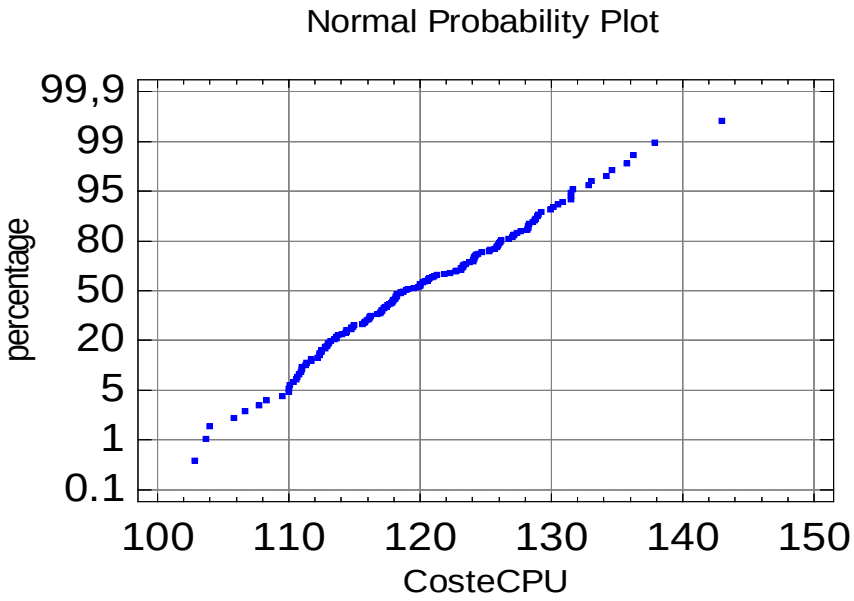
Surname, name:	
Group: **1E**	Signature:
Indicate with a tick mark the partials examined	1st ☐ 2nd ☐

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.
5. **Do not unstaple any page of the exam** (do not remove the staple).
6. The exam consists of 6 questions, 5 ones corresponding to the first partial (80%) and one about the second partial (20%). The lecturer will correct those partial exams indicated by the student with a tick mark in this page. **All questions of each partial exam 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: **3 hours**

Statistics final EXAM
June 10th 2013

1. (1st Partial) Certain bank uses a database with financial information. In order to improve the performance of queries made to this database, it was decided to register for each query the CPU cost (variable *CosteCPU*), which is the time (in milliseconds) required by each query in certain input/output operation with this database. In a preliminary analysis, 200 queries performed over a month were randomly selected, and the CPU cost of each query was registered. The values obtained are plotted on the Normal Probability Plot shown below.



Answer the following questions according to this plot:

a) Define in detail what is the population and the random variable involved in this study. (**2.5 points**)

b) What can be concluded about the distribution of the random variable under study? Justify your answer. (**2.5 points**)

c) What parameters of dispersion and position are the most convenient to characterize the data under study? Justify your answer. (**2.5 points**)

d) Obtain from the plot the estimated value of the population mean and standard deviation of the data under study. Indicate the procedure applied to obtain such values. (**2.5 points**)

2. (1st Partial) The employees of certain company are distributed as follows: 10% perform management and administrative tasks, 10% are sales representatives, and 80% are plant workers. An auditory about absenteeism in the workplace was carried out. Those employees who miss their workplace repeatedly for unjustified reasons are regarded as absent. The percentage of absent employees was 15% in the first group, 8% for sales representatives and 5% for plant workers.

a) Indicate and describe the events involved, and the probabilities associated to the events. (**1 point**)

b) What is the percentage of absent employees in the company? (**3 points**)

c) Certain employee who is regarded as absent is called by the manager. What is the probability of being a sales representative? (**3 points**)

d) If an employee is randomly selected certain day, what is the probability of being a non-absent sales representative? (**3 points**)

3. (1st Partial) Chisco Ltd manufactures two types of network cards: RA and RB. Each type is sold in packs of 100 units. Due to problems in the manufacturing process, it turns out that 2% of RA cards are defective, and 1% of RB cards are defective. The company ELECTONYS uses network cards RA and RB marketed by Chisco Ltd in certain equipment.

a) If two packs are randomly selected (one of RA cards and another pack of RB cards), what is the probability to find in total more than two defective cards? **(6 points)**

b) If one pack of RB cards is randomly selected, the probability to find ***k*** or less defective cards is < 99%. What is the value ***k***? **(4 points)**

4. (1st Partial) In certain library, the time used by customers to perform book searches in the computers is distributed as a Normal variable with a mean of 200 s and with a standard deviation of 30 s. If a customer executes a book search, what is the probability to take between 150 and 200 seconds?

5. (1st Partial) Certain company wants to know if the differences in the number of errors made by two programmers (one working in the mornings and the other working in the afternoons) are statistically significant. To study this issue, a sample of 61 programs performed by one programmer is randomly taken, and another sample of 61 programs from the other. The following data are obtained from the samples:

SUMMARY STATISTICS:

	PROGRAMMER-1	PROGRAMMER-2
Count	61	61
Average	9,11475	11,8852
Median	9,0	12,0
Mode	10,0	
Variance	7,06995	10,0699
Standard deviation	2,65894	3,17332
Minimum	3,0	5,0
Maximum	17,0	20,0
Range	14,0	15,0
Stnd. skewness	1,90818	0,448924
Stnd. kurtosis	1,74821	-0,512362

Regarding the average number of errors made by the programmers, are the differences statistically significant, considering $\alpha=0.05$?

6. (2nd Partial) In the framework of a project to evaluate the performance of a new file administration software for certain operative system, two variables were measured for 21 files of similar characteristics: time of execution (ms) and time of disk access (ms). A regression analysis was performed as part of the study, and the following results were obtained:

Regression Analysis - Linear model: Y = a + b*X

Dependent variable: Y
Independent variable: X

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	0,0161212	0,735938	_____	
Slope	3,32307	0,190529	_____	

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	55,6913	1	55,6913	304,20	0,0000
Residual	_____	___	_____		
Total (Corr.)	59,1697	___			

According to the results of the regression analysis, answer the following questions justifying conveniently the replies:

a) Propose and estimate a regression model to predict the time of execution as a function of the time of disk access (α=0.05). **(4 points)**

b) If the time o

c) f disk access for a certain file is 3 ms, what is the expected average time of execution? **(3 points)**

d) Calculate the value of the Residual Variance associated to the regression model. What is the practical interpretation of this parameter? **(3 puntos)**

SOLUTION

1a) The population is formed by all possible queries that can be performed to the database. The random variable is the time (in milliseconds) required by each query in certain input/output operation with this database.

1b) Since all points follow a straight line in the normal probability plot, it can be assumed that data follow a Normal distribution. No outliers are observed. The plot indicates that the median is 120, and the mean will also be about 120 because the distribution is symmetric. The minimum is 103, the maximum is 143 and consequently the range is 40. The interquartile range is approximately 124-114 = 10.

1c) Since values follow a Normal distribution which is symmetric and no outliers are observed, the mean is coincident with the median and will be the most convenient parameters of position. The variance will be a good parameter of dispersion, as well as the standard deviation. Also the interquartile range, though this parameter is more useful in the case of asymmetric data.

1d) Percentile 50 (vertical scale in the figure) corresponds to CPU cost = 120. This value is the median, which is coincident in this case with the mean because the distribution is symmetric. In a Normal distribution, the interval [m-2s ; m+2s] accounts for 95% of the values, and consequently 2.5% of the values are below (m-2s). This value is the percentile 2.5%, which according to the plot is about 105. If m-2s = 105, it turns out that s = (120-105)/2 **≈ 7.5**

2a) Event D: the employee performs management and administrative tasks.
C: the employee is a sales representative
Pw: the employee is a plant worker A: the employee is regarded as absent
P(D)=0.1 ; P(C)=0.1; P(Pw)=0.8 ; P(A/D)=0.15 ; P(A/C)=0.08; P(A/Pw)=0.05

2b) According to the Total Probability theorem, the percentage is **6.3%**:
P(A)=P(D)·P(A/D)+P(C)·P(A/C)+P(Pw)·P(A/Pw) = 0.1·0.15+0.1·0.08+0.8·0.05 = 0.063

2c) $P(C/A)=\dfrac{P(C)\cdot P(A/C)}{P(A)}=\dfrac{0.1\cdot 0.08}{0.063}=$ **0.127**

2d) $P\left(\overline{A}\cap C\right)=P(C)\cdot P\left(\overline{A}/C\right)=P(C)\cdot\left[1-P(A/C)\right]=0.1\cdot(1-0.08)=$ **0.092**
The following procedure is wrong because the events are not independent:
$P\left(\overline{A}\right)\cdot P(C)=(1-0.063)\cdot 0.1=0.0937\neq P\left(\overline{A}\cap C\right)$
Also note that $P\left(\overline{A}/C\right)=1-P(A/C)$ but it turns out that $P\left(C/\overline{A}\right)\neq 1-P(C/A)$

3a) The following random variables are defined:
X_A: number of defective cards in a pack of 100 units of RA cards.
X_B: number of defective cards in a pack of 100 units of RB cards.
$X_A\approx Bi(n=100;p=0.02)$; $X_B\approx Bi(n=100;p=0.01)$
As *n* is high and p is low, the binomial distributions can be approximated by Poisson distributions: $X_A\approx Ps(\lambda=n\cdot p=2)$; $X_B\approx Ps(\lambda=n\cdot p=1)$
$P\left[(X_A+X_B)>2\right]=P\left[Ps(\lambda=2+1)>2\right]=1-P\left[Ps(\lambda=3)\leq 2)\right]=$ (with the abacus)= **0.58**

(alternative calculation) $=1-e^{-3}\cdot\left(\dfrac{3^0}{0!}+\dfrac{3^1}{1!}+\dfrac{3^2}{2!}\right)=1-8.5\cdot e^{-3}=$ **0.577**

Alternative calculation: As n=100 in both cases: $(X_A + X_B) \approx Bi(n = 200;\ p = 0.015)$

$$P[(X_A + X_B) > 2] = 1 - \binom{200}{0} p^0 \cdot 0.985^{200} - \binom{200}{1} \cdot 0.015^1 \cdot 0.985^{199} - \binom{200}{2} \cdot 0.015^2 \cdot 0.985^{198} =$$

$$= 1 - 0.985^{200} - 200 \cdot 0.015^1 \cdot 0.985^{199} - 100 \cdot 199 \cdot 0.015^2 \cdot 0.985^{198} = \mathbf{0.578}$$

3b) Using the Poisson abacus: $P(X_B \le k) < 0.99$; $P[Ps(\lambda = 1) \le k] < 0.99$

Reading in the abacus with a probability 0.99 (horizontal line) and λ=1 (vertical line), both lines are intersected in a point intermediate of curves 3 and 4, so it is uncertain if the solution is k=3 or k=4.

If k=4, from the abacus: $P[Ps(\lambda = 1) \le 4] = 0.996 > 0.99$ The condition is not satisfied.

If k=3, from the abacus: $P[Ps(\lambda = 1) \le 3] = 0.98 < 0.99$ So, the solution is **k=3**.

4) The time T follows a distribution: $T \approx N(m = 200; \sigma = 30)$

$$P(150 < T < 200) = P(T < 200) - P(T < 150) = P[N(200;30) < 200] - P[N(200;30) < 150] =$$

$$= 0.5 - P[N(0;1) < (150 - 200)/30] = 0.5 - P[N(0;1) < -1.667] = 0.5 - P[N(0;1) > 1.667] =$$

=(table) = 0.5 - 0.048 = **0.452**

5) As the standardized skewness and kurtosis coefficients belong to the interval [-2; 2], there is not enough evidence to conclude that they are different from zero in the population. Thus, it can be assumed that data follow a Normal distribution and we can apply the following equations, also assuming that $\sigma_1 = \sigma_2$.

$$H_0 : m_1 = m_2 \ ; \quad H_1 : m_1 \neq m_2 \ ; \qquad t^{\alpha/2}_{n_1+n_2-2} = t^{0.025}_{61+61-2} = t^{0.025}_{120} = 1.98$$

$$S_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{1}{61} + \frac{1}{61}} \cdot \sqrt{\frac{60 \cdot 7.07 + 60 \cdot 10.07}{61 + 61 - 2}} = 0.5301$$

H_0 will be rejected if the following condition is satisfied, which is actually the case:

$$\left| \frac{\overline{x}_1 - \overline{x}_2}{S_{\overline{x}_1 - \overline{x}_2}} \right| > t^{0.025}_{120} ; \quad \left| \frac{9.1147 - 11.885}{0.5301} \right| = 5.226 > 1.98$$

Thus, the answer is YES, there is enough evidence to affirm that the observed differences in the average number of errors are statistically significant.

6a) The proposed model is: $T_{execution} = 0.01612 + 3.323 \cdot T_{disk\_access}$

Model estimation: this equation will be useful in practice if it can be guaranteed at the population level that the correlation between both variables (y = a+b·x) actually exists, which implies that the slope *b* is different from zero. The hypothesis test is:

$H_0 : b = 0$; $H_1 : b \neq 0$ The p-value of the global significance test ("analysis of variance") is p=0.0000. The same p-value will correspond to the slope because there is only one variable in the model. As p-value < 0.05, the null hypothesis is rejected: the proposed model is valid. Alternative procedure: the null hypothesis is rejected because $t = b_i / s_{b_i} = 3.323/0.1905 = 17.4 > \left(t^{\alpha/2}_{n-1-I} = t^{0.025}_{21-1-1} = 2.093\right)$

6b) $E(T_{exec} / T_{disk} = 3) = 0.01612 + 3.323 \cdot 3 = \mathbf{9.985}$

6c) Df_{total} = 21-1=20; Df_{resid} = 20-1=19; $SS_{resid} = SS_{total} - SS_{model}$ =59.1697-55.6913=3.478

$MS_{resid} = SS_{res}/Df_{res}$ = 3.4784/19 = **0.1831** = residual variance

Practical interpretation: it is the variance of the conditional distribution (i.e., the distribution of Y when X takes a particular value), which is assumed to be constant (homoscedasticity). This value is used to compute conditional probabilities.