Grado en Ingeniería Informática

Estadística

SEGUNDO PARCIAL

2 de junio de 2014

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. En un estudio sobre un tipo de procesador, se han realizado 9 pruebas registrándose las velocidades de ejecución de un determinado tipo de operación. El valor medio de los datos ha resultado 52,02 y la desviación típica 0,82.

a) Calcula un intervalo de confianza para la media (nivel de confianza 95%). ¿Se puede admitir una media m=53? *(3,5 puntos)*

b) Asumiendo que la media poblacional fuese 53, ¿cuál es la probabilidad de haber obtenido una media muestral superior a 52,02? *(3 puntos)*

c) ¿Se puede admitir una desviación típica σ=1,6? Utiliza un riesgo de primera especie α=1%. *(3,5 puntos)*

2. Se realiza un experimento para ensayar el efecto de dos factores (modelo de procesador y memoria RAM) en el tiempo (en milisegundos) que se tarda en realizar una búsqueda en una base de datos de grandes dimensiones. Se ensayan tres modelos de procesador y dos memorias (10MB y 20MB), realizándose dos repeticiones de cada una de las posibles combinaciones. Los datos obtenidos se indican en la siguiente tabla. Se asume que los datos son normales.

Memoria RAM	modelo 1	modelo 2	modelo 3
10 MB	102; 108	171; 176	119; 125
20 MB	92; 97	157; 164	117; 123

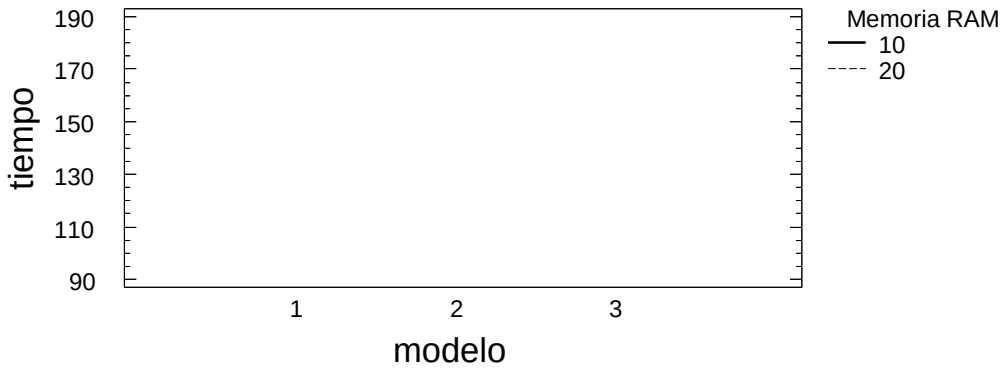
Los resultados obtenidos con Statgraphics son los siguientes:

Analysis of Variance for TIEMPO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:memoria_RAM	216,75	[illegible]		
B:modelo	9453,5	[illegible]		
INTERACTIONS				
AB	[illegible]	[illegible]		
RESIDUAL	[illegible]	[illegible]	17,25	
TOTAL (CORRECTED)	9840,25	[illegible]		

a) Determinar si el efecto simple de alguno de los factores o de la interacción es estadísticamente significativo, considerando α=5%. *(4 puntos)*

b) Dibuja el gráfico de medias de la interacción, justificando la respuesta. A la vista de éste y teniendo en cuenta los resultados del ANOVA, interpreta la información que contiene el gráfico. *(3,5 puntos)*



c) Teniendo en cuenta los resultados del ANOVA, determinar las condiciones operativas óptimas con las cuales se minimiza el tiempo de búsqueda en la base de datos, considerando α=1%. Calcular el tiempo medio previsto en dichas condiciones. *(2,5 puntos)*

3. Una compañía discográfica está investigando la posibilidad de pronosticar las ventas (miles de discos/mes) a partir de la inversión realizada en publicidad (miles euros/mes). Para ello se ha estudiado un grupo de datos históricos de ventas e inversiones y se ha llevado a cabo un análisis de regresión lineal. Algunos de los resultados obtenidos aparecen a continuación (α=0,01):

Simple Regression - Ventas vs. Publicidad
Dependent variable: Ventas (miles de discos/mes)
Independent variable: Publicidad (miles euros/mes)
Linear model: Y = a + b*X
Coefficients

Parameter	*Estimate*	*Standard Error*	*T Statistic*	*P-Value*
Intercept	134,14	7,53658		0,0000
Slope	0,0961245	0,00963236		

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	433688,	1	433688,	99,59	0,0000
Residual	862264,	198	4354,87		
Total (Corr.)	1,29595E6	199			

Correlation Coefficient = 0,578488

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

a) Estima los parámetros del modelo de regresión, estudia su significación y plantea dicho modelo. *(3 puntos)*

b) ¿Qué volumen medio tendrán las ventas mensuales previstas si la inversión en publicidad en un cierto mes es de 10.000 euros? *(2 puntos)*

c) Calcula el Coeficiente de Determinación del modelo. ¿Qué representa en la práctica este coeficiente? *(2,5 puntos)*

d) Explica qué representa el Cuadrado Medio Residual (CMR) e indica cuál es su valor en este ejemplo. *(2,5 puntos)*

SOLUCIÓN DEL SEGUNDO PARCIAL

1a) El intervalo de confianza para la velocidad media con nivel de confianza 95%, se calcula con la fórmula:

$$[\overline{X} - t_8^{\alpha/2=0,025}\frac{s}{\sqrt{N}}, \overline{X} + t_8^{\alpha/2=0,025}\frac{s}{\sqrt{N}}] \qquad [52,02 - 2,306\frac{0,82}{\sqrt{9}}, 52,02 + 2,306\frac{0,82}{\sqrt{9}}]$$

El intervalo es: [**51,39, 52,65**] Como el valor m=53 no esté en el intervalo, **no** es admisible considerar m=53.

1b) $P(\bar{x} > 52,02) = P\left(\frac{\bar{x}-m}{s/\sqrt{n}} > \frac{52,02-m}{s/\sqrt{n}}\right) = P\left(t_8 > \frac{52,02-53}{0,82/\sqrt{9}}\right) = P(t_8 > -3,58) \approx$

$\approx$1-0,005 = **0,995** (valor exacto con Statgraphics: 0,996)

1c) El intervalo para la desviación típica con α=1% es:

$$\left[\sqrt{(N-1)\frac{s^2}{g_2}}, \sqrt{(N-1)\frac{s^2}{g_1}}\right] \qquad \left[\sqrt{(9-1)\frac{0,82^2}{g_2}}, \sqrt{(9-1)\frac{0,82^2}{g_1}}\right]$$

Donde g_1=**1,344** (en la tabla de la distribución Chi-cuadrado en probabilidad de cola derecha 0,995 con 8 grados de libertad) y g_2=**21,955** (en la misma tabla pero con probabilidad 0,005).
El intervalo resulta: [**0,49, 2**]. Como el valor σ=1,6 está dentro del intervalo, se puede admitir esa desviación típica en la población.

2a) Grados de libertad totales = 12 - 1 = 11
Grados de libertad del factor memoria RAM = 2 niveles - 1 = 1
Grados de libertad del factor modelo = 3 variantes - 1 = 2
Grados de libertad de la interacción: 1 · 2 = 2
Grados de libertad residuales, se obtienen por diferencia: 11 - 1 - 2 - 2 = 6

$SC_{residual} = CM_{resid} \cdot gl_{resid} = 17,25 \cdot 6 = 103,5$
$SC_{interac} = SC_{total} - SC_{res} - SC_{RAM} - SC_{modelo} = 9840,25-103,5-216,75-9453,5 = 66,5$
$F_{ratioRAM} = (SC/gl)/CM_{res} = (216,75/1) / 17,25 = 12,57$
$F_{ratio\_modelo} = (SC/gl)/CM_{res} = (9453,5/2) / 17,25 = 274,01$

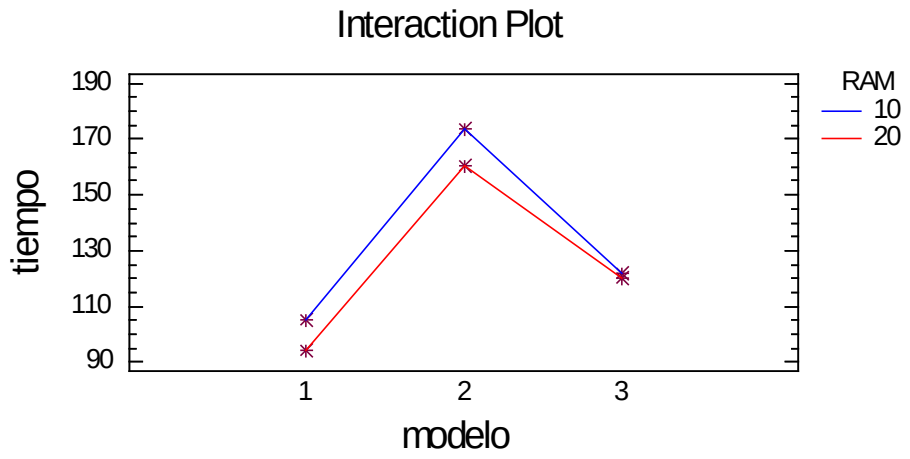
Considerando α=0,05, el efecto simple del factor memoria RAM es estadísticamente significativo ya que el F-ratio (12,57) es mayor al valor crítico de tablas ($F_{1;6}$) que vale 5,99.
El efecto simple del factor modelo es estadísticamente significativo ya que el F-ratio (274,01) es mayor al valor crítico de tablas ($F_{2;6}$) que vale 5,14.
El efecto de la interacción NO es estadísticamente significativo ya que el F-ratio (1,93) es menor al valor crítico de tablas ($F_{2;6}$) que vale 5,14.

La tabla resumen completa es la siguiente (se muestra también el p-valor aunque éste sólo puede calcularse con Statgraphics).

Analysis of Variance for TIEMPO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:memoria_RAM	216,75	1	216,75	12,57	0,0121
B:modelo	9453,5	2	4726,75	274,01	0,0000
INTERACTIONS					
AB	66,5	2	33,25	1,93	0,2257
RESIDUAL	103,5	6	17,25		
TOTAL (CORRECTED)	9840,25	11			

2b) Valores medios de cada tratamiento (a partir de la tabla de los datos):
(102+108)/2=105 (171+176)/2= 173,5 (119+125)/2=122
(92+97)/2= 94,5 (157+164)/2=160,5 (117+123)/2=120
Estos valores se representan en el gráfico, obteniéndose:



La interacción no resulta estadísticamente significativa, de modo que no hay suficiente evidencia para afirmar que el efecto de la memoria RAM sea distinto para cada uno de los tres modelos. Así pues, a nivel poblacional, hay que admitir que el tiempo con RAM=10 es significativamente mayor que con RAM=20, independientemente del tipo de modelo. Respecto al modelo, que es un factor cualitativo, el tiempo para modelo=2 resulta también significativamente mayor que para modelo=1. En caso de modelo=3, el tiempo es intermedio, pero a partir del gráfico no puede afirmarse si las diferencias respecto a los otros dos modelos resultan estadísticamente significativas, ya que no se tienen los intervalos LSD.

2c) Si se considera α=1%, el factor modelo es estadísticamente significativo por ser la F-ratio mayor al valor crítico de una $F_{2;6}$ (274,01 >> 10,92). Sin embargo, el factor memoria RAM no es significativo, por ser la F-ratio menor al valor crítico de una $F_{1;6}$ (12,57 < 13,75). Por tanto, no hay evidencia suficiente para afirmar que se tarde más tiempo a nivel poblacional con RAM=10 que con RAM=20. Así pues, en las condiciones operativas óptimas se tiene en cuenta sólo el factor "modelo". A la vista del gráfico de la interacción, se obtendrá un menor tiempo con el modelo 1. El tiempo medio previsto con el modelo 1 será: (102+108+92+97)/4 = **99,75**.

Solución problema 3:

3a) El modelo de regresión (Y=a+b·X) tiene dos parámetros: la ordenada en el origen (a) y la pendiente (b). El valor estimado de ambos parámetros se obtiene a partir de la tabla de resultados: ordenada (*intercept*) = 134,14; pendiente (*slope*) = 0,09612.

El parámetro a (*Intercept*) resulta estadísticamente significativo (es decir, distinto de cero a nivel poblacional) ya que su p-valor (*p-value*) es menor que 0,01. El p-valor asociado a la pendiente (*slope*) no se indica en la tabla de "coefficients", pero éste coincide con el p-valor de la tabla "analysis of variance" (test de significación global del ajuste) que es menor de 0,01 por lo que también resulta significativo.

Así pues, el modelo que habría que plantear es el siguiente:

Y= 134,14 + 0,09612·X ←→ **Ventas = 134,14 + 0,09612·Publicidad**

3b) Si la inversión es de 10.000 euros, como las unidades son en miles de euros, hay que sustituir la variable "publicidad" por 10. El volumen medio estimado de las ventas será:

Ventas = 134,14+0,09612·10 = **135,101** miles de discos/mes = 135.101 discos/mes

3c) $R^2 = \frac{SCE_{(modelo)}}{SCT_{(total)}} \cdot 100 = \frac{433688}{1295950} \cdot 100 = 33{,}465\%$

En el caso particular de modelos de regresión simple, como es el caso, el Coeficiente de Determinación también se puede obtener a partir del Coeficiente de Correlación como: $\mathbf{R^2 = (r_{xy})^2 \cdot 100}$ = $(0{,}578488)^2 \cdot 100 \cong$ **33,5%**
$\mathbf{R^2}$ permite valorar la **calidad del ajuste**. Mide el porcentaje de la variación de la v.a. dependiente (las ventas mensuales de discos) explicado o provocado por los cambios en la v.a. independiente (la inversión mensual en publicidad).

3d) El **CMR** es una estimación de la **Varianza Residual** (σ^2_R), es decir la varianza de los residuos (diferencia entre valor observado y valor predicho por el modelo de regresión). Esta varianza representa el orden de magnitud del efecto conjunto de todos los factores (variables aleatorias) no considerados en el modelo (recta de regresión), incluidos las anomalías y factores no controlables.
Mean Square → **CMR = 4354,87** unidades^2

En el caso particular de modelos de regresión simple, como es el caso que nos ocupa, la varianza residual también se habría podido obtener como:

$\mathbf{S^2_{residual} = S^2_y(1-r^2_{xy})}$ = $6512{,}32 \cdot (1-(0{,}578488)^2) \cong$ **4332,98** unidades^2