Grado en Ingeniería Informática

Estadística

EXAMEN FINAL

24 de junio de 2016

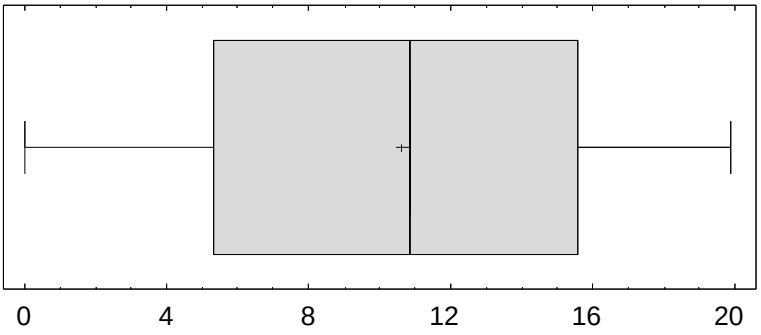
Apellidos y nombre:	
Grupo:	Firma:

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 (40%) y 3 del segundo (60%). 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) El tiempo de búsqueda de un fichero en determinada base de datos documental es una variable aleatoria cuya distribución se desconoce. Se obtiene el tiempo de búsqueda en milisegundos (ms) para cierta muestra aleatoria de ficheros. A partir de los datos obtenidos se obtiene la siguiente representación gráfica.



a) Indicar la población en estudio, cuál es la variable aleatoria y de qué tipo es dicha variable. *(3 puntos)*

b) Calcula al menos 5 parámetros descriptivos que puedan obtenerse a partir de este gráfico, indicando de qué tipo es cada uno de ellos. *(3 puntos)*

c) ¿Es posible deducir el número de ficheros (tamaño de la muestra) a partir del cual se ha obtenido este gráfico? *(1 punto)*

d) Asumiendo que los datos siguen un modelo uniforme, calcular la probabilidad de que el tiempo de búsqueda sea superior a 8 ms. *(3 puntos)*

2. (1er Parcial) La factoría ARTUDITU está interesada en controlar la calidad de los transistores que fabrica. En este proceso se conoce que el 1% de los transistores fabricados son defectuosos.

a) En un control rutinario se toma una muestra al azar de 15 transistores. ¿Cuál es la probabilidad de encontrar más de dos transistores defectuosos? *(3 puntos)*

b) Si se toma al azar una muestra de 15 transistores y se sabe que alguno de ellos es defectuoso, ¿cuál es la probabilidad de que haya exactamente dos defectuosos? *(3 puntos)*

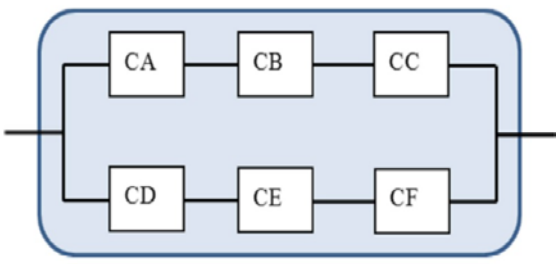
c) Si al cabo de un mes se extraen cien muestras al azar, cada una de tamaño 15, ¿cuál es la probabilidad de que en un mes se encuentren más de diez transistores defectuosos? *(4 puntos)*

3. (1er Parcial) Cierto dispositivo está formado por seis componentes idénticos (CA, CB, CC, CD, CE, CF) cuyo tiempo de funcionamiento hasta el fallo se distribuye exponencialmente. Se conoce que la fiabilidad de los componentes a las 1000 horas de funcionamiento es de 0,9. Los componentes presentan un funcionamiento independiente los unos de los otros. Responde las siguientes preguntas justificando detalladamente las respuestas, y definiendo todas las variables y sucesos empleados.

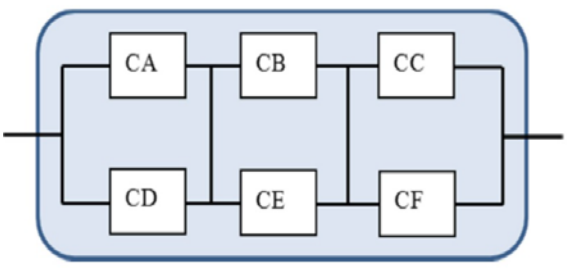
a) Calcula el tiempo medio de vida de los componentes mencionados. *(3 puntos)*

b) Dos estudiantes (EST1 y EST2) tienen opiniones diferentes respecto al tipo más adecuado de montaje de los componentes con el fin de aumentar la fiabilidad del dispositivo resultante. A continuación se muestra la propuesta de montaje de los seis componentes por parte de cada estudiante:

Propuesta EST1:



Propuesta EST2:



Calcula la fiabilidad de cada propuesta. ¿Cuál de las dos es más adecuada teniendo en cuenta el objetivo perseguido? *(7 puntos)*

4. (2º Parcial) Responde a cada una de las siguientes preguntas justificando convenientemente la respuesta.

a) Dada una población normal con desviación típica σ=5, extraemos una muestra de tamaño 16. ¿Cuál es la probabilidad de que la media muestral y la media poblacional difieran en menos de 2 unidades? *(3 puntos)*

b) En cierto proceso se mide de forma rutinaria un parámetro de calidad. Se realizan ciertos cambios operativos en el proceso con el objetivo de mejorar la calidad. Para estudiar la efectividad de dichos cambios, se ha tomado una muestra aleatoria de 5 unidades, obteniéndose una media muestral de 20,5 y una varianza muestral de 1,2. Obtener un intervalo de confianza para la media poblacional, considerando α=0,05. ¿Es razonable admitir que la media poblacional del parámetro es de 22 tras los cambios efectuados en el proceso? *(3,5 puntos)*

c) ¿Es admisible asumir que la varianza poblacional del parámetro, después de los cambios operativos realizados, es de 9,0 unidades^2? *(3,5 puntos)*

5. (2º Parcial) Se desea estudiar el efecto que la configuración (tres posibles: A, B y C) y el tamaño de memoria Caché (3 niveles: bajo, medio, alto) tienen sobre el rendimiento medio de un sistema informático. Cada tratamiento se ha ensayado tres veces.

a) Analiza qué efectos son estadísticamente significativos a partir del cuadro resumen del ANOVA (utiliza α=5%), teniendo en cuenta que: SC_{total}=11039,2; SC_{config}=604,47 ; $SC_{caché}$=8890,4 ; $SC_{residual}$=690,005. *(5 puntos)*

b) Asumiendo que se cumple la hipótesis de homocedasticidad, ¿cuánto vale la estimación de la varianza de cada una de las poblaciones en estudio? *(2 puntos)*

c) En general, ¿qué información adicional a la proporcionada por la tabla resumen del ANOVA puede obtenerse a partir de una representación gráfica de los intervalos LSD? *(3 puntos)*

6. (2º Parcial) En un almacén logístico se dispone de un sistema automático para la gestión de stocks. Para los últimos 100 pedidos gestionados por el almacén, se conoce el número de unidades del pedido y el tiempo (minutos) que se tarda en procesar dicho pedido. Si se construye un modelo de regresión lineal para predecir este tiempo, los resultados obtenidos con Statgraphics son los siguientes:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: TIEMPO
Independent variable: UNIDADES
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept      -0,886722       0,212955        -4,1639         0,0001
Slope          0,0321159     0,00259249
------------------------------------------------------------------------------


                           Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                      39,9565
Residual                   25,5156
------------------------------------------------------------------------------
Total (Corr.)              65,4721     99
```

a) A partir de la información proporcionada por Statgraphics, ¿cuál es el modelo de regresión estimado? *(2 puntos)*

b) ¿Existe un efecto lineal estadísticamente significativo del número de unidades sobre el tiempo medio de proceso? Considerar un riesgo de primera especie del 1%. *(2 puntos)*

c) Obtener el coeficiente de determinación. Indicar cómo se modificará dicho coeficiente si el tiempo se expresa en segundos en lugar de minutos.

(2 puntos)

d) Sabiendo que el tiempo medio de proceso previsto para un pedido de 80 unidades es 1,68 minutos, ¿cuál es la probabilidad de que un pedido de 80 unidades elegido al azar sea procesado en un tiempo superior a 2 minutos?

(4 puntos)

SOLUCIÓN

1a) La población en estudio está formada por el conjunto de ficheros contenidos en la base de datos documental. La variable aleatoria es el tiempo (medido en milisegundos) que se tarda en encontrar un fichero al azar buscado en dicha base de datos. Esta variable es unidimensional, cuantitativa y de tipo continuo (ya que el tiempo se mide en una escala continua).

1b) - Rango = máximo - mínimo = 20 - 0 = 20.
- Intervalo intercuartílico (IIC) = 3er cuartil - 1er cuartil = 15,6 - 5,2 = 10,4.
- Mediana = 10,9 (línea vertical dentro de la caja)
- Media ≈ 10,7 (punto en forma de cruz dentro de la caja)
- Tercer cuartil = 15,6 (extremo derecho de la caja).

Los dos primeros parámetros indicados informan sobre la dispersión de los datos, y los otros tres informan sobre la posición. Además de estos 5 parámetros, otros estadísticos útiles que se pueden obtener del gráfico son:
- Primer cuartil = 5,2 (extremo izq. de la caja), es un parámetro de posición.
- Máximo ≈ 20 (parámetro de posición).
- Coeficiente de asimetría ≈ 0 ya que la forma del gráfico es prácticamente simétrica. Es un parámetro de forma.

1c) No, a partir de un gráfico box-whisker (caja y bigotes) no es posible deducir el número de datos a partir de los cuales se ha construido.

1d) X ≈ U (0; 20); $P(X \le x) = (x-a)/(b-a)$;

$$P(X>8) = 1 - P(X<8) = 1 - (8-0)/(20-0) = 1 - (8/20) = \mathbf{0,6}$$

2a) v.a. X: nº de transistores defectuosos en una muestra de 15. Esta variable sigue una distribución binomial de parámetros: X ≈ Bi (n=15, p=0,01). Aplicando la función de probabilidad de este tipo de distribución:

$$P(X>2) = 1 - P(X \le 2) = 1 - P(X=0) - P(X=1) - P(X=2) =$$

$$= 1 - \binom{15}{0} \cdot 0{,}01^0 \cdot 0{,}99^{15} - \binom{15}{1} \cdot 0{,}01^1 \cdot 0{,}99^{14} - \binom{15}{2} \cdot 0{,}01^2 \cdot 0{,}99^{13} =$$

$$= 1 - 0{,}8601 - 0{,}1303 - 0{,}0092 = \mathbf{0{,}000416}$$

2b) Nos piden la siguiente probabilidad condicional:

$$P[(X=2)/(X>0)] = \frac{P[(X=2) \cap (X>0)]}{P(X>0)} = \frac{P(X=2)}{P(X>0)} = \frac{0{,}00921}{1-0{,}8601} = 0{,}0658$$

2c) El hecho de tomar 100 muestras consecutivas de 15 transistores y contar el número de defectuosos en el total de ellas, es equivalente a tomar una única muestra de tamaño 100·15. Por tanto, la variable aleatoria Y: nº de transistores defectuosos encontrados en 100 muestras de tamaño 15 seguirá una distribución binomial de parámetros: Y ≈ Bi (n=1500, p=0,01). Dado que el valor de n es muy elevado y la probabilidad es pequeña, puede aproximarse a una distribución Poisson: Y ≈ Ps (λ = n·p= 15). A partir del ábaco de Poisson, trazando una línea vertical para λ = 15 que corta la curva 10, en este punto de corte se lee en el eje vertical una probabilidad de 0.12. Así pues:

$$P(Y>10) = 1 - P(Y \le 10) = 1 - 0{,}12 = \mathbf{0.88}$$

Otra forma alternativa de resolver el problema es a partir del teorema central del límite. La variable Y es la suma de los transistores defectuosos encontrados en el total de 100 muestras: $Y = X_1 + ... + X_{100}$. Dado que el número de sumandos es elevado, podemos aplicar el teorema central del límite de modo que Y puede asumirse con distribución normal.

$X_i \approx Bi(n=15, p=0{,}01)$; $E(X)=n\cdot p=0{,}15$; $\sigma_X^2 = n\cdot p\cdot(1-p) = 15\cdot 0{,}01\cdot 0{,}99 = 0{,}1485$

$$E(X_1+...+X_{100}) = E(X_1)+...+E(X_{100}) = 100\cdot E(X) = 100\cdot 0{,}15 = 15$$

$$\sigma^2(X_1+...+X_{100}) = \sigma^2(X_1)+...+\sigma^2(X_{100}) = 100\cdot \sigma^2(X) = 100\cdot 0{,}1485 = 14{,}85$$

$$P(Y>10) \approx P\left[N\left(15, \sqrt{14{,}85}\right) \geq 10{,}5\right] = P\left[N(0;\ 1) \geq \frac{10{,}5-15}{\sqrt{14{,}85}}\right] = P\left[N(0;\ 1) \geq -1{,}17\right] = \mathbf{0{,}88}$$

3a) Variable aleatoria T: tiempo de funcionamiento (horas) hasta el fallo del componente. Si la fiabilidad del componente a las 1000 h es 0,9, por definición de fiabilidad esto implica que P(T>1000)=0,9.

$P(T>1000) = e^{-\alpha\cdot 1000} = 0{,}9$; $\alpha = -(\ln 0{,}9)/1000$.

Media de una variable exponencial: $E(X) = 1/\alpha = -1000/(\ln 0{,}9) = \mathbf{9491{,}2}$ horas.

3b) Suceso A: el componente A sigue en funcionamiento al cabo de 1000 h. P(A)=0,9. De modo análogo se definen los sucesos B, C, D, E y F.

Cálculo de la fiabilidad del montaje propuesto por EST-1:
La fiabilidad del montaje a las 1000 h, es decir la probabilidad de que funcione más de 1000 h, es la probabilidad de que funcione la rama superior o bien (suceso unión) de que funcione la rama inferior. Para que funcione cada rama, tienen que funcionar todos los componentes (suceso intersección). Así pues:

Fiabilidad = $P\left[(A\cap B\cap C)\cup(D\cap E\cap F)\right] =$

$$= P\left[(A\cap B\cap C)\right] + P\left[(D\cap E\cap F)\right] - P\left[(A\cap B\cap C)\cap(D\cap E\cap F)\right] =$$

Asumiendo independencia de sucesos, tenemos que:

$$= P(A)\cdot P(B)\cdot P(C) + P(D)\cdot P(E)\cdot P(F) - P(A)\cdot P(B)\cdot P(C)\cdot P(D)\cdot P(E)\cdot P(F) =$$

$$= 0{,}9^3 + 0{,}9^3 - 0{,}9^6 = \mathbf{0{,}9266}$$

Cálculo de la fiabilidad del montaje propuesto por EST-2:
El montaje es equivalente a 3 subcircuitos en serie, cada uno de los cuales tiene dos componentes en paralelo. La fiabilidad del montaje a las 1000 h será la probabilidad de que funcionen los tres subcircuitos (suceso intersección). Para que funcione cada subcircuito, tiene que funcionar el componente de la rama superior o el inferior (suceso unión). Dado que el funcionamiento es independiente y la fiabilidad es la misma para todos los componentes:

Fiabilidad = $P\left[(A\cup D)\cap(B\cup E)\cap(C\cup F)\right] = P(A\cup D)\cdot P(B\cup E)\cdot P(C\cup F) =$

$$= \left[P(A\cup D)\right]^3 = \left[P(A)+P(D)-P(A)\cdot P(D)\right]^3 = \left[0{,}9+0{,}9-0{,}9^2\right]^3 = \mathbf{0{,}9703}$$

La propuesta del 2º estudiante es más adecuada ya que su fiabilidad es mayor.

4a) Sabiendo que $\bar{x} \approx N\left(m\ ;\ \sigma/\sqrt{n}\right)$ se obtiene: $\bar{x} - m \approx N\left(0\ ;\ \sigma/\sqrt{n}\right)$.

$$P\left(\left|\bar{x}-m\right| < 2\right) = P\left[\left(\bar{x}-m\right) \in [-2;\ 2]\right] = 1-2\cdot P\left[\left(\bar{x}-m\right) > 2\right] = 1-2\cdot P\left[N(0;\sigma/\sqrt{n}) > 2\right] =$$

$$= 1-2\cdot P\left[N(0;1) > \frac{2-0}{5/\sqrt{16}}\right] = 1-2\cdot P\left[N(0;1) > 1{,}6\right] = 1-2\cdot 0{,}0548 = \mathbf{0{,}8904}$$

4b) El intervalo de confianza de la media poblacional (m) se calcula a partir de la media muestral (20.5) y la varianza muestral (1.2):

$\bar{x} \pm t_{n-1}^{\alpha/2} \cdot s/\sqrt{n} = 20{,}5 \pm t_4^{0{,}025} \cdot \sqrt{1{,}2}/\sqrt{5} = 20{,}5 \pm 2{,}776 \cdot 0{,}4899 = 20{,}5 \pm 1{,}36$

$m \in [19{,}14;\ 21{,}86]$. No es razonable admitir que la media poblacional sea 22 para α=0,05 ya que este valor está fuera de este intervalo de confianza obtenido.

4c) Obtenemos en primer lugar un intervalo de confianza para la varianza de la población: $\sigma^2 \in \left[(n-1)\cdot s^2/g_2;\ (n-1)\cdot s^2/g_1\right]$ siendo g_1 y g_2 los valores críticos de una distribución gi cuadrado con 4 grados de libertad que comprenden el 95% de valores de dicha distribución. A partir de la tabla de esta distribución se obtiene g_1=0,484, g_2=11,143. Así pues, el intervalo de confianza será:

$\sigma^2 \in \left[4\cdot 1{,}2/11{,}143;\ 4\cdot 1{,}2/0{,}484\right] = [0{,}43;\ 9{,}92]$

Es admisible asumir que la varianza poblacional del parámetro es σ^2=9 para α=0,05 ya que este valor está dentro del intervalo obtenido.

5a) Tres posibles configuraciones combinadas con 3 niveles de memoria da lugar a 9 tratamientos. Cada uno se ensaya 3 veces, por lo que en total tenemos 27 datos experimentales, de modo que los grados de libertad totales serán 27-1=26. Cada factor tiene 3 niveles (dos grados de libertad), la interacción tiene 4 grados de libertad (el producto 2·2). El cuadrado medio se obtiene dividiendo la suma de cuadrados entre los grados de libertad, y la F-ratio se obtiene dividiendo el cuadrado medio correspondiente entre el cuadrado medio residual.

OV	Suma de cuadrados	gl	Cuadrado medio	F-Ratio	P-Valor
Config	604.47	2	302,235	7,88	<0,05
Caché	8890,4	2	4445,2	115,97	<0,05
Config*Caché	854,325	4	213,58	5,57	<0,05
RESIDUAL	690,005	18	38,33		
TOTAL	11039,2	26			

El valor crítico de una distribución F con 2 y 18 grados de libertad para α=5% vale 3,55. Como la F-ratio de Configuración es mayor que este valor crítico, este factor tiene un efecto estadísticamente significativo sobre el rendimiento medio. Para memoria caché también F-ratio=115,97 > 3,55, por lo que este factor también influye significativamente sobre el rendimiento medio.

El valor crítico de una F con 4 y 18 g.l. para α=5% vale 2,93. Como la F-ratio de la interacción es 5,57 >2,93, puede decirse que el efecto de la interacción también resulta estadísticamente significativo.

5b) Cada uno de los tratamientos corresponde a una población distinta. Asumiendo que se cumple la hipótesis de homocedasticidad, esto implica que la varianza de todas las poblaciones será la misma, la cual equivale a la varianza de los residuos (varianza residual). En ANOVA, dicha varianza coincide con el cuadrado medio residual, que según la tabla obtenida anteriormente vale **38,33**.

5c) 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.

En caso de que el factor sea cuantitativo con más de dos niveles, el gráfico de medias con intervalos LSD aporta una idea sobre el posible efecto lineal o cuadrático del factor sobre la variable respuesta, pero conviene emplear regresión para estudiar la significación estadística de dicho efecto.

6a) Teniendo en cuenta los valores estimados de la pendiente (slope) y la ordenada en el origen (intercept), la ecuación del modelo de regresión será:
Tiempo = -0,8867 + 0,03212·unidades

6b) Calculamos el cociente entre el valor estimado de la pendiente y su error estándar: 0,03212 / 0,002592 =12,38. Si fuera cierta la hipótesis nula de que la pendiente de la recta es cero a nivel poblacional, este cociente seguirá una distribución t de Student con N-1-I = 100-1-1 = 98 grados de libertad. Pero el valor 12,38 es muy poco frecuente para esta distribución, con lo cual se rechaza la hipótesis nula: hay suficiente evidencia para afirmar que la pendiente de la recta es distinta de cero a nivel de la población, y por tanto, el efecto lineal es estadísticamente significativo.

6c) $R^2 = \frac{SC_{modelo}}{SC_{total}} = \frac{39{,}9565}{65{,}4721} = \mathbf{0{,}6103}$ = 61,03%

Este coeficiente no se modifica si las variables se expresan en otras unidades, ya que eso dicha modificación no altera el grado de correlación entre las variables, que es lo que se cuantifica por medio de este coeficiente.

6d) La distribución condicional de Y cuando X vale 10 será de tipo normal, siendo la media 1,68 (dato del enunciado) y la varianza será la residual. Dicha varianza se estima a partir del cuadrado medio residual, por lo que hay que completar la tabla. El modelo tiene un grado de libertad porque solamente hay una variable incluida. Así pues, 25,516 / 98 = 0,2604.

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	39,9565	1			
Residual	25,5156	98	0,2604		
Total (Corr.)	65,4721	99			

$$Y/X=10 \approx N(1{,}68;\ \sqrt{0{,}2604}); \qquad P(Y>2/X=10) = P\left[N(1{,}68;\sqrt{0{,}2604})>2\right] =$$
$$= P\left[N(0;\ 1) > (2-1{,}68)/\sqrt{0{,}2604}\right] = P\left[N(0;\ 1) > 0{,}627\right] = \mathbf{0{,}266}$$