PROBABILIDAD Y ESTADÍSTICA
SEGUNDO PARCIAL 2005

DATOS DEL ESTUDIANTE

No. de Parcial	Apellidos y Nombres	Cédula

- **La duración total de la prueba es de 3 horas y media.**
- **Publicación de resultados: miércoles 20 de julio - 18:00 hs.**
- **Muestra de parciales: viernes 22 de julio - 16:00 hs.**

Nota: En todas las pruebas de hipótesis utilice el siguiente criterio de decisión: se acepta la hipótesis nula si el p-valor es superior a 0.1

Problema 1 (35 puntos)

Un sistema de información está formado por dos servidores. Cuando un cliente accede al sistema es asignado con probabilidad p al servidor A y con probabilidad $(1-p)$ al servidor B. De esta manera la variable aleatoria X que representa el tiempo de servicio de un cliente tiene una distribución cuya densidad es:

$$f_X(x) = \begin{cases} p\,\dfrac{1}{\theta_1}\,e^{-\frac{x}{\theta_1}} + (1-p)\,\dfrac{1}{\theta_2}\,e^{-\frac{x}{\theta_2}} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \tag{1}$$

donde θ_1 y θ_2 son dos parámetros positivos.

a) **(5 puntos)** Probar que $\mathbf{E}(X) = p\,\theta_1 + (1-p)\,\theta_2$ y que $\mathbf{E}(X^2) = 2\,p\,\theta_1^2 + 2\,(1-p)\,\theta_2^2$.
Sugerencia: Identifique distribuciones conocidas para evitar cálculos.

b) **(12 puntos)** Suponga que 100 clientes acceden al sistema. Se obtiene así una muestra de tiempos de servicios $X_1, X_2 \ldots, X_{100}$ *i.i.d.* con distribución correspondiente a la densidad definida en (1). Suponga además que $\theta_1 = 1$ y $\theta_2 = \frac{1}{2}$.

i) Si $p = \frac{1}{3}$, estime la probabilidad de que el tiempo total de servicio, esto es $T = \sum_{i=1}^{100} X_i$, sea superior a 77.

ii) Estime el menor valor de p para el cual se cumple $\mathbf{P}\,(T \leq 75) \leq \frac{1}{2}$.

c) **(8 puntos)** Suponga ahora que 1000 clientes son asignados con igual probabilidad a cualquiera de los servidores. Esto es $X_1, X_2 \ldots, X_{1000}$ son variables *i.i.d.* con distribución correspondiente a la densidad definida en (1) con $p = \frac{1}{2}$.

Sabiendo que $\theta_1 > \theta_2$ y que

$$\sum_{i=1}^{1000} X_i = 4000 \quad \text{y} \quad \sum_{i=1}^{1000} X_i^2 = 32080$$

estime los parámetros θ_1 y θ_2 por el método de los momentos.

d) **(10 puntos)** Se dispone ahora de una muestra aleatoria **(no hay que verificar este supuesto)** de tiempos de servicios

0.19	0.05	1.34	0.98	0.57	0.11	0.37	1.35

i) Calcule la función de distribución F_X de una variable aleatoria X con densidad definida en (1).

ii) Verificar mediante una prueba de ajuste si los datos corresponden a la distribución calculada en la parte anterior con $\theta_1 = 1$, $\theta_2 = \frac{1}{2}$ y $p = 0.2$.

Ayuda: Puede ser de utilidad la siguiente tabla de valores

X_i	0.19	0.05	1.34	0.98	0.57	0.11	0.37	1.35
e^{-X_i}	0.827	0.951	0.262	0.375	0.565	0.896	0.691	0.259
e^{-2X_i}	0.684	0.905	0.069	0.141	0.320	0.802	0.477	0.067

Problema 2 (25 puntos)

Los siguientes datos corresponden a los pesos (medidos en kilogramos) de doce boxeadores en actividad del Palermo Boxing Club:

99.0	79.8	77.4	87.8	59.0	71.0	51.4	92.1	69.4	68.9	70.1	90.2

Ayuda: Pueden ser de utilidad los siguientes datos: $\overline{X}_n = 76.34$ y $\sigma_n = 13.51$.

a) **(5 puntos)** ¿Es razonable suponer que los datos son aleatorios? (Realice una sola prueba de hipótesis)

b) **(8 puntos)** ¿Es razonable suponer que los datos tienen distribución normal? (Realice una sola prueba de hipótesis)

c) **(12 puntos)** Asumiendo que los datos tienen distribución normal $N(\mu, \sigma^2)$,

i) ¿Es razonable suponer que $\mu = 75$?

ii) ¿Es razonable suponer que $\sigma = 12$?

SOLUCIÓN

Problema 1

a) De la definición de valor esperado:

$$\mathbf{E}(X) = \int_{\mathbf{R}} x f_X(x)dx = p\int_0^{\infty} x\,\frac{1}{\theta_1}\,e^{-\frac{x}{\theta_1}}dx \;+\; (1-p)\int_0^{\infty} x\,\frac{1}{\theta_2}\,e^{-\frac{x}{\theta_2}}dx.$$

Notemos que la primera integral corresponde al valor esperado de una variable X_1 con distribución exponencial de parámetro $\frac{1}{\theta_1}$ y que la segunda integral correponde al valor esperado de una variable X_2 con distribución exponencial de parámtero $\frac{1}{\theta_2}$, de manera que

$$\mathbf{E}(X) = p\mathbf{E}(X_1) + (1-p)\mathbf{E}(X_2) = p\theta_1 + (1-p)\theta_2.$$

Análogamente:

$$\mathbf{E}(X^2) = \int_{\mathbf{R}} x^2 f_X(x)dx = p\int_0^{\infty} x^2\,\frac{1}{\theta_1}\,e^{-\frac{x}{\theta_1}}dx + (1-p)\int_0^{\infty} x^2\,\frac{1}{\theta_2}\,e^{-\frac{x}{\theta_2}}dx = p\mathbf{E}(X_1^2)+(1-p)\mathbf{E}(X_2^2)$$

donde X_1 y X_2 son las variables definidas arriba. Ahora, si Y es una variable exponencial de parámetro λ, entonces $\mathbf{E}(Y^2) = 2\,\frac{1}{\lambda^2}$. Por lo tanto:

$$\mathbf{E}(X^2) = 2p\theta_1^2 + 2(1-p)\theta_2^2.$$

b) Es fácil verificar que si X es una variable aleatoria con distribución correspondiente a la densidad definida en (1), con $\theta_1 = 1$ y $\theta_2 = \frac{1}{2}$, entonces

$$\mathbf{E}(X) = \frac{1}{2}(1+p) \quad \text{y} \quad \mathbf{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}^2(X) = \frac{3}{2}p + \frac{1}{2} - \frac{1}{4}(1+p)^2. \tag{2}$$

i) Si $p = \frac{1}{3}$, resulta $\mu = \mathbf{E}(X) = \frac{2}{3}$ y $\sigma^2 = \mathbf{Var}(X) = \frac{5}{9}$. Para estimar $\mathbf{P}(T > 77)$ usamos el Teorema Central del Límite:

$$\begin{aligned}\mathbf{P}(T > 77) &= \mathbf{P}\left(\sum_{i=1}^{100} X_i > 77\right) = \mathbf{P}\left(\sqrt{100}\frac{(\overline{X}_{100}-\mu)}{\sigma} > 30\frac{(\frac{77}{100}-\frac{2}{3})}{\sqrt{5}}\right)\\ &\simeq \mathbf{P}(N(0,1) > 1.39) = 1 - \Phi(1.39) = 0.08227.\end{aligned}$$

ii) Usando (2) y el Teorema Central del Límite, resulta:

$$\begin{aligned}\mathbf{P}(T \le 75) &= \mathbf{P}\left(\sqrt{100}\frac{[\overline{X}_{100} - \frac{1}{2}(1+p)]}{\sqrt{\frac{3}{2}p + \frac{1}{2} - \frac{1}{4}(1+p)^2}} \le 10\frac{[\frac{75}{100} - \frac{1}{2}(1+p)]}{\sqrt{\frac{3}{2}p + \frac{1}{2} - \frac{1}{4}(1+p)^2}}\right)\\ &\simeq \mathbf{P}\left(N(0,1) \le 10\frac{[\frac{75}{100} - \frac{1}{2}(1+p)]}{\sqrt{\frac{3}{2}p + \frac{1}{2} - \frac{1}{4}(1+p)^2}}\right).\end{aligned}$$

De modo que $\mathbf{P}(T \leq 75) \leq \frac{1}{2}$ si y sólo si

$$\frac{75}{100} - \frac{1}{2}(1+p) \leq 0,$$

esto es, si y sólo si $p \geq \frac{1}{2}$. Así que el menor valor de p para el cual se cumple la desigualdad es $p = \frac{1}{2}$.

c) Para aplicar el método de los momentos consideramos las igualdades: $\overline{X} = \mathbf{E}(X)$ y $\overline{X^2} = \mathbf{E}(X^2)$. Usando los resultado de la parte a) con $p = \frac{1}{2}$ y los datos de la letra, queda el siguiente sistema de dos ecuaciones con dos incógnitas:

$$\frac{1}{2}\theta_1 + \frac{1}{2}\theta_2 = 4 \quad , \quad \theta_1^2 + \theta_2^2 = 32.080.$$

De la primera igualdad: $\theta_2 = 8 - \theta_1$. Llevando esto a la segunda igualdad se obtiene la siguiente ecuación cuadrática para θ_1:

$$\theta_1^2 - 8\theta_1 + 15.96 = 0,$$

cuyas soluciones son 3.8 y 4.2. Es fácil ver que la solución $\theta_1 = 4.2$ corresponde a $\theta_2 = 3.8$ y viceversa. De manera que la estimación de parámetros por el método de los momentos, compatible con $\theta_1 > \theta_2$, es

$$\hat{\theta}_1 = 4.2 \quad \text{y} \quad \hat{\theta}_2 = 3.8.$$

d)

i)

$$F_X(t) = \int_{-\infty}^{t} f(x)dx = \begin{cases} 1 - p\, e^{-\frac{t}{\theta_1}} - (1-p)\, e^{-\frac{t}{\theta_2}} & \text{si } t > 0 \\ 0 & \text{si } t \leq 0 \end{cases}$$

ii) Poniendo $\theta_1 = 1$, $\theta_2 = \frac{1}{2}$ y $p = 0.2$ en la expresión anterior, obtenemos la función de distribución:

$$F_0(t) = \begin{cases} 1 - 0.2\, e^{-t} - 0.8\, e^{-2t} & \text{si } t > 0 \\ 0 & \text{si } t \leq 0 \end{cases}$$

Para verificar si los datos corresponden a la función de distribución F_0, implementamos la prueba de Kolmogorov-Smirnov. Para esto resulta de utilidad la tabla de valores dada en la Ayuda.

X_i^*	$\frac{i}{n}$	$F_0(X_i^*)$	$\lvert\frac{i}{n} - F_0(X_i^*)\rvert$	$\lvert\frac{i}{n} - F_0(X_{i+1}^*)\rvert$
0.05	0.125	0.0859	0.0391	0.0859
0.11	0.250	0.1789	0.0711	0.0.539
0.19	0.375	0.2875	0.0875	0.0.375
0.37	0.500	0.4802	0.0198	0.1052
0.57	0.625	0.6311	0.0061	0.1311
0.98	0.750	0.8123	0.0623	**0.1873**
1.34	0.825	0.8928	0.0178	0.1428
1.35	1	0.8944	0.1056	0.0194

De esta manera el estadístico del test es $D_n = 0.1873$ y de la tabla de Kolmogorov-Smirnov se obtiene $\alpha^* > 0.2$; por lo tanto aceptamos que la distribución de los datos es F_0.

Problema 2

a) Para estudiar si los datos pueden suponerse aleatorios realizamos el *Test de rachas de subidas y bajadas* o el *Test de Spearman.*

Test de Rachas

Recordamos que se definen las variables auxiliares U_i de la siguiente manera: $U_i = 1$ si $X_i \leq X_{i+1}$ y $U_i = 0$ en caso contrario.

X_i	U_i
99.0	
79.8	0
77.4	0
87.8	1
59.0	0
71.0	1
51.4	0
92.1	1
69.4	0
68.9	0
70.1	1
90.2	1

Hay 8 rachas, de manera que el estadístico del test es $R = 8$. Como $R = 8 > \frac{2(12)-1}{3} = 7.6667$, planteamos

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, X_2, ..., X_n$ son *iid*	"hay muchas rachas"

Usando la tabla tenemos que el p-valor es $\alpha^* = 0.5547 > 0.1$. Conclusión: No se rechaza $\mathbf{H}_0$.

Test de rangos de Spearman

La tabla muestra los datos ordenados X_i^*, los índices i (esto es, el lugar que ocupaban los datos originalmente) y los rangos $R(X_i)$.

X_i^*	i	$R(X_i)$
51.4	7	1
59.0	5	2
68.9	10	3
69.4	9	4
70.1	11	5
71.0	6	6
77.4	3	7
79.8	2	8
87.8	4	9
90.2	12	10
92.1	8	11
99.0	1	12

Se obtiene $\sum_{i=1}^{12}(R(X_i) - i)^2 = 366$, y por lo tanto el estadístico del test resulta

$$r_s = 1 - 6\frac{\sum_{i=1}^{12}(R(X_i) - i)^2}{12(12^2 - 1)} = -0.27972.$$

La tabla del test para $n = 12$ muestra que $\alpha^* > 0.10$. Por lo tanto la muestra pasa el test. Es razonable, entonces, suponer que la muestra es aleatoria.

b) Aplicamos las pruebas de normalidad:

Test de D'Agostino

El estadístico del test es

$$DA_n = \frac{\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)X_i^*}{n^2\sigma_n},$$

que en este caso da $DA_n = 0.282224$ y de la tabla se obtiene $\alpha^* > 0.20$. Por lo tanto la muestra pasa el test.

Test de Shapiro-Wilk

El estadístico del test es

$$W_n = \frac{\left(\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i(X_{n-i+1}^* - X_i^*)\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Orden inverso	Orden	a_i	$a_i(X_{n-i+1}^* - X_i^*)$
99.0	51.4	0.5475	26.061
92.1	59.0	0.3325	11.006
90.2	68.9	0.2347	4.999
87.8	69.4	0.1586	2.918
79.8	70.1	0.0922	0.894
77.4	71.0	0.0303	0.194

Por lo tanto $W_n = \frac{(46.07)^2}{2190.41} = 0.969$ y de la tabla se obtiene que el p-valor es $\alpha^* > 0.5$.

Conclusión: No se rechaza $\mathbf{H}_0$.

c) Asumimos aquí que los datos son normales.

i) Construimos un intervalo de confianza al 90% para μ con σ desconocido (el intervalo es exacto, pues los datos son normales):

$$I = \left[\overline{X}_n - \frac{s_n}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1)\, ,\, \overline{X}_n + \frac{s_n}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) \right]$$

Sabemos que

$$\overline{X}_n = 76.34 \quad , \quad t_{0.05}(11) = 1.796 \quad \text{y} \quad s_n = \sqrt{\frac{n}{n-1}}\sigma_n = 14.11$$

de manera que el intervalo de confianza queda

$$I = [69.017\, ,\, 83.666].$$

Como el valor 75 cae dentro de ese intervalo concluimos que es razonable suponer $\mu = 75$.

De otra manera: Implementamos un test paramétrico para datos normales. Como $\overline{X}_n = 76.34$, consideramos las siguientes hipótesis

$\mathbf{H}_0$	$\mathbf{H}_1$
$\mu = 75$	$\mu > 75$

El estadístico del test es: $T = \sqrt{n}\left(\frac{\overline{X}_n - 75}{s_n}\right) = 0.329$ y la región crítica:

$$\mathcal{R}_\alpha = \{T \geq t_\alpha(n-1)\}.$$

Usando la tabla de la distribución t-Student se obtiene $0.25 < \alpha^* < 0.4$ y por lo tento se acepta H_0.

ii) Construimos un intervalo de confianza al 90% para σ^2 con μ desconocido (el intervalo es exacto, pues los datos son normales):

$$I = \left[\frac{(n-1)}{\chi^2_{\frac{\alpha}{2}}(n-1)} s_n^2\, ,\, \frac{(n-1)}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} s_n^2 \right]$$

Sabemos que

$$s_n^2 = 199.11 \quad , \quad \chi^2_{0.05}(11) = 19.68 \quad \text{y} \quad \chi^2_{0.95}(11) = 4.57$$

de manera que el intervalo de confianza queda

$$I = [111.293\, ,\, 479.265].$$

Como el valor $(12)^2 = 144$ cae dentro de ese intervalo concluimos que es razonable suponer $\sigma = 12$.

De otra manera: Implementamos un test paramétrico para datos normales. Como $s_n^2 = 199.11$, consideramos las siguientes hipótesis

$\mathbf{H}_0$	$\mathbf{H}_1$
$\sigma^2 = 144$	$\sigma^2 > 144$

El estadístico del test es: $E = \frac{(n-1)s_n^2}{12^2} = 15.210$ y la región crítica:

$$\mathcal{R}_\alpha = \{E \geq \chi^2_\alpha(n-1)\}.$$

Usando la tabla de la distribución χ^2 se obtiene $0.10 < \alpha^* < 0.50$ y por lo tanto se acepta H_0.