Grado en Ingeniería Informática

Estadística

EXAMEN FINAL

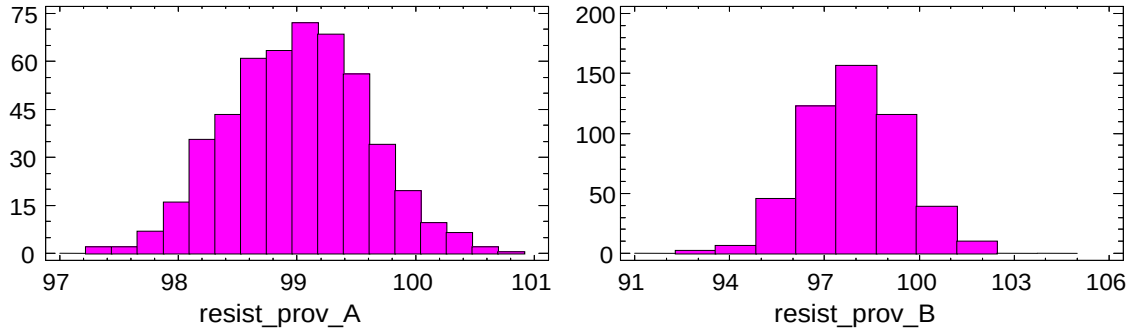
15 de junio de 2012

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 a cada parcial. 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) Una empresa que fabrica equipos informáticos utiliza un cierto componente electrónico de resistencia 100 ohmios, el cual puede ser suministrado por dos proveedores distintos (A o B). Para estudiar si existen diferencias en cuanto a la resistencia de los componentes suministrados por cada proveedor, se toman 500 componentes del proveedor A y otros 500 del proveedor B. Tras medir la resistencia de estos componentes, se obtienen los siguientes histogramas:



A la vista de estos histogramas, responde a las siguientes preguntas justificando convenientemente las respuestas.

a) ¿Qué indica la escala vertical? ¿Por qué es tan diferente en los dos casos? (**2,5 puntos**)

La escala vertical es frecuencia absoluta: número de datos contenido en cada intervalo del histograma. Esta escala es mucho mayor en el histograma del proveedor B porque éste tiene muchos menos intervalos (menos barras). Teniendo en cuenta que ambos histogramas se han construido con 500 datos, al dividir el rango de variación de la longitud en un menor número de intervalos aparecen más datos en cada uno de ellos, aumentando por tanto la frecuencia absoluta.

b) ¿En cuál de los dos proveedores hay más dispersión en los valores de resistencia? ¿Por qué? (**2,5 puntos**)

Rango de A ≈ 101 – 97 = 4 ohms

Rango de B ≈ 102,5 – 92,5 = 10 ohms

Como los rangos son tan distintos y teniendo en cuenta que en ambos casos el modelo normal parece adecuado, el proveedor B tendrá mayor variabilidad que el A (es decir, mayor desviación típica, varianza e intervalo intercuartílico).

c) ¿Crees que la técnica utilizada es adecuada para detectar datos anómalos? ¿Qué otras técnicas utilizarías? (**2,5 puntos**)

El histograma es en general una técnica poco adecuada para detectar datos anómalos, ya que un solo dato bastante extremo daría lugar a una barra de altura unitaria, que fácilmente puede pasar desapercibida. Para la detección de

datos anómalos es más conveniente el uso del diagrama box-whisker o el papel probabilístico normal.

d) Indica qué parámetros son los más adecuados en este caso para cuantificar la posición de la resistencia y da el valor aproximado para cada proveedor. (**2,5 puntos**)

Como se trata de una distribución simétrica, el parámetro más adecuado para cuantificar la posición de la resistencia, sería la media, que ocupa el valor central de la figura que encierra el histograma. Para cada proveedor la media estimada a partir del histograma, es aproximadamente:

Prov. A: media≈99 ohms

Prov. B: media≈98 ohms

2. (1er Parcial) Una cibernauta ha constatado que diariamente recibe en promedio 5 visitas en un blog que ha creado.

a) ¿Cuál es la variable aleatoria considerada? ¿Qué distribución sigue? (**2 puntos**)

Sea la variable cuantitativa discreta *X*={número de visitas diarias en el blog}. El problema ya nos dice que el promedio de esta variable es de 5 visitas/día. Al distribuirse *X* como una variable de Poisson deducimos que el parámetro λ es de 5 visitas/día (ya que, por definición, debe coincidir con el promedio).

b) Calcula la probabilidad de que reciba diariamente al menos 2 visitas. (**3 puntos**)

$P(X\geq 2)=1-P(X\leq 1)=1-0{,}04=0{,}96$ donde 0,04 se obtiene en el ábaco de Poisson con $\lambda=5$ y valor de x=1.

c) Calcula la probabilidad de que reciba en una semana al menos 40 visitas (utiliza la aproximación a la distribución normal). (**5 puntos**)

Sea Y={número de visitas en una semana en el blog}. Y se distribuirá, por ser la suma de 7 variables de Poisson, con el mismo modelo con $\lambda=7x5=35$. Por ser $\lambda>9$, Y se puede aproximar con la distribución normal con media m=35 y $\sigma=(35)^{1/2}=5{,}92$. Por tanto, aplicando dicha aproximación y la corrección de continuidad:

$$P(Y\geq 40)=P(N(35, 5{,}92)>39{,}5)=P(N(0,1)>\frac{39{,}5-35}{5{,}92})=P(N(0,1)>0{,}76)=0{,}2236$$

3. (1er Parcial) Un dispositivo está formado por tres componentes idénticos y de funcionamiento independiente. La vida de estos componentes se distribuye de forma exponencial, y la fiabilidad de los mismos a las 120 horas de funcionamiento es del 80%.

a) Calcula la fiabilidad de un componente a las 400 horas. **(4 puntos)**

va X_i = {vida del componente i (horas)} ~ Exp(α)

Fiabilidad del componente i a las 120 horas del 80% →

$P(X_i \geq 120) = 0,8 = e^{-120\alpha}$ → **$\alpha = 0,0018$**

Fiabilidad del componente i a las 400 horas → $P(X_i \geq 400) = e^{-400x0,0018} = 0,487$

b) Sabiendo que el dispositivo funciona si lo hacen al menos 2 de sus componentes, calcular la fiabilidad del dispositivo a las 400 horas de funcionamiento. **(6 puntos)**

Sucesos

A = El componente 1 funciona correctamente al menos 400 h

B = El componente 2 funciona correctamente al menos 400 h

C = El componente 3 funciona correctamente al menos 400 h

D = El **Dispositivo** funciona correctamente al menos 400 h →

D = (A∩B) ∪ (A∩C) ∪ (B∩C)

va X_D = {vida del **Dispositivo** (horas)}

Fiabilidad del **Dispositivo** a las 400 horas → $P(X_{\mathbf{D}} \geq 400)$ =

P((A∩B) ∪ (A∩C) ∪ (B∩C)) = P(A∩B) + P(A∩C) + P(B∩C) - P(A∩B∩A∩C) - P(A∩B∩B∩C) - P(A∩C∩B∩C) + P(A∩B∩A∩C∩B∩C) = P(A∩B) + P(A∩C) + P(B∩C) - P(A∩B∩C) - P(A∩B∩C) - ~~P(A∩B∩C) + P(A∩B∩C)~~ = P(A∩B) + P(A∩C) + P(B∩C) - 2P(A∩B∩C) = P(A) P(B) + P(A) P(C) + P(B) P(C) – 2 P(A) P(B) P(C) = 3 $(0,487)^2$ - 2 $(0,487)^3$ = 0,711 – 0,231 = **<u>0,48</u>**

También se puede resolver de la siguiente manera:

Sea Y=nº de componentes en 3 del dispositivo que duran al menos 400 h

Y seguirá distribución binomial con n=3 y p=0,487.

Fiabilidad del dispositivo a las 400 h= P(Y≥2)=P(B(3,0,487)≥2)=

$= \binom{3}{2} 0,487^2(1-0,487)^1 + \binom{3}{3} 0,487^3(1-0,487)^0$=**<u>0,48</u>**

4. (2º Parcial) Las tablas *hash* son estructuras de datos, que llevan asociadas claves y que permiten la búsqueda de elementos con costes computacionales bajos sin que dichos elementos hayan tenido que ser previamente ordenados al ser introducidos.
Para conseguir dicha búsqueda eficiente se dispone de la función de *hash*, cuya misión no es otra, que la de ofrecer la posición del elemento buscado a través de una clave.
Un programador propone una función de *hash* y se dispone a evaluarla. Para ello realiza una búsqueda aleatoria de 30 elementos dentro de la tabla *hash* y mide el tiempo (en ms) que se tarda en encontrar cada elemento.
A continuación se muestran algunos de los resultados estadísticos del estudio:

```
Media = 23,17
Desviación típica = 5,72
Asimetría tipi. = 1,01
Curtosis tipificada = 0,13
```

A la vista de estos resultados, se pide:

a) ¿Es aceptable una media poblacional de 37 ms en el tiempo de búsqueda con un riesgo α=0,05? Utiliza el *test t*. (**3 puntos**)

$$H_0: m = 37$$
$$H_1: m \neq 37$$

$$t_{calc} = \left| \frac{23{,}17 - 37}{5{,}72/\sqrt{30}} \right| = |-13{,}24| > t_{29}^{\alpha=0{,}05} = 2{,}045 \; luego \; aceptamos \; H1$$

b) ¿Se puede aceptar que la desviación típica del tiempo de búsqueda en la población sea de 3 ms, con un nivel de confianza del 95%? (**3 puntos**)

$$H_0: \sigma = 3$$
$$H_1: \sigma \neq 3$$

Intervalo de confianza para σ $\left[\sqrt{\frac{(N-1)s^2}{g_2}}, \sqrt{\frac{(N-1)s^2}{g_1}}\right]$

$$g_1 = 16{,}047$$
$$g_2 = 45{,}722$$

$$\sigma = 3 \notin \left[\sqrt{29\frac{5{,}72^2}{45{,}722}} \quad \sqrt{29\frac{5{,}72^2}{16{,}047}}\right] = [4{,}55 \quad 7{,}69]$$

Por tanto no es admisible la H_0: σ=3, con un nivel de confianza del 95%. La desviación típica σ≠3.

Un segundo programador propone otra función de *hash*, y para evaluarla realiza la búsqueda al azar de 12 elementos en la tabla *hash*, midiendo, sobre la misma máquina que el programador anterior, el tiempo en ms que se tarda en realizar la búsqueda de cada uno de los 12 elementos y obtiene los siguientes resultados:

```
Media = 19,32
Intervalo de confianza para la diferencia de medias 95%
[-0,1516;7,8516]
```

c) ¿Puede afirmarse con un nivel de significación α=0,05 que existen diferencias significativas entre los tiempos medios de búsqueda, entre ambas funciones *hash*? ¿Y con un nivel de confianza del 99%? (**2 puntos**)

El intervalo de confianza para la diferencia de medias, con α=0,05, contiene el valor 0. Por tanto, no puede afirmarse con dicho nivel de significación que existe diferencia significativa entre los tiempos medios de búsqueda entre ambas funciones *hash*. Si aumenta el nivel de confianza al 99%, el intervalo de confianza sería más amplio, por lo que seguiría conteniendo el valor 0. Lo que implica que con el 99% de confianza tampoco puede afirmarse que hay diferencia significativa entre los tiempos medios de búsqueda con las dos funciones *hash*.

d) ¿Qué interpretación práctica tiene este Intervalo de Confianza? (**2 puntos**)

El intervalo de confianza es una estimación del verdadero valor de la diferencia que hay entre las dos medias poblaciones con ambas funciones *hash*. El nivel de confianza es la probabilidad de que el intervalo contenga dicho verdadero valor. Permite contrastar la hipótesis nula de igualdad de medias.

5. (2º Parcial) Una empresa de software matemático está interesada en averiguar cuál de dos programas de integración numérica es más rápido. Para ello se realiza un experimento con un total de 12 funciones a integrar, cuatro de tipo 1, cuatro de tipo 2 y otras cuatro de tipo 3. Con cada una de ellas se midió el tiempo (en milisegundos) que el programa tarda en ejecutar el procedimiento de integración, utilizando el programa A en 6 de ellas, y el B en las otras 6.

a) Completa el cuadro resumen del ANOVA e indica qué efectos son estadísticamente significativos (α = 0,05). Justifica los cálculos realizados. (**3 puntos**)

Analysis of Variance for Tiempo - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Integral	555,167	**2**	277,583	90,03	**<0,05**
B:Programa	**70,0833**	**1**	70,0833	22,73	0,0031
INTERACTIONS					
AB	41,1667	**2**	20,5833	**6,6757**	**<0,05**
RESIDUAL	**18,5**	**6**	3,08333		
TOTAL (CORRECTED)	684,917	**11**			

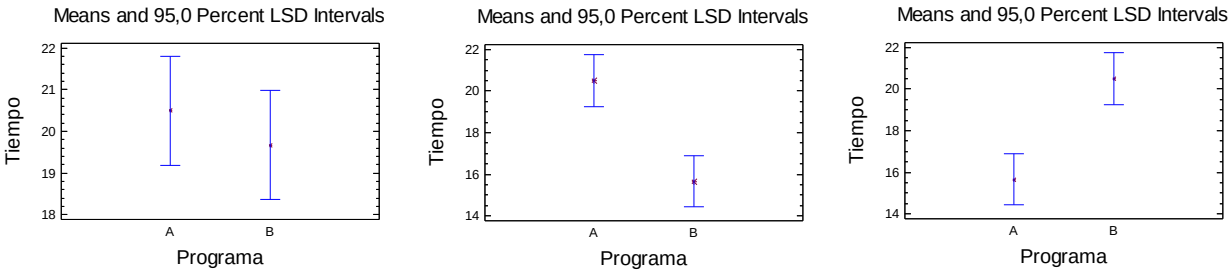
Grados de libertad de tipo de integral=3-1=2
Grados de libertad de programa (g.l.p.)= 2-1=1
Suma de Cuadrados de programa=Cuadrado medio programa x g.l.p.=70,0833
Grados de libertad de la interacción=2x1=2
Grados de libertad totales=12-1=11
Grdaos de libertad residuales (g.l.r.)=11-2-1-2=6
Suma de Cuadrados residual= Cuadrado medio residual x g.l.r.=
=3,0833x6=18,4998
Esta suma de cuadrados también se puede obtener:
Suma de Cuadrados residual= SCtotal-SCintegral-SCprograma-SCinterac.=
=684,917-555,167-70,0833-41,1667=18,5
F-ratio de la interacción=CMinteracción/CMresidual=20,5833/3,08333=6,6757

El efecto del tipo de integral es estadísticamente significativo, porque Fratio=90,03 es > F de tabla con 2 y 6 grados de libertad, para α=0,05 (Ftabla=5,14). El p-value será < 0,05
El efecto del programa es estadísticamente significativo porque su p-value es < 0,05.
El efecto de la interacción estadísticamente significativo porque F-ratio = 6,67 es > Ftabla=5,14. Su p-value será < 0,05.

b) En general, ¿qué información adicional a la proporcionada por la tabla resumen del ANOVA, da una representación gráfica de los intervalos LSD? (**2 puntos**)

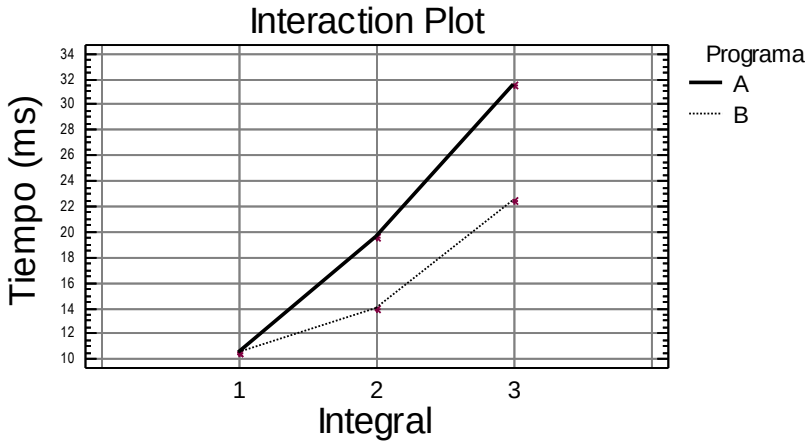
Los intervalos LSD sirven para interpretar estadísticamente el efecto significativo de factores cualitativos con más de dos variantes. Si en la tabla del ANOVA el efecto no es significativo, los intervalos LSD no aportan información adicional (se verá que todos los intervalos están solapados y por tanto no hay ninguna diferencia significativa entre las medias). Si en el ANOVA el efecto es significativo, esto indica que al menos dos de las medias difieren entre sí. Con la representación de los intervalos LSD se aporta en este caso la información adicional respecto a qué medias son las que difieren.

c) Indica cuál de los tres gráficos siguientes se corresponde con el análisis del enunciado. Justifica la respuesta apoyándote en la figura del apartado d. (**2,5 puntos**)



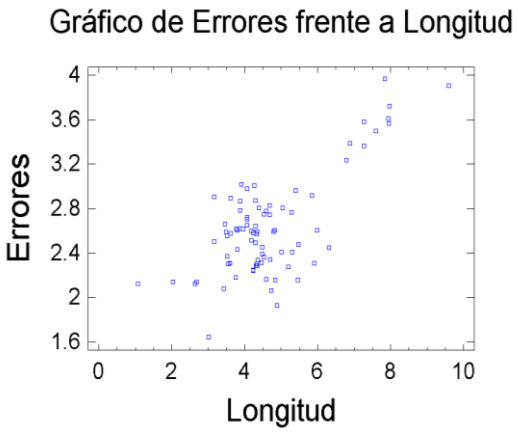
El efecto de programa es significativo en la tabla de ANOVA. Por tanto los intervalos LSD de sus dos variantes no se solapan, y tal y como se ve en el gráfico del apartado d, con A hay mayor tiempo en promedio que con B. Esto indica que el gráfico del centro es el que se corresponde con los análisis del enunciado.

d) A la vista del siguiente gráfico de medias describe el efecto de ambos factores sobre el tiempo medio que se tarda en ejecutar el procedimiento de integración. (**2,5 puntos**)



Con el tipo de integral 1 no hay diferencia entre los programas A y B. Con el tipo 2 hay más tiempo en promedio con A que con B. Ocurre lo mismo con el tipo de integral 3, aunque en este caso la diferencia entre programas es mayor en promedio que con el tipo 2. Visto de otra forma: con el programa A hay más diferencia entre el tipo de integral 1 y el 2, que con el programa B. Ocurre lo mismo con la diferencia entre los tiempos medios de los tipos de integral 2 y 3, ya que esta diferencia es algo mayor con el programa A que con el B.

6. (2º Parcial) La conexión entre dos sistemas informáticos utiliza un enlace de alta capacidad que sin embargo sufre problemas de ruido e interferencia debidos a distintos factores. Con el objeto de estudiar este problema se pretende evaluar el efecto concreto que tiene la longitud media de las tramas de bits entre ambos sistemas sobre el número de errores detectados durante la transmisión de determinados ficheros de prueba de gran tamaño. Se ha recopilado la siguiente información referida a la transmisión de 82 ficheros entre los dos sistemas, para cada uno de los cuales se han anotado la longitud media de las tramas utilizadas en la transmisión y el número medio de errores detectados en esas tramas.



Resumen Estadístico

	Longitud	Errores
Frecuencia	82	82
Media	4.65893	2.63361
Mediana	4.332	2.5896
Varianza	2.04982	0.202711
Desviación típica	1.43172	0.450234
Mínimo	1.072	1.6458
Máximo	9.592	3.965
Rango	8.52	2.3192
Primer cuartil	3.864	2.3062
Tercer cuartil	5.032	2.8041
Rango intercuar.	1.168	0.4979
Asimetría tipi.	3.87271	3.65235
Curtosis tipificada	3.38387	2.05077

Coeficiente de Correlación = 0.722933

a) A partir del gráfico de dispersión anterior, describe la naturaleza de la relación entre las dos variables objeto de estudio. (**1 punto**)

La naturaleza de la relación entre los datos es lineal intermedia y positiva.

b) Teniendo en cuenta la tabla del resumen estadístico anterior determina los parámetros de la recta de regresión correspondiente a la relación entre las dos variables objeto de estudio, y plantea la ecuación del modelo de regresión simple. (**3 puntos**)

$b = r_{xy}\, S_y/S_x = 0.722933*0.450234/1.43172 = 0.227341$

$a = \bar{y} - b\bar{x} = 2.63361 - 0.227341*4.65893 = 1.57444$

La ecuación del modelo de regresión simple sería por tanto:

Promedio de errores/longitud = 1.57444 + 0.227341*Longitud

c) Calcula el número promedio de errores para una longitud de trama igual a 6. (**1 punto**)

Promedio de errores/(Longitud = 6) = 1.57444 + 0.227341*6 = 2.9385

Complementando los resultados anteriores, se ha hecho un análisis mediante Statgraphics que se presenta parcialmente en las siguientes tablas:

Análisis de Regresión - Modelo Lineal Y = a + b*X

Variable dependiente: Errores
Variable independiente: Longitud

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
Ordenada		0.118335	13.3049	
Pendiente		0.0242919	9.35872	0.0000

Análisis de la Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	8.5814	1	8.5814	87.59	
Residuo	7.83818	80	0.0979773		
Total (Corr.)	16.4196	81			

d) ¿Es el modelo planteado significativo globalmente? ¿Y cada uno de sus parámetros? Utiliza un nivel de significación $\alpha = 0.01$. **(3 puntos)**

La significación del modelo se puede valorar a partir del F-ratio de la tabla del ANOVA:

F-ratio = 87.59 > $F_{1,80}^{\alpha = 0.01}$ = 6.96.

Por tanto, el modelo globalmente considerado es significativo y explica la relación lineal entre ambas variables objeto de estudio. El p-valor es por tanto en este caso inferior a α= 0.01. Esto es equivalente a afirmar que la pendiente difiere significativamente de cero. Este resultado también se extrae de la primera tabla, ya que el p-valor de dicha pendiente es < 0,01.

Para ver si la ordenada es significativa, se compara el valor absoluto de su estadístico t=13,3049 con la t de tabla con 80 grados de libertad para $\alpha/2$=0,005. Esta t de tabla está entre 2,617 y 2,66. Como 13,3049 es mayor que cualquier valor en este rango, se concluye que la ordenada difiere también significativamente de cero.

e) Indica qué representa y para qué sirve el Cuadrado Medio Residual del ANOVA del modelo de regresión. **(2 puntos)**

El Cuadrado Medio Residual del ANOVA del modelo de regresión, es una estimación de la varianza de Y condicionada a un valor de la variable X. Cuantifica el efecto conjunto de todos aquellos efectos y factores que no incluye el modelo. Con dicho valor se pueden calcular probabilidades y construir intervalos para la variable Y condicionada a valores determinados de la X.