Grados en Informática
Métodos Estadísticos Examen Septiembre 2018

- **Tiempo: 2 horas 30 minutos**.
- Dejar DNI encima de la mesa. **Apagar y guardar el MÓVIL**.
- **El alumno debe realizar todos los ejercicios.**

APELLIDOS, NOMBRE:
DNI: Grupo: Titulación:

---

1. La siguiente tabla muestra el ruido introducido (Y en dB) en una comunicación sin cable, respecto a la distancia (X en Km.) entre emisor y receptor. Hallar:

x_i	1	2	2	4	4	4	5	5
y_i	1	1	2	2	4	5	4	5
n_i	3	4	1	2	4	3	4	1

   a) Ajustar una curva de la forma $y = \frac{1}{a+\frac{b}{x}}$ y estudiar la bondad del ajuste.

   b) Indicar si este ajuste es mejor que el lineal.

   c) Hallar la mediana de $Y/_{X<4}$ *(1+0.75+0.75=2.5 Puntos)*

2. La creación de un nuevo software, se compone de varias fases (desarrollo es cascada): 1) Especificación, 2) Diseño, 3) Implementación, 4) Verificación y 5) Mantenimiento. Cada fase solo puede iniciarse a la conclusión de la anterior. Si prescindimos de la última fase (mantenimiento), el tiempo total T (en días) desde que se toma la decisión de crear el nuevo software hasta su puesta en el mercado puede descomponerse en $T = t_1 + t_2 + t_3 + t_4$, donde se estima que: $t_1 \rightsquigarrow N(4, 1)$, $t_2 \rightsquigarrow N(10, 2)$, $t_4 \rightsquigarrow N(15, 2)$. La fase 3 será realizada por 3 programadores A, B y C, por lo que estará terminada cuando termine el último de ellos, así: $t_3 = \max\{t_{3a}, t_{3b}, t_{3c}\}$, donde: $t_{3a} \rightsquigarrow N(55, 5)$, $t_{3b} \rightsquigarrow N(60, 4)$ y $t_{3c} \rightsquigarrow N(60, 3)$. Se supone que los tiempos empleados en cada fase son independientes. Hallar:

   a) La distribución que sigue $\xi = t_1+t_2+t_4$ y la probabilidad de que tarde menos de 30 días ($P(\xi < 30)$).

   b) El tiempo T_0 tal que $P(\xi < T_0) = 0.95$

   c) Probabilidad de que t_3 dure menos de 70 días (los 3 programadores hayan concluido sus partes respectivas).

   d) Probabilidad de que tras 70 días con la tarea 3, tan solo 1 programador haya terminado su tarea (quedando inconclusas 2 de ellas). *(0.5+0.5+0.5+0.5=2 Puntos)*

3. Un servicio técnico debe atender las averías producidas en un modelo de lavadora. Sospecha que una deficiente puesta en marcha es el motivo del alto porcentaje de averías. Solo existen 2 empresas A y B que realicen esa puesta en marcha. Durante el año 2017 se atendieron 260 averías de este tipo, que corresponden al 5.2 % de las instaladas en total. Este tipo de avería representa el 10 % de las instaladas por A y el 2 % de las instaladas por B. Hallar:

   a) Porcentaje de lavadoras que instala A.

   b) ¿Se observan diferencias significativas en los porcentajes de averías para A y B?. Usar $\alpha = 0.01$. *(0.75+0.75=1.5 Puntos)*

4. Se desea observar la influencia de un programa de entrenamiento en el tiempo de realización de una tarea, para ello, se toma una muestra de 30 individuos y se mide el tiempo empleado antes del programa, resultando una media de 5.25 minutos y $s_1 = 1.88$ minutos, mientras que una muestra de 35 individuos, tras el entrenamiento, muestran una media de 2.37 minutos y $s_2 = 1.45$ minutos. Contrastar al 5 % las afirmaciones:

   a) Los individuos tras el programa de entrenamiento tardan menos tiempo.

   b) La realización del programa reduce el tiempo medio en 3 minutos. *(0.75+0.75=1.5 Puntos)*

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

5. **(E.D.) (MATLAB)** Dada la tabla bidimensional:

   a) Ajustar una curva de la forma $Y = a + b\sqrt{X} + cX$ a los datos de la tabla:

x_i	0	1	1	2	2	3	4
y_i	0	1	0	2	3	4	6
n_i	2	5	2	2	6	4	5

   b) Hallar el coeficiente de determinación del ajuste realizado.

(0.4+0.3=0.7 Puntos)

6. **(MATLAB)** Un mecanismo está formado por 3 componentes A, B y C, que funcionan de forma independiente. Los tiempos hasta que se produce una avería en un componente concreto, siguen exponenciales de medias 10, 6 y 7, respectivamente. El mecanismo funciona (F) si lo hacen al menos 2 de ellos. ($F = (A \wedge B) \vee (A \wedge C) \vee (B \wedge C)$).

   Calcular mediante simulación con 10000 iteraciones

   a) Media y varianza del tiempo hasta el fallo de alguno de los 3 componentes.

   b) La probabilidad de que el mecanismo funcione tras 10 años.

(0.4+0.4=0.8 Puntos)

7. El tiempo (en décimas de segundo) empleado en el repostaje por 2 equipos diferentes A y B de Formula 1 durante el año 2018 ha sido:

   Secuencia A: $t_A = \{114, 123, 94, 116, 112, 101, 107, 112, 135, 129\}$

   Secuencia B: $t_B = \{83, 154, 117, 104, 116, 102, 103, 129\}$

   a) ¿Puede concluirse, al 1 % de significación, que el equipo B realiza el repostaje en menos tiempo que el A?

   b) Si el tiempo medio para todos los equipos es de 113 décimas de segundo. ¿Puede concluirse, al 1 %, que el tiempo empleado por el equipo A es diferente?

(0.5+0.5 Puntos)

SOLUCIONES:

Problema 1:

Para realizar el ajuste $y = \frac{1}{a+\frac{b}{x}}$, hacemos: $\frac{1}{y} = a + b\frac{1}{x}$. Haciendo el cambio: $X = \frac{1}{x}$, $Y = \frac{1}{y}$, queda: $Y = a + bX$, por lo que será ajustar una recta a X e Y.

x_i	y_i	n_i	X_i	Y_i	n_iX_i	$n_iX_i^2$	$n_iX_iY_i$	n_iY_i	n_iy_i	$n_iy_i^2$	y_i^{est}	r_i	n_ir_i	$n_ir_i^2$
1	1	3	1.00	1.00	3.00	3.0000	3.00	3.0	3	3	0.8617	0.1383	0.4148	0.0573
2	1	4	0.50	1.00	2.00	1.0000	2.00	4.0	4	4	1.6039	-0.6039	-2.4157	1.4589
2	2	1	0.50	0.50	0.50	0.2500	0.25	0.5	2	4	1.6039	0.3961	0.3961	0.1569
4	2	2	0.25	0.50	0.50	0.1250	0.25	1.0	4	8	2.8170	-0.8170	-1.6340	1.3350
4	4	4	0.25	0.25	1.00	0.2500	0.25	1.0	16	64	2.8170	1.1830	4.7320	5.5979
4	5	3	0.25	0.20	0.75	0.1875	0.15	0.6	15	75	2.8170	2.1830	6.5490	14.2964
5	4	4	0.20	0.25	0.80	0.1600	0.20	1.0	16	64	3.3191	0.6809	2.7238	1.8547
5	5	1	0.20	0.20	0.20	0.0400	0.04	0.2	5	25	3.3191	1.6809	1.6809	2.8256
		22			8.75	5.0125	6.14	11.3	65	247			12.4468	27.5827

Las ecuaciones normales serán:

$$\begin{pmatrix} 22 & 8.75 \\ 8.75 & 5.0125 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 11.3 \\ 6.14 \end{pmatrix} \Rightarrow a = 0.0865, b = 1.0739 \Rightarrow$$

$$Y = 0.0865 + 1.0739X \Rightarrow \mathbf{y} = \frac{\mathbf{1}}{\mathbf{0.0865} + \frac{\mathbf{1.0739}}{\mathbf{x}}}$$

Para la bondad del ajuste voy a calcular la varianza de y y la del los residuos: $V_y = \frac{247}{22} - \left(\frac{65}{22}\right)^2 \approx 2.4979 \Rightarrow \sigma_y \approx 1.5805$

$V_r = \frac{27.5827}{22} - \left(\frac{12.4468}{22}\right)^2 \approx 0.9337, \Rightarrow R^2 = 1 - \frac{0.9337}{2.4979} \approx 0.6262$

NOTA: SSE=27.5827

1-b: Para comparar con el lineal no es necesario realizar el ajuste lineal, basta con calcular el coeficiente de correlación lineal (r).

$r = \frac{cov}{\sigma_x \sigma_y}$

Ya hemos calculado $\sigma_y = 1.5805$, calculemos lo restante:

x_i	y_i	n_i	n_ix_i	$n_ix_i^2$	$n_ix_iy_i$
1	1	3	3	3	3
2	1	4	8	16	8
2	2	1	2	4	4
4	2	2	8	32	16
4	4	4	16	64	64
4	5	3	12	48	60
5	4	4	20	100	80
5	5	1	5	25	25
		22	74	292	260

Así, $V_x = \frac{292}{22} - \left(\frac{74}{22}\right)^2 \approx 1.9587 \Rightarrow \sigma_x \approx 1.3995$ y $Cov = \frac{260}{22} - \frac{74}{22}\frac{65}{22} \approx 1.8802$.

Resultando $r = \frac{cov}{\sigma_x \sigma_y} = \frac{1.8802}{1.3995(1.5805)} \approx 0.85. \Rightarrow r^2 \approx 0.7225$

Luego es mejor el ajuste lineal.

1-c: Filtrando los datos $X < 4$ y agrupando en Y, tenemos la tabla:

y_i	n_i
1	7
2	1
	8

Al ser $N = 8$ par, será la media entre el que ocupa el lugar 4º y 5º, pero en este caso ambos valen 1. Así pues, **Mediana=1**.

Problema 2:

La variable $\xi = t_1 + t_2 + t_4$ es la suma de 3 normales independientes, luego $\xi \rightsquigarrow N(4+10+15, \sqrt{1^2+2^2+2^2}) = N(29, 3)$.

Nos piden $P(\xi < 30) = P(z < \frac{30-29}{3}) = 0.6306$

2-b: Nos piden T_0, tal que: $P(\xi < T_0) = 0.95, \Rightarrow P(z < \frac{T_0-29}{3}) = 0.95, \Rightarrow P(z \geq \frac{T_0-29}{3}) = 0.05, \Rightarrow P\left(z \geq \frac{T_0-29}{3}\right) = 1.645, \rightarrow T_0 = 29 + 1.645(3) = 33.935$

2-c: Que los 3 programadores hayan terminado su parte de t_3 antes de 70 días (se está suponiendo independencia), será: $P(t_3 < 70) = P\left((t_{3a} < 70) \wedge (t_{3b} < 70) \wedge (t_{3c} < 70)\right) = P(\text{máx}\{t_{3a}, t_{3b}, t_{3c}\} < 70) =$

$$= P(t_{3a} < 70)P(t_{3b} < 70)P(t_{3c} < 70) = P\left(z < \frac{70-55}{5}\right) P\left(z < \frac{70-60}{4}\right) P\left(z < \frac{70-60}{3}\right) =$$

$$= P(z < 3)P(z < 2.5)P(z < 3.3333) \approx (0.9987)(0.9938)(0.9996) \approx \mathbf{0.9929}$$

2-d: Ahora nos piden:
$P = P(t_{3a} < 70)P(t_{3b} \geq 70)P(t_{3c} \geq 70) + P(t_{3a} \geq 70)P(t_{3b} < 70)P(t_{3c} \geq 70) +$
$+ P(t_{3a} \geq 70)P(t_{3b} \geq 70)P(t_{3c} < 70) \approx (0.9987)(0.0062)(0.0004) + (0.0013)(0.9938)(0.0004) +$
$+(0.0013)(0.0062)(0.9996) \approx \mathbf{0.000011615}$

Problema 3:

Llamemos p a la proporción instalada por A, $q = 1 - p$ será la proporción instalada por B.

Por el teorema de la probabilidad total $P(averia) = P(A) * P(averia/A) + P(B) * P(averia/B) \Rightarrow 0.52 = p * 0.10 + (1-p) + 0.02 \Rightarrow p = 0.4 \Rightarrow q = 0.6$

Luego A instala el 40 % y B el 60 %.

3-b:

El total de lavadoras instaladas, llamemosle N, debe verificar que su 5.2 % sea 260: $N * \frac{5.2}{100} = 260, \Rightarrow N = 5000$

De ellas A instala el 40 % es decir $N_A = 5000 * 0.4 = 2000$ y $N_B = 3000$

El apartado nos pide realizar un contraste de diferencia de proporciones:

$H_0 : p(averia/A) = p(averia/B) \qquad H_a : p(averia/A) \neq p(averia/B)$

Consultando las tablas, la región crítica es: $\frac{|p_1 - p_2|}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} > z_{\frac{\alpha}{2}}$, donde $z_{\frac{\alpha}{2}} = z_{0.005} = 2.5758$.

Además $P(averia/A) = 0.10$, $P(averia/B) = 0.02$, tenemos: $E = \frac{|0.10-0.02|}{\sqrt{\frac{0.1(0.9)}{2000} + \frac{0.02(0.98)}{3000}}} = 11.1441 > 2.5758$, por lo que estaremos en la región crítica y las proporciones son diferentes. (Hipótesis alternativa).

Problema 4:

Se trata de un contraste de diferencias de medias: $H_0 : \mu_1 \leq \mu_2 \qquad H_a : \mu_1 > \mu_2$

La región crítica es (contraste de muestras grandes y varianzas desconocidas): $E = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > z_\alpha$, donde $z_\alpha = z_{0.05} = 1.645$, mientras que $E = \frac{5.25-2.37}{\sqrt{\frac{1.88^2}{30} + \frac{1.45^2}{35}}} = 6.8285 > 1.645$, luego estamos en la región crítica (hipótesis alternativa), por lo que concluimos que **se reduce el tiempo de realización de la tarea**.

4-b:

Que se reduzca el tiempo medio en 3 minutos, significa debemos contrastar que $\mu_1 = \mu_2 + 3$.

La hipótesis nula será $H_0 : \mu_1 = \mu_2 + 3$ y la alternativa $\mu_1 \neq \mu_2 + 3$

La región crítica será $E = \frac{|\bar{x}_1 - (\bar{x}_2 + 3)|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > z_{\frac{\alpha}{2}}$, donde $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

$E = \frac{|5.25-(2.37+3)|}{\sqrt{\frac{1.88^2}{30} + \frac{1.45^2}{35}}} = 0.2845 \not> 1.96$ por lo que nos quedamos con la hipótesis nula y **aceptamos que se reduce en 3 minutos.**

Problema 5:

La base será $\mathcal{B} = \{1, \sqrt{x}, x\}$ por lo que las ecuaciones normales son:

$$\left.\begin{array}{rcllll} \sum_i n_i y_i & = & aN & +b\sum_i n_i\sqrt{x_i} & +c\sum_i n_i x_i \\ \sum_i n_i y_i\sqrt{x_i} & = & a\sum_i n_i\sqrt{x_i} & +b\sum_i n_i x_i & +c\sum_i n_i x_i\sqrt{x_i} \\ \sum_i n_i y_i x_i & = & a\sum_i n_i x_i & +b\sum_i n_i\sqrt{x_i} x_i & +c\sum_i n_i x_i^2 \end{array}\right\}$$

El programa en MATLAB será:

```
x=[0 1 1 2 2 3 4], y=[0 1 0 2 3 4 6], n=[2 5 2 2 6 4 5], N=sum(n)
A=[N sum(n.*sqrt(x)) sum(n.*x);
 sum(n.*sqrt(x)) sum(n.*x) sum(n.*x.*sqrt(x));
 sum(n.*x) sum(n.*x.*sqrt(x)) sum(n.*x.^2)]
B=[sum(n.*y);sum(n.*y.*sqrt(x));sum(n.*y.*x)]
sol=A\B, a=sol(1),b=sol(2),c=sol(3)
yest=a+b*sqrt(x)+c*x
res=y-yest
Vr=sum(n.*res.^2)/N-(sum(n.*res)/N)^2
Vy=sum(n.*y.^2)/N-(sum(n.*y)/N)^2
R2=1-Vr/Vy
```

Problema 6:

```
disp('Problema 6')
nit=10000
ta=exprnd(10,1,nit);
tb=exprnd(6,1,nit);
tc=exprnd(7,1,nit);
tm=min([ta;tb;tc]);
MED=mean(tm)
VAR=var(tm,1)
disp('6-b')
c1=(ta>10);
c2=(tb>10);
c3=(tc>10);
c=(c1&c2)|(c1&c3)|(c2&c3);
p=sum(c)/nit
disp('Otra forma')
cc=c1.*c2+c1.*c3+c2.*c3;
p2=sum(cc>0)/nit
disp('Teoría')
pa=1-expcdf(10,10),pb=1-expcdf(10,6),pc=1-expcdf(10,7)
prob=pa*pb+pc*(pa*(1-pb)+(1-pa)*pb)
```

Problema 7:

```
disp('Problema 7')
A=[114 123 94 116 112 101 107 112 135 129]
B=[83 154 117 104 116 102 103 129]
[H1,P1]=ttest2(A,B,0.01,'right')
[H2,P2]=ttest(A,113,0.01,'both')
```