Grado en Ingeniería Informática

Estadística

SEGUNDO PARCIAL

7 de junio de 2016

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. El tiempo que tarda en compilarse un cierto tipo de programas es una variable aleatoria que fluctúa uniformemente entre 5 y 30 milisegundos. Se quieren compilar consecutivamente 50 programas.

a) ¿Qué tipo de distribución sigue la variable: tiempo total que tardarían en compilarse los 50 programas? Justifica la respuesta. Calcular la media y la varianza de dicha variable. *(4 puntos)*

b) Calcular aproximadamente la probabilidad de que el tiempo total sea superior a un segundo. *(3 puntos)*

c) Calcular entre qué límites estará comprendido aproximadamente el 95% de valores del tiempo total de compilación de los 50 programas. *(3 puntos)*

2. Teniendo en cuenta que el rendimiento diario de un sistema informático (en MIPS) puede considerarse una variable aleatoria con distribución normal, contesta justificadamente las siguientes preguntas.

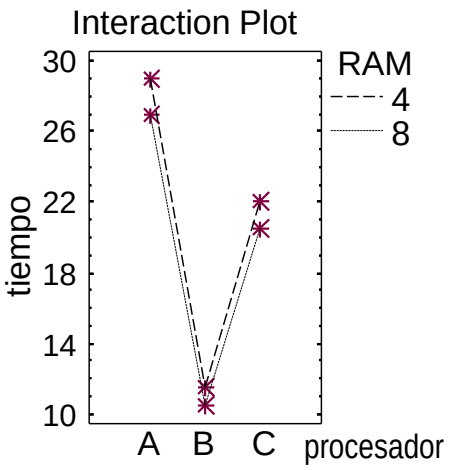
a) Se toman los valores del rendimiento del sistema correspondientes a 20 días al azar. Asumiendo que la desviación típica en la población vale $\sigma = 7$ MIPS, ¿cuál es la probabilidad de obtener una varianza muestral menor de 70 $MIPS^2$? *(4 puntos)*

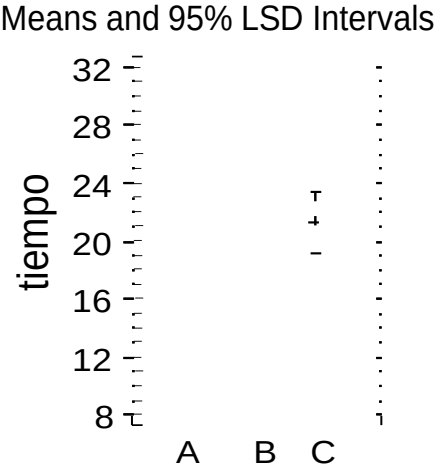
b) Se ha obtenido una muestra de valores de rendimiento correspondientes a 25 días tomados al azar. La media muestral ha resultado ser de 85 MIPS, con una desviación típica muestral de 6,5 MIPS. Considerando un riesgo de primera especie del 5%, ¿se puede admitir un rendimiento medio en la población de 95 MIPS? Justifica convenientemente la respuesta. *(4 puntos)*

c) ¿Qué se entiende por “riesgo de primera especie” de un test de hipótesis? *(2 puntos)*

3. 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. Para ello, se ejecuta el algoritmo en 12 ordenadores de diferentes características, 4 de ellos con un procesador de tipo A, 4 de tipo B y 4 de tipo C. La mitad de ellos tienen 4 GB de memoria RAM y la otra mitad, 8 GB. Los valores experimentales obtenidos de tiempo (variable "t" en microsegundos) se indican a continuación. Para analizar los valores obtenidos se ha utilizado un ANOVA, proporcionando las gráficas que se muestran a continuación, y las siguientes sumas de cuadrados: SC_{total} = 610,92; SC_{proc} = 586,17; SC_{RAM} = 6,75; $SC_{proc \cdot RAM}$ = 0,5.

PR.	RAM	t
A	4	31
A	4	27
A	8	28
A	8	26
B	4	12
B	4	11
B	8	10
B	8	11
C	4	21
C	4	23
C	8	22
C	8	19





a) Estudiar si el efecto simple de los dos factores y el de su interacción resulta estadísticamente significativo (α=0,05). *(4 puntos)*

b) A la vista del gráfico de la interacción, teniendo en cuenta los resultados obtenidos en el apartado anterior y considerando α=0,05, ¿cómo afecta la memoria RAM al tiempo de ejecución del algoritmo? *(2 puntos)*

c) En el gráfico de intervalos LSD se muestra el intervalo correspondiente al tipo de procesador C. Dibuja los intervalos LSD para los otros dos tipos de procesador, justificando convenientemente los cálculos. *(2 puntos)*

d) El procesador de tipo A tiene una velocidad de 1 GHz; el de tipo B, 3 GHz y el de tipo C, 2 GHz. Teniendo en cuenta toda la información disponible, ¿puede hablarse de un efecto lineal o cuadrático de la velocidad del procesador sobre el tiempo de ejecución de dicho algoritmo? *(2 puntos)*

4. El tiempo (en microsegundos) que tarda un programa informático en realizar cierta operación con matrices es función de dos parámetros: velocidad del procesador (GHz) y tamaño de la matriz. Se dispone de 50 matrices que han sido procesadas con dicho programa. La matriz de varianzas-covarianzas que relaciona las tres variables se indica a continuación, así como el resultado de dos modelos de regresión lineal: uno que relaciona tiempo y velocidad, y el segundo modelo que relaciona tiempo con tamaño.

	Tiempo	Velocidad	Tamaño
Tiempo	1,46648	-1,16058	4,49981
Velocidad	-1,16058	1,07253	-1,85217
Tamaño	4,49981	-1,85217	60,3879

Regression Analysis - Linear model: Y = a + b*X
Dependent variable: TIEMPO
Independent variable: VELOCIDAD

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	7,31367	0,258654	28,2759	0,0000
Slope	-1,08209	0,0639626	[illegible]	[illegible]

Regression Analysis - Linear model: Y = a + b*X
Dependent variable: TIEMPO
Independent variable: TAMAÑO

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	2,02794	0,317818	6,3808	0,0000
Slope	0,074515	0,0197547	[illegible]	[illegible]

a) ¿Se puede afirmar que la variable Tiempo presenta una correlación más fuerte con el tamaño que con la velocidad? *(2 puntos)*

b) A la vista de los resultados obtenidos, indica cuáles son los dos modelos de regresión estimados a partir de los análisis realizados. *(1 punto)*

c) ¿Hay suficiente evidencia para afirmar que el efecto lineal del tamaño es estadísticamente significativo? ¿Y el efecto lineal de la velocidad? Considerar α=0,05. *(2 puntos)*

d) ¿Cuál es el tiempo medio esperado al realizar una operación con una matriz, considerando una velocidad de 2 GHz? *(1 punto)*

e) Calcular la varianza residual del primer modelo de regresión planteado. *(2 puntos)*

f) En caso de que la velocidad sea de 2 GHz, calcular la probabilidad de que el tiempo de cálculo sea superior a 5 microsegundos. *(2 puntos)*

SOLUCIÓN

1a) La variable aleatoria de tiempo (X) sigue una distribución: U(5; 30).
E(X)=(5+30)/2 = 17,5; $\sigma^2 = (b-a)^2/12 = (30-5)^2/12 = 52{,}08$
Tiempo total: Y = X_1+ X_2 + ... + X_{50}. Según el Teorema Central del Límite, la suma de N variables aleatorias, sea cual sea su distribución, tiende a una distribución normal si N es suficientemente grande y las variables son independientes. Por tanto, el tiempo total seguirá una distribución normal, cuya media y varianza serán:

$$E(X_1+...+X_{50}) = E(X_1)+...+E(X_{50}) = 50\cdot E(X) = 50\cdot 17{,}5 = \mathbf{875}$$
$$\sigma^2(X_1+...+X_{50}) = \sigma^2(X_1)+...+\sigma^2(X_{50}) = 50\cdot\sigma^2(X) = 50\cdot 52{,}08 = \mathbf{2604{,}2}$$

1b) Ya que las unidades de X son milisegundos, P(Y > 1 seg.) = P (Y>1000)

$$P(Y>1000) = P\left[N\left(875,\sqrt{2604}\right)>1000\right] = P\left[N(0;1) > \frac{1000-875}{\sqrt{2604}}\right] = P[N(0;1)>2{,}45] = \mathbf{0{,}0071}$$

1c) En una distribución normal, el intervalo m ± 1,96·s comprende el 95% de los valores. Así pues, el intervalo será: $875 \pm 1{,}96\cdot\sqrt{2604} =$ **[775 ; 975]**.

También puede ser correcto redondeando a 2: $875 \pm 2\cdot\sqrt{2604} = [773\ ;\ 977]$.

2a) Sabiendo que $(n-1)\cdot s^2/\sigma^2 \approx \chi^2_{n-1}$, en este caso n=20 (tamaño de muestra) y σ=7, de modo que la probabilidad que nos piden se calcula del siguiente modo:

$$P(s^2<70) = P\left(s^2\cdot\frac{n-1}{\sigma^2} < 70\cdot\frac{n-1}{\sigma^2}\right) = P\left(\chi^2_{19} < 70\cdot\frac{19}{7^2}\right) = P(\chi^2_{19}<27{,}14) \approx \mathbf{0{,}9}$$

2b) La hipótesis nula que hay que contrastar es H_0: m=95, siendo n=25, s=6,5.

$\left|\frac{\bar{x}-m_0}{s/\sqrt{n}}\right| = \left|\frac{85-95}{6{,}5/5}\right| = 7{,}7$ Este valor tiene que comparase con: $t^{\alpha/2}_{n-1} = t^{0.025}_{24} = 2{,}064$

El valor de la t-calculada es mayor que el valor crítico de una t de Student con 24 grados de libertad, lo que implica que 7,7 no es un valor frecuente de esta distribución. Por tanto, se rechaza la hipótesis nula: no se puede admitir un rendimiento medio en la población de 95 MIPS, considerando α=0,05.

Otro método alternativo consiste en comprobar que 95 está fuera del intervalo de confianza de la media poblacional, que resulta ser: [82,32; 87,68].

2c) Por definición, el riesgo de primera especie (también llamado nivel de significación) de un test de hipótesis es la probabilidad de rechazar la hipótesis nula cuando es cierta.

3a) La tabla del ANOVA se construye en base a los siguientes cálculos:
1) Grados de libertad totales = 11 (12 datos menos uno)
2) Grados de libertad del factor proveedor = 3 variantes - 1 = 2.
3) Grados de libertad del factor RAM = 2 niveles - 1 = 1
4) Grados de libertad de la interacción = 2·1 = 2
5) Grados de libertad residuales (obtenidos por diferencia) = 11 - 2 - 1 - 2 = 6

6) $SC_{resid} = SC_{total} - SC_{prov.} - SC_{RAM} - SC_{prov \cdot RAM} = 586,17 - 6,75 - 0,5 = 17,5$
7) Cuadrado medio: $CM_{proveedor} = SC_{prov} / gl_{prov} = 586,17 / 2 = 293,08$
8) $CM_{RAM} = SC_{RAM} / gl_{RAM} = 6,75 / 1 = 6,75$
9) $CM_{prov \cdot RAM} = SC_{prov \cdot RAM} / gl_{prov \cdot RAM} = 0,5 / 2 = 0,25$
10) $F_{prov} = CM_{prov}/CM_{res} = 293,08 / 2,917 = 100,49$
11) $F_{RAM} = CM_{RAM}/CM_{res} = 6,75 / 2,917 = 2,31$
12) $F_{prov \cdot RAM} = CM_{prov \cdot RAM}/CM_{res} = 0,25 / 2,917 = 0,09$

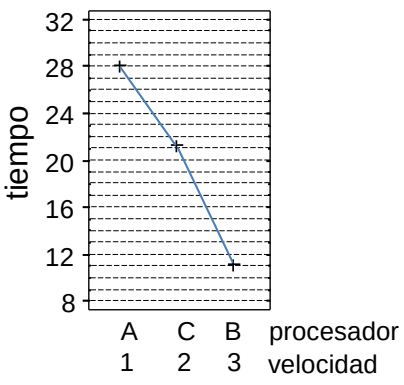
Source	Sum of Squares	Df	Mean Square	F-Ratio
A:PROVEEDOR	586,17	2	293,085	100,49
B:RAM	6,75	1	6,75	2,31
INTERACTION	0,5	2	0,25	0,09
RESIDUAL	17,5	6	2,91667	
TOTAL (CORRECTED)	610,92	11		

El efecto del factor proveedor es estadísticamente significativo para α=0,05 porque su F-ratio = 100,49 es mayor que el valor crítico de tablas: $F_{2;6}^{0.05}$=5,14.
Sin embargo, el efecto del factor RAMr **no** es estadísticamente significativo porque su F-ratio = 2,31 es menor que el valor crítico de tablas: $F_{1;6}^{0.05}$=5,99.
El efecto de la interacción tampoco es estadísticamente significativo porque su F-ratio = 0,09 es menor que el valor crítico: $F_{2;6}^{0.05}$=5,14.

3b) El efecto de la interacción no resulta estadísticamente significativo, lo cual es coherente con las líneas paralelas observadas en el gráfico de la interacción. Dicho gráfico muestra también que los valores medios de tiempo observados para RAM=8 son menores que para RAM=4. Sin embargo, el efecto de memoria RAM **no** resulta estadísticamente significativo, lo que implica que no hay suficiente evidencia para afirmar que los ordenadores con RAM=8 ejecuten el algoritmo con menor tiempo en promedio que los ordenadores con RAM=4.

3c) La anchura de los intervalos LSD depende del número de datos. Dado que hay 4 valores para cada proveedor, los tres intervalos tendrán la misma anchura que en el caso del proveedor C, el cual se muestra en la figura, y que está comprendido aproximadamente entre 19,1 y 23,3 (es decir, punto medio ±2,1). El punto intermedio es la media muestral.
Media de valores para A: (31+27+28+26)/4 = 28
Media de valores para B: (12+11+10+11)/4 = 11
Intervalos LSD para A: 28 ±2,1 [25,9 ; 30,1]; proveedor B: 11 ±2,1 [8,9 ; 13,1]

3d) Los valores medios de tiempo obtenidos para los procesadores A, B, and C son: 28, 11 y 21.25, respectivamente. Teniendo en cuenta la velocidad del procesador, el gráfico de la derecha indica el valor medio en función de la velocidad. Se puede ajustar una línea recta razonablemente bien a estos valores medios, lo cual indica que la velocidad del procesador ejerce un efecto lineal en el tiempo.



4a) Si llamamos Y: tiempo, X_1: velocidad y X_2: tamaño, las varianzas corresponden a los valores de la diagonal principal de la matriz de varianzas-covarianzas: $s_Y^2 = 1{,}4665$; $s_{X_1}^2 = 1{,}0725$; $s_{X_2}^2 = 60{,}388$

También se obtiene de dicha matriz: $\text{cov}(X_1, Y) = -1{,}1606$; $\text{cov}(X_2, Y) = 4{,}4998$

$$R^2_{X_1,Y} = r^2 = \frac{\text{cov}^2(X_1,Y)}{s^2_{X_1} \cdot s^2_Y} = \frac{-1{,}1606^2}{1{,}0725 \cdot 1{,}4665} = 0{,}8564$$

$$R^2_{X_2,Y} = r^2 = \frac{\text{cov}^2(X_2,Y)}{s^2_{X_2} \cdot s^2_Y} = \frac{4{,}4998^2}{60{,}388 \cdot 1{,}4665} = 0{,}229$$

No puede afirmarse que el *tiempo* presente una correlación más fuerte con el *tamaño* que con la *velocidad*, ya que el coeficiente de determinación (R^2) entre *tiempo* y *tamaño* (X_2) es menor que en el caso de *velocidad*.

4b) Teniendo en cuenta los valores estimado que se indican en la tabla de resultados, las ecuaciones que nos piden son las siguientes:
Primer modelo: Tiempo = 7,314 - 1,082 · velocidad
Segundo modelo: Tiempo = 2,028 + 0,0745 · tamaño

4c) En el primer modelo, el efecto lineal de *velocidad* será estadísticamente significativo si la pendiente de la línea recta es distinta de cero a nivel poblacional. Así pues, dada la ecuación $Y=\beta_0+ \beta_1 \cdot X$, la hipótesis nula será $\beta_1=0$ frente a la alternativa H_1: $\beta_1 \neq 0$. Dado que $b_i / s_{b_i} \approx t_{N-1-I} \approx t_{48}$ siendo I=1 (modelo con una variable explicativa), N=50 y $\alpha=0{,}05$, se obtiene que: $b_i / s_{b_i} = -1{,}082/0{,}064 = -16{,}9$ el cual es un valor poco frecuente de una distribución t_{48}. Por ello, se rechaza H_0: hay suficiente evidencia para afirmar que el efecto lineal de *velocidad* resulta estadísticamente significativo.

En el segundo modelo: $b_i / s_{b_i} = 0{,}0745/0{,}0197 = 3{,}77$ el cual es un valor poco frecuente de una distribución t_{48} (95% de valores comprendidos entre -2.011 y 2.011). Por lo tanto, el efecto simple de *tamaño* también es significativo.

4d) El valor pedido se obtiene sustituyendo velocidad=2 en el primer modelo:
E (tiempo / velocidad=2) = 7,314 - 1,082 · 2 = **5,15**

4e) La varianza residual se calcula como:
$s^2_{res} = s^2_Y \cdot (1 - r^2_{X_1,Y}) = 1{,}4665 \cdot (1 - 0{,}8564) =$ **0,2106**

4f) Cuando velocidad=2, la distribución condicional del *tiempo* será de tipo normal, siendo la media obtenida por el modelo ajustado (5,15) y la varianza será la de los residuos (0,199). A partir de esta distribución:

$$P(t > 5) = P\left[N\left(5{,}15; \sqrt{0{,}21}\right) > 5\right] = P\left[N(0;1) > \frac{5 - 5{,}15}{\sqrt{0{,}21}}\right] = P[N(0;1) > -0{,}327] = 1 - 0{,}37 = \mathbf{0{,}63}$$