Grados en Ingeniería Informática

Estadística

Examen **Convocatoria Septiembre 2017**

- A resolver en **2 horas**
- Dejar DNI encima de la mesa.
- **Apagar y guardar el MÓVIL**.

APELLIDOS, NOMBRE:
DNI: Especialidad: Grupo:

---

1. Un productor de patatas desea comprobar cómo influye la cantidad suministrada (N) de un nuevo abonado en $g. \times$ planta, con el tamaño/categoría de las mismas (T) medido como unidades/15 Kg. (unidades por saco de 15 Kg.). En el estudio se obtuvo la siguiente tabla de frecuencias:

$N\backslash T$	60 ó menos	$(60, 90]$	$(90, 110]$	más de 110
24	0	1	2	4
30	1	4	6	4
36	4	2	2	0

Calcular:

(a) Calcular la moda y la mediana de la variable T=tamaño del fruto.

(b) Calcular la dispersión relativa (no dependiente de la unidad de medida) de la variable $N/_{T\in(90,110]}$.

(c) Hallar el coeficiente de correlación lineal e interpretarlo.

(1+0.5+0.5= 2 Puntos)

2. La duración de un emisor de televisión sigue una distribución exponencial ξ de media $\mu = 10$ en años. Se lanzan al espacio 3 emisores que se supone van a funcionar de forma independiente. Hallar:

(a) Un intervalo de confianza del 95% para la duración de cada emisor. Es decir, el intervalo I de extremos los cuantiles 0.025 y 0.975.

(b) El tiempo T para el cual $P(\max\{\xi_i\} \leq T) = 0.05$, es decir, exista un 95% de probabilidad de que se esté recibiendo alguna señal (alguno funcione) después del instante T.

(c) Probabilidad de que tras 12 años, se hayan estropeado exactamente 2 emisores.

(d) Probabilidad de que si se estropea un emisor en el lanzamiento, alguno de los otros siga funcionando tras 12 años.

(e) Se tiene la sospecha de que en lugar de seguir ξ una exponencial, sigue una normal de media $\mu = 10$ y $\sigma = 2$. ¿Cuál es ahora la probabilidad de que tras 12 años se hayan estropeado exactamente 2 emisores?

(0.75+0.75+0.5+0.5+0.75= 3.25 Puntos)

3. Se decide realizar un estudio en una plantación de fresas variedad A, para analizar si el cultivo de otra variedad B puede resultar más rentable. Se toman 46 parcelas de igual superficie y se plantan 26 con la variedad A y las 20 de restantes con la B. Tomando como unidad monetaria "mil euros", se obtuvo un beneficio neto medio por parcela de 5.7 y de 5.9 respectivamente, mientras que $\sum_{i\in A} x_i^2 = 869.74$ y $\sum_{i\in B} x_i^2 = 734.2$, donde x_i representa el beneficio neto de da cada parcela.

(a) Encontrar un intervalo de confianza para el beneficio neto por parcela con la variedad B, al nivel del 90%.

(b) Saber si se puede concluir que el beneficio neto medio por parcela plantada con la variedad B es estrictamente superior al plantado con la variedad A, al nivel del 90%.

(c) ¿Se puede afirmar, al 95%, que la variedad B produce resultados con mayor dispersión que los de la variedad A?

(0.75+0.75+0.75 = 2.25 Puntos)

4. **PRÁCTICAS (Sólo instrucciones)**

La cantidad de basura C recogida por el barco dedicado a la limpieza de playas en determinado municipio (por año y en Tm.), resultó ser:

T en años	2008	2009	2010	2012	2013	2014	2015	2016
C en Tm.	17	23	23	28	30	35	37	44

Se pide:

(a) Ajustar la recta de tendencia C/T y el coeficiente de correlación lineal.

(b) Estimar, mediante el modelo, la cantidad que se prevé recoger durante el año $t = 2017$.

(0.5+0.2=0.7 Puntos)

5. **PRÁCTICAS (Sólo instrucciones)**

La demora D en la llegada de un metro desde la parada X (con hora de paso a las 8:00) a TEATINOS (llegada a 8:00+D), se compone del tiempo de retraso (ξ_1) en llegar el tren a la parada X, más el que invierte en el recorrido ξ_2. ($D = \xi_1 + \xi_2$)

La variable $\xi_1 = \max\{0, \xi_0\}$ donde ξ_0 sigue una normal N(2,1.5), (en minutos) esto es, si el metro llega adelantado a su hora de salida se espera para salir puntual, mientras que ξ_2 sigue una normal de media 21 minutos y 3 de desviación.

(a) Estimar mediante simulación (50000 iteraciones) la probabilidad de llegar pasadas las 8:30.

(b) Queremos contrastar que la duración del trayecto ξ_2 del metro es en realidad superior a 21 minutos de media, (sin fiarnos tampoco de que la desviación sea 3) para ello medimos durante 14 días el tiempo del trayecto obteniendo los tiempos 22, 21, 20, 24, 27, 25, 25, 26, 21, 20, 27, 26, 21, 25. Hacer el contraste al nivel del 5%.

(c) Contrastar con esos datos si la desviación típica es 3.

(0.8+0.5+0.5=1.8 Puntos)

Problema 1:
a:

T_i	n_i	h_i	a_i	N_i
60 o menos	5	30	0.1667	5
$(60, 90]$	7	30	0.2333	12
$(90, 110]$	10	20	0.5000	22
Mas de 110	8	20	0.4000	30
	30			

Moda: $\Delta_1 = 0.5 - 0.2333 = 0.2667$, $\Delta_2 = 0.5 - 0.4 = 0.1 \Rightarrow M_o = 90 + \frac{0.2667}{0.2667+0.1}20 = 104.5455$
Mediana: N=30 $\Rightarrow \frac{N}{2} = 15$ El intervalo mediano es el (90,110] y la mediana es:

$$M_e = L_i + \frac{\frac{N}{2} - N_{i-1}}{n_i}a_i = 90 + \frac{15-12}{10}20 = 96$$

1-b: La medida de dispersión que no es afectada por la unidad de medida es el coeficiente de variación. Nos fijamos solo en la columna (90,110]

x_i	$x'_i = \frac{x_i-30}{6}$	n_i	$n_i x'_i$	$n_i {x'}_i^2$
24	-1	2	-2	2
30	0	6	0	0
36	1	2	2	2
		10	0	4

$V' = 4/10 - 0^2 = 0.4 \Rightarrow \sigma' = \sqrt{0.4} \Rightarrow \sigma = 6\sqrt{0.4}$ $\bar{x'} = 0 \Rightarrow \bar{x} = 30 + 6(0) = 30$
Por tanto, la medida de dispersión pedida vale: $CV = \frac{6\sqrt{0.4}}{30} \approx 0.1265$

1-c:
Calculamos la media y varianza de la N:
$\bar{N} = \frac{24*7+30*15+36*8}{30} = 30.2$, $m_2(N) = \frac{24^2*7+30^2*15+36^2*8}{30} = 930, \Rightarrow V_N = 930 - 30.2^2 = 17.96$
$\bar{T} = \frac{45*5+75*7+100*10+120*8}{30} = 90.3333$, $m_2(T) = \frac{45^2*5+75^2*7+100^2*10+120^2*8}{30} = 8823.3333,$
$\Rightarrow V_T = 8823.3333 - (90.3333)^2 \approx 663.2222$
$m_{1,1} = \frac{45*(30+36*4)+75*(24+30*4+36*2)+100*(24*2+30*6+36*2)+120*(24*4+30*4)}{30} = 2665$
$\Rightarrow$ Cov $= 2665 - (30.2)(90.3333) = -63.0667$
y el coeficiente de correlación lineal queda: $r = \frac{-63.0667}{\sqrt{663.2222*17.96}} \approx -0.5779$
Luego existe una correlación lineal inversa (no muy fuerte) entre el nuevo abonado y el número de unidades por saco, es decir, aumenta el tamaño del fruto al abonar.

Problema 2:
a: Si la media es 10, entonces $\lambda = \frac{1}{10} = 0.1$ y la función de distribución es $F(x) = 1 - e^{-0.1x}$
El cuantil 0.025 (a) verifica: $P(\xi \leq a) = F(a) = 0.025 \Rightarrow 1 - e^{-0.1a} = 0.025 \Rightarrow a = 0.2532$
El cuantil 0.975 (b) verifica: $P(\xi \leq b) = F(b) = 0.025 \Rightarrow 1 - e^{-0.1b} = 0.975 \Rightarrow b = 36.8888$
luego el intervalo es $I = [a, b] = [0.2532, 36.8888]$

b: $P(\max \xi_i \leq T) = P((\xi_1 \leq T) \wedge (\xi_2 \leq T) \wedge (\xi_3 \leq T)) = \{\text{independencia}\} =$
$= P((\xi_1 \leq T) \cdot P(\xi_2 \leq T) \cdot P(\xi_3 \leq T)) = (1 - e^{-0.1T})^3$ que lo igualamos a 0.05:
$(1 - e^{-0.1T})^3 = 0.05 \Rightarrow 1 - e^{-0.1T} = \sqrt[3]{0.05} \approx 0.3684 \Rightarrow e^{-0.1T} = 0.6316 \Rightarrow 4.595$ años.

c: La probabilidad de que se estropee cada emisor en 12 años es: $F(12) = P(\xi_i \leq 12) = 1 - e^{-0.1(12)} = 1 - e^{-1.2} \approx 0.6988$
Por tanto: $P(E = 2) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} 0.6988^2 0.3012 \approx 0.4412$

d: Consideramos solo los dos que superan el lanzamiento.
$P(\text{funcione alguno}) = 1 - P(\text{fallen todos}) = 1 - F(10)^2 = 1 - (1 - e^{-0.1}(12))^2 = 1 - 0.6988^2 \approx 0.5117$

e: Ahora se trata de una normal N(10,2).
Calculamos la probabilidad de que uno de ellos falle antes o igual a 12 años: $P(\xi \leq 12) = P(z \leq \frac{12-10}{2}) = P(z \leq 1) = 0.8413$, por lo que:
$P(E = 2) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} 0.8413^2 0.1587 \approx 0.3370$

Problema 3:
a: Tenemos los datos $n_A = 26$, $\bar{x_A} = 5.7 \Rightarrow \sum_{i \in A} x_i = 5.7(26) = 148.2$

$\sum_{i\in A} x_i^2 = 869.74 \Rightarrow s_A^2 = \frac{26}{25}(\frac{869.74}{26} - 5.7^2) \approx 1$
$n_B = 20$, $\bar{x_B} = 5.9$, $s_B^2 = \frac{20}{19}\left(\frac{734.2}{20} - 5.9^2\right) \approx 2$
El intervalo pedido será: $I = [\bar{x}_B \pm t_{0.05,19}\frac{s_B}{\sqrt{20}}] = [5.3532, 6.4468]$, pues $t_{0.05,19} = 1.729$ y $s_B = \sqrt{2}$.

b: Se trata de un contraste para la diferencia de media, varianzas desconocidas y muestra grande.
$H_0 : \mu_B \leq \mu_A$ contra $H_a : \mu_B > \mu_A$
La región crítica es: $E = \frac{\bar{x_B} - \bar{x_A}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} > z_\alpha$, que en nuestro caso resulta:

$E = \frac{5.9-5.7}{\sqrt{\frac{1}{26}+\frac{2}{20}}} = 0.5373$ que es menor que $z_{0.10} = 1.2816$, por lo que no se cumple la condición de región crítica y por tanto aceptamos la hipótesis nula, esto es, **no existe suficiente evidencia para concluir que la media de B es mayor que la media de A.**

c: Se trata de un contraste unilateral de la varianza.
$H_0 : \sigma_B^2 \leq \sigma_A^2$ contra que $H_a : \sigma_B^2 > \sigma_A^2$ que tiene por región crítica:
$\frac{s_B^2}{s_A^2} > F_{\alpha,n_A-1,n_B-1}$, $\Rightarrow \frac{2}{1} = 2$ mientras que de las tablas se obtiene:
$F_{0.05,15,25} = 2.089$ y $F_{0.05,24,25} = 1.964$ por lo que interpolamos para obtener $F_{0.05,19,25} = 2.089 + 4/9(1.964 - 2.089) = 2.0334$

Así pues el estadístico $E = \frac{s_B^2}{s_A^2} = 2$ no está en la región crítica (aunque por muy poco) y nos quedamos con la hipótesis nula: **No existe bastante evidencia para afirmar que la dispersión de B es mayor que la de A**.

Problema 4: Prácticas

```
disp('a:')
x=2008:2016
y=[17 23 23 28 30 35 37 44]
p=polyfit(x,y,1), a=p(2), b=p(1)
disp('La recta es C=a+b*T')
% Podría haberse calculado con:
% N=8, A=[N sum(x);sum(x) sum(x.^2)], B=[sum(y);sum(x.*y)], sol=A\B, a=sol(1), b=sol(2)
N=8,mx=sum(x)/N,m2x=sum(x.^2)/N, Vx=m2x-mx^2
my=sum(y)/N,m2y=sum(y.^2)/N, Vy=m2y-my^2
cov=sum(x.*y)/N-mx*my
r=cov/sqrt(Vx*Vy)

disp('b:')
C2017=a+b*2017
```

Problema 5: Prácticas

```
NIT=50000
x0=normrnd(2,1.5,NIT,1);
x1=max(0,x0);
x2=normrnd(21,3,NIT,1); % Para el apartado b
disp('a:')
c=(x1+x2>30); p=sum(c)/NIT
Ip=[p-1.96*sqrt(p*(1-p)/NIT),p+1.96*sqrt(p*(1-p)/NIT)]

disp('b:')
x=[22,21,20,24,27,25,25,26,21,21,27,26,21,25]
alfa=0.05
N=length(x);
[H,p]=ttest(x,21,0.05,'right')
disp('Si H=1 indica que tarda mas de 21 min.')

disp('c:')
[Hc,pc]=vartest(x,9,0.05,'both')
```