etsinf

Grado en Ingeniería Informática

Estadística

EXAMEN FINAL

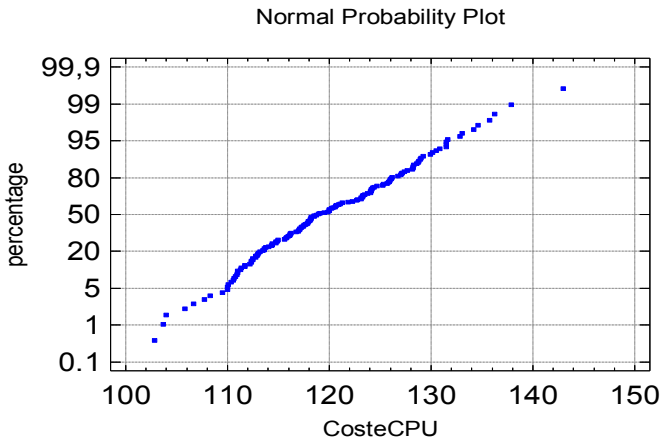
10 de junio de 2013

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, 5 correspondientes al primer parcial (80%) y una del segundo (20%). 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) Se pretende mejorar el rendimiento de las consultas realizadas por diversas sucursales bancarias a una base de datos (BD) con información financiera (FDB). Con este objetivo se ha decidido registrar, para cada consulta, el *Coste CPU*, que representa el tiempo empleado por cada consulta en efectuar una operación de E/S en la FDB. En un primer análisis, se han seleccionado 200 consultas al azar realizadas a lo largo de un mes, anotándose el *Coste CPU* de cada consulta. Los resultados obtenidos (en milisegundos para facilitar los cálculos), representados sobre Papel Probabilístico Normal, se muestran a continuación. A partir del gráfico adjunto, se pide:



A partir del gráfico adjunto, se pide:

a) Definir detalladamente la población y la variable aleatoria implicadas en el estudio. (**2,5 puntos**)

b) ¿Qué se puede decir sobre la distribución de la variable aleatoria analizada? (**2,5 puntos**)

c) ¿Qué parámetros de dispersión y posición serán los más adecuados para caracterizar los datos analizados?. Justifica tu respuesta. (**2,5 puntos**)

d) Obtener a partir del gráfico, el valor de la media y de la desviación típica poblacional de los datos analizados indicando el procedimiento seguido. (**2,5 puntos**)

2. (1er Parcial) Los trabajadores de una empresa se distribuyen del siguiente modo: un 10% de Directivos y Administrativos, un 10% de Comerciales y un 80% de Trabajadores de Planta. Durante un periodo de control se ha observado que el absentismo laboral es del 15% para los trabajadores del primer grupo, el 8% para los del segundo grupo y el 5% para los del tercero.

a) Indica y describe los sucesos implicados, así como las probabilidades asociadas a los mismos. (**1 punto**)

b) ¿Cuál es el porcentaje de absentismo global en la empresa? (**3 puntos**)

c) Sabiendo que un trabajador es absentista, ¿cuál es la probabilidad de que éste sea un comercial? (**3 puntos**)

d) Si un día se escoge un trabajador al azar de la empresa ¿qué probabilidad hay de que esté no esté ausente y que además sea un comercial? (**3 puntos**)

3. (1er Parcial) Chisco Ltd fabrica dos tipos de tarjetas de red: RA y RB que comercializa en lotes de 100. Se sabe que en el proceso de fabricación de las tarjetas RA se produce un 2% de tarjetas con algún defecto, mientras que en las de tipo RB el promedio de tarjetas defectuosas es de 1.
La empresa ELECTONYS utiliza tarjetas de red RA y RB comercializadas por Chisco Ltd en los equipos que monta.

a) Calcula la probabilidad de que en 2 lotes de tarjetas (un lote de cada tipo) se obtenga, en total, más de 2 tarjetas con algún defecto. **(6 puntos)**

b) Calcula el número de tarjetas de red tipo RB (**x**) de forma que la probabilidad de encontrar como máximo **x** tarjetas defectuosas sea, como mucho, del 99%. **(4 puntos)**

4. (1er Parcial) El tiempo empleado por cada grupo de visitantes que usan los ordenadores de búsqueda instalados en una biblioteca se distribuye como una variable normal de media 200 s y desviación típica 30 s. ¿Cuál es la probabilidad de que el tiempo empleado por un grupo de visitantes en la búsqueda de documentación esté comprendido entre 150 y 200 segundos?

5. (1er Parcial) Una empresa desea saber si hay una diferencia significativa entre el número de errores que realizan sus dos programadores, el que trabaja por la mañana y el que lo hace por la tarde. Para ello, se coge una muestra aleatoria de 61 programas realizados por cada uno de ellos, y se obtienen los siguientes datos:

Resumen Estadístico

	PROG 1	PROG 2
Frecuencia	61	61
Media	9,11475	11,8852
Mediana	9,0	12,0
Moda	10,0	
Varianza	7,06995	10,0699
Desviación típica	2,65894	3,17332
Mínimo	3,0	5,0
Máximo	17,0	20,0
Rango	14,0	15,0
Asimetría tipificada	1,90818	0,448924
Curtosis tipificada	1,74821	-0,512362

¿Hay diferencias significativas en el número medio de errores en los programas realizados por los dos programadores? Utiliza $\alpha = 0,05$.

6. (2º Parcial) En el ámbito de un estudio sobre un nuevo sistema de archivos para un sistema operativo se está analizando la relación entre el tiempo de ejecución (ms) de 21 proceso de un determinado tipo y el tiempo medio de acceso a disco (ms) obtenido durante la ejecución de cada proceso.
Parte del estudio consistió en realizar un estudio de Regresión, obteniendo los siguientes resultados:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------
Dependent variable: Y
Independent variable: X
------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------
Intercept      0,0161212      0,735938      __________
Slope            3,32307      0,190529      __________
------------------------------------------------------------------------

                           Analysis of Variance
------------------------------------------------------------------------
Source      Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------
Model             55,6913      1      55,6913     304,20       0,0000
Residual          _______     ___     ________
------------------------------------------------------------------------
Total (Corr.)     59,1697     ___
```

A partir de los resultados anteriores, contesta a las siguientes preguntas justificando las respuestas:

a) Plantea y estima un modelo de regresión que permita estimar el tiempo de respuesta de los procesos, a partir del tiempo medio de acceso a disco para este sistema de archivos (α=0,05). **(4 puntos)**

b) ¿Qué tiempo medio de respuesta cabe esperar para un proceso, cuando el tiempo medio de acceso a disco es de 3 ms? **(3 puntos)**

c) Indica qué valor tiene la Varianza Residual asociada al modelo y cuál es su interpretación práctica. **(3 puntos)**

SOLUCIÓN

1a) La población está formada por todas las posibles consultas que pueden realizarse con la base de datos. La variable aleatoria es el tiempo (en milisegundos) empleado por cada consulta en efectuar una operación de E/S con esta base de datos.

1b) Dado que todos los puntos se ajustan a una línea recta en el papel probabilístico normal, puede asumirse que los datos siguen una distribución normal. No se observan datos anómalos. El gráfico indica que la mediana es 120, y la media será cercana a 120 dado que la distribución es simétrica. El mínimo es 103, el máximo es 143 y por tanto el rango es 40. El rango intercuartílico es aproximadamente 124-114 = 10.

1c) Dado que los valores siguen una distribución normal (que es simétrica) y no se observan datos anómalos, la media coincide con la mediana y será el parámetro de posición más adecuado. La varianza será un buen parámetro de dispersión, al igual que la desviación típica. También el rango intercuartílico, aunque este parámetro es más útil en caso de distribuciones asimétricas.

1d) El percentil 50 (leyendo en la escala vertical) corresponde a coste CPU = 120. Este valor es la mediana, que coincide en este caso con la media dada la simetría de la distribución. En una distribución normal, el intervalo [m-2s ; m+2s] comprende el 95% de los datos, y por tanto el 2,5% de los valores están por debajo de (m-2s). Este valor es el percentil 2,5%, el cual según el gráfico es 105 aproximadamente. Si m-2s = 105, se deduce que: s = (120-105)/2 ≈ **7,5**

2a) Suceso D: el trabajador es directivo o administrativo. C: el trabajador es comercial.
Pw: el empleado trabaja en planta A: el trabajador está ausente
P(D)=0.1 ; P(C)=0.1; P(Pw)=0.8 ; P(A/D)=0.15 ; P(A/C)=0.08; P(A/Pw)=0.05

2b) Aplicando el Teorema de la Probabilidad Total, el porcentaje es **6,3%**:
P(A)=P(D)·P(A/D)+P(C)·P(A/C)+P(Pw)·P(A/Pw) = 0.1·0.15+0.1·0.08+0.8·0.05 = 0.063

2c) $P(C/A)=\dfrac{P(C)\cdot P(A/C)}{P(A)}=\dfrac{0.1\cdot 0.08}{0.063}=$**0,127**

2d) $P\left(\overline{A}\cap C\right)=P(C)\cdot P\left(\overline{A}/C\right)=P(C)\cdot\left[1-P(A/C)\right]=0.1\cdot(1-0.08)=$**0,092**
El siguiente procedimiento no es correcto porque los sucesos no son independientes:
$P\left(\overline{A}\right)\cdot P(C)=(1-0.063)\cdot 0.1=0.0937\neq P\left(\overline{A}\cap C\right)$
Conviene recordar que: $P\left(\overline{A}/C\right)=1-P(A/C)$, pero: $P\left(C/\overline{A}\right)\neq 1-P(C/A)$

3a) Se definen las siguientes variables aleatorias:
X_A: nº de tarjetas defectuosas en un lote de 100 unidades de tarjetas RA
X_B: nº de tarjetas defectuosas en un lote de 100 unidades de tarjetas RB.
$X_A\approx Bi(n=100;p=0.02)$; $X_B\approx Bi(n=100;p=0.01)$
Dado que *n* es elevado y p es bajo, la distribución binomial puede aproximarse por una distribución Poisson: $X_A\approx Ps(\lambda=n\cdot p=2)$; $X_B\approx Ps(\lambda=n\cdot p=1)$
$P\left[(X_A+X_B)>2\right]=P\left[Ps(\lambda=2+1)>2\right]=1-P\left[Ps(\lambda=3)\leq 2)\right]=$(uso del ábaco)= **0,58**

(cálculo alternativo) = $1-e^{-3}\cdot\left(\dfrac{3^0}{0!}+\dfrac{3^1}{1!}+\dfrac{3^2}{2!}\right)=1-8.5\cdot e^{-3}=$**0,577**

Cálculo alternativo: al ser n=100 en ambos casos: $(X_A+X_B)\approx Bi(n=200;p=0.015)$

$$P[(X_A + X_B) > 2] = 1 - \binom{200}{0} p^0 \cdot 0.985^{200} - \binom{200}{1} \cdot 0.015^1 \cdot 0.985^{199} - \binom{200}{2} \cdot 0.015^2 \cdot 0.985^{198} =$$
$$= 1 - 0.985^{200} - 200 \cdot 0.015 \cdot 0.985^{199} - 100 \cdot 199 \cdot 0.015^2 \cdot 0.985^{198} = \mathbf{0.578}$$

3b) A partir del ábaco de Poisson: $P(X_B \le k) < 0.99$; $P[Ps(\lambda = 1) \le k] < 0.99$

Leyendo en el ábaco con una probabilidad de 0,99 (línea horizontal) y λ=1 (línea vertical), ambas líneas se cortan en un punto intermedio entre las curvas 3 y 4, de modo que no está claro si la solución es k=3 o bien k=4:

Si k=4, con el ábaco: $P[Ps(\lambda = 1) \le 4] = 0.996 > 0.99$ No se cumple la condición.

Si k=3, con el ábaco: $P[Ps(\lambda = 1) \le 3] = 0.98 < 0.99$ Por tanto, la solución es **k=3**.

4) El tiempo T sigue una distribución: $T \approx N(m = 200; \sigma = 30)$

$$P(150 < T < 200) = P(T < 200) - P(T < 150) = P[N(200;30) \le 200] - P[N(200;30) < 150] =$$
$$= 0.5 - P[N(0;1) < (150 - 200)/30] = 0.5 - P[N(0;1) < -1.667] = 0.5 - P[N(0;1) > 1.667] =$$

=(tabla) = 0.5 - 0.048 = **0,452**

5) Dado que los coeficientes estandarizados de asimetría y cursos están comprendidos en el intervalo [-2; 2], no hay suficiente evidencia para afirmar que sean distintos de cero a nivel poblacional. Por tanto, puede asumirse que los datos siguen una distribución Normal, y asumiendo que $\sigma_1 = \sigma_2$, pueden aplicarse las ecuaciones:

$$H_0 : m_1 = m_2 \; ; \quad H_1 : m_1 \neq m_2 \; ; \quad t^{\alpha/2}_{n_1+n_2-2} = t^{0.025}_{61+61-2} = t^{0.025}_{120} = 1.98$$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{1}{61} + \frac{1}{61}} \cdot \sqrt{\frac{60 \cdot 7.07 + 60 \cdot 10.07}{61 + 61 - 2}} = 0.5301$$

H_0 será rechazada si se cumple la siguiente condición, como sucede de hecho:

$$\left| \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}} \right| > t^{0.025}_{120} \; ; \quad \left| \frac{9.1147 - 11.885}{0.5301} \right| = 5.226 > 1.98$$

Por tanto, la respuesta es SÍ, hay suficiente evidencia para afirmar que las diferencias observadas en el número medio de errores son estadísticamente significativas.

Nota: este tipo de problemas quedó fuera del temario a partir del curso 2014-2015

6a) El modelo propuesto es: $T_{ejecución} = 0.01612 + 3.323 \cdot T_{acceso\_disco}$

Estimación del modelo: esta ecuación será útil en la práctica si se puede garantizar que existe correlación entre ambas variables a nivel poblacional, lo que implica que la pendiente *b* sea distinta de cero. La hipótesis nula a contrastar es:

$H_0 : b = 0$; $H_1 : b \neq 0$ El p-valor del test de significación global ("analysis of variance") es muy bajo (p=0.0000). Este mismo p-valor corresponderá a la pendiente porque sólo hay una variable en el modelo. Dado que p-valor < 0,05, se rechaza la hipótesis nula, concluyéndose que el modelo propuesto es válido.

Procedimiento alternativo: se rechaza la hipótesis nula porque:

$$t = b_i / s_{b_i} = 3.323/0.1905 = 17.4 > \left(t^{\alpha/2}_{n-1-I} = t^{0.025}_{21-1-1} = 2.093\right)$$

6b) $E(T_{ejec} / T_{disco} = 3) = 0.01612 + 3.323 \cdot 3 =$ **9,985**

6c) gl_{total} = 21-1=20; gl_{resid} = 20-1=19; $SC_{resid} = SC_{total} - SC_{modelo}$ =59.1697-55.6913=3.478

$CM_{resid} = SC_{res}/gl_{res}$ = 3.4784/19 = **0.1831** = varianza residual

Interpretación práctica: es la varianza de la distribución condicional (es decir, la distribución de Y cuando X toma un valor particular), que se asume constante (hipótesis de homocedasticidad). Este valor se emplea para calcular probabilidades condicionales.