Grado en Ingeniería Informática

Estadística

SEGUNDO PARCIAL

1 de junio de 2015

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 el rendimiento de un sistema informático se ha obtenido una muestra de 14 valores del tiempo de acceso a los datos almacenados (en segundos). La media muestral ha resultado 5,4 s y la desviación típica muestral vale 1,9 s. Se asume que los datos siguen una distribución Normal.

a) Calcula un intervalo de confianza para la media del tiempo de acceso con un nivel de confianza del 95%. ¿Qué interpretación práctica tiene la probabilidad 0.95 asociada al correspondiente intervalo de confianza? *(3 puntos)*

b) ¿Puede admitirse que el tiempo medio de acceso en la población es de 4 s con un riesgo de primera especie de 0,05? Justifica la respuesta, indicando cuál es el test de hipótesis que se plantea. *(1,5 puntos)*

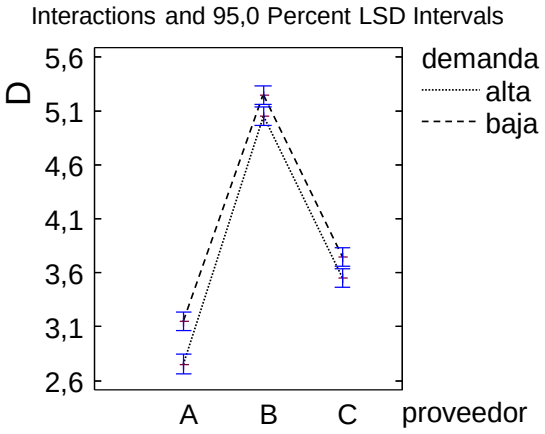
c) A partir de los cálculos realizados en el apartado a), ¿puede admitirse que el tiempo medio de acceso en la población es de 4 s con un riesgo de primera especie de 0,1? Justifica tu respuesta. *(1 punto)*

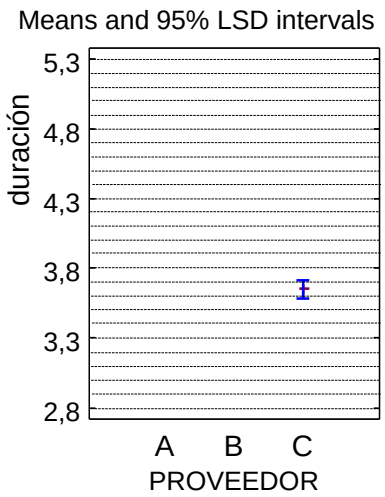
d) ¿Puede admitirse que la muestra procede de una población con una desviación típica (σ) de 2 s? (α=5%). *(3 puntos)*

e) ¿Qué tipos de decisiones erróneas podemos tomar al plantear un test de hipótesis? ¿Qué nombres reciben estos tipos de errores? *(1,5 puntos)*

2. Para estudiar si la duración (D) de las baterías (en años) es distinta en función de tres proveedores disponibles (A, B y C), se toma una muestra de 4 baterías de cada uno y se ensayan en condiciones similares, dos de cada muestra en condiciones de demanda energética alta y las otras dos en demanda energética baja. Los datos resultantes se muestran en la siguiente tabla.

PR	DEM	D
A	alta	2,7
A	alta	2,8
A	baja	3,2
A	baja	3,1
B	alta	5,0
B	alta	5,1
B	baja	5,2
B	baja	5,3
C	alta	3,5
C	alta	3,6
C	baja	3,7
C	baja	3,8





Analysis of Variance for DURACION - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:PROVEEDOR				
B:DEMANDA	0,213333			
INTERACTIONS				
AB				2,67
RESIDUAL			0,005	
TOTAL (CORRECTED)	10,3767			

a) Completa la tabla del ANOVA, indicando los cálculos realizados. ¿Qué efectos son estadísticamente significativos? (considerar α=0,05).

(3,5 puntos)

b) ¿Los resultados obtenidos en el apartado anterior son coherentes con el gráfico de la interacción mostrado? Justifica la respuesta. *(1,5 puntos)*

c) A la vista de los resultados obtenidos, ¿cómo afecta la demanda energética a la duración de las baterías? *(1,5 puntos)*

d) En el gráfico de intervalos LSD se muestra solamente el intervalo correspondiente al proveedor C. Dibuja los otros dos intervalos, justificando los cálculos realizados. ¿Puede hablarse de un efecto lineal o cuadrático de alguno de los factores? *(2,5 puntos)*

e) Describir cómo se podría estudiar en este caso la existencia de datos anómalos. *(1 punto)*

3. Se ha realizado un estudio de regresión con el fin de evaluar los tiempos empleados por un nuevo algoritmo de ordenación de vectores. En el estudio se han considerado vectores entre 500 y 800 elementos. Así, se han tenido en cuenta diferentes tamaños de problema (variable TAMAÑO, número de elementos del vector) y se ha registrado el tiempo medio requerido en la ordenación de diferentes instancias (variable TPO_ORDEN, expresado en microsegundos). El ajuste obtenido mediante Statgraphics a partir del modelo de regresión lineal simple planteado ha sido el siguiente:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: TPO_ORDEN
Independent variable: TAMAÑO
------------------------------------------------------------------------------
                            Standard          T
Parameter       Estimate      Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept       -323,203      216,543                       0,1476
Slope            1,83038     0,319413
------------------------------------------------------------------------------
                          Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square   F-Ratio    P-Value
------------------------------------------------------------------------------
Model                    1,1833E6
Residual                 936897,0
------------------------------------------------------------------------------
Total (Corr.)           2,12019E6     27
```

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

a) ¿Cuál es el modelo de regresión estimado con el fin de predecir el tiempo de ordenación de un vector en función de su tamaño? ¿Te parece lógico el valor estimado de los parámetros del modelo teniendo en cuenta su interpretación? Justifica convenientemente tu respuesta. *(2 puntos)*

b) Justifica si el efecto lineal del tamaño resulta estadísticamente significativo (α=0,05). *(2 puntos)*

c) ¿Entre qué límites fluctuará, en el 95% de los casos, el tiempo de ordenación de un vector de 800 elementos utilizando dicho algoritmo? *(2,5 puntos)*

d) Calcula el valor del coeficiente de correlación entre el tiempo de ordenación de un vector y su tamaño. *(1,5 puntos)*

e) Además del tiempo de ordenación, en el estudio se dispone de otras dos variables continuas (Y_2 e Y_3) medidas sobre la misma población de vectores. A partir de estas tres variables (Y_1, Y_2 e Y_3) se han obtenido las siguientes covarianzas muestrales: $cov(Y_1, Y_2) = 24,16$; $cov(Y_1, Y_3) = 18,39$; $cov(Y_2, Y_3) = -57,12$. ¿Puede afirmarse que la correlación muestral más fuerte se da entre las variables Y_2 e Y_3? Razona la respuesta. *(2 puntos)*

SOLUCIÓN

1a) Intervalo de confianza para la media poblacional (m) con un nivel de confianza del 95% (α=0,05):

$[\overline{X} - t_{13}^{0,025}\frac{s}{\sqrt{N}}, \overline{X} + t_{13}^{0,025}\frac{s}{\sqrt{N}}]$; $[5{,}4 - 2{,}160\frac{1{,}9}{\sqrt{14}};\ 5{,}4 + 2{,}160\frac{1{,}9}{\sqrt{14}}]$; **[4,303; 6,497]**

La probabilidad 0.95 asociada al intervalo de confianza obtenido significa que si se toman 100 muestras de la población y de cada una de ellas se calcula el intervalo de confianza, en promedio 95 de estos intervalos contendrán el verdadero valor de la media poblacional, y 5 de ellos no lo contendrán. Por tanto, existe una probabilidad de 0.95 de que el verdadero valor desconocido de la media poblacional esté dentro del intervalo que se ha calculado.

1b) El test de hipótesis plantea si el tiempo medio de acceso en la población es de 4 s, frente a la hipótesis alternativa de que sea distinto de 4:
H_0: m=4 ; H_1: m≠4. Dado que el valor 4 no está incluido en el intervalo calculado en el apartado anterior (obtenido con 1-α=0.95), **no** puede admitirse que m=4 (considerando α=0,05).

1c) El valor 4 queda fuera del intervalo de confianza obtenido en el apartado a) con α=0,05. Considerando α=0,10, el intervalo se hace más estrecho, de modo que seguro que el valor 4 quedará también fuera del intervalo, por lo cual no puede admitirse m=4. El intervalo será más estrecho porque el valor crítico de la t de Student con α=10% vale $t_{13}^{0,05}$ = 1,771 en lugar de $t_{13}^{0,025}$ = 2,160.

1d) Intervalo de confianza para σ: $\left[\sqrt{(n-1)\cdot s^2/g_2};\sqrt{(n-1)\cdot s^2/g_1}\right]$ siendo g_1 el valor crítico que cumple P($\chi^2_{13} > g_1$)=0,975 y siendo g_2 el valor crítico tal que P($\chi^2_{13} > g_2$)=0,025. A partir de tablas se obtiene: g_1 =5,009 ; g_2 =24,736. Siendo n-1=13: $\left[\sqrt{13\cdot 1{,}9^2/24{,}736};\sqrt{13\cdot 1{,}9^2/5{,}009}\right] = [1{,}377;\ 3{,}061]$

Como el valor σ=2 está dentro del intervalo obtenido, puede admitirse que la muestra procede de una población con σ=2.

1e) En un test de hipótesis se pueden cometer dos tipos de decisiones erróneas:
- Rechazar la hipótesis nula cuando realmente es cierta: se denomina error de primera especie, de tipo I o error α.
- Aceptar la hipótesis nula cuando realmente es falsa: se denomina error de segunda especie, error de tipo II o error β.

2a) Para completar la tabla del ANOVA los cálculos son los siguientes:
1) Grados de libertad totales = 11 (12 datos que hay en total menos uno)
2) Grados de libertad del factor proveedor = 3 variantes - 1 = 2.
3) Grados de libertad del factor demanda = 2 niveles - 1 = 1
4) Grados de libertad de la interacción = 2·1 = 2
5) Grados de libertad residuales (por diferencia) = 11 - 2 - 1 - 2 = 6

6) $SC_{resid} = CM_{resid} \cdot gl_{res} = 0{,}005 \cdot 6 = 0{,}03$
7) $CM_{A\cdot B} / CM_{res} = \text{F-ratio}_{A\cdot B}$; $CM_{A\cdot B} = 2{,}67 \cdot 0{,}005 = 0{,}013333$
8) $SC_{A\cdot B} = CM_{A\cdot B} \cdot gl_{A\cdot B} = 0{,}01333 \cdot 2 = 0{,}02667$
9) $SC_{prov} = SC_{total} - SC_{res} - SC_{AB} - SC_{dem} = 10{,}377 - 0{,}03 - 0{,}0267 - 0{,}213 = 10{,}1067$
10) $CM_{prov} = SC_{prov} / gl_{prov} = 10{,}1067 / 2 = 5{,}0533$
11) $CM_{dem} = SC_{dem} / gl_{dem} = 0{,}2133 / 1 = 0{,}2133$
12) $F_{prov} = CM_{prov}/CM_{res} = 5{,}0533 / 0{,}005 = 1010{,}67$
13) $F_{dem} = CM_{dem}/CM_{res} = 0{,}2133 / 0{,}005 = 42{,}67$

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:PROVEEDOR	10,1067	2	5,05333	1010,67
B:DEMANDA	0,213333	1	0,213333	42,67
INTERACTIONS				
AB	0,02667	2	0,0133333	2,67
RESIDUAL	0,03	6	0,005	
TOTAL (CORRECTED)	10,3767	11		

El efecto de proveedor es estadísticamente significativo para α=0,05 porque su F-ratio= 1010,67 es superior al valor crítico de tablas: $F_{2;6}^{0,05}$=5,14.
El efecto de demanda también es significativo porque su F-ratio=42,67 es superior al valor crítico de tablas: $F_{1;6}^{0,05}$=5,99.
Sin embargo, el efecto de la interacción no es estadísticamente significativo porque su F-ratio=2,67 resulta inferior al valor crítico de tablas: $F_{2;6}^{0,05}$=5,14.

2b) Los resultados obtenidos en el apartado anterior son los siguientes:
- La suma de cuadrados del factor proveedor es la mayor de todas, lo que indica que este factor es el principal responsable de la variabilidad de los datos. Esto es coherente con el gráfico de la interacción, pues existen diferencias muy marcadas sobre todo entre los proveedores A y B.
- El efecto del factor demanda es estadísticamente significativo, pero su suma de cuadrados es bastante inferior a la de proveedor. Esto es coherente con el gráfico, pues al pasar de demanda baja a alta disminuye la duración, pero estas diferencias son claramente menores a las observadas entre los proveedores. Además, el hecho de que no todos los intervalos LSD se solapen entre demanda alta y baja también sugiere que el factor demanda resulta estadísticamente significativo.
- La interacción no resulta estadísticamente significativa: es coherente con el gráfico ya que las rectas correspondientes a demanda alta o baja son prácticamente paralelas.

2c) Cuando la demanda energética es alta, la duración de las baterías es significativamente menor. Este efecto de la demanda es independiente del tipo de proveedor, ya que la interacción no resulta estadísticamente significativa. Teniendo en cuenta que la media de los datos con demanda energética alta es 3,78 años y la duración media con demanda baja es 4,05 años (la diferencia es

0,27), puede concluirse que cuando la demanda energética es baja la duración aumenta en promedio 0,27 años para cualquiera de los tres proveedores.

2d) La anchura de los intervalos LSD depende del número de datos y del nivel de confianza. Dado que hay 4 datos de cada proveedor y el nivel de confianza es el mismo, los tres intervalos tendrán la misma anchura que en el caso del proveedor C mostrado en la figura (aproximadamente ±0,05).
Media de los datos de A: (2,7+2,8+3,2+3,1)/4 = 2,95
Media de los datos de B: (5+5,1+5,2+5,3)/4 = 5,15
Intervalos LSD del proveedor A: 2,95 ±0,05; proveedor B: 5,15 ±0,05

- Proveedor: se trata de un factor cualitativo, por lo que **no** puede hablarse de efecto lineal ni cuadrático (de hecho, reordenando el orden de los proveedores cambia totalmente la forma del gráfico).
- Demanda: se trata de un factor cuantitativo pero no puede saberse si el efecto es lineal o cuadrático porque sólo se han ensayado dos niveles. Para ello sería necesario disponer de información a más de dos niveles.

2e) Para estudiar la existencia de datos anómalos en un ANOVA, el método más eficiente consiste en guardar los residuos (correspondientes al modelo donde se incluyen todos los efectos estadísticamente significativos) y, posteriormente, hay que representar dichos residuos en un papel probabilístico normal. Todos los valores que se alejen claramente de la recta, bien por ser anormalmente altos o bajos, pueden considerarse como anómalos.

3a) Teniendo en cuenta los valores estimados indicados en la tabla de resultados, la ecuación del modelo de regresión es la siguiente:
Tpo_orden = -323,203 + 1,83038 · TAMAÑO
Aunque el p-valor de la constante no resulta estadísticamente significativo, es incorrecto despreciarla y emplear el modelo: Tpo_orden = 1,83038·TAMAÑO

- El valor estimado de la pendiente de la recta (1,83) resulta lógico ya que a mayor tamaño del vector, más tiempo se necesitará para su ordenación.
- El valor estimado de la constante del modelo (-323,203) podría parecer ilógico, ya que cuando el tamaño tiende a cero el tiempo sería negativo lo cual es absurdo. Pero esta situación no puede darse ya que el modelo se ha ajustado dentro de un rango de valores de tamaño entre 500 y 800 elementos. Cuando el tamaño es 500 el tiempo estimado es 592 μs, lo que garantiza que los tiempos predichos son siempre positivos en el rango de tamaños ensayados.

3b) El tamaño ejerce un efecto lineal estadísticamente significativo en el tiempo si la pendiente de la recta es distinta de cero a nivel de la población. Por tanto, si se ajusta un modelo: $Y=\beta_0+\beta_1\cdot X$, se trata de contrastar la hipótesis nula H_0: $\beta_1=0$ frente a la alternativa H_1: $\beta_1\neq 0$. Teniendo en cuenta que $b_i/s_{b_i} \approx t_{N-1-I} \approx t_{26}$ siendo I=1 (modelo con una variable explicativa), N=28 (nº total de observaciones = grados de libertad totales + 1) y $\alpha=0,05$, resulta que: $b_i/s_{b_i} = 1,83/0,319 = 5,73$ el cual es un valor poco frecuente de la distribución t_{26}, por lo que se rechaza la hipótesis nula. Por tanto, existe suficiente evidencia para afirmar que el efecto lineal es estadísticamente significativo.

OTRO MÉTODO ALTERNATIVO: En regresión lineal simple, el p-valor asociado a la pendiente coincide exactamente con el p-valor del test de significación global del modelo (ANOVA). Completando dicha tabla:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1,1833E6	1	1,1833E6	32,8380	
Residual	936897,0	26	36034,5		
Total (Corr.)	2,12019E6	27			

El F-ratio es superior al valor crítico de tablas: $F_{1;26}^{0,05}$=4,23. Por tanto, se rechaza la hipótesis nula y se concluye que el efecto lineal es significativo.

3c) Cuando el tamaño es 800, la distribución condicional del tiempo será una distribución normal cuya media viene dada por la ecuación del modelo y cuya varianza será la varianza de los residuos, que coincide con el cuadrado medio residual: $CM_{res} = SC_{res} / gl_{res} = 936897 / 26 = 36034{,}5$

Valor medio $\text{tiempo}_{\text{tamaño=800}}$ = -323,203 + 1,83038 · 800 = 1141,1

En una distribución Normal, m±1,96·s comprende el 95% de los datos. Así pues, el intervalo que nos piden será: $1141{,}1 \pm 1{,}96 \cdot \sqrt{36034{,}5}$ = **[769 ; 1513]**

No obstante, si el valor 1,96 se redondea a 2 el resultado puede considerarse también correcto.

3d) $r = \sqrt{R^2} = \sqrt{\frac{SC_{\text{mod}}}{SC_{total}}} = \sqrt{\frac{1{,}1833 \cdot 10^6}{2{,}1202 \cdot 10^6}} = \mathbf{0{,}747}$

El signo del coeficiente de correlación es positivo porque la pendiente es positiva.

3e) Para poder determinar si la correlación muestral es más fuerte o más débil, necesariamente hay que calcular el coeficiente de correlación a partir de la covarianza: $r = \text{cov} / (s_X \cdot s_Y)$. Pero esto no es posible ya que no se conocen las desviaciones típicas de las variables Y_2 e Y_3. Por tanto, **no** puede afirmarse que la correlación muestral más fuerte corresponda a las variables Y_2 e Y_3.