Grado en Ingeniería Informática

Estadística

SEGUNDO PARCIAL

7 de junio de 2017

Apellidos, nombre:	
Grupo:	Firma:

Instrucciones

1. Rellenar la información de cabecera del examen.
2. Responder a cada pregunta en la hoja correspondiente.
3. Justificar todas las respuestas.
4. No se permiten anotaciones personales en el formulario.
5. No se permite tener teléfonos móviles encima de la mesa. Sobre la mesa sólo se permite el DNI, calculadora, útiles de escritura, las tablas y el formulario.
6. No desgrapar las hojas.
7. Todas las preguntas puntúan lo mismo (sobre 10).
8. Se debe firmar en las hojas que hay en la mesa del profesor al entregar el examen. Esta firma es el justificante de la entrega del mismo.
9. Tiempo disponible: **2 horas**

1. Se desea estudiar la distribución del tiempo de ejecución de cierto programa de gestión de stocks. Para ello se llevan a cabo 15 ejecuciones de este programa, obteniéndose un tiempo medio de 47,07 ms y una varianza de 2,97 ms^2.

a) Cierto estudio afirma que el tiempo medio esperado para este tipo de programa es de 47,8 ms. Considerando α=0,05, ¿qué dos procedimientos estadísticos pueden emplearse para estudiar si es admisible esta hipótesis? *(1 punto)*

b) Utilizando uno de estos procedimientos, determina si es admisible la afirmación realizada en el estudio con un nivel de confianza del 95%. *(2,5 puntos)*

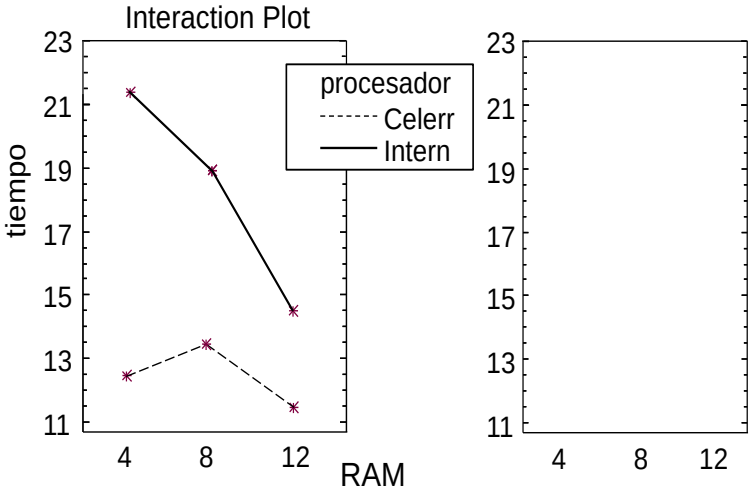
c) ¿Qué pasaría en este caso en concreto si en lugar de considerar un nivel de confianza del 95% hubiera sido del 99%? En general, ¿qué consecuencias podría tener el disminuir el nivel de confianza con respecto a la conclusión final en este tipo de análisis? *(2,5 puntos)*

d) Ese mismo estudio afirma que la variabilidad asociada al tiempo de ejecución puede cuantificarse con σ = 5 ms. ¿Es admisible esta afirmación a partir de los datos disponibles, considerando un riesgo de primera especie del 5%? *(2,5 puntos)*

e) Para poder contestar a los apartados anteriores, ¿qué hipótesis has tenido que asumir respecto a la distribución de la población muestreada? *(1,5 puntos)*

2. Se pretende estudiar cómo afecta el tipo de procesador y el tamaño de la memoria RAM sobre el tiempo que tarda en ejecutarse cierto algoritmo que trabaja con matrices de gran tamaño generadas a partir de búsquedas en internet. Con este objetivo, se toma una muestra aleatoria de 24 matrices: 8 de ellas se procesan con un equipo de 4 GB de RAM, otras 8 con un equipo de 8 GB, y otras 8 matrices con uno de 12 GB. La mitad de las matrices son tratadas con un ordenador dotado de un procesador Celerr© y la otra mitad con un procesador Intern©, tal como se indica en la tabla inferior. Los valores experimentales obtenidos de tiempo (medido en milisegundos), indicados a continuación, se han analizado con ANOVA.

RAM	Proc.	tiempo
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) Estudia qué efectos resultan estadísticamente significativos a partir del cuadro resumen del ANOVA (utiliza α=5%), teniendo en cuenta que: SC_{total} = 409,625; SC_{RAM} = 77,25; SC_{proces} = 210,042; SC_{resid} = 88,75.

(3,5 puntos)

b) Teniendo en cuenta los resultados obtenidos en el apartado anterior y considerando α=0,05, interpreta el gráfico de la interacción: ¿cómo afecta la memoria RAM y el tipo de procesador al tiempo de ejecución del algoritmo? Justifica la respuesta. *(1,5 puntos)*

c) En el gráfico vacío de la derecha, dibuja el gráfico de intervalos LSD para el factor "memoria RAM" con un nivel de confianza del 95%, sabiendo que la anchura de dichos intervalos es $\overline{x_i} \pm 1{,}17$ ms. Interpretar la naturaleza del efecto del factor RAM sobre el tiempo medio de ejecución. *(2 puntos)*

d) A la vista de todos los resultados obtenidos, ¿cuál sería la condición operativa óptima para minimizar el tiempo de ejecución, considerando α=5%?

(1 punto)

e) Si el algoritmo se ejecuta en las condiciones operativas óptimas establecidas en el apartado anterior, y asumiendo que se cumple la hipótesis de normalidad y homocedasticidad, ¿cuál es la probabilidad de que el tiempo de ejecución sea inferior a 13 ms? *(2 puntos)*

3. Una pequeña empresa de reprografía, que cuenta con una única impresora, desea estudiar la relación existente entre el tiempo de impresión de un trabajo (Y) y el número de páginas del trabajo (X) a imprimir con el fin de gestionar mejor los pedidos y la asignación de tareas. Para ello se ha recogido una muestra de trabajos a lo largo de una semana y para cada uno de ellos se han anotado el número de páginas del trabajo y el tiempo de impresión del mismo. Tras algunos análisis estadísticos, se han obtenido los siguientes resultados:

Resumen Estadístico

	X	*Y*
Recuento	75	75
Promedio	5,44	57,6227

Covarianzas

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

Coeficientes Lineal: Y = a + b*X

Parámetro	*Estimación*	*Error estándar*	*Estadístico t*	*p-valor*
Constante		2,6272	5,14444	0,0000
Pendiente		0,428601		

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*F-ratio*	*p-valor*
Modelo	39343,3	1	39343,3	357,86	0,0000
Residuo	8025,61	73	109,94		
Total (Corr.)	47368,9	74			

A la vista de los resultados anteriores, responde a las siguientes preguntas justificando convenientemente la respuesta:

a) Calcula el Coeficiente de Correlación entre las variables estudiadas y describe la relación entre las mismas. *(1,5 puntos)*

b) ¿Qué forma tendría el diagrama de dispersión entre X e Y? Dibuja una representación aproximada. *(1 punto)*

c) Estima los parámetros del modelo de regresión lineal que permite estudiar el tiempo de impresión de un trabajo en función del número de páginas de éste. *(1,5 puntos)*

d) Estudia la significación estadística de cada uno de los parámetros (α=0,05) del modelo. ¿Cuál es la ecuación del modelo de regresión estimado a partir de los datos? *(2 puntos)*

e) De acuerdo con el modelo planteado, ¿qué tiempo de impresión se espera obtener, en promedio, para un trabajo comprendido por 6 páginas? *(1 punto)*

f) ¿Qué parámetro permite cuantificar la calidad del ajuste del modelo obtenido a nuestros datos? Calcula el valor de dicho parámetro e interpreta su significado en este estudio. *(1,5 puntos)*

g) Calcula la varianza residual e indica qué representa este valor en el estudio realizado. *(1,5 puntos)*

SOLUCIÓN

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:

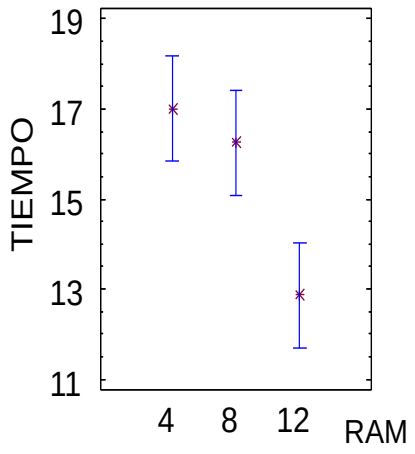
- Dado que el factor RAM resulta estadísticamente significativo, a la vista del gráfico se deduce que el tiempo medio a nivel poblacional con el procesador Celerr© es menor que para el otro procesador.
- Dado que el factor procesador también resulta significativo, al aumentar la RAM en promedio desciende 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: $r=\frac{\text{cov}}{\sqrt{s_x^2}\cdot\sqrt{s_y^2}}=\frac{65{,}573}{\sqrt{8{,}088\cdot 640{,}12}}=0{,}911$

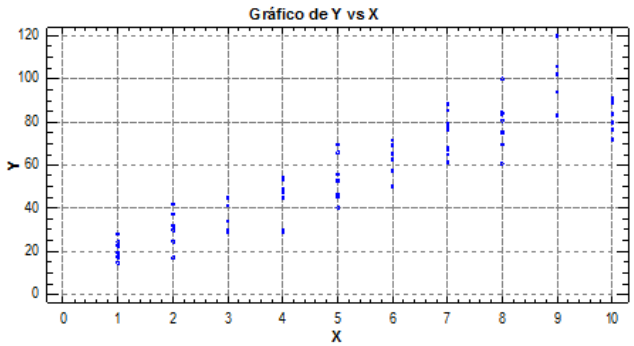
Dado que el valor obtenido es positivo y cercano a 1 podemos considerar que existe una relación lineal positiva y fuerte entre el *tiempo de impresión de un trabajo* (Y) y el *número de páginas del trabajo* (X) a imprimir.

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.