Grado en Ingeniería Informática

Estadística

EXAMEN FINAL

19 de junio de 2014

Apellidos y nombre:	
Grupo:	Firma:

Marcar las casillas de los parciales presentados: P1 ☐ P2 ☐

Instrucciones

1. **Rellenar** la cabecera del examen: **nombre, grupo y firma**.
2. Responder a cada pregunta en la hoja correspondiente.
3. **Justificar todas las respuestas**.
4. No se permiten anotaciones personales en el formulario. Sobre la mesa sólo se permite el DNI, calculadora, útiles de escritura, las tablas y el formulario.
5. **No desgrapar** las hojas.
6. El examen consta de 6 preguntas, 3 correspondientes al primer parcial (50%) y 3 del segundo (50%). El profesor corregirá los parciales que el alumno haya señalado en la cabecera del examen. **En cada parcial, todas las preguntas puntúan lo mismo** (sobre 10).
7. 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.
8. Tiempo disponible: **3 horas**

1. (1er Parcial) Un centro comercial vende tres modelos de portátiles (A, B y C). En la siguiente tabla de frecuencias se recoge las ventas de cada uno de ellos durante el año 2013, en función de la edad del cliente: cliente joven (edad < 30), adulto (de 30 a 50) y cliente senior (edad > 50). Se asume que cada cliente recogido en la tabla compró solamente un portátil.

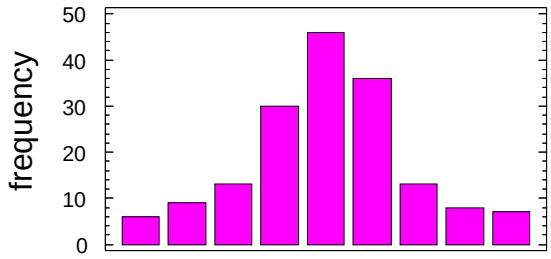
	modelo A	modelo B	modelo C	Row Total
joven	30	46	36	112
	26,79%	41,07%	32,14%	66,67%
adulto	9	13	8	30
	30,00%	43,33%	26,67%	17,86%
senior	6	13	7	26
	23,08%	50,00%	26,92%	15,48%
Column Total	45	72	51	168
	26,79%	42,86%	30,36%	100,00%

Responde a las siguientes preguntas justificando convenientemente tu respuesta.

a) Calcular la probabilidad de que un cliente elegido al azar sea joven y haya comprado un portátil del modelo A o B. *(3 puntos)*

b) A partir de la información contenida en la tabla, ¿son independientes los sucesos *ser cliente joven* y *haber comprado un portátil del modelo A*?
(4 puntos)

c) La siguiente figura muestra un gráfico de barras construido a partir de los datos reflejados en la tabla de frecuencias (se ha omitido la información del eje horizontal). ¿Qué conclusión se deduce de este gráfico teniendo en cuenta la información proporcionada en la tabla de frecuencias? *(3 puntos)*



2. (1er Parcial) El tiempo que tarda un servidor informático en ejecutar una solicitud (acceso) sigue una distribución exponencial, siendo la mediana 3 milisegundos.

a) ¿Cuál es la probabilidad de que una solicitud sea ejecutada en menos de 2 milisegundos? *(3 puntos)*

b) Calcular el percentil 95 de la variable aleatoria objeto de estudio. *(2 puntos)*

c) Si se seleccionan al azar 10 accesos consecutivos al servidor, ¿cuál es la probabilidad de que el tiempo total sea superior a 50 milisegundos? *(5 puntos)*

3. (1er Parcial) Un programa informático registra diariamente el número de fallos que se producen en las máquinas de una industria. En promedio se producen 3 fallos por día.

a) Calcular la probabilidad de que se produzcan más de 20 fallos en 6 días. *(5 puntos)*

b) ¿Cuál es la probabilidad de que en una semana (6 días laborables) se produzcan exactamente 15 fallos? *(5 puntos)*

4. (2º Parcial) Como parte de un estudio sobre el rendimiento académico de los alumnos de 1er curso del Grado en Ingeniería Informática, se han analizado las calificaciones de 111 alumnos en el 2º parcial de Estadística. Se ha obtenido una media muestral de 5,54 y un intervalo de confianza al 95% para la calificación media poblacional de [5,2 ; 5,9].

Suponiendo que los 111 alumnos evaluados constituyen una muestra representativa de todos los alumnos de primero y que la calificación obtenida por un alumno sigue una distribución normal, se pide:

a) ¿Qué interpretación tiene el intervalo de confianza obtenido? *(3,5 puntos)*

b) ¿Puede aceptarse que la nota media de los alumnos de primer curso es 6,2 para un nivel de confianza del 95%? ¿Y para un nivel de confianza del 90%? Justifica todas tus respuestas. *(3 puntos)*

c) Teniendo en cuenta que la desviación típica obtenida en las calificaciones de los 111 alumnos ha resultado 1,88, calcular un intervalo de confianza al 95% para la desviación típica poblacional σ. *(3,5 puntos)*

5. (2º Parcial) 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 (10, 20 y 30) y tres memorias (10 MB, 20 MB y 30 MB), realizándose dos repeticiones de cada una de las posibles combinaciones. Se supone que todos los datos son normales.

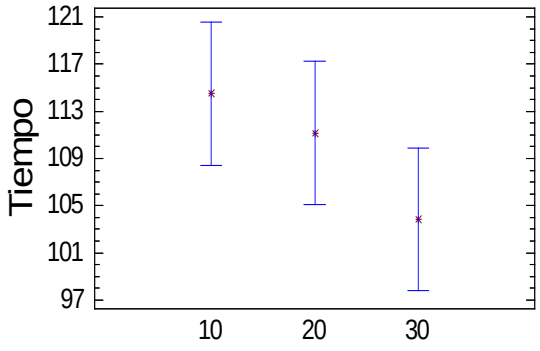
Los resultados obtenidos con el Statgraphics son los siguientes:

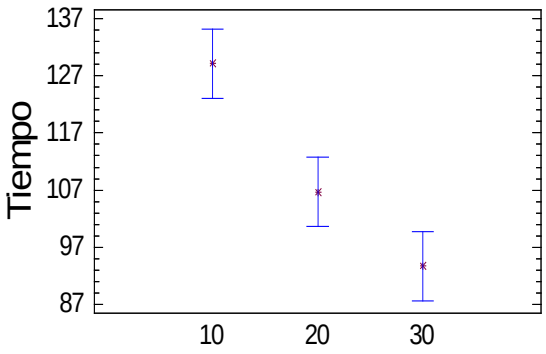
Analysis of Variance for TIEMPO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F- Ratio
MAIN EFFECTS				
A:Memoria	3871,0			
B:Modelo	357,333			
INTERACTIONS				
AB				
RESIDUAL			86,3889	
TOTAL (CORRECTED)	5028,5			

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

b) En los dos gráficos siguientes se han representado los intervalos LSD de los dos factores analizados considerando un nivel de confianza del 95%. Indica, justificando tu respuesta, a qué factor corresponde cada una de estas dos representaciones. *(3 puntos)*





c) ¿Cuáles serían las condiciones operativas óptimas para minimizar el tiempo de búsqueda en la base de datos? (considerar α=10%). *(3 puntos)*

6. (2º Parcial) Se ha medido durante 13 días el tiempo medio de respuesta de un sistema informático (en segundos) y la carga media en el mismo (en consultas por minuto). Con los datos obtenidos se ha realizado un ajuste de regresión lineal.

Multiple Regression Analysis

Dependent variable: T_respuesta

Parameter	Estimate	Standard Error	T Statistic
CONSTANT	0,0747486	0,190814	
CARGA	0,789228	0,084857	

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	17,1757			
Residual				
Total	17,5169			

a) Plantea el modelo estimado y analiza qué parámetros del mismo son estadísticamente significativos (α=5%). *(4 puntos)*

b) ¿Entre qué límites estará aproximadamente en el 95% de los casos el tiempo de respuesta cuando la carga sea igual a 4? *(6 puntos)*

SOLUCIÓN DEL EXAMEN

EJERCICIO - 1

1a) (30+46)/168 = 0,4524 = 45,24%

1b) Suceso A: el cliente ha comprado un portátil modelo A.
Suceso B: el cliente es joven.
P(A/B) = 30/112 = 0,2679 ; P(A) = 45/168 = 0,2679
P(B/A) = 30/45 = 0,6667 ; P(B) = 112/168 = 0,6667
Se cumple P(A/B) = P(A), y además se cumple que P(B/A) = P(B), de modo que se concluye que los sucesos A y B son independientes.

1c) El gráfico de barras muestra los valores de las 9 frecuencias absolutas de la tabla. La principal conclusión es que hay tres frecuencias bastante superiores a las restantes, que son: 46 (cliente joven - modelo B), 36 (cliente joven - modelo C) y 30 (cliente joven - modelo A). Esto se debe a que la proporción de clientes jóvenes es la mayoritaria (66,67%).

EJERCICIO - 2

2a) $P(X<x)=1-e^{-\alpha\cdot x}$; $P(X<3)=0{,}5=1-e^{-\alpha\cdot 3}$; $\alpha=-(\ln 0{,}5)/3=0{,}231$
$P(X<2)=1-e^{-0{,}231\cdot 2}=\mathbf{0{,}37}$

2b) El valor del percentil 95 (Z_{95}) es aquel que cumple: $P(X<Z_{95})=0{,}95$
$P(X<Z_{95})=1-e^{-0{,}231\cdot Z_{95}}=0{,}95$; $Z_{95}=-[\ln(1-0{,}95)]/0{,}231=\mathbf{12{,}96\ ms}$

2c) Variable aleatoria X: tiempo (ms) del servidor A en ejecutar una solicitud
X~exp(α) ; $E(X)=1/\alpha=4{,}328$; $\sigma^2(X)=1/\alpha^2=18{,}732$
Variable aleatoria Y: tiempo total (ms) de 10 accesos consecutivos
$Y=X_1+X_2+...+X_{10}$; $E(Y)=E(X_1+...+X_{10})=E(X_1)+...+E(X_{10})=10\cdot E(X)=43{,}28$
$\sigma^2(Y)=\sigma^2(X_1+...+X_{10})=\sigma^2(X_1)+...+\sigma^2(X_{10})=10\cdot\sigma^2(X)=10\cdot 18{,}732=187{,}32$
Asumiendo que Y es una distribución Normal por el teorema central del límite:

$$P(Y>50)=P\left[N\left(m=43{,}28;\sigma=\sqrt{187{,}32}\right)>50\right]=P\left[N(0;1)>\frac{50-43{,}28}{\sqrt{187{,}32}}\right]=P[N(0;1)>0{,}49]=\mathbf{0{,}312}$$

EJERCICIO - 3

3a) Variable X: nº de fallos en un día. X~Ps(λ); E(X)= λ =3
Variable Y: nº de fallos en 6 días. $Y=X_1+...+X_6$; $Y\approx Ps(\lambda=\lambda_{X_1}+...+\lambda_{X_6})$;
$Y\approx Ps(\lambda=3\cdot 6=18)$; $P(Y>20)=1-P(Y\le 20)=$(ábaco) = 1 - 0,73=**0,27**

3b) $P(Y=15)=e^{-18}\cdot\frac{18^{15}}{15!}=\mathbf{0{,}0786}$

EJERCICIO - 4

4a) Si tomáramos 100 muestras diferentes, pero del mismo tamaño, de la misma población y calculáramos el intervalo de confianza (IC) para la nota media, éste cambiaría también, al hacerlo la media y desviación típica muestrales. Sin embargo, en 95 de los 100 IC calculados estaría el verdadero valor de la nota media de Estadística de los alumnos de primero.

Así pues, el intervalo de confianza obtenido recoge el conjunto de hipótesis nulas compatibles con nuestros datos (muestra) para un determinado nivel de significación. La probabilidad de que la calificación media poblacional (para todos los alumnos de primer curso) se encuentre contenida entre 5,2 y 5,9 es del 95%.

4b) Dado que el valor 6,2 no está incluido en el intervalo de confianza al 95%, la hipótesis de calificación media para los alumnos de primer curso = 6,2 no sería aceptable para un nivel de confianza del 95%.
El intervalo de confianza al 90% sería más estrecho que el anterior con lo que, obviamente, tampoco incluiría al valor 6,2. Por tanto, no sería aceptable una calificación media de los alumnos de primer curso de 6,2 para un nivel de confianza del 90%.

4c) $$IC_{\sigma}^{95\%} = \left[\sqrt{(n-1)\frac{s^2}{g_2}}, \sqrt{(n-1)\frac{s^2}{g_1}}\right] = \left[\sqrt{110\frac{1{,}88^2}{140{,}9}}, \sqrt{110\frac{1{,}88^2}{82{,}9}}\right] = [1{,}66;\ 2{,}16]$$

$$Tabla\ \ g_1 / P\left(\chi^2_{n-1} \le g_1\right) = P\left(\chi^2_{111-1} \le g_1\right) = \frac{\alpha}{2} = \frac{0{,}05}{2} = 0{,}025 \rightarrow g_1 = 82{,}86$$

$$Tabla\ \ g_2 / P\left(\chi^2_{n-1} \ge g_2\right) = P\left(\chi^2_{111-1} \ge g_2\right) = \frac{\alpha}{2} = \frac{0{,}05}{2} = 0{,}025 \rightarrow g_2 = 140{,}92$$

EJERCICIO - 5

5a) Nº total de datos = 9 tratamientos x 2 repeticiones = 18
Grados de libertad totales = 18 - 1 = 17
Grados de libertad del factor memoria RAM = 3 niveles - 1 = 2
Grados de libertad del factor modelo = 3 variantes - 1 = 2
Grados de libertad de la interacción: 2 · 2 = 4
Grados de libertad residuales, se obtienen por diferencia: 17 - 2 - 2 - 4 = 9

$SC_{residual} = CM_{resid} \cdot gl_{resid} = 86{,}3889 \cdot 9 = 777{,}5$
$SC_{interac} = SC_{total} - SC_{res} - SC_{RAM} - SC_{modelo} = 5028{,}5 - 777{,}5 - 3871 - 357{,}33 = 22{,}67$
$F_{ratioRAM} = (SC/gl)/CM_{res} = (3871/2) / 86{,}39 = 22{,}4$
$F_{ratio\_modelo} = (SC/gl)/CM_{res} = (357{,}33/2) / 86{,}39 = 2{,}07$
$F_{ratio\_interac} = (SC/gl)/CM_{res} = (22{,}67/4) / 86{,}39 = 0{,}07$

Considerando α=0,05, el efecto simple del factor memoria RAM es estadísticamente significativo ya que el F-ratio (22,4) es mayor al valor crítico de tablas ($F_{2;9}$) que vale 4,26.
El efecto simple del factor modelo NO es estadísticamente significativo ya que el F-ratio (2,07) es menor al valor crítico de tablas ($F_{2;9}$) que vale 4,26.

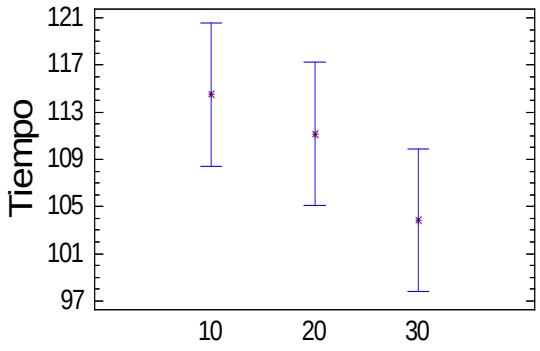
El efecto de la interacción NO es estadísticamente significativo ya que el F-ratio (0,07) es menor al valor crítico de tablas ($F_{4;9}$) que vale 3,63.

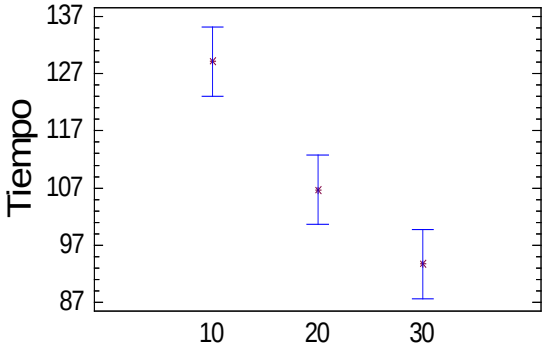
La tabla resumen completa es la siguiente (se muestra también el p-valor aunque éste sólo puede calcularse con Statgraphics).

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	3871,0	2	1935,5	22,40	0,0003
B:Modelo	357,333	2	178,667	2,07	0,1824
INTERACTIONS					
AB	22,6667	4	5,66667	0,07	0,9907
RESIDUAL	777,5	9	86,3889		
TOTAL (CORRECTED)	5028,5	17			

5b) La figura de la derecha corresponde al factor memoria ya que el primer intervalo LSD no se solapa con los otros dos, lo que indica que el efecto es estadísticamente significativo. Dado que éste es el único factor significativo, según el apartado anterior, necesariamente dicha figura debe corresponder al factor memoria.
En cambio, en la figura de la izquierda todos los intervalos LSD se solapan, lo que sugiere que el efecto simple del factor no resulta estadísticamente significativo. Por tanto, la figura de la izquierda debe corresponder al factor modelo que no es significativo considerando un nivel de significación del 5%.





5c) El gráfico de medias corresponde a un nivel de confianza del 95% (α=5%). Dado que el factor modelo no resulta estadísticamente significativo, daría lo mismo trabajar con cualquiera de los tres modelos. La media de tiempo con RAM=30 es la menor de las tres, por lo que el tiempo medio obtenido con RAM=30 será significativamente menor que los otros, de modo que ésta será la condición operativa óptima para α=5%.
En este caso se pide considerar α=10%, lo que daría lugar a intervalos LSD más estrechos, pero la conclusión sería la misma ya que:
- En el caso del factor memoria (figura derecha), ninguno de los tres intervalos se solapará si estos son más estrechos.
- En el caso del factor modelo (figura izquierda), aunque los intervalos sean algo más estrechos es razonable suponer que el de 20 se solape con el de 30.

EJERCICIO - 6

6a) A partir de los valores estimados de los parámetros que se muestran en la tabla, el modelo estimado será: E(T_resp/carga) = 0,0747+0,789· carga

Las dos t-calc (t-statistic) se obtienen como: estimate/standard_error:
Para la ordenada en el origen (constante): t-calc=0,07475/0,1908 = 0,39
Para la pendiente: t-calc= 0,7892/0,08486 = 9,3
Valor crítico de tablas: $t_{11}^{0,025}$=2,201. La ordenada no es significativa porque |0,39|<2,201. La pendiente es significativa ya que |9,3|>2,201.

También se puede estudiar la significación de la pendiente con el ANOVA:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio
Model	17,1757	1	17,1757	554,05
Residual	0,341	11	0,031	
Total	17,5169	12		

La F-tabla=$F_{1,11}^{0,05}$=4,84. Como F-ratio>F-tabla, se concluye que el efecto lineal de carga sobre el tiempo medio de respuesta es estadísticamente significativo.

6b) (T_resp/carga=4) = 0,0747+0,789·4 = 3,232 ; $S_{residual}=(0,031)^{1/2}$= 0,176
T_resp/carga=4 ~ N(m=3,232; σ=0,176)

Teniendo en cuenta que en una Normal el intervalo m±2σ comprende aproximadamente el 95% de los datos, el intervalo de la distribución Normal en este caso que comprende el 95% de los valores será:

Límites del 95%: 3,232 ± 2·0,176 = [**2,88 ; 3,82**]