Grados en Informática

Métodos Estadísticos Examen Junio 2012

- **Tiempo: 2 horas 30 minutos**.
- Dejar DNI encima de la mesa. **Apagar y guardar el MÓVIL**.

APELLIDOS, NOMBRE:
DNI: Grupo: Titulación:

1. El tiempo que tarda una determinada máquina en cortar un raíl del metro (variable Y) es función de la cantidad de impurezas del acero dada en % (variable X). Se sospecha que la función que liga ambas variables es del tipo : $Y = \frac{1}{b^X + 0.5}$

   a) Ajustar dicha función a los datos:

X	1	0.25	0.125	0.05
Y	0.4	1	1.2	4/3

   b) ¿Es este ajuste mejor o peor que el lineal?

   c) ¿Que variable es más dispersa X ó Y? Responder mediante el parámetro adecuado.

   (1+0.75+0.25=2.25 Puntos)

2. Se desea separar las cerezas de la variedad Burlat en 3 calidades: "Extra", "primera" y "segunda" categoría según su tamaño. Suponemos que el tamaño de las mismas (diámetro) sigue una distribución normal y establecemos, en principio, como límites de clase 22 mm. y 30 mm., encontrando que el 32.28 % de la producción es menor que 22 mm. y el 25.78 % es mayor de 30 mm., hallar:

   a) La media y varianza de los diámetros de las cerezas.

   b) Posteriormente deseamos afinar más y deseamos que el 40 % de la producción sea de categoría "primera" y el 20 % de categoria "extra". ¿Cuáles deben ser ahora los límites de cada clase.

   (0.75+0.5=1.25 Puntos)

3. La variable aleatoria ξ tiene por función de distribución: $F(x) = \begin{cases} \frac{x^2}{6} & x \in [0,2] \\ a(3x - \frac{x^2}{2} - 4) + \frac{2}{3} & x \in (2,3] \\ 0 & x \leq 0 \\ 1 & x \geq 3 \end{cases}$

   a) Determinar el valor de a.

   b) Hallar $P(\xi > 2.7)$.

   c) Hallar la probabilidad de que de 100 variables aleatorias independientes siguiendo esa distribución, 2 ó más sean mayores que 2.7.

   (0.5+0.25+0.5=1.25 Puntos)

4. *a*) Ajustar una distribución exponencial a los datos:

x_i	$[0,10]$	$(10,14]$	$[14,18]$	$[18,24)$	Más de 24	
n_i	1508	831	377	202	82	3000

   b) Analizar la bondad del ajuste.

   (0.5+0.75=1.25 Puntos)

5. Una empresa de autobuses ha diseñado una campaña para mejorar la puntualidad de los mismos. Para ello, ha implementado un curso de 4 semanas, donde un monitor va asesorando al conductor durante la conducción.

   Para ver su eficacia, se miden los desfases en los tiempos de llegada previstos y reales a determinadas paradas, antes y después del curso, obteniéndose los valores:

   Antes: Analizados 75 desfases, Suma=750 min., Suma de cuadrados=20836 min.

   Después: Analizados 125 desfases, Suma=750 min., Suma de cuadrados=6597 min.

a) Analizar al 5 % si el curso ha mejorado la puntualidad. pues los desfases se han reducido.

b) Analizar al mismo nivel si se han reducido las variaciones (varianza).

(0.75+0.75=1.5 Puntos)

========== Entregar en folio aparte ==========

Indicar, tan solo, las órdenes necesarias (MATLAB o lenguaje equivalente) para resolverlos, pero sin usar calculadora ni tablas.

6. Dada la tabla bidimensional:

   a) Ajustar una parábola de la forma $y = a + bx + cx^2$ a los datos de la tabla:

$X\backslash Y$	0	1	2	3	4
$(-\infty, 0]$	0	0	25	25	0
$(0, 5]$	0	25	20	5	20
$(5, 10]$	15	15	5	0	12
$(10, 15]$	42	2	0	0	0

   b) Hallar el coeficiente de determinación del ajuste realizado.

(0.5+0.25=0.75 Puntos)

7. *a*) Estimar, mediante el método de Montecarlo con 10000 iteraciones, la probabilidad de que al sumar los cuadrados de 10 variables aleatorias ξ_i que siguen una N(0,2), se obtenga un valor mayor que 30 $P\left(\sum_{i=1}^{10} |\xi_i| > 30\right)$.

   b) Hallar el intervalo de confianza al 95 % para esta probabilidad.

(0.5+0.25=0.75 Puntos)

8. Dos algoritmos están diseñados para resolver aproximadamente un problema NP (no polinómico) en tiempo real, por ejemplo, obtener el clique máximo de un grafo. Se quiere comparar la eficiencia de la solución obtenida, para ello se resuelven un conjunto de problemas por ambos métodos, obteniéndose los valores:

Grafo:	1	2	3	4	5	6	7	8	9	10
Sol. Met 1:	12	10	9	6	15	12	18	9	8	13
Sol. Met 2:	13	10	8	8	18	11	19	11	9	15

   a) Contrastar al nivel α =3 % que el segundo método es mejor, pues aumenta el tamaño medio del clique obtenido.

   b) Contrastar al mismo nivel, que la media del segundo método es 12.5.

(0.5+0.5=1 Punto)

Soluciones:

Problema 1:

El ajuste pedido es $Y = \frac{1}{b^X + 0.5} \Rightarrow b^X + 0.5 = \frac{1}{Y} \Rightarrow b^X = \frac{1}{Y} - 0.5 \Rightarrow XLn(b) = Ln(\frac{1}{Y} - 0.5) \Rightarrow \mathbf{BX} = \mathbf{y}$, donde $B = Ln(b)$ e $y = Ln\left(\frac{1}{Y} - 0.5\right)$

Como en el ajuste $y = BX$ el único elemento de la base es X, tenemos:

$$\langle y, X\rangle = B\langle X, X\rangle \Rightarrow \sum_i y_i X_i = B\sum_i X_i^2$$

Formamos la tabla:

X_i	Y_i	y_i	$y_i X_i$	X_i^2	y_i^{exp}	r_i	r_i^2	Y_i^2
1	0.4	0.6931	0.6931	1	0.5446	−0.1446	0.0209	0.16
0.25	1	−0.6931	−0.1733	0.0625	0.6349	0.3651	0.1333	1
0.125	1.2	−1.0986	−0.1373	0.0156	0.6507	0.5493	0.3018	1.44
0.05	4/3	−1.3863	−0.0693	0.0025	0.6602	0.6731	0.4531	1.7778
1.425	3.9333		0.3132	1.0806		1.4430	0.9091	4.3778

El valor de B se obtiene como: $B = \frac{0.3132}{1.0806} = 0.2898 \Rightarrow b = e^B = 1.3362$ y el ajuste queda:

$$\mathbf{Y} = \frac{\mathbf{1}}{\mathbf{1.3362^X + 0.5}}$$

1-b:

Debemos comparar R^2 de este ajuste con r^2 del lineal.

Para hallar $R^2 = 1 - \frac{V_r}{V_y}$ calculamos los y_i estimados por el ajuste (columna 6) y los residuos (columna 7).

Varianza residual: $V_r = \frac{\sum_i r_i^2}{4} - \left(\frac{\sum_i r_i}{4}\right)^2 = \frac{0.9091}{4} - \left(\frac{1.443}{4}\right)^2 = 0.0971$

Varianza de Y: $V_Y = \frac{\sum_i Y_i^2}{4} - \left(\frac{\sum_i Y_i}{4}\right)^2 = \frac{4.3778}{4} - \left(\frac{3.9333}{4}\right)^2 = 0.1275$

Razón de determinación: $R^2 = 1 - \frac{V_r}{V_y} = 1 - \frac{0.0971}{0.1275} = 0.2382$

Para el lineal, calculamos V_X, cov(X,Y) y r^2:

Varianza de X: $V_x = \frac{1.0806}{4} - \left(\frac{1.425}{4}\right)^2 = 0.1432$

$m_{1,1} = \sum_i \frac{X_i Y_i}{4} = \frac{0.4+0.25+0.15+0.0667}{4} = 0.2167 \Rightarrow cov(X,Y) = 0.2167 - \frac{1.425}{4}\frac{3.9333}{4} = -0.1336$

$r^2 = \frac{cov^2}{V_X V_Y} = \frac{(-0.1336)^2}{(0.1432)(0.1275)} = 0.978$

Luego es mucho mejor el ajuste lineal.

1-c: El parámetro apropiado es el coeficiente de variación:

$CV(X) = \frac{\sigma_X}{|\bar{X}|} = \frac{\sqrt{0.1432}}{0.3563} = 1.0624 \qquad CV(Y) = \frac{\sigma_Y}{|\bar{Y}|} = \frac{\sqrt{0.1275}}{0.9833} = 0.3631$

Luego tiene menos dispersión la variable Y.

Problema 2: Nos dicen que ξ sigue una normal $N(\mu, \sigma)$ y que $P(\xi < 22) = 0.3228$ luego $P\left(z < \frac{22-\mu}{\sigma}\right) = 0.3228 \Rightarrow P\left(z > -\frac{22-\mu}{\sigma}\right) = 0.3228 \Rightarrow -\frac{22-\mu}{\sigma} = 0.46$ (en las tablas) $\mu - 0.46\sigma = 22$.

También nos dicen que $P(\xi > 30) = 0.2578 \Rightarrow P\left(z > \frac{30-\mu}{\sigma}\right) = 0.2578 \Rightarrow \frac{30-\mu}{\sigma} = 0.65$ (en las tablas) $\mu + 0.65\sigma = 30$.

Resolviendo: $\left.\begin{array}{l}\mu - 0.46\sigma = 22\\ \mu + 0.65\sigma = 30\end{array}\right\} \Rightarrow \mu = 25.3153,\ \sigma = 7.2072$

2-b

$P(\xi > a) = 0.20 \Rightarrow P\left(z > \frac{a-25.3153}{7.2072}\right) = 0.2 \Rightarrow \frac{a-25.3153}{7.2072} \approx 0.842 \Rightarrow a = 25.3153 + 0.842(7.2072) = 31.37$

$P(\xi < b) = 0.40 \Rightarrow P\left(z < \frac{b-25.3153}{7.2072}\right) = 0.4 \Rightarrow -\frac{b-25.3153}{7.2072} \approx 0.253 \Rightarrow b = 25.3153 - 0.253(7.2072) = 23.48$

Problema 3

Debe verificarse que $F(3) = 1$, por tanto: $a(9 - 4.5 - 4) + 2/3 = 1 \Rightarrow a = \frac{2}{3}$

3-b: $P(\xi > 2.7) = 1 - F(2.7) = 1 - \left[\frac{2}{3}(3(2.7) - \frac{2.7^2}{2} - 4) + \frac{2}{3}\right] = 0.03$

3-c: Ahora $\eta \to B(100, 0.03) \to P(3)$ (Binomial que se aproxima por una Poisson).

Se pide $P(\eta \geq 2) = 1 - P(\eta < 2) = 1 - P(\eta = 0) - P(\eta = 1) = 1 - e^{-3}\frac{3^0}{0!} - e^{-3}\frac{3^1}{1!} = 1 - 4e^{-3} = 0.8009$

Problema 4:

Sabemos que el parámetro λ de la exponencial es el inverso de la esperanza matemática. Calculamos la media de los datos:

x_i	$[0,10]$	$(10,14]$	$[14,18]$	$[18,24)$	Más de 24	
n_i	1508	831	377	202	82	3000
x_i'	5	12	16	21	27	
$n_i x_i'$	7540	9972	6032	4242	2214	30000
p_i	0.6321	0.1213	0.0813	0.0746	0.0907	
$e_i = np_i$	1896.3	363.9	243.9	223.8	272.1	
n_i^2	2274064	690561	142129	40804	6724	
$\frac{n_i^2}{e^i}$	1199.2	1897.7	582.7	182.3	24.7	3886.6

Luego la media estimada es $\bar{x} = \frac{30000}{3000} = 10 \Rightarrow \lambda = 0.1$

La función de distribución de la exponencial es $F(x) = 1 - e^{-\lambda x} = 1 - e^{-0.1x}$

Las probabilidades de cada clase, supuesta que la función de probabilidad sea E(10), se calculan mediante la función de distribución:

$P(\xi \in [0,10]) = F(10) = 1 - e^{-0.1(10)} = 0.6321$ $P(\xi \in (10,14) = F(14) - F(12) = 0.1213, \ldots$

Las frecuencias esperadas son $np_i = 3000p_i$

Ahora calculamos el estadístico (observadas $o_i = n_i$) $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} = \sum_i \frac{o_i^2}{e^i} - n = 886.6$

Buscamos en tablas $\chi^2_{0.05,3} = 9.348$. Como el valor experimental es mayor rechazamos la hipótesis nula y **la exponencial no es un buen ajuste a los datos**.

Problema 5:

Con los datos calculamos $\bar{x}_A = \frac{750}{75} = 10$ y $V_A = \frac{20836}{75} - 10^2 = 177.8133 \Rightarrow s_A^2 = \frac{75}{74}177.8133 = 180.2162$

$\bar{x}_D = \frac{750}{125} = 6$ y $V_D = \frac{6597}{125} - 6^2 = 16.776 \Rightarrow s_D^2 = \frac{125}{124}16.776 = 16.9113$

5-a: Se trata de un contraste unilateral de la media, varianzas desconocidas y muestras grandes.

$H_0 : \mu_A \leq \mu_D$

$H_a : \mu_A > \mu_D$ La región crítica es: $E = \frac{\bar{x}_A - \bar{x}_D}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_D^2}{n_D}}} > z_\alpha \Rightarrow$

$z_{0.05} = 1.645$ y $E = \frac{10-6}{\sqrt{\frac{180.2162}{75} + \frac{16.9113}{125}}} = 2.5107$

Como $E > 1.645$ se rechaza la hipótesis nula y se acepta la alternativa, por lo que se han reducido los desfases.

5-b: Ahora se trata de un contraste unilateral de la igualdad de varianzas.

$H_0 : \sigma_A^2 \leq \sigma_D^2$

$H_0 : \sigma_A^2 > \sigma_D^2$. La región crítica es $\frac{s_A^2}{s_D^2} > F_{\alpha;n_A-1;n_D-1} = F_{0.05;74;124}$

$\frac{s_A^2}{s_D^2} = \frac{180.2162}{16.9113} = 10.6566$, mientras que buscamos en las tablas de la F de Fisher-Snedecor: $F_{0.05;74;124}$ y vemos: $F_{0.05;60;120} = 1.429$, $F_{0.05;60;10000} = 1.318$, $F_{0.05;120;120} = 1.352$, $F_{0.05;120;10000} = 1.222$.

El valor buscado resultaría de interpolar entre esos 4 valores, lo que podemos evitar pues resulta evidente que su valor es inferior a 1.429 y por tanto se verifica que $\frac{s_A^2}{s_D^2} = 10.6566$ es mayor que $F_{0.05;74;124}$ y por tanto se rechaza la hipótesis nula y se acepta la alternativa, es decir, existen menos desviaciones en los desfases tras el curso.

NOTA: Para interpolar el valor pedido tendríamos que hacerlo en 3 fases:

1) Calculamos $F_{0.05,74,120} \approx 1.429 + \frac{74-60}{120-60}(1.352 - 1.429) = 1.411$

2) Calculamos $F_{0.05,74,10000} \approx 1.318 + \frac{74-60}{120-60}(1.222 - 1.318) = 1.2956$

3) Con estos 2 interpolamos de nuevo: $F_{0.05;74;124} \approx 1.411 + \frac{124-120}{10000-120}(1.2956 - 1.411) = 1.411$

Matlab obtiene $F_{0.05;74;124} = 1.3978$, que resulta ser mucho más preciso, mediante **finv(0.95,74,124)**.

Soluciones MATLAB:

```
disp('Problema 6')
x=[-2.5 -2.5 2.5 2.5 2.5 2.5 7.5 7.5 7.5 7.5 12.5 12.5]
y=[ 2    3   1   2   3   4   0   1   2   4    0   1]
n=[25   25  25  20   5  20  15  15   5  12   42   2]
N=sum(n)
B=[sum(y.*n);sum(y.*x.*n);sum(y.*x.^2.*n)]
A=[  N          sum(x.*n)    sum(x.^2.*n);
  sum(x.*n)    sum(x.^2.*n) sum(x.^3.*n);
  sum(x.^2.*n) sum(x.^3.*n) sum(x.^4.*n)]
sol=A\B
a=sol(1),b=sol(2),c=sol(3)
% La solución es y=a+bx+cx^2
yest=a+b*x+c*x.^2
% Tambien vale p=[c b a], yest=polyval(p,x)
r=y-yest
r2=r.^2
Vr=sum(r2.*n)/N-(sum(r.*n)/N)^2
Vy=sum(y.^2.*n)/N-(sum(y.*n)/N)^2
R2=1-Vr/Vy


disp('Problema 7')
NIT=10000
cont=0;
for k=1:NIT
    x=2*randn(1,10); %Genero 10 normales N(0,2)
    xi=sum(abs(x)); %Sumo sus valores absolutos
    if xi>30, cont=cont+1;end %Si verifica la condición incremento contador
end
p=cont/NIT
I=[p-1.96*sqrt(p*(1-p)/NIT),p+1.96*sqrt(p*(1-p)/NIT)]

disp('Problema 8')
M1=[12 10 9 6 15 12 18 9 8 13]
M2=[13 10 8 8 18 11 19 11 9 15]
[H1,P1]=ttestm(M1,M2,0.03,'left')
[H2,P2]=ttestm(M2,12.5,0.03)
```