Grado en Ingeniería Informática

Estadística

EXAMEN FINAL

19 de junio de 2015

Apellidos y nombre:	
Grupo:	Firma:

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

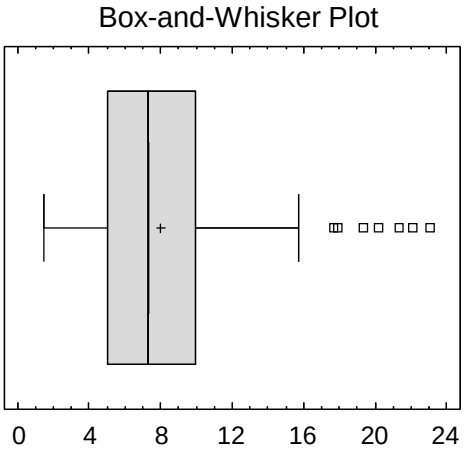
Instrucciones

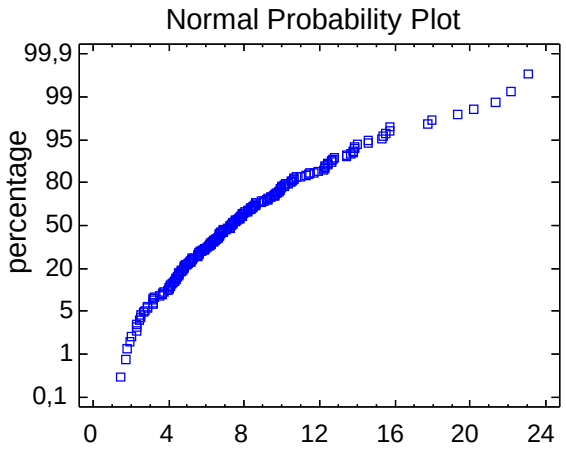
1. **Rellenar** la cabecera del examen: **nombre, grupo y firma**.
2. Responder a cada pregunta en la hoja correspondiente.
3. **Justificar todas las respuestas**.
4. No se permiten anotaciones personales en el formulario. Sobre la mesa sólo se permite el DNI, calculadora, útiles de escritura, las tablas 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) Para estudiar el funcionamiento de un sistema informático se ha registrado durante 200 días a las 9:00 de la mañana el tiempo en milisegundos (ms) que tardó dicho sistema en ejecutar un programa de prueba (*benchmark*) que opera con datos de clientes medidos en tiempo real.

a) Indica cuál es la población en estudio y cuál es la variable aleatoria que se está considerando. *(3 puntos)*

b) Los datos se han representado en un diagrama Box-Whisker y en un papel probabilístico Normal, los cuales se indican a continuación. A la vista de estos gráficos, ¿qué parámetro de dispersión consideras más adecuado para describir las características de la muestra? ¿Por qué? Calcula dicho parámetro. Calcula también el parámetro de posición que sería el más adecuado. *(3,5 puntos)*





c) Sabiendo que cierto día el tiempo de ejecución fue superior a 10 ms, ¿cuál es la probabilidad de que dicho tiempo haya sido mayor de 15 ms? *(3,5 puntos)*

2. (1er Parcial) Una biblioteca almacena en un registro informático el número total de libros que cada usuario ha tomado prestados en cada mes. Los datos disponibles indican que el valor medio de esta variable aleatoria es de 3 para cierto usuario.

a) Si dicho usuario ha tomado prestado un libro en los primeros días de cierto mes, ¿cuál es la probabilidad de que tome al menos otro libro más (es decir, dos o más en total) en ese mismo mes? *(3,5 puntos)*

b) ¿Cuál es la probabilidad de que dicho usuario tome prestados menos de 9 libros en el plazo de tres meses consecutivos? *(3 puntos)*

c) ¿Cuál es la probabilidad de que dicho usuario haya tomado prestados más de 30 libros durante un periodo de un año? *(3,5 puntos)*

3. (1er Parcial) El tiempo que utilizan los usuarios de una biblioteca consultando en un terminal fluctúa exponencialmente con mediana 15 minutos.

a) Calcular la media del tiempo de consulta en dicho terminal. *(5 puntos)*

b) Cuando un usuario va a usar el terminal, encuentra que está ocupado por otro usuario que lleva ya 10 minutos. ¿Cuál es la probabilidad de que tenga que esperar más de 30 minutos antes de que el terminal quede libre? *(5 puntos)*

4. (2º Parcial) Una industria elabora resistencias eléctricas empleadas en circuitos electrónicos de equipos informáticos. El valor de la resistencia (variable R) es un parámetro de calidad que se controla de forma periódica. El departamento de control de calidad asume que este parámetro sigue una distribución Normal de media 20 Ω y desviación típica 0,2.

a) Se produce un cambio en el proceso que podría haber afectado al valor medio de la resistencia. Para estudiarlo, se toma una muestra aleatoria de 5 resistencias, obteniéndose los siguientes valores: 20,7; 19,8; 20,4; 20,1; 20,5. La varianza de esta muestra es 0,125 Ω^2. Considerando un riesgo de primera especie del 5%, ¿existe suficiente evidencia para afirmar que el cambio producido en el proceso ha alterado la resistencia media a nivel poblacional?
(5 puntos)

b) Teniendo en cuenta los datos del apartado anterior y considerando α=0,05, ¿existe suficiente evidencia para afirmar que el cambio ocurrido haya alterado la desviación típica poblacional (es decir, que ésta sea distinta de 0,2)?
(5 puntos)

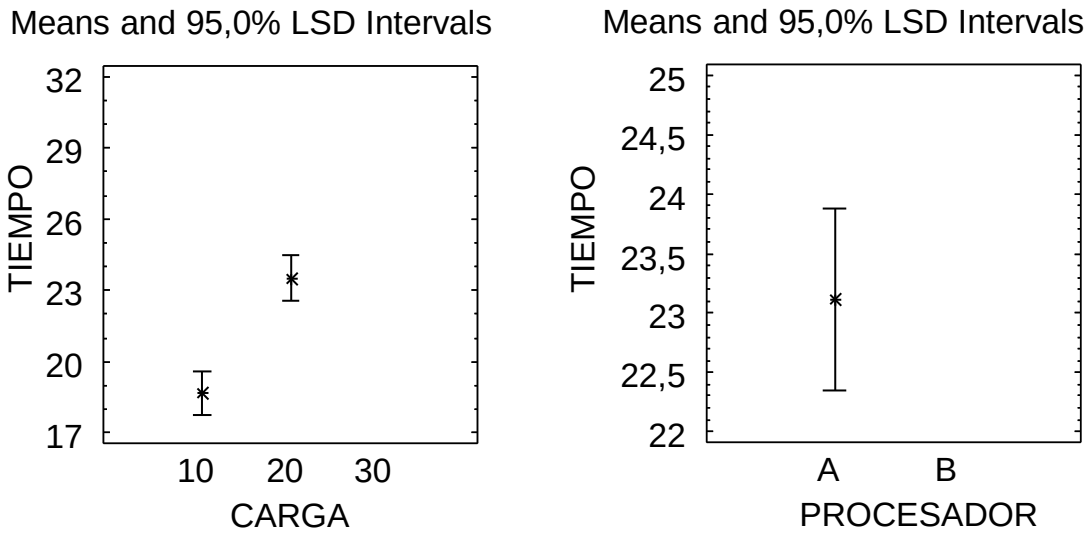
5. (2º Parcial) Se desea estudiar el efecto que tienen el tipo de procesador y la carga de trabajo sobre el tiempo de utilización de la CPU, en la ejecución de cierto tipo de procedimientos. Para ello se han probado dos procesadores (A y B) combinados con tres cargas (10, 20 y 30). Cada tratamiento se repitió tres veces, obteniéndose los siguientes resultados experimentales:

PROCESADOR	A	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B	B
CARGA	10	10	10	20	20	20	30	30	30	10	10	10	20	20	20	30	30	30
TIEMPO	15	20	17	23	25	24	28	27	29	19	22	19	22	24	23	29	28	30

Analizando estos datos con ANOVA se obtienen los siguientes valores: $SC_{proces.} = 3{,}5556$; $SC_{carga} = 290{,}111$; $SC_{proc\ x\ carga} = 10{,}111$; $SC_{total} = 330{,}444$.

a) Si tomamos un riesgo de primera especie de 0,05, ¿qué factores resultan estadísticamente significativos? *(5 puntos)*

b) Completa los siguientes gráficos de intervalos LSD, justificando convenientemente los cálculos. *(2,5 puntos)*



c) A la vista de los resultados obtenidos de los gráficos anteriores, ¿qué información se deduce? ¿Es coherente esta información con las conclusiones obtenidas en el apartado a)? *(2,5 puntos)*

6. (2º Parcial) Se quiere estudiar la relación lineal que hay entre dos variables: el tamaño (en MB) de cierto tipo de archivos y su tiempo de descarga desde la red (en segundos). Para ello se ha tomado una muestra de 25 archivos, en la que se ha medido el tamaño de cada uno de ellos y el tiempo que han tardado en descargarse, el cual es función del tamaño. Al ajustar el modelo de regresión lineal obtenemos los siguientes resultados con Statgraphics:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Y
Independent variable: X
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept                     25,4095       0,465623         0,6459
Slope            1,24544      0,06573
------------------------------------------------------------------------------

                          Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares    Df   Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                                       149323,0
Residual                 9566,21
------------------------------------------------------------------------------
Total (Corr.)
```

a) Escribe la ecuación del modelo, indicando las variables dependiente e independiente. ¿Son significativos los parámetros considerando α = 5%?

(4,5 puntos)

b) ¿Qué porcentaje de la variabilidad observada en el tiempo de descarga está explicado por el efecto lineal del tamaño? ¿Cómo se denomina al parámetro que aporta esta información? *(3 puntos)*

c) Si queremos descargar un archivo de 40 MB, ¿cuántos minutos tardaremos en promedio aproximadamente? *(2,5 puntos)*

SOLUCION DEL EXAMEN

1a) La población está formada por el conjunto de todos los posibles días en los cuales se puede ejecutar dicho programa. También sería correcto considerar que la población está formada por el conjunto de todas las posibles ejecuciones que se llevan a cabo una vez al día de dicho programa.

La variable aleatoria es el tiempo (en milisegundos) que tarda el sistema informático en ejecutar el programa de prueba.

1b) La forma de ambos gráficos indica que la distribución es asimétrica positiva, sin que se detecten claramente datos anómalos. En estos casos, el parámetro de dispersión más adecuado es el intervalo intercuartílico, ya que éste es más robusto que el rango o la varianza frente a la presencia de valores extremos. Es decir, en una distribución asimétrica positiva existe cierta probabilidad de encontrar algún valor extremo (bastante elevado), que alterará notablemente el rango y la varianza. Sin embargo, el intervalo intercuartílico no quedará alterado. Por la misma razón, la mediana es un parámetro de posición más adecuado que la media.
Primer cuartil: Q1 = 5 (extremo izquierdo de la caja en el diagrama Box-Whisker)
Tercer cuartil: Q3 = 10 (extremo derecho de la caja)
Intervalo intercuartílico = Q3 - Q1 = 10 - 5 = **5 ms**
Mediana = 7,3 aproximadamente (línea central de la caja).

1c) $$P\left[(X>15)/(X>10)\right]=\frac{P\left[(X>15)\cap(X>10)\right]}{P(X>10)}=\frac{P(X>15)}{P(X>10)}=\frac{0{,}05}{0{,}25}=0{,}2$$

P(X>10)=0,25 por ser 10 el valor del tercer cuartil
P(X>15) = 1-P(X<15) = 1 - 0,95 = 0,05 (el valor 0,95 se obtiene a partir del papel probabilístico normal, leyendo en la escala vertical para X=15).

2a) La variable aleatoria X: "nº de libros tomados prestados en un mes" es discreta con valor mínimo cero y valor máximo no acotado. Por tanto, puede asumirse que sigue una distribución de tipo Poisson. El valor medio de esta distribución coincide con el parámetro, así que: X~Ps(λ=3). Las probabilidades para valores concretos se obtienen aplicando la función de probabilidad: $P(X=x)=e^{-\lambda}\cdot\lambda^{x}/x!$

$$P\left[(X\geq 2)/(X>0)\right]=\frac{P\left[(X\geq 2)\cap(X>0)\right]}{P(X>0)}=\frac{P(X\geq 2)}{P(X>0)}=\frac{1-P(X=0)-P(X=1)}{1-P(X=0)}=$$

$$=\frac{1-e^{-3}\cdot(3^0/0!)-e^{-3}\cdot(3^1/1!)}{1-e^{-3}\cdot 3^0/0!}=\frac{1-4\cdot e^{-3}}{1-e^{-3}}=\frac{0{,}8009}{0{,}9502}=0{,}843$$

2b) La variable aleatoria Y: "nº de libros tomados prestados en 3 meses" es la suma de 12 variables aleatorias: $Y=X_{1º}+X_{2º}+X_{3º}$ (es decir, el nº de libros tomados en el primer

mes, el segundo y el tercero). La suma de variables Poisson es a su vez una variable del mismo tipo cuyo parámetro es: $\lambda_Y = \lambda_{X1}+\lambda_{X2}+\lambda_{X3} = 3+3+3 = 9$

$$P(Y<9) = P(Y\le 8) = P\left[Ps(\lambda=9)\le 8\right] = \mathbf{0,456}$$

El resultado se obtiene a partir del ábaco de Poisson con λ=9 leyendo en la curva "8".

2c) La variable aleatoria Z: "nº de libros tomados prestados en 12 meses (1 año)" es la suma de tres variables: $Y=X_{1º}+X_{2º}+...+X_{12º}$ (es decir, el nº de libros tomados a lo largo de los 12 meses). La suma de variables Poisson es a su vez una variable del mismo tipo cuyo parámetro es: $\lambda_Z = \lambda_{X1}+...+\lambda_{X12} = 3+...+3 = 12\cdot 3 = 36$. No se puede emplear el ábaco de Poisson por ser λ>30 de modo que hay que aproximar a una Normal:

$$P(Z>30) = P\left[Ps(\lambda=36)>30\right] \approx P\left[N\left(m=36,\ \sigma^2=36\right)>30{,}5\right] =$$
$$= P\left[N(0;\ 1) > (30{,}5-36)/\sqrt{36}\right] = P\left[N(0;1) > -0{,}917\right] = 1-0{,}18 = 0{,}82$$

3a) $T\sim exp(\alpha)$; $P(T>t) = e^{-\alpha\cdot t}$; $P(T>15) = 0{,}5 = e^{-\alpha\cdot 15}$; $\ln(0{,}5) = -15\alpha$;
$\alpha = -(\ln 0{,}5)/15 = 0{,}0462$; $E(T) = 1/\alpha = 1/0{,}0462 =$ **21,64** minutos

3b) $P\left[(T>40)/(T>10)\right] = P(T>30)$ por la propiedad de falta de memoria de la distribución exponencial: $P(T>30) = e^{-\alpha\cdot 30} = e^{-0{,}0462\cdot 30} =$ **0,25**

4a) La hipótesis nula que se plantea es H_0: m=20 frente a la alternativa H_1: m≠20.

Media muestral: $\bar{x} = (20{,}7+19{,}8+20{,}4+20{,}1+20{,}5)/5 = 20{,}3$

$$\frac{\bar{x}-m_0}{s/\sqrt{n}} = \frac{20{,}3-20}{\sqrt{0{,}125}/\sqrt{5}} = 1{,}897$$

Este estadístico de contraste sigue una distribución t de Student con 4 grados de libertad (n-1). Según tablas, el 95% de valores de esta distribución (α=0,05) varían entre -2,776 y 2,776. Así pues, el valor obtenido 1,897 es un valor frecuente de esta distribución, por lo que se acepta la hipótesis nula: **no** existe suficiente evidencia para afirmar que el cambio producido haya alterado la resistencia media.

4b) La hipótesis nula que se plantea es H_0: σ=0,2 frente a la alternativa H_1: $\sigma\neq 0{,}2$ lo cual es equivalente a considerar H_0: σ^2=0,04 frente a H_1: $\sigma^2\neq 0{,}04$.

Intervalo de confianza para la varianza poblacional:

$$IC_{\sigma^2} = \left[(n-1)\cdot s^2/g_2;\ (n-1)\cdot s^2/g_1\right] = \left[4\cdot 0{,}125/11{,}143;\ 4\cdot 0{,}125/0{,}484\right] = \left[0{,}045;1{,}03\right]$$

Los parámetros g_1=0,484 y g_2=11,143 son los valores críticos de una distribución χ^2 con 4 grados de libertad obtenidos de tablas que abarcan el 95% de valores de dicha distribución, es decir: $P\left(\chi^2_4 > 11{,}143\right) = 0{,}025$; $P\left(\chi^2_4 < 0{,}484\right) = 0{,}025$.

El valor 0,04 está fuera del intervalo obtenido, por lo que existe suficiente evidencia (considerando α=0,05) para afirmar que el cambio ocurrido ha aumentado la desviación típica poblacional.

5a) La tabla del ANOVA completada es la siguiente:

Analysis of Variance for TIEMPO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:PROCESADOR	3,55556	1	3,55556	1,60
B:CARGA	290,111	2	145,056	65,28
INTERACTIONS				
AB	10,1111	2	5,05556	2,28
RESIDUAL	26,6667	12	2,22222	
TOTAL (CORRECTED)	330,444	17		

Grados de libertad totales = n-1 = 18 datos - 1 = 17.
Grados de libertad de procesador = 2 variantes -1 = 1; Gr. lib. carga = 3 niveles -1 = 2
Gr. lib. interacción = 1 · 2 = 2 ; Grados de libertad residuales = 17 -1 -2 -2 = 12
$SC_{resid} = SC_{total} - \Sigma SC$ = 330,444 - 3,5556 - 290,111 - 10,111 = 26,667
El cuadrado medio se obtiene dividiendo las sumas de cuadrados entre los grados de libertad, y la F-ratio se obtiene dividiendo el cuadrado medio entre 2,2222.

El efecto del factor procesador no es estadísticamente significativo para α=0,05 porque su F-ratio= 1,60 es inferior al valor crítico de tablas: $F_{1;12}^{0,05}$=4,75.
El efecto del factor carga es significativo porque su F-ratio=65,28 es muy superior al valor crítico de tablas: $F_{2;12}^{0,05}$=3,89.
El efecto de la interacción no es significativo porque su F-ratio=2,28 resulta inferior al valor crítico de tablas: $F_{2;12}^{0,05}$=3,89.

5b) Para carga=30, el valor medio de tiempo = (28+27+29+29+28+30)/6 = 28,5
En el gráfico de la izquierda el intervalo LSD que hay que dibujar estará centrado en 28,5 y tendrá la misma anchura que los otros dos (28,5±0,97), ya que el número de datos es el mismo en los tres casos.
Para procesador B, el valor medio = (19+22+19+22+24+23+29+28+30)/9 = 24
En el gráfico de la derecha el intervalo LSD para procesador D estará centrado en 24 y tendrá la misma anchura que para el procesador A: 24±0,77 (de 23,23 a 24,77).

5c) El intervalo LSD para carga=30 no se solapa con los otros dos, lo cual es coherente con el hecho de que el factor carga resulta estadísticamente significativo. Para carga=10 el tiempo medio es 18,7. Para carga=20 es 23,5 y para carga=30 es 28,5. Entre 10 y 20 el incremento medio de tiempo es 4,8 unidades (23,5 - 18,7), mientras que entre 20 y 30 el incremento es prácticamente el mismo (28,5 - 23,5 = 5), lo cual indica que el efecto de la carga en el tiempo medio es lineal.

El intervalo LSD para procesador B (obtenido con 1-α=0,95) se solapa con el intervalo de procesador A, lo cual implica que hay que aceptar la hipótesis nula H_0: $m_A = m_B$ (es decir, no existen diferencias estadísticamente significativas entre el tiempo medio obtenido con los procesadores A y B). Esto es coherente con el hecho de que el factor procesador no resulte estadísticamente significativo para α=0,05.

6a) La constante del modelo, cuyo valor no aparece en la tabla, se calcula teniendo en cuenta que: $t_{statistic} = b_i / s_{b_i}$; 0,4656 = b_i /25,409; b_i = 0,4656·25,409 = 11,83

El tiempo de descarga es función (es decir, depende) del tamaño del archivo. Así pues, el tiempo es la variable dependiente y el tamaño es la variable independiente. El valor estimado de la pendiente (slope) según la tabla es 1,2454. La ecuación del modelo será:
Y=11,83 + 1,2454 X ; Tiempo = 11,83 + 1,2454 · tamaño

Siendo el modelo teórico $Y=\beta_0+\beta_1 \cdot X$, en este caso la constante del modelo no es estadísticamente significativa ya que su p-valor (0,6459) es superior a 0,05 (riesgo de primera especie), por lo que hay que aceptar la hipótesis nula H_0: $\beta_0=0$. Pero aunque no sea significativa, no es correcto despreciarla y emplear el modelo: tiempo = 1,2454· tamaño.

En relación a la pendiente, hay que contrastar la hipótesis nula H_0: $\beta_1=0$ frente a la alternativa H_1: $\beta_1 \neq 0$. Teniendo en cuenta que $b_i / s_{b_i} \approx t_{N-1-I} \approx t_{23}$ siendo I=1 (modelo con una variable explicativa), N=25 (nº total de observaciones) y α=0,05, resulta que: $b_i / s_{b_i} = 1{,}2454 / 0{,}06573 = 18{,}9$ el cual es un valor muy poco frecuente de la distribución t_{23}, por lo que se rechaza la hipótesis nula. Por tanto, la pendiente de la recta también es un parámetro estadísticamente significativo (es decir, distinto de cero a nivel poblacional).

6b) Este parámetro se denomina coeficiente de determinación (R^2) y se calcula como:

$$R^2 = \frac{SC_{modelo}}{SC_{total}} = \frac{SC_{modelo}}{SC_{modelo} + SC_{resid}} = \frac{149323}{149323 + 9566{,}21} = 0{,}9398 = 93{,}98\%$$

Ya que: SC_{modelo} / gr. lib. = CM_{modelo} ; SC_{modelo} = gr. lib.· CM_{mod} = 1 · 149323
(un grado de libertad por haber sólo una variable explicativa en el modelo).

6c) La predicción del modelo para tamaño=40 será:
Tiempo = 11,83 + 1,2454 ·40 = 61,65 segundos ≈ **1 minuto**