etsinf

DEIOAC VALENCIA

Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 14th 2011

Surname, name:

Group: **1E** Signature:

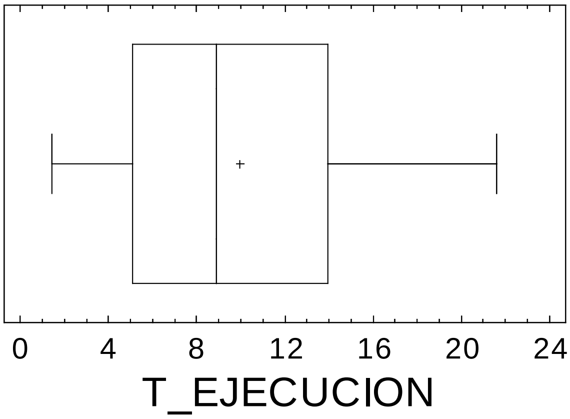
Indicate with a tick mark the partials examined

1st □ 2nd □ 3rd □

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 3 blocks (each one corresponding to one partial exam), with 3 questions in each block (9 in total). The lecturer will correct those blocks indicated by the student with a tick mark in this page. All questions of each block 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**

(1st Partial) The following plot has been obtained with a sample of 100 data corresponding to the time of execution (in seconds) of a certain program.



According to this plot, answer the following questions justifying conveniently the reply.

a) What is the random variable under study? Indicate what is the type of this variable.

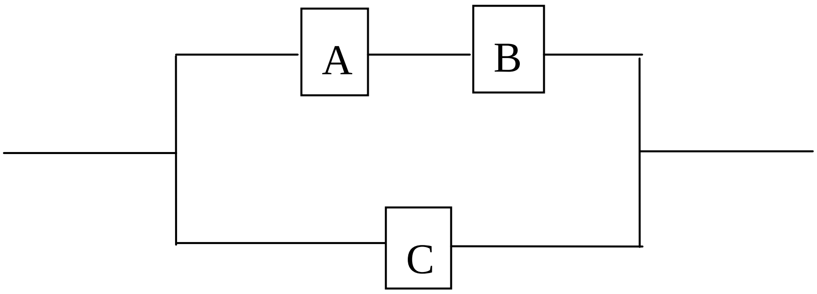
b) What is the name of this kind of plot?

c) Calculate the interquartile range. Indicate if it is a parameter of position or dispersion.

d) What parameters of position can be obtained from this plot? Which one would be the most convenient to describe this sample?

e) If the variance is calculated with the 100 data, what will be the units?

2. (1st Partial) One company manufactures devices comprised by 3 electronic components A, B and C assembled as indicated in the following scheme:



Components of type A have a reliability of 90% after 5000 hours of operation. The reliability of type B and C components after 5000 hours is 80% and 75%, respectively.

a) Define clearly all events implied in this problem. Indicate, justifying conveniently the reply, if such events can be considered as independent, complementary and/or exclusive.

b) Calculate the reliability of the device at 5000 hours of operation.

3. (1st Partial) In a certain industrial process, it is known that 1% of LCD screens manufactured are defective.

a) If different samples of 20 screens are taken, what is the probability to find a sample with at least 2 defective screens?

b) If screens are shipped in sets of 600 units, what is the probability to find a set with more than 10 defective screens?

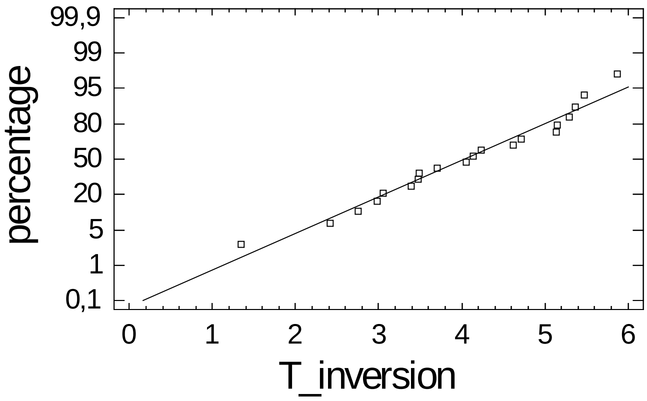
c) If 5 sets of 600 screens have arrived at a department store, what is the probability to find in total more than 30 defective screens?

4. (2nd Partial) A certain technical assistance service knows that the duration of phone calls received by technicians follows an exponential distribution with a median of 4 minutes.

a) What is the probability of a technician to spend more than 5 minutes at the phone with a call?

b) Knowing that a technician has already spent 5 minutes at the phone with a call, what is the probability of that phone call to have a duration lower than 7 minutes?

5. (2nd Partial) In order to study the efficiency of certain algorithm for matrix inversion, the time taken by the algorithm to invert a matrix (in milliseconds) has been sampled using 20 matrices of similar characteristics. The resulting 20 values are represented in the following plot:



Observe the plot and answer the following questions, justifying all the replies:

a) Indicate what is the name of this type of plot. **(2 points)**

b) What type of distribution can be assumed for the variable under study? Justify the reply. **(2 points)**

c) Estimate approximately the parameters of the distribution according to the plot. **(3 points)**

d) Taking into account the parameters estimated in the previous section, calculate the probability of spending less than 25 ms to invert 5 matrixes.
(3 points)

6. (2nd Partial) One company of digital printing has a master computer that manages a printer. This master receives files from many computers. It is known that the time required by the master to print a file follows a Normal distribution with average 6 and standard deviation 1.8 seconds. The company considers that this time is excessive, and hires a computer technician to change the operative system. In order to study if this change has been efficient, the company measures the time required to print 5 files randomly selected, resulting the following values (in seconds): 8; 17; 10; 11; 15. Considering $\alpha = 0.05$, indicate which of the following sentences is true, justifying conveniently the reply.

a) The average time of printing is significantly different after the change of operative system, and consequently that change has resulted efficient.

b) The company should complain to the computer technician because the average time after the change of operative system has not resulted significantly different to the average time prior to that change.

c) The average time of printing has not resulted significantly different, but the company cannot complain to the computer technician because the sample size is too small.

d) The company should complain to the computer technician for other reasons.

7. (3rd Partial) In the subject of Statistics, one group of 80 students carries out in March one test comprised by 4 exercises of probability (each one scoring 0.1 points), obtaining the following results: 15 students got a mark of 0.1; 30 students scored 0.2; 25 students got a mark of 0.3; 10 students scored 0.4. The average mark was 0.2375.

Afterwards, the same 80 students did in June another test with 4 exercises of distributions (each one scoring 0.1 points), obtaining the following results: 10 students got a mark of 0.1; 25 students scored 0.2; 30 students got a mark of 0.3; 15 students obtained a mark of 0.4. The average mark of these 80 data is 0.2625 and the standard deviation is 0.0933.

The lecturer affirms that the average marks in March and June differ significantly. Is there enough evidence to support this statement?

a) Yes, considering a significance level $\alpha = 0.15$, but no considering $\alpha = 0.10$.

b) Yes, considering a significance level $\alpha = 0.10$, but no considering $\alpha = 0.05$.

c) Yes, considering a significance level $\alpha = 0.05$, but no considering $\alpha = 0.01$.

d) No, considering a significance level $\alpha = 0.15$.

e) Yes, considering a significance level $\alpha = 0.01$.

Note: use the hypothesis test for two Normal populations, assuming that the population variances are equal.

8. (3rd Partial) In a certain computer, the time (in milliseconds) used by the CPU (denoted as TCPU) has been recorded for 48 programs of similar characteristics. Data have been analyzed using a linear regression model in order to predict the average TCPU as a function of variable X: S_memory – 70 (size of available memory in Kb minus 70 Kb). Results obtained are shown below:

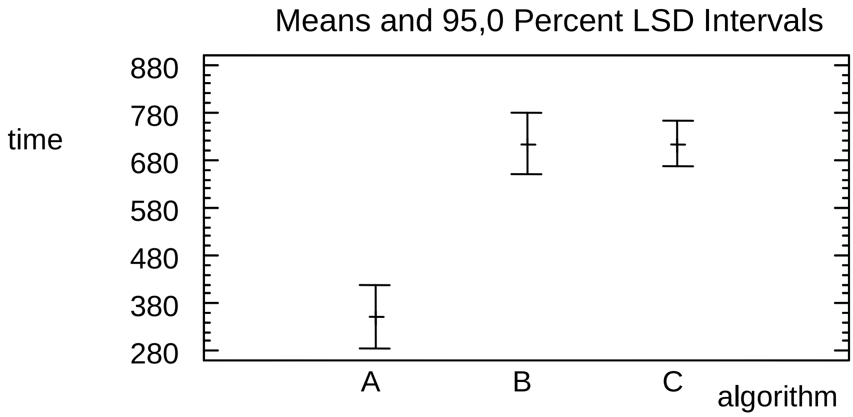
```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: TCPU
Independent variable: S_memory-70
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic      P-Value
-----------------------------------------------------------------------------
Intercept      42,3705         0,214873      [redacted]     [redacted]

Slope          0,573997        0,019460      [redacted]     [redacted]
-----------------------------------------------------------------------------

Analysis of Variance
-----------------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                        538,112      1      538,112     870,00       <0,05
Residual                     28,4519     46      0,61852
-----------------------------------------------------------------------------
Total (Corr.)                566,564     47
```

a) Write the mathematical equation of the estimated regression model. Complete the above table (fill the gaps in black) and analyze if the model parameters are statistically significant. Use α=5%.

b) Estimate the coefficient of determination of this model and interpret its value.

c) Obtain the interval comprising 99% of the cases for the time used by the CPU (TCPU) when the size of available memory is 80 Kb.

9. (3rd Partial) One consulting company has developed three algorithms (A, B and C) for performing complex operations with large data matrices. In order to determine which algorithm is the most efficient, 12 matrices of similar size are randomly selected. Algorithm A is applied to three of them; algorithm B is applied to 3 matrices, and C to the remaining 6 ones. Results obtained, measured in milliseconds (variable "time") are indicated in the table below. Data are analyzed with ANOVA using Statgraphics, resulting the following plot:



algorithm A	algorithm B	algorithm C
373 365 312	739 711 695	615 844 711 648 809 663

a) Which of the following statements is true?

a.1) According to the plot, it can be deduced that the p-value of the ANOVA test is higher than 0.05.

a.2) According to the plot, it can be deduced that the p-value of the ANOVA test is lower than 0.05.

a.3) It is not possible to deduce from this plot the p-value of the ANOVA test (i.e. statements a.1 and a.2 are uncertain).

a.4) The p-value depends on the significance level of the ANOVA test, which cannot be deduced from the plot.

Justify your reply:

b) Taking into account that $\bar{x}_A$ = 350, $\bar{x}_B = \bar{x}_C$ = 750, which algorithm is the least convenient?

b.1) Algorithm B, because the longitude of its LSD interval is larger than in the case of C which suggests that B has more probability to reach higher values of time.
b.2) Algorithm C, because the longitude of its LSD interval is lower than in the case of B which suggests that its standard deviation is lower.
b.3) Algorithms B or C.
b.4) Any of the three, because the null hypothesis H_0: $m_A = m_B = m_C$ should be accepted.

Justify your reply:

c) One hypothesis in ANOVA is that the population of time data follows a Normal model in each of the three algorithms tested. How could we verify if this hypothesis is admissible?

c.1) The hypothesis of normality is admissible because LSD intervals are symmetric.
c.2) We should calculate the residuals of ANOVA and study if they follow reasonably a Normal model.
c.3) There are not enough data to study if the Normal model is admissible.

Justify your reply:

SOLUTION OF THE FINAL EXAM - 14th JUNE 2011

1a) Random variable: time of execution (seconds) every time that the program is executed. The type of variable is: quantitative and continuous (values higher than zero).

1b) The plot is a box and whisker plot.

1c) Quartiles: Q1=5 (left edge of the box); Q3=14 (right edge);
Interquartile range = Q3 - Q1 = 14 - 5 = 9
It is a parameter of dispersion because if this parameter is high indicates a high dispersion (variability) of the values.

1d) Two parameters of position can be obtained: mean = 10 (cross inside the box) and median=9 (vertical line inside the box). In this case, the median is more convenient than the mean because the distribution is positively skewed.

1e) The units would be seconds^2 because: $s^2 = \frac{\sum\left(x_i - \overline{x}\right)^2}{n-1} = \frac{\sum\left(\text{sec} - \text{sec}\right)^2}{cons\tan t} = \text{sec}^2$

2a) Event A: component of type A works after 5000 hours of operation.
Event B: component of type B works after 5000 hours of operation.
Event C: component of type C works after 5000 hours of operation.
Event D: the whole device works after 5000 hours of operation.

Independent: events A, B and C are assumed to be independent, which implies that the failure in one component does not affect the performance of the others. However, event D is not independent of A, B or C because $P\left(D/\overline{C}\right) \neq P(D)$.

Complementary: two events X and Y are complementary if P(X)=1-P(Y). In this case, none of the events are complementary.

Exclusive: two events X and Y are exclusive if $P\left(X \cap Y\right) = 0$. In this case A, B and C are not exclusive because it is possible that $A \cap B$ or $A \cap C$ can work after 5000 h. D is not exclusive with respect to A, B or C for the same reason.

2b) $P(D) = P\left[(A \cap B) \cup C\right] = P(A \cap B) + P(C) - P\left(A \cap B \cap C\right) =$
(assuming that the events are independent)
$= P(A) \cdot P(B) + P(C) - P(A) \cdot P(B) \cdot P(C) = 0.9 \cdot 0.8 + 0.75 - 0.9 \cdot 0.8 \cdot 0.75 = \mathbf{0.93}$

3a) X: number of defective screens in a sample of 20 units; $X \approx Bi(20; p = 0.01)$

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \binom{20}{0} \cdot 0.01^0 \cdot 0.99^{20} + \binom{20}{1} \cdot 0.01^1 \cdot 0.99^{19} = \mathbf{0.0169}$$

3b) Y: number of defective screen in a set of 600 units. $Y \approx Bi(600; p = 0.01)$
It can be approximated to a Poisson distribution:
$Y \approx Ps(\lambda = n \cdot p = 600 \cdot 0.01 = 6)$; $P(Y > 10) = 1 - P(Y \leq 10) = 1 - 0.96 = \mathbf{0.04}$

The value 0.96 is obtained from the Poisson abacus: the value λ=6 crosses the curve 10 at a value in the vertical scale of 0.96.

3c) Z: number of defective screen in 5·600 = 3000 screens; $Z \approx Bi(3000; p = 0.01)$
It can be approximated to a Poisson distribution:
$Z \approx Ps(\lambda = n \cdot p = 3000 \cdot 0.01 = 30)$; $P(Z > 30) = 1 - P(Z \le 30) = 1 - 0.55 = \mathbf{0.45}$
The value 0.55 is obtained from the Poisson abacus: the value λ=30 crosses the curve 30 at a value in the vertical scale of 0.55.

4a) X: duration of a phone call; $X \approx \exp(\alpha)$; $P(X < x) = 1 - e^{-\alpha \cdot x}$
$P(X < 4) = 0.5 = 1 - e^{-\alpha \cdot 4}$; $0.5 = e^{-\alpha \cdot 4}$; $\alpha = -(\ln 0.5)/4 = 0.1733$
$P(X > 5) = 1 - P(X \le 5) = 1 - \left(1 - e^{-0.1733 \cdot 5}\right) = \mathbf{0.42}$

4b) $P(X < 7/X > 5) = 1 - P(X > 7/X > 5) =$ (property of lack of memory)=
$= 1 - P(X > (7-5)) = P(X < 2) = 1 - e^{-0.1733 \cdot 2} = \mathbf{0.293}$

5a) The plot is a Normal Probability Plot.

5b) Given that points follow approximately a straight line, we can assume that data follow approximately a Normal distribution.

5c) The normal distribution has two parameters: mean and standard deviation. The median is 4, because according to the plot, P(time<4)=50% which is the definition of median. The shape of the plot indicates a symmetric distribution, which implies that the mean will be around 4. In a normal distribution, 2.5% of the values are below m-2s, and 2.5% of the values are above m+2s. Reading at the vertical scale 2.5%, the value is approximately 1.5, so that: 1.5=m-2s; 1.5=4-2s; s=(4-1.5)/2 = 1.25. $X \approx N(m = 4; s = 1.25)$

5d) X: time to invert 1 matrix; Y: time to invert 5 matrixes; $Y = X_1 + X_2 + ... + X_5$
$E(Y) = E(X_1) + ... + E(X_5) = 5 \cdot E(X) = 5 \cdot 4 = 20$
$\sigma^2(Y) = \sigma^2(X_1) + ... + \sigma^2(X_5) = 5 \cdot \sigma^2(X) = 5 \cdot 1.25^2 = 7.81$
$$P(Y < 25) = P\left[N\left(20, \sqrt{7.81}\right) < 25\right] = P\left[N(0;1) < \frac{25-20}{\sqrt{7.81}}\right] = P[N(0;1) < 1.79] = \mathbf{0.963}$$

6) $\bar{x} = (8+17+10+11+15)/5 = 12.2$; H$_0$: m=6; H$_1$: m≠6
$$\frac{\bar{x} - m}{\sigma/\sqrt{n}} \approx N(0;1) \ ; \ \frac{12.2-6}{1.8/\sqrt{5}} = 7.7 > (Z_{\alpha/2} = 1.96)$$
As 7.7 is an infrequent value of the N(0; 1), the conclusion is to reject the null hypothesis: the population mean is higher than 6 because the sample mean is 12.2>6. The solution is **d**: the company should complain because the time has increased significantly, which is something undesirable.

7) June: $\overline{x_J} = 0.2625$; $s_J = 0.0933$; March: $\overline{x_M} = 0.2375$

$$s_M^2 = \frac{15\cdot(0.1-0.2375)^2 + 30\cdot(0.2-0.2375)^2 + 25\cdot(0.3-0.2375)^2 + 10\cdot(0.4-0.2375)^2}{80-1} = 0.0087$$

$s_M = \sqrt{s_M^2} = 0.0933$; $H_0 : m_M = m_J$; $H_1 : m_M \neq m_J$

$$S = \sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2}} = \sqrt{\frac{79\cdot 0.0933^2 + 79\cdot 0.0933^2}{80+80-2}} = 0.0933$$

$$S_{\overline{x_1}-\overline{x_2}} = S\cdot\sqrt{(1/n_1)+(1/n_2)} = 0.0933\sqrt{(1/80)+(1/80)} = 0.0148$$

$$\left|\frac{\overline{x_1}-\overline{x_2}}{S_{\overline{x_1}-\overline{x_2}}}\right| = \left|\frac{0.2375-0.2625}{0.0148}\right| = 1.689 \approx t_{n_1+n_2-2} \approx t_{158}$$

From the t-table: $t_{158}^{0.05} = 1.65$; $t_{158}^{0.025} = 1.97$

If α=0.1, $1.689 > t_{158}^{0.05}$ and the conclusion is to reject H$_0$ ($m_M \neq m_J$). The answer is yes, the average marks in March and June differ significantly.

If α=0.05, $1.689 < t_{158}^{0.05}$ and the conclusion is to accept H$_0$: $m_M = m_J$. The answer is no, the average marks in March and June do not differ significantly.

Thus, the correct answer is **b**.

8a) Equation of the regression model: TCPU=42.3705+0.574·(S_memory-70)

For the intercept: $t_{stat} = b_i / s_{b_i} = 42.3705/0.2149 = 197.2$; $t_{stat} \approx t_{N-1-I} \approx t_{46}$

Being N=48 values, and I=1 (one explicative variable included in the model)

If α=0.0005, $t_{46}^{0.0005} = 3.515 << 197.2 \Rightarrow p-value < 0.0005$

For the slope: $t_{stat} = b_i / s_{b_i} = 0.574/0.01946 = 29.5$; $t_{stat} \approx t_{N-1-I} \approx t_{46}$

If α=0.0005, $t_{46}^{0.0005} = 3.515 << 29.5 \Rightarrow p-value < 0.0005$

Using α=5% (indicated in the statement), as the p-value < 0.05, it can be concluded that both parameters are statistically significant. The completed table is:

```
-----------------------------------------------------------------------------
                                  Standard           T
Parameter       Estimate            Error        Statistic      P-Value
-----------------------------------------------------------------------------
Intercept      42,3705          0,214873          197.2         <0.0005

Slope          0,573997         0,019460           29.5         <0.0005
-----------------------------------------------------------------------------
```

8b) $R^2 = \frac{SS_{explained}}{SS_{total}}\cdot 100 = \frac{538.11}{566.56}\cdot 100 = 94.98\%$

Interpretation: 94.98% of the variability (variance) of TCPU is explained by the model (i.e., is explained by the variability of the size of available memory).

8c) $TCPU = 42.3705 + 0.574\cdot(S\_memory - 70)$

$E(TCPU / S\_memory = 80) = 42.3705 + 0.574\cdot(80-70) = 48.11$

$\sigma_{resid}^2 = MS_{residuals} = 0.61852$ (value from the table); $\sigma_{resid} = 0.7865$

In a distribution N(0; 1), the interval [-2.57, 2.57] comprises 99% of the values. In this case, when the size of memory is 80, the 99% interval will be:

$48.11 \pm 2.57\cdot 0.7865 \Rightarrow$ **[46.09, 50.13]**

9a) The top of the plot indicates “95% LSD intervals”, which implies that 95% is the confidence level = 1-α. Thus, α=0.05. Given that the LSD interval of A does not overlap with the one of B nor C, the conclusion is to reject the null hypothesis and, therefore, it can be concluded that p-value < 0.05. As a conclusion, the correct answer is **a.2**.
9b) The longitude (amplitude) of LSD intervals only depends on the number of data. As the LSD intervals of B and C overlap, the conclusion is to accept H_0: m_B=m_C. Thus, any of the algorithms is the least convenient (the correct solution is **b.3**).

9c) The sentence c.1 is false because LSD intervals are always symmetric; c3 is also false because it is possible to study the normality of 12 values by means of a normal probability plot. The true answer is **c2** as it is the usual way to check the hypothesis of normality.