Grados en Ingeniería Informática

Métodos Estadísticos

Examen **Convocatoria Septiembre 2016**

- A resolver en **2 horas y 15 minutos.**
- Dejar DNI encima de la mesa.
- **Apagar y guardar el MÓVIL**.

APELLIDOS, NOMBRE:
DNI: Especialidad: Grupo:

---

1. Un estudio entre el precio de venta (P en euros) y la cantidad diaria vendida diaria (V en unidades), proporciona la tabla:

Precio.\ Venta	$[0, 100]$	$(100, 150]$	$(150, 200]$	$(200, \infty)$
$[2, 4]$	0	0	5	5
$(4, 6]$	0	4	6	4
$(6, 10]$	10	2	0	0

   (a) Ajustar una recta de "Venta" sobre "Precio". ($V = a + bP$, a y b constantes)

   (b) Ajustar una función del tipo $VP = c$ (c constante, $V = f(P)$).

   (c) Indicar mediante el parámetro apropiado ¿Cuál es el mejor ajuste?

   (d) Estimar la moda de la marginal de V.

(0.5+0.5+0.5+0.75=2.25 Puntos)

2. La duración (en horas) de un modelo de componente incluido en la pantalla de un ordenador sigue una variable aleatoria exponencial de media $\mu = 100000 = 10^5$ horas,

   (a) Si cada pantalla necesita que funcionen los 5 componentes que posee de forma independiente, hallar la probabilidad de que una pantalla necesite ser reparada en periodo de garantía (18 meses $\approx$ 13140 horas).

   (b) Un centro escolar compra 90 ordenadores. ¿Cuál es la probabilidad de que se averíen 16 o más de ellos en periodo de garantía?

   (c) Se realiza una prueba de duración con una versión mejorada del componente, determinándose que el 10% de ellos se habían averiado antes de las 328940.73 horas de funcionamiento. Hallar la nueva media y la nueva función de distribución.

(1+0.75+1=2.75 Puntos)

3. El control de calidad de una empresa desea medir el tiempo de acceso a la memoria DDR3-1600 (en nanosegundos (ns.)), para ello se seleccionan al azar 200 unidades recién fabricadas y se obtienen los datos $\sum_{i=1}^{200} t_i = 1140\ ns.$ y $\sum_{i=1}^{200} t_i^2 = 108386\ ns.$, siendo t_i el tiempo de acceso y suponiendo que los datos provienen de una distribución normal, hallar:

   (a) Un intervalo de confianza para el tiempo de acceso al 90%.

   (b) Saber si puede ser aceptada la hipótesis de que el tiempo medio de acceso es inferior o igual a 5.5 ns. al nivel del 99%.

   (c) Realizando una variación en la fabricación se realizó una prueba con 400 memorias, obteniéndose una media de 5.4 ns. y una cuasivarianza de 1.21 ns. ¿Podemos aceptar que el nuevo tipo obtiene mejores (menores) tiempos de acceso al 98% de confianza que el modelo anterior?.

(0.75+0.75+1=2.5 Puntos)

4. **PRÁCTICAS (Sólo instrucciones)**

La población estimada de lince ibérico (Lynx pardinus) en libertad proporciona los datos:

Año	1960	1984	1993	1999	2005	2007	2012	2015
Número	5000	1100	1035	750	150	235	312	404

Se pide:

(a) Ajustar una parábola al número de linces en función del año.

(b) Teniendo en cuenta que desde el año 2002 existe un programa de cría en cautividad y reintroducción. Estimar una recta de regresión $N = a + bT$ usando solamente los datos del siglo XXI.

(c) Estimar SSE (cuadrados de la diferencia entre estimados y reales) para ambos métodos, pero usando solamente los datos del siglo XXI.

(0.25+0.25+0.5 = 1 Punto)

5. **PRÁCTICAS (Sólo instrucciones)**

Para determinar la velocidad instantánea de un coche, un medidor realiza 2 disparos de radar, midiendo el cociente entre el espacio recorrido y el tiempo entre los sucesivos disparos. Si el tiempo entre disparos Δt sigue una normal de media 0.1 s. y desviación 0.0002 s., mientras que el espacio recorrido Δe sigue una uniforme continua de media 3.4 m. y amplitud 0.2 m.

Hallar mediante simulación con 10000 iteraciones:

(a) Media y desviación típica de $v = 3600\frac{\Delta e}{\Delta t}$ Km/h.

(b) Probabilidad de que la medición rebase los 120 Km/h.

NOTA: La constante 3600 sale de pasar de m/ms a Km/h. *(0.25+0.25=0.5 Puntos)*

6. **PRÁCTICAS (Sólo instrucciones)**

Un arqueólogo conoce, para cada especie de homínido que la longitud del hueso frontal sigue una distribución normal, aunque de parámetros diferentes para cada especie. Para la especie A que se sabe habitaba la región de la excavación, estos parámetros son: $\mu_A = 12.0$ y $\sigma_A = 1.1$

El arqueólogo ha encontrado 37 fósiles de huesos frontales proporcionando los valores:

x={ 9.4010, 14.8865, 8.8196, 13.9024, 11.8101, 15.0917, 6.5978, 11.0058, 8.4804, 18.7700, 13.5630, 14.9474, 8.8545, 10.3285, 10.8188, 14.2461, 10.8053, 13.2539, 6.3705, 10.6154, 9.4410, 7.5574, 12.7699, 12.2050, 11.5837, 8.1658, 14.3187, 12.3754, 10.7523, 11.5572, 10.8450, 7.1245, 10.7859, 9.4216, 9.0520, 8.6090, 10.1661 }

(a) Contrastar que pertenecen a la especie A al nivel de significación $\alpha = 0.025$ en 2 fases:

   i. Contrastar que la varianza es compatible con los valores conocidos para la especie A.

   ii. Contrastar que la media es compatible con los valores conocidos para la especie A (suponer rechazado el contraste de la varianza).

(b) Suponiendo rechazada la hipótesis de que los huesos encontrados pertenezcan a la especie A, estimar intervalos de confianza con $\alpha = 0.025$ para los valores de la media y varianza de la nueva especie.

(0.25+0.25+0.5=1 Punto)

SOLUCIONES:
Problema 1:
1-a: Vamos a usar la fórmula: $V - \bar{V} = b(P - \bar{P})$, para lo que construimos la tabla:

P_i	V_i	n_i	n_iP_i	$n_iP_i^2$	n_iV_i	$n_iV_i^2$	$n_iP_iV_i$
3	175	5	15	45	875	153125	2625
3	225	5	15	45	1125	253125	3375
5	125	4	20	100	500	62500	2500
5	175	6	30	150	1050	183750	5250
5	225	4	20	100	900	202500	4500
8	50	10	80	640	500	25000	4000
8	125	2	16	128	250	31250	2000
		36	196	1208	5200	911250	24250

Así $\bar{P} = \frac{196}{36} = 5.4444$, $\bar{V} = \frac{5200}{36} = 144.4444$, $\sigma_P^2 = \frac{1208}{36} - 5.4444^2 = 3.9136$, $\sigma_V^2 = \frac{911250}{36} - 144.4444^2 \approx 4448.3025$ y $m_{1,1} = \frac{\sum_i n_iP_iV_i}{N} = \frac{24250}{36} = 673.6111 \Rightarrow cov(P,V) = 673.6111 - 5.4444 \cdot 144.44444 \approx -112.8086$ mostrando el valor negativo de la misma como al aumentar el precio disminuye la cantidad vendida.

Usando las fórmulas del tema 2 pág. 45: $b = \frac{Cov(P,V)}{\sigma_P^2} = \frac{-112.8086}{3.9136} = -28.8249$, $a = \bar{V} - b\bar{P} = 301.3801$, resultando la recta: $\mathbf{V = 301.3801 - 28.8249P}$

1-b: Debemos ajustar $V = \frac{c}{P}$ luego la base consta de un único vector $\mathcal{B} = \{\frac{1}{P}\} = \{X\}$ (donde $X_i = \frac{1}{P_i}$) quedando la única ecuación:

$$\left\langle \frac{1}{P}, V \right\rangle = c \left\langle \frac{1}{P}, \frac{1}{P} \right\rangle \Rightarrow \sum_i n_i \frac{1}{P_i} V_i = c \sum_i n_i \frac{1}{P_i}\frac{1}{P_i} \Rightarrow c = \frac{\sum_i n_iX_iV_i}{\sum_i n_iX_i^2}$$

Formamos la tabla (donde $V_i^{est} = \frac{c}{P_i} = cX_i$ y $r_i = V_i - V_i^{est}$):

P_i	V_i	n_i	$X_i = 1/P_i$	n_iX_i	$n_iX_i^2$	$n_iX_iV_i$	V_i^{est}	r_i	n_ir_i	$n_ir_i^2$
3	175	5	0.3333	1.6667	0.55556	291.67	224.256	−49.256	−246.28	12131.0
3	225	5	0.3333	1.6667	0.55556	375.00	224.256	0.744	3.72	2.8
5	125	4	0.2000	0.8000	0.16000	100.00	134.554	−9.554	−38.22	365.1
5	175	6	0.2000	1.2000	0.24000	210.00	134.554	40.446	242.68	9815.3
5	225	4	0.2000	0.8000	0.16000	180.00	134.554	90.446	361.78	32722.0
8	50	10	0.1250	1.2500	0.15625	62.50	84.096	−34.096	−340.96	11625.5
8	125	2	0.1250	0.2500	0.03125	31.25	84.096	40.904	81.81	3346.2
		36			1.85861	1250.42			64.53	70008.0

Obteniendo $\mathbf{c} = \frac{\sum_\mathbf{i} \mathbf{n_iX_iV_i}}{\sum_\mathbf{i} \mathbf{n_iX_i^2}} = \frac{\mathbf{1250.42}}{\mathbf{1.85861}} \approx \mathbf{672.7694}$

1-c: Para ver cuál es el mejor voy a usar el coeficiente de determinación R^2:

Hiperbólico: Calculamos las columnas $V_i^{est} = \frac{672.7694}{P_i}$, $r_i = V_i - V_i^{est}$, la varianza residual de los residuos: $V_r = \frac{\sum_i n_ir_i^2}{N} - \left(\frac{\sum_i n_ir_i}{N}\right)^2 = \frac{7000.8}{36} - \left(\frac{64.53}{36}\right)^2 \approx 1941.452$

y el coeficiente $R^2 = 1 - \frac{Vr}{\sigma_V^2} = 1 - \frac{1941.452}{4448.302} = 0.56355$ ya que $\sigma_V^2 = \frac{911250}{36} - \left(\frac{5200}{36}\rangle\right)^2 \approx 4448.3047$

Lineal: Debemos calcular r^2 que coincide para el caso de la recta con R^2, pero puede calcularse mediante $r = \frac{cov}{\sqrt{\sigma_V^2\sigma_P^2}} \Rightarrow r^2 = \frac{cov^2}{\sigma_V^2\sigma_P^2} \approx \frac{-112.8086^2}{4448.30247 \cdot 3.91358} \approx 0.730998$

Resultando ser mejor el ajuste lineal.

1-d:
Formamos la tabla: (n_i = Frecuencias absolutas, a_i = amplitud, h_i = alturas)

Int.	n_i	a_i	h_i
$[0, 100]$	10	100	0.10
$(100, 150]$	6	50	0.12
$(150, 200]$	11	50	0.22
$(200, \infty)$	9	50	0.18

La máxima altura corresponde al intervalo $(150, 200]$. La moda es: $\mathbf{Mo = 150 + \frac{0.10}{0.10+0.04}50 = 185.7143}$ donde $\Delta_1 = 0.22 - 0.12$ y $\Delta_2 = 0.22 - 0.18$ en la fórmula del Tema 1 pág. 30.

Problema 2:

a: Nos dicen que la media es 100000 por lo que $\lambda = 1/100000 = 10^{-5}$ (Tema 5b pag. 47) y la función de distribución es $F(x) = 1 - e^{-10^{-5}x}$ para $x > 0$ y 0 para $x \leq 0$.

La probabilidad de que un componente dure menos de las 13140 horas es $F(13140) = 1 - e^{-10^{-5}(13140)} \approx 0.1213$. Así P(funcione tras 13140 horas)=1-0.1213=0.8787

Por tanto la probabilidad de que todos funcionen (considerando independencia) es igual a:

P(funcionen todos)$= 0.8787^5 = 0.5238 \Rightarrow P(\text{alguno falle}) = 1 - 0.5238 = \mathbf{0.4762}$

b: El número de averiados X seguirá una binomial de parámetros $n = 90$ y $p = 1 - 0.5238 = 0.4762$. Como $n > 30$ y tanto $np = 42.858$ como nq son mayores de 5, podrá aproximarse por una normal de media $\mu = np = 42.858$ y desviación $\sigma = \sqrt{npq} = \sqrt{90(0.4762)(0.5238)} = 4.738$.

$\mathbf{P(X \geq 16)}$ = (Corrección continuidad) $= P(X^* > 15.5) = P\left(z > \frac{15.5-42.858}{4.738}\right) = P(z > -5.7741) = 1 - P(z > 5.7741) \approx \mathbf{1 - 4.0241(10)^{-9} \approx 1}$

c: Nos dicen que $P(\xi < 328940.73) = 0.1 \Rightarrow F(328940.73) = 1 - e^{-\lambda(328940.73)} = 0.1 \Rightarrow e^{-328940.73\lambda} = 0.9$ $\Rightarrow -328940.73\lambda = \text{Ln}(0.9)$ obteniendose $\lambda \approx 3.203(10)^{-7}$ y la nueva media será $\mu' = \frac{1}{\lambda} \approx \mathbf{3122049.4}$ horas.

La nueva función de distribución es: $\mathbf{F'(x) = 1 - e^{-3.203\ 10^{-7}x}}$ si $\mathbf{x > 0}$ y $\mathbf{F(x) = 0}$ en el resto.

Problema 3:

a: En todo el problema necesitaremos $\bar{t}$ (media muestral) y s^2 (cuasivarianza): $\bar{t} = \frac{\sum_{i=1}^{200} t_i}{200} = \frac{1140}{200} = 5.7$, $s^2 = \frac{n}{n-1}\left(\frac{\sum_{i=1}^{200} t_i^2}{n} - (\bar{t})^2\right) = \frac{200}{199}\left(\frac{108386}{200} - 5.7^2\right) = 512 \Rightarrow s = \sqrt{512} \approx 22.6274$.

El intervalo de confianza para el tiempo de acceso (muestra grande, varianza desconocida) es:

$$I = \left[\bar{t} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right] = \left[5.7 \pm 1.645 \frac{22.6274}{\sqrt{200}}\right] = [5.7 \pm 2.632] = [3.068, 8.332]$$

pues $\alpha = 0.1 \Rightarrow z_{\frac{\alpha}{2}} = z_{0.05} \approx 1.645$.

b: Se trata de un contraste unilateral de la media con muestra grande y varianza desconocida:

$H_0 : \mu \leq 5.5$ contra $H_a : \mu > 5.5$

La región crítica es: $E = \frac{\mu - \mu_0}{s/\sqrt{n}} > z_\alpha$

$E = \frac{5.7-5.5}{22.6274/\sqrt{200}} = 0.125$ que no es mayor que $z_{0.99} = 2.32$, por lo que **admitimos que puede cumplirse que la media sea igual o menor a 5.5**.

c: Ahora es un contraste unilateral de diferencias de medias, muestras grandes y varianzas desconocidas.

$H_0 : \mu_1 \leq \mu_2$ contra $H_a : \mu_1 > \mu_2$

La región crítica es: $E = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + s_2^2 n_2}} > z_\alpha$

$E = \frac{5.7-5.4}{\sqrt{\frac{512}{200} + \frac{1.21}{400}}} = 0.1874$ que no es mayor que $z_{0.02} = 2.054$, por lo que **no se puede asegurar que obtiene tiempos de acceso menores** (los resultados no son concluyentes).

Parte de Matlab

Problema 4:

```
clc, clear all, format compact,disp('Examen Septiembre 2016')
disp('Problema 4')
N=[5000 1100 1035 750 150 235 312 404]
T=[1960 1984 1993 1999 2005 2007 2012 2015]
disp('4-a')
pa=polyfit(T,N,2)
disp('4-b')
NN=N(5:8) % NN=[150 235 312 404]
TT=T(5:8) % TT=[2005 2007 2012 2015]
pb=polyfit(TT,NN,1)
a=pb(2),b=pb(1)
disp('4-c')
NNest_a=polyval(pa,TT)
NNest_b=polyval(pb,TT)
SSEa=sum((NN-NNest_a).^2)
SSEb=sum((NN-NNest_b).^2)
```

Problema 5:

```
clc, clear all, format compact
disp('Problema 5')
disp('5-a')
N=10000
DeltaT=normrnd(0.1,0.0002,1,N);
DeltaE=unifrnd(3.4-0.2/2,3.4+0.2/2,1,N);
V=3600*DeltaE./DeltaT;
m=mean(V),desv=sqrt(var(V))
disp('5-b')
cond=(V>120);
p=sum(cond)/N
```

Problema 6:

```
clc, clear all, format compact
disp('Problema 6')
disp('6-a1')
x=[9.4010, 14.8865, 8.8196, 13.9024, 11.8101, 15.0917,  6.5978, 11.0058, 8.4804, 18.7700,...
13.5630, 14.9474, 8.8545, 10.3285, 10.8188, 14.2461, 10.8053, 13.2539, 6.3705, 10.6154,...
9.4410, 7.5574, 12.7699, 12.2050, 11.5837, 8.1658, 14.3187, 12.3754, 10.7523, 11.5572,...
10.8450, 7.1245, 10.7859, 9.4216,  9.0520,  8.6090, 10.1661]
[Ha,pa]=vartest(x,1.1^2,0.025,'both')
disp('6-a2')
[Hb,pb]=ttest(x,12,0.025,'both')
disp('6-b')
m=mean(x), s=sqrt(var(x)), n=length(x);
z=norminv(1-0.025/2)
Im=[m-z*s/sqrt(n),m+z*s/sqrt(n)]
Iv=[(n-1)*s^2/chi2inv(1-0.025/2,n-1),(n-1)*s^2/chi2inv(0.025/2,n-1)]
```