Grados en Informática
Métodos Estadísticos Examen Junio 2017

- **Tiempo: 2 hora 15 minutos**.
- Dejar DNI encima de la mesa. **Apagar y guardar el MÓVIL**.

APELLIDOS, NOMBRE:
DNI: Grupo: Titulación:

---

1. **(ED)** Se ha medido el rendimiento medio (R en Tm/Ha) para la uva moscatel según la altitud de la plantación (H en Hm.). Se obtuvo la tabla:

$R \backslash H$	$[0,1]$	$(1,2]$	$(2,4]$	Más de 4 Hm.
$[0,2]$	0	0	2	2
$(2,6]$	0	2	2	0
$(6,10]$	2	2	0	0
$(10,\infty)$	3	1	0	0

Se pide:

(a) Ajustar la recta $R = a + bH$ de mínimos cuadrados y estudiar la bondad de dicho ajuste.

(b) Ajustar una función de la forma $R = \frac{c}{d+H^2}$ y analizar la bondad del ajuste.

(c) Estimar la producción media y su varianza para los cultivos situados por encima de 200m=2 Hm.

(0.75+1+0.5=2.25 Puntos)

2. El tiempo en años que tarda un componente en ser desechado sigue una variable aleatoria continua ξ, tiene por función de distribución:

$$F_\xi(x) = P(\xi \leq x) = \begin{cases} 0 & x < 1 \\ K(x-1)^2 & 1 \leq x \leq 12 \\ 1 & x > 12 \end{cases}$$

(a) Hallar el cuartil 1 (Q_1) y la media.

(b) Hallar la probabilidad de que un componente tarde menos de 10 años en desecharse sabiendo que ha durado más de 5 años.

(1+0.5=1.5 Puntos)

3. Una empresa conoce que el número de elementos averiados ξ en un lote de 500 productos de una marca determinada, sigue una distribución de Poisson. Tras un gran número de experiencias ha llegado a la conclusión de que la probabilidad de tener algún elemento defectuoso en el lote es: $P(\xi \geq 1) = 0.613259$. Se pide:

(a) Calcular las media y varianza de ξ.

(b) Calcular la probabilidad de encontrar más de 1 elemento defectuoso en el lote.

(c) Si compramos 50 lotes de 500. ¿Cuál será la probabilidad de encontrar más de 5 lotes conteniendo más de un elemento defectuoso.

(0.75+0.5+0.5=1.75 Puntos)

4. Se quiere realizar un estudio sobre la cantidad de sal común (NaCl) contenida en la paella servida en los restaurantes de la provincia, encontrándose que para una muestra de 10 elementos en restaurantes de Málaga capital (A) se obtiene un contenido medio de sal de 0.5 g. y una cuasivarianza (S^2) de 0.01 g. por cada 100 g. de paella. Por otra parte, para 16 restaurantes de la costa (B) se obtuvo una media de 0.45 g. y una cuasivarianza de 0.0064 g. por cada 100 g.

Además existe una recomendación que indica que el contenido medio no debe sobrepasar 0.38 g. por cada 100 g. Contrastar (suponiendo normalidad y con $\alpha = 0.10$):

(a) Que el contenido medio de sal en los restaurantes A es mayor de 0.38.

(b) Que el contenido medio de sal en los restaurantes A es mayor que en los B.

(0.75+1.25=2 Puntos)

5. **(ED) (Indicar las órdenes MATLAB necesarias)** Hallar el sesgo, la curtosis y el coeficiente de variación a los datos:

$$\{5.2, 6.7, 6.3, 7.0, 4.3, 5.7, 6.8, 7.1, 5.3, 6.4, 5.3, 4.9, 7.1\}$$

(0.75 Puntos)

6. **(Indicar las órdenes MATLAB necesarias)** En una dehesa se está estudiando la producción de corcho de determinados alcornoques de los que se conoce, por datos anteriores, que la producción (por árbol) seguía una Normal de media 1.1 quintales y varianza 4.

   (a) Suponiendo esos datos correctos. ¿Cuál es la probabilidad de elegir 5 árboles al azar y todos produzcan más de 1 quintal?

   (b) Queremos saber si el valor de la varianza sigue siendo fiable actualmente y si la producción ha aumentado; para ello tomamos una muestra de 26 árboles elegidos aleatoriamente que nos proporcionan los resultados siguientes:

   Producción={0, 0.0254, 1.6903, 3.7775, 0.6115, 2.1252, 1.2810, 3.1802, 0.0598, 1.6460, 2.3814, 3.1565, 3.7838, 1.7215, 0, 0.5502, 0.0987, 4.9240, 0.7294, 2.7579, 1.3279, 2.8567, 0.5183, 0, 0, 2.2904}

   Obtener conclusiones al 90%.

   NOTA: Suponer que el valor 4 de la varianza ya no resulta fiable tras el primer contraste.

   (0.5+(0.5+0.75)=1.75 Puntos)

SOLUCIONES:

Problema 1:

1-a: Llamando $y = R$ y $x = H$, quedando la recta como $y = a + bx$. Vamos a hallarla por la fórmula: $y - \bar{y} = \frac{cov}{V_x}(x - \bar{x})$, donde:

$N = 16$, $\bar{x} = \frac{5(0.5)+5(1.5)+4(3)+2(5)}{16} = 2$, $m_{2,0} = \frac{5(0.5^2)+5(1.5^2)+4(3^2)+2(5^2)}{16} = 6.1563$

$V_x = 6.1563 - 2^2 = 2.1563$, $\sigma_x = 1.4684$

$\bar{y} = \frac{4(1)+4(4)+4(8)+4(12)}{16} = 6.25$, $m_{0,2} = \frac{4(1^2)+4(4^2)+4(8^2)+4(12^2)}{16} = 56.25$, $V_y = 56.25 - 6.25^2 = 17.1875$, $\sigma_y = 4.1458$

$m_{1,1} = \frac{2(3)(1)+2(5)(1)+2(1.5)(4)+2(3)(4)+2(0.5)(8)+2(1.5)(8)+3(0.5)(12)+1(1.5)(12)}{16} = 7.5$, $Cov = m_{1,1} - \bar{x}\bar{y} = 7.5 - 2(6.25) = -5$

la pendiente $m_{Y/X} = \frac{Cov}{V_x} = \frac{-5}{2.1563} = -2.3188$ quedando la recta: $y - 6.25 = -2.3188(x - 2) \Rightarrow$
$R = -2.3188H + 10.8877$

NOTA: Podría haberse obtenido mediante las ecuaciones normales: $16a + 32b = 100$ y $32a + 98.5b = 120$.

Por otra parte $r = \frac{Cov}{\sigma_x \sigma_y} = \frac{-5}{(1.4684)(4.1458)} = -0.8213$, $r^2 = 0.6746$ y $Vr = V_y(1-r^2) = 17.1875(1-0.6746) = 5.5933$ cualquiera de ellas es una medida de la bondad del ajuste realizado.

1-b: Linealizamos $R = \frac{c}{d+H^2}$ mediante: $d + H^2 = \frac{c}{R} \Rightarrow \frac{1}{R} = \frac{d}{c} + \frac{1}{c}H^2$.

Hacemos el cambio $X = H^2$, $Y = \frac{1}{R}$ quedando: $Y = A + BX$ con $A = \frac{d}{c}$ y $B = \frac{1}{c}$

h_i	r_i	n_i	X_i	Y_i	n_iX_i	$n_iX_i^2$	n_iY_i	$n_iX_iY_i$
3.0	1	2	9.00	1	18.00	162.0000	2.0000	18.0000
5.0	1	2	25.00	1	50.00	1250.0000	2.0000	50.0000
1.5	4	2	2.25	0.25	4.50	10.1250	0.5000	1.1250
3.0	4	2	9.00	0.25	18.00	162.0000	0.5000	4.5000
0.5	8	2	0.25	0.125	0.50	0.1250	0.2500	0.0625
1.5	8	2	2.25	0.125	4.50	10.1250	0.2500	0.5625
0.5	12	3	0.25	0.0833	0.75	0.1875	0.2500	0.0625
1.5	12	1	2.25	0.0833	2.25	5.0625	0.0833	0.1875
		16			98.50	1599.625	5.8333	75.5000

Resolvemos el sistema:

$$\begin{pmatrix} N & \sum_i n_iX_i \\ \sum_i n_iX_i & \sum_i n_iX_i^2 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum_i n_iY_i \\ \sum_i n_iX_iY_i \end{pmatrix} \Rightarrow \begin{pmatrix} 16 & 98.5 \\ 98.5 & 1599.625 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 5.8333 \\ 74.5 \end{pmatrix} \Rightarrow$$

$$A \approx 0.1254, B = 0.03885 \Rightarrow c = \frac{1}{B} \approx 25.7391 \text{ y } d = cA \approx 3.2278$$

La función ajustada queda:

$$\mathbf{R} = \frac{\mathbf{25.7391}}{\mathbf{3.2278 + H^2}}$$

Para la bondad del ajuste calcularemos el coeficiente R^2:

Calcularemos las columnas $r_i^{est} = \frac{25.7391}{3.2278+H^2}$, residuos $e_i = r_i - r_i^{est}$:

h_i	r_i	n_i	r_i^{est}	e_i	n_ie_i	$n_ie_i^2$
3.0	1	2	2.104966	−1.104966	−2.209933	2.441901
5.0	1	2	0.911835	0.088165	0.176329	0.015546
1.5	4	2	4.698805	−0.698805	−1.397610	0.976657
3.0	4	2	2.104966	1.895034	3.790067	7.182306
0.5	8	2	7.400978	0.599022	1.198044	0.717655
1.5	8	2	4.698805	3.301195	6.602390	21.795777
0.5	12	3	7.400978	4.599022	13.797066	63.453010
1.5	12	1	4.698805	7.301195	7.301195	53.307449
		16			29.257549	149.890300

Resultando $V_r = \frac{149.8903}{16} - \left(\frac{29.257549}{16}\right)^2 \approx 5.593297$ y $R^2 = 1 - \frac{V_r}{V_y} = 1 - \frac{5.593297}{17.1875} \approx 0.649491$

1-c: Filtramos aquellos cultivos que cumplen la condición de estar situados por encima de los 200 m., quedando la tabla:

R_i	n_i	r_i	n_ir_i	$n_ir_i^2$
$[0, 2]$	4	1	4	4
$(2, 6]$	2	4	8	32
	6		12	36

siendo r_i la marca de clase del intervalo.

Así, $\bar{R}/_{H>2} = \frac{12}{6} = 2$ y $V_R/_{H>2} = \frac{36}{6} - 2^2 = 2$

Problema 2:

2-a: Lo primero será determinar la constante k. Al tratarse de una distribución continua la F(x) deberá serlo también y por tanto $F(12) = 1 \Rightarrow k((12-1)^2 = 1 \Rightarrow k = \frac{1}{121}$

El cuartil 1 verifica $F(Q_1) = 0.25 \Rightarrow \frac{1}{121}(Q_1 - 1)^2 = 0.25 \Rightarrow Q_1 = 6.5$

Para calcular la media debemos hallar la función de densidad, así derivamos y $f(x) = \frac{2}{121}(x-1)$ si $x \in [1, 12]$ y $f(x) = 0$ en otro caso.

$\mu = E(x) = \int_1^{12} x \frac{2}{121}(x-1)dx = \frac{2}{121}\left[\frac{x^3}{3} - \frac{x^2}{2}\right]_1^{12} = \frac{25}{3}$

2-b:

Se pide $P(\xi < 10/\xi > 5) = \frac{P(5<\xi<10)}{P(\xi>5)} = \frac{F(10)-F(5)}{1-F(5)} = \frac{81/121-16/121}{1-16/121} = \frac{13}{21}$, ya que $F(5) = \frac{1}{121}(5-1)^2 = \frac{16}{121}$ y $F(10) = \frac{1}{121}(10-1)^2 = \frac{81}{121}$

Problema 3:

3-a:

Partimos de que $P(\xi \geq 1) = 0.613259 \Rightarrow P(\xi = 0) = 1 - 0.613259 = 0.386741 = e^{-\lambda}\frac{\lambda^0}{0!} \Rightarrow e^{-\lambda} = 0.386741 \Rightarrow \lambda = -Ln(0.386741) \approx 0.95$

Y como en la distribución de Poisson la media $\mu = \lambda = 0.95$ y la varianza también es λ, pues $V = 0.95$

3-b:

$P(\xi > 1) = 1 - P(\xi = 0) - P(\xi = 1) = 1 - e^{-0.95} - e^{-0.95}\frac{0.95^1}{1!} = 1 - e^{-0.95}(1.95) \approx 0.245855$

3-c:

Llamamos ν al n^o. de lotes conteniendo más de 1 defectuoso, la variable ν seguirá una Binomial de parámetros n=50 y p=0.245855. ($\nu \to B(50, 0.245855)$).

Nos piden $P(\nu > 5)$, pero al ser $n > 30$ deberemos aproximar. Tenemos $np = 50(0.245855) \approx 12.29275 > 5$ ($nq = 50(1 - 0.245855) > 5$), por lo que lo haremos mediante una normal de media $\mu = 12.29275$ y desviación típica $\sigma = \sqrt{50(0.245855)(1-0.245855)} \approx 3.04475$

Entonces (y empleando la corrección de continuidad) $P(\nu > 5) \approx P(\nu' > 5.5) = P(z > \frac{5.5-12.29275}{3.04475}) = P(z > -2.2309751) = 1 - P(z > 2.2309751) \approx 0.987158$

Problema 4:

4-a:

Nos piden el contraste:

$H_0 : \bar{x}_A \leq 3.8$ y $H_a : \bar{x}_A > 3.8$ que se trata de un contraste unilateral de la media, varianza desconocida y muestras pequeñas. Usaremos el estadístico $E = \frac{\bar{x}_A - \mu_0}{s/\sqrt{n}}$ y la región crítica será $E > t_\alpha, n-1 = t_{0.10,9} = 1.383$

Como $E = \frac{0.5-0.38}{0.1/\sqrt{10}} \approx 3.7994733 > 1.383$ rechazamos la hipótesis nula y concluimos que contiene de media mayor cantidad de sal que la recomendada.

4-b: Ahora debemos contrastar $H_0 : \mu_A \leq \mu_B$ contra $H_a : \mu_A > \mu_B$

Nos vamos a las tablas y debemos discernir entre los casos de varianzas iguales y varianzas distintas de un contraste unilateral de diferencias de median varianzas desconocidas y muestras pequeñas, por lo que debemos realizar un contraste de igualdad de varianzas.

CONTRASTE PREVIO:

$H_0 : \sigma_A^2 = \sigma_B^2$

$H_a : \sigma_A^2 \neq \sigma_B^2$ para el que la región crítica es: $\frac{s_A^2}{s_B^2} \notin [F_{1-\frac{\alpha}{2},n_A-1,n_B-1}; F_{\frac{\alpha}{2},n_A-1,n_B-1}]$

En nuestro caso $\alpha = 0.1$, $n_A = 10$, $n_B = 16$, luego el intervalo es: $F_{0.95,9,15}, F_{0.05,9,15}] = [\frac{1}{F_{0.05}15,9}; F_{0.05,9,15}] = [\frac{1}{3.006}; 2.588] = [0.3327; 2.588]$.

Mientras que $E = \frac{s_A^2}{s_B^2} = \frac{0.01}{0.0064} = 1.5625 \in [0.3864; 3.006]$ y por tanto nos quedamos con la hipótesis nula de igualdad de varianzas.

Regresamos al contraste pedido para varianzas iguales:

$H_0 : \mu_A \leq \mu_B$

$H_a : \mu_A > \mu_B$ para el que la región crítica es: $E = \frac{\bar{x}_A - \bar{x}_B}{s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} > t_{\alpha,n_A+n_B-2}$

En nuestro caso $t_{\alpha,n_A+n_B-2} = t_{0.01,24} = 1.318$

Calculamos $s_p^2 = \frac{9(0.01)+15(0.0064)}{24} = 0.00775$ y $s_p = \sqrt{0.00775} \approx 0.088$, por lo que $E = \frac{0.5-0.45}{0.088\sqrt{\frac{1}{10}+\frac{1}{16}}} \approx 1.4089 > 1.318$. Por tanto, rechazaremos la hipótesis nula, concluyendo que el contenido medio de sal en los restaurantes tipo A es mayor que en los del tipo B.

Problema 5:

```
format compact,clc,clear all
disp('Problema 5')
x=[5.2, 6.7, 6.3, 7.0, 4.3, 5.7, 6.8, 7.1, 5.3, 6.4, 5.3, 4.9, 7.1]
N=length(x)
m1=sum(x)/N,m2=sum(x.^2)/N,m3=sum(x.^3)/N,m4=sum(x.^4)/N
V=m2-m1^2,s=sqrt(V)
mu3=m3-3*m2*m1+2*m1^3
mu4=m4-4*m3*m1+6*m2*m1^2-3*m1^4
sesgo=mu3/s^3
curtosis=mu4/s^4-3
```

Problema 6:

```
pm1=1-normcdf(1,1.1,2) % Probabilidad de que un árbol produzca mas de 1 quintal.
Pa=pm1^5 % Probabilidad de que 5 árboles independiente produzcan mas de 1 quintal cada uno.
disp('6-b:')
X=[0, 0.0254, 1.6903, 3.7775, 0.6115, 2.1252, 1.2810, 3.1802, 0.0598,...
1.6460, 2.3814, 3.1565,   3.7838, 1.7215, 0, 0.5502, 0.0987, 4.9240,...
0.7294, 2.7579, 1.3279, 2.8567, 0.5183, 0, 0, 2.2904]
[H,P]=vartest(X,4,0.1,'both')
[H2,P2]=ttest(X,1.1,0.1,'right')
```

Comentario: En la nota nos dicen que el valor 4 de la varianza no resulta fiable, es decir, en el contraste bilateral (b1) de la varianza, nos ha salido la hipótesis alternativa, eso nos hace que en el (b2) contrastemos la media sin conocer el valor de la varianza (ttest) en lugar de con varianza conocida (ztest).