Bachelor Degree in Computer Engineering

Statistics **group E (English)**

SECOND PARTIAL EXAM

June 7th 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. One computer engineer wants to study the distribution of the execution time of certain program designed for the management of stocks. For this purpose, the program was run 15 times, resulting an average time of 47.07 ms and a variance of 2.97 ms^2

a) Certain study states that the average time expected for this type of program is 47.8 ms. Considering α=0.05, what two different statistical procedures can be applied to study if this hypothesis is acceptable? *(1 point)*

b) By using one of these procedures, determine if the statement made in that study is acceptable, with a confidence level of 95%. *(2.5 points)*

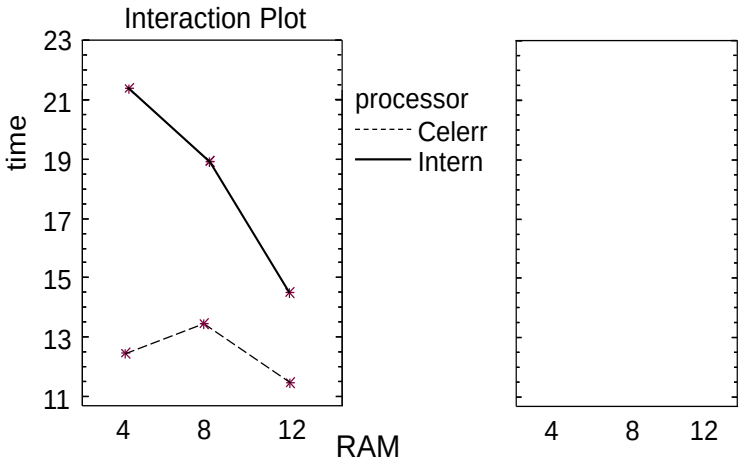
c) What would happen in this particular case if a confidence level of 99% is considered instead of 95%? In general, what would be the consequences regarding the final conclusion in this type of analysis if the confidence level decreases? *(2.5 points)*

d) The same study states that the variability associated with the execution time can be quantified with σ = 5 ms. Is this statement admissible based on the data available, considering a type I risk of 5%? *(2.5 points)*

e) In order to answer the questions displayed above, what hypothesis needs to be assumed about the distribution of the population sampled? *(1.5 points)*

2. One experiment is carried out to study the effect of processor type and RAM memory on the execution time of certain algorithm that operates with large matrices generated from searches on internet. With this objective, a sample of 24 matrices was randomly taken: 8 of them were processed using 4 GB of RAM, other 8 matrices using 8 GB, and the other 8 ones using 12 GB. Half of these matrices were treated with a computer equipped with a Celerr© processor and the other half with an Intern© processor, as indicated below. The values of time (measured in milliseconds) obtained experimentally, which are indicated in the table, have been analyzed using ANOVA.

RAM	Proc.	time
4	Celerr	10; 12; 13; 15
4	Intern	23; 25; 20; 18
8	Celerr	12; 13; 15; 14
8	Intern	22; 15; 21; 18
12	Celerr	12; 13; 10; 10
12	Intern	14; 16; 15; 13



a) The summary table of ANOVA reveals that: SS_{total} = 409.625; SS_{RAM} = 77.25; SS_{proces} = 210.042; SS_{resid} = 88.75. Taking into account this information, what effects are statistically significant, considering α = 5%? *(3.5 points)*

) Taking into account the results obtained in the previous section and considering α = 0.05, interpret the interaction plot: how does the RAM memory and processor type affect to the execution time of the algorithm? Justify your reply. *(1.5 points)*

c) Draw the LSD intervals inside the empty graph (on the right of the interaction plot) for the factor "RAM memory" with a confidence level of 95%, knowing that the width of these intervals is $\overline{x_i} \pm$ 1.17 ms. For this factor, describe the nature of the effect of RAM on the mean execution time.

(2 points)

d) Taking into account all the results obtained, what would be the optimum operational condition that should be adopted to minimize the execution time, considering α=5%? *(1 point)*

e) If the algorithm runs under the optimum operational conditions established in the previous question, and assuming the hypotheses of normality and homoscedasticity, what is the probability to take less than 13 ms for the execution? *(2 points)*

3. Certain small reprography service equipped with a single printer, in order to better manage the orders and assignments, wants to study the relationship between the time required to print a document (Y) and the number of pages of the document (X). For this purpose, a sample of documents to print was randomly taken during one week, and for each one, the number of pages was recorded as well as the time for printing. After some statistical analyses, the following results were obtained:

Summary statistics

	X	*Y*
Count	75	75
Average	5,44	57,6227

Covariances

	X	Y
X	8,08757	65,5735
Y	65,5735	640,121

Coefficients Linear: Y = a + b*X

Parameter	*Estimate*	*Standard Error*	*t statistic*	*p-value*
Constant		2,6272	5,14444	0,0000
Slope		0,428601		

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-ratio*	*p-value*
Model	39343,3	1	39343,3	357,86	0,0000
Residual	8025,61	73	109,94		
Total (Corr.)	47368,9	74			

Based on these results, answer the following questions justifying the reply:

a) Calculate the Coefficient of Correlation between both variables studied and describe the relationship between them. *(1.5 points)*

b) How would look like the scatterplot between X and Y? Draw an approximate representation. *(1 point)*

c) Estimate the parameters of the linear regression model that relates the time required to print a document as a function of the number of pages. *(1.5 points)*

d) Study the statistical significance of each parameter of the model ($\alpha = 0.05$). What is the equation of the regression model estimated from the data? *(2 points)*

e) According to the proposed model, what is the average time expected to print a document comprised of 6 pages? *(1 point)*

f) What parameter quantifies the *quality of fitting* of the regression model to our experimental data? Calculate the value of this parameter and provide an interpretation in this study. *(1.5 points)*

g) Calculate the residual variance. What does this value represent in the present study? *(1.5 points)*

SOLUTION

1a) Se pretende estudiar la hipótesis nula de que el tiempo medio a nivel poblacional es de 47,8 ms frente a la hipótesis alternativa de que sea distinto. Es decir H_0: m = 47,8; H_1: m ≠ 47,8. Los dos procedimientos son:
1.- Calcular el estadístico de contraste (t calculada) que sigue una distribución t de Student (con n-1 grados de libertad), y comprobar si es mayor o menor (en valor absoluto) al valor crítico de tablas.
2.- Calcular el intervalo de confianza para la media poblacional, y verificar si el valor supuesto de la hipótesis nula está dentro o fuera de dicho intervalo.

1b) Valor del estadístico de contraste: $t_{calc} = \dfrac{\bar{x} - m_0}{s/\sqrt{n}} = \dfrac{47{,}07 - 47{,}8}{\sqrt{2{,}97}/\sqrt{15}} = -1{,}641$

Para un nivel de confianza del 95% (es decir, α=0,05), el valor crítico de tablas vale: $t_{n-1}^{\alpha/2} = t_{14}^{0{,}025} = 2{,}145$. Dado que el estadístico de contraste es menor en valor absoluto al valor crítico, se acepta H_0: m = 47,8 (no hay evidencia suficiente para rechazarla). Por tanto, es admisible la afirmación realizada en el estudio.

1c) Si el nivel de confianza (1-α) aumenta (es decir, si α disminuye), el valor crítico de tablas aumenta, de modo que en este caso en concreto la conclusión sería la misma para el test de hipótesis, ya que $t_{calc} < t_{n-1}^{\alpha/2}$.

Al disminuir 1-α no se modifica el estadístico de contraste, pero en general la conclusión final puede ser distinta. En este caso, por ejemplo, si el nivel de confianza disminuye al 80%, el valor de t_{calc} supera (en valor absoluto) al valor crítico: $1{,}64 > (t_{n-1}^{\alpha/2} = t_{14}^{0{,}1} = 1{,}345)$ de modo que se rechazaría la hipótesis nula.

1d) Se pretende estudiar la hipótesis nula H_0: $\sigma^2 = 5^2 = 25$. El estadístico de contraste sigue una distribución χ^2_{14} (ya que n=15). El 95% de valores de esta distribución están comprendidos entre 5,629 y 26,119 (valores críticos obtenidos de la tabla). El intervalo de confianza para la varianza poblacional es:

$$\sigma^2 \in \left[\frac{(n-1)\cdot s^2}{26{,}119}; \frac{(n-1)\cdot s^2}{5{,}629}\right] = \left[\frac{14\cdot 2{,}97}{26{,}119}; \frac{14\cdot 2{,}97}{5{,}629}\right] = \left[1{,}59;\ 7{,}39\right]$$

Dado que 25 está fuera de este intervalo, se rechaza la hipótesis nula: a partir de los datos disponibles no es admisible afirmar que σ= 5.

1e) La fórmula del estadístico de contraste empleada en los apartados anteriores asume que se ha realizado un muestreo aleatorio simple (es decir, todos los individuos de la población tienen la misma probabilidad de pertenecer a la muestra) y que la población muestreada tiene una distribución de tipo normal, lo cual implica también que no existen datos anómalos.

2a) Grados de libertad totales = N-1; G.l. de cada factor = nº variantes-1; G.l. interacción = 2·1=2; Cuadrado medio= SC / g.l.; F-ratio = CM / CM_{resid}
$SC_{interac} = SC_{total} - SC_{RAM} - SC_{proc} - SC_{resid}$.

	SC	gr. lib.	CM	F-ratio
RAM	**77,25**	2	38,62	$7{,}83 > (F_{2;18}^{0,05} = 3{,}55)$
Procesador	**210,04**	1	210,04	$42{,}60 > (F_{1;18}^{0,05} = 4{,}41)$
Interacción	33,58	2	16,79	$3{,}41 < (F_{2;18}^{0,05} = 3{,}55)$
Residual	**88,75**	18	4,93	
Total	**409,62**	23		

En la tabla resumen del ANOVA mostrada, en la columna de la derecha se indican los valores críticos de la tabla F para α=0,05. El efecto simple del factor RAM y procesador resultan estadísticamente significativos, ya que la F-ratio es mayor que el valor crítico. Esto no sucede con el efecto de la interacción por ser menor al valor crítico. No obstante, si se hubiera considerado α=0,1 la interacción resultaría significativa: $3{,}41 > (F_{2;18}^{0,1} = 2{,}62)$.

2b) El gráfico de la interacción hay que interpretarlo teniendo en cuenta que el efecto de la interacción no resulta estadísticamente significativo para α=0,05. Por tanto, a pesar de que las líneas no son paralelas entre los dos tipos de procesadores con los datos resultantes de la muestra, a nivel poblacional hay que asumir que el efecto de la memoria RAM es el mismo para los dos procesadores (es decir, sería como si las líneas fuesen paralelas a nivel poblacional). Hay que interpretar el efecto simple de cada factor por separado:

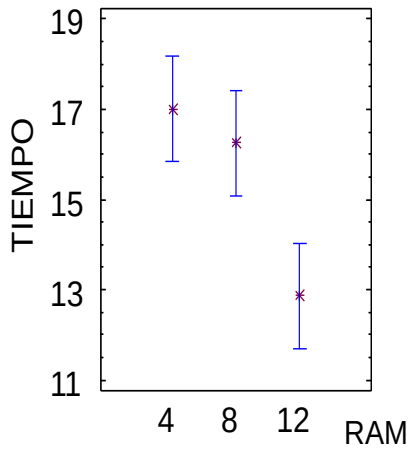
- Dado que el factor RAM resulta estadísticamente significativo, a la vista del gráfico se deduce que el tiempo medio a nivel poblacional con el procesador Celerr© es menor que para el otro procesador.
- Dado que el factor procesador también resulta significativo, al aumentar la RAM en promedio desciende el tiempo de ejecución, con un efecto cuadrático (tal como se deduce del apartado 2c).

2c) Los intervalos LSD se obtienen a partir del valor medio obtenido con cada RAM, teniendo en cuenta que la anchura en este caso es de ± 1,17 :

$$LSD_{RAM=4} = [(10+12+13+15+23+25+20+18)/8] \pm 1{,}17 = [15{,}83;\ 18{,}17]$$

$$LSD_{RAM=8} = [(12+13+15+14+22+15+21+18)/8] \pm 1{,}17 = [15{,}08;\ 17{,}42]$$

$$LSD_{RAM=12} = [(12+13+10+10+14+16+15+13)/8] \pm 1{,}17 = [11{,}71;\ 14{,}05]$$



Dado que RAM es un factor cuantitativo, nos piden describir cómo varía el tiempo de ejecución al aumentar la memoria RAM. El gráfico indica que el tiempo disminuye de forma cuadrática al aumentar la RAM, ya que los tres valores medios no están alineados.
Puede hablarse también de un efecto lineal negativo (pendiente negativa) y un efecto cuadrático negativo (curvatura hacia abajo). No es posible determinar en este caso si tanto el efecto lineal como el cuadrático resultan estadísticamente significativos, para lo cual sería necesario emplear regresión múltiple.

2d) Dado que el factor procesador resulta estadísticamente significativo, para minimizar el tiempo de ejecución hay que emplear un procesador Celerr© con 12 GB de RAM (al menos), ya que en estas condiciones (a la vista del gráfico anterior) el tiempo es significativamente menor.

2e) El tiempo medio obtenido experimentalmente con 12 GB de RAM y procesador Celerr© es: (12+13+10+10)/4 = 11,25. Este valor será el tiempo medio esperado trabajando en dichas condiciones, asumiendo que la interacción resulta estadísticamente significativa (es decir, considerando α=0,1). Si se asume la hipótesis de normalidad y homocedasticidad, la distribución será normal, cuya varianza se estima a partir del cuadrado medio residual, de valor 4,93 (apartado 2a). En estas condiciones:

$$P\left[N\left(11{,}25;\sqrt{4{,}93}\right)<13\right]=P\left[N(0;1)<\frac{13-11{,}25}{\sqrt{4{,}93}}\right]=P\left[N(0;1)<0{,}788\right]=1-0{,}2153=0{,}785$$

3a) A partir de la matriz de varianzas-covarianzas se deducen las varianzas: s^2_x = 8,088 ; s^2_y=640,121 y también la covarianza: $cov_{x,y}$ = 65,573.

Coeficiente de correlación: $r=\frac{\text{cov}}{\sqrt{s_x^2}\cdot\sqrt{s_y^2}}=\frac{65{,}573}{\sqrt{8{,}088\cdot 640{,}12}}=0{,}911$

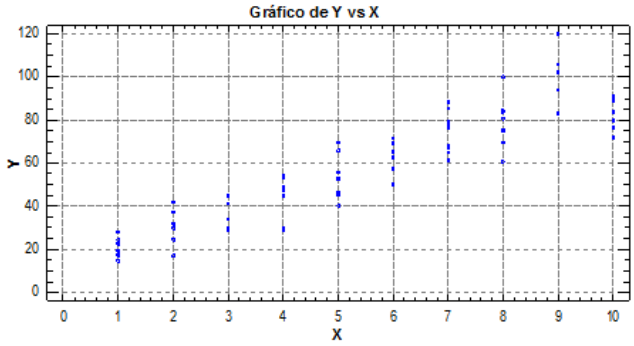
Dado que el valor obtenido es positivo y cercano a 1 podemos considerar que existe una relación lineal positiva y fuerte entre el *tiempo de impresión de un trabajo* (Y) y el *número de páginas del trabajo* (X) a imprimir.

3b) De acuerdo con la naturaleza de la relación entre Y y X descrita en el apartado anterior (relación positiva fuerte), se deduce que los puntos estarán bastante ajustados a la recta de regresión (de pendiente positiva). Por otra parte, teniendo en cuenta que las distribuciones marginales tanto de X como de Y son conocidas asumiendo la hipótesis de normalidad, el 95% de sus valores estarán comprendidos en el intervalo [media ± 2·s], de modo que:

Variación del 95% de valores de X: $5{,}44\pm 2\cdot\sqrt{8{,}088}$ ≈ de 0 a 11

Variación del 95% de valores de Y: $57{,}6\pm 2\cdot\sqrt{640{,}1}$ ≈ de 7 a 108

Teniendo en cuenta toda esta información, una representación aproximada del diagrama de dispersión sería el siguiente:



3c) Estimación de los parámetros del modelo:

$$b = r \cdot \frac{\sqrt{s_y^2}}{\sqrt{s_x^2}} = 0{,}911 \cdot \frac{\sqrt{640{,}1}}{\sqrt{8{,}088}} = 8{,}108 \quad ; \quad a = \overline{Y} - b \cdot \overline{X} = 57{,}623 - 8{,}108 \cdot 5{,}44 = 13{,}51$$

3d) Ordenada en el origen: puesto que el p-valor asociado a este parámetro es prácticamente cero, y por tanto menor que α (0,05), se rechaza la hipótesis nula, de modo que podemos admitir que la ordenada en el origen es estadísticamente significativa (es decir, distinta de cero a nivel poblacional).

Pendiente: dado que el p-valor del test de significación global del ajuste es también casi nulo, podemos admitir que existe un efecto a nivel poblacional del número de páginas sobre el tiempo de impresión medio.

La ecuación del modelo de regresión para estimar el tiempo de impresión (Y) en función del número de páginas (X) sería: $Y = 13{,}5155 + 8{,}108 \cdot X$

3e) El tiempo de impresión esperado, en promedio (medido en unidades de tiempo), para un trabajo que consta de 6 páginas será:
$E(Y/X=6) = 13{,}5155 + 8{,}108 \cdot 6 =$ **62,16**

3f) La calidad del ajuste del modelo se cuantifica a través del coeficiente de determinación (R^2), el cual se calcula como:
$R^2 = 100 \cdot SC_{modelo} / SC_{total} = 100 \cdot 39343{,}3 / 47368{,}9 =$ **83,06%**
Este parámetro indica que el 83,1% de la variabilidad de Y (tiempo de impresión) está explicada por el modelo. Al ser un valor relativamente alto, podríamos calificar el ajuste como bueno. Este parámetro es útil para comparar la calidad del ajuste en modelos alternativos.

3g) La varianza residual se estima con el cuadrado medio residual, que vale **109,94**. Asumiendo que se cumple la hipótesis de homocedasticidad, este valor representa la varianza de la distribución condicional de Y para un valor cualquiera de X. En el estudio realizado, al obtener la recta de regresión, la varianza residual estima el orden de magnitud del efecto conjunto de los factores no considerados en el estudio.