PROBABILIDAD Y ESTADÍSTICA
EXAMEN FEBRERO 2005

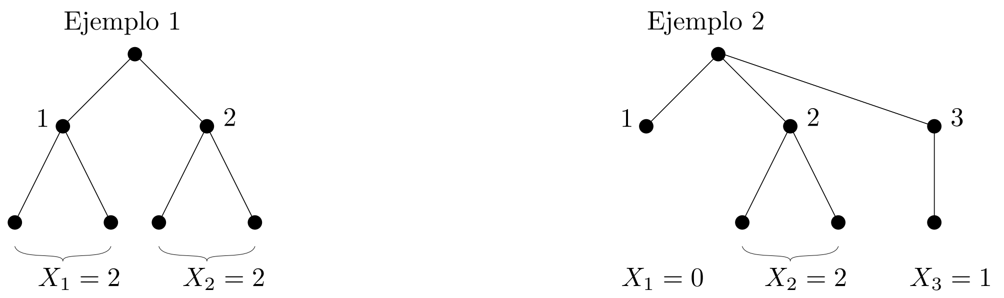
DATOS DEL ESTUDIANTE

N de Examen	Nombre	Cédula

- La duración del examen es de **3 horas y media**.
- El puntaje mínimo para salvar el examen es de **50 puntos**.
- **Publicación de resultados:** Viernes 4 de marzo a las 18:00 hs.
- **Muestra de exámenes:** Martes 8 de marzo a las 17:00 hs.

Problema 1 (45 puntos)

Proceso de Ramificación: Considere un individuo (el progenitor) que a lo largo de su vida tiene un número N de hijos, donde N puede valer 1, 2 ó 3, con $\mathbf{P}(N = i) = \frac{1}{3}$, para $i = 1, 2, 3$. Estos hijos forman la *primera generación*. Suponga ahora que cada uno de estos N individuos engendra, independientemente de los demás, un número aleatorio de hijos. Se denota X_j al número de hijos que tiene el individuo j, para $j = 1, \ldots, N$. Estos individuos forman la *segunda generación* (ver figura). Se supone que las variables aleatorias X_j, $j = 1, \ldots, N$ son independientes e idénticamente distribuidas con $X_j \sim Bin(2, p)$, $0 < p < 1$. Finalmente, se denota con Z al número total de individuos en la segunda generación; esto es, $Z = X_1 + \ldots + X_N$.



a) **(8 puntos)** Calcule $\mathbf{P}(Z = 0)$; esto es, la probabilidad de que en la segunda generación no haya individuos.

   Sugerencia: Condicione sobre los posibles valores de la variable N.

b) **(10 puntos)** Calcule $\mathbf{P}(Z = 1)$; esto es, la probabilidad de que en la segunda generación haya exactamente un individuo.

c) **(15 puntos)**

   i) Calcule la distribución de la variable Z para cada valor de N fijo (esto es, para cada $N = i$ con $i = 1, 2, 3$)

ii) Pruebe que el valor esperado para el número de individuos en la segunda generación es:

$$\mathbf{E}(Z) = \mathbf{E}(N)\mathbf{E}(X_1) = 4p.$$

Sugerencia: Condicione sobre los posibles valores de la variable N.

iii) Construya un estimador consistente para el parámetro p en base a las observaciones del número total de individuos en la segunda generación.

d) **(12 puntos)** Suponga ahora que el progenitor tiene $N = 81$ hijos y que cada uno de estos individuos engendra a su vez nuevos hijos de manera independiente. Como antes, se denota con $X_j \sim Bin(2, p)$ el número de hijos que tiene el individuo j, para $j = 1, \ldots, 81$, y con Z el número total de individuos en la segunda generación. Estime el valor de p para el cual se cumple $\mathbf{P}(Z \leq 100) = \frac{1}{2}$.

Problema 2 (15 puntos)

Considere la siguiente función

$$f(x) = \begin{cases} ax & \text{si } 0 \leq x \leq 1 \\ a & \text{si } 1 < x \leq b \\ 0 & \text{en otro caso} \end{cases}$$

donde $b > 1$ y $1 < a < 2$.

a) **(3 puntos)** Hallar b en función de a para que f sea una densidad de probabilidad.

b) **(5 puntos)** Calcule el valor esperado y la mediana para una variable aleatoria X con densidad f.

c) **(7 puntos)** Suponga que se tiene una muestra $X_1, X_2 \ldots, X_n$ de variables i.i.d. con la misma distribución que X. Construya un estimador consistente para el parámetro a.

Problema 3 (40 puntos)

Un artículo publicado en 1993 sugiere que el tiempo de reacción para una respuesta en tránsito a una señal de frenado se puede modelar con una distribución normal de media $\mu = 1.25\, s$ y desviación estándar $\sigma = 0.46\, s$. Para investigar si dicho modelo es correcto se consideran los siguientes datos correspondientes a tiempos de reacción ante señales de frenado:

0.98	2.25	1.19	1.30	1.73	1.27	1.20	0.86	1.38	0.63	1.57	1.99

Nota: En todas las pruebas de hipótesis use el siguiente criterio de decisión: se acepta la hipótesis nula si el p-valor es superior a 0.10.

a) **(10 puntos)** ¿Es razonable suponer que la muestra es aleatoria? (Realice dos pruebas de hipótesis.)

b) **(14 puntos)** ¿Es razonable suponer que la muestra es gaussiana? (Realice una sola prueba de hipótesis.)

Sugerencia: Pueden ser de utilidad los siguientes valores: $\overline{X_n} = 1.3625$ y $\sigma_n^2 = 0.197$.

c) **(16 puntos)** Suponiendo que la muestra sea efectivamente gaussiana y tratando de especificar los parámetros:

i) ¿Es razonable suponer que la desviación estándar de la muestra es $\sigma = 0.46$? (Realice una sola prueba de hipótesis.)

ii) ¿Es razonable suponer que la media es $\mu = 1.25$? (Realice una sola prueba de hipótesis.)

PROBABILIDAD Y ESTADÍSTICA
EXAMEN FEBRERO 2005: SOLUCIÓN

Problema 1

a) Condicionando sobre los posibles valores de N se tiene:

$$\begin{aligned}
\mathbf{P}(Z=0) &= \mathbf{P}(Z=0|N=1)\mathbf{P}(N=1)+\mathbf{P}(Z=0|N=2)\mathbf{P}(N=2)+\mathbf{P}(Z=0|N=3)\mathbf{P}(N=3)\\
&= \mathbf{P}(X_1=0)\times\frac{1}{3}+\mathbf{P}(X_1=0,X_2=0)\times\frac{1}{3}+\mathbf{P}(X_1=0,X_2=0,X_3=0)\times\frac{1}{3}\\
&= \mathbf{P}(X_1=0)\times\frac{1}{3}+\mathbf{P}(X_1=0)\mathbf{P}(X_2=0)\times\frac{1}{3}+\mathbf{P}(X_1=0)\mathbf{P}(X_2=0)\mathbf{P}(X_3=0)\times\frac{1}{3}\\
&= \frac{1}{3}\left[(1-p)^2+(1-p)^4+(1-p)^6\right].
\end{aligned}$$

b) De manera análoga:

$$\begin{aligned}
\mathbf{P}(Z=1) &= \mathbf{P}(Z=1|N=1)\mathbf{P}(N=1)+\mathbf{P}(Z=1|N=2)\mathbf{P}(N=2)+\mathbf{P}(Z=1|N=3)\mathbf{P}(N=3)\\
&= \mathbf{P}(X_1=1)\times\frac{1}{3}+\mathbf{P}(\{X_1=1,X_2=0\}\cup\{X_1=0,X_2=1\})\times\frac{1}{3}\\
&+ \mathbf{P}(\{X_1=1,X_2=0,X_3=0\}\cup\{X_1=0,X_2=1,X_3=0\}\cup\{X_1=0,X_2=0,X_3=1\})\times\frac{1}{3}\\
&= \frac{1}{3}\left[2p(1-p)+4p(1-p)^3+6p(1-p)^5.\right]
\end{aligned}$$

c) i) Usando propiedades bien conocidas para la suma de variables binomiales independientes, resulta:

- Para $N=1$, $Z=X_1 \Rightarrow Z\sim Bin(2,p)$.
- Para $N=2$, $Z=X_1+X_2 \Rightarrow Z\sim Bin(4,p)$.
- Para $N=3$, $Z=X_1+X_2+X_3 \Rightarrow Z\sim Bin(6,p)$.

ii) Los posibles valores para la variable Z son 0,1,2,3,4,5,6. De manera que:

$$\begin{aligned}
\mathbf{E}(Z) &= \sum_{k=0}^{6} k\mathbf{P}(Z=k)\\
&= \sum_{k=0}^{6} k\left[\mathbf{P}(Z=k|N=1)\mathbf{P}(N=1)+\mathbf{P}(Z=k|N=2)\mathbf{P}(N=2)+\mathbf{P}(Z=k|N=3)\mathbf{P}(N=3)\right],
\end{aligned}$$

por lo tanto

$$\begin{aligned}
\mathbf{E}(Z) &= \mathbf{P}(N=1)\sum_k k\mathbf{P}(Z=k|N=1)+\mathbf{P}(N=2)\sum_k k\mathbf{P}(Z=k|N=2)\\
&+ \mathbf{P}(N=3)\sum_k k\mathbf{P}(Z=k|N=3).
\end{aligned}$$

Usando la información del item anterior sobre la distribución de la variable Z cuando N está fijo e interpretando las sumatorias como valores esperados de Z para N fijo, tenemos:

$$\begin{aligned}
\mathbf{E}(Z) &= \mathbf{P}(N=1)\,\mathbf{E}(X_1) + \mathbf{P}(N=2)\,\mathbf{E}(X_1+X_2) + \mathbf{P}(N=3)\,\mathbf{E}(X_1+X_2+X_3) \\
&= \mathbf{P}(N=1)\mathbf{E}(X_1) + \mathbf{P}(N=2)\,2\,\mathbf{E}(X_1) + \mathbf{P}(N=3)\,3\,\mathbf{E}(X_1) \\
&= [\mathbf{P}(N=1) + 2\,\mathbf{P}(N=2) + 3\,\mathbf{P}(N=3)]\ \mathbf{E}(X_1) \\
&= \mathbf{E}(N)\,\mathbf{E}(X_1).
\end{aligned}$$

Finalmente, usando

$$\mathbf{E}(N) = 1\,\mathbf{P}(N=1) + 2\,\mathbf{P}(N=2) + 3\,\mathbf{P}(N=3) = \frac{1}{3} + 2\frac{1}{3} + 3\frac{1}{3} = 2$$

y el hecho que si $X_1 \sim \mathrm{Bin}(2,p)$ entonces $\mathbf{E}(X_1) = 2p$, resulta:

$$\mathbf{E}(Z) = 4p.$$

ii) Por la Ley Fuerte de los Grandes Números, sabemos que si $Z_1, Z_2, \cdots, Z_n$ son observaciones del número total de hijos de la segunda generación, entonces:

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \longrightarrow \mathbf{E}(Z) = 4p.$$

De modo que $\widehat{p} = \frac{1}{4}\overline{Z}_n$ es un estimador consistente para el parámetro p.

d) Para $N = 81$, se tiene que $Z = X_1 + X_2 + \cdots + X_{81}$. Aplicando el Teorema Central del Límite a las variables $X_i \sim \mathrm{Bin}(2,p)$, resulta

$$\begin{aligned}
\mathbf{P}(Z \leq 100) &= \mathbf{P}(X_1 + \cdots + X_{81} \leq 100) \\
&= \mathbf{P}\left(\frac{\sqrt{81}\left(\frac{X_1+\cdots+X_{81}}{81} - 2p\right)}{\sqrt{2p(1-p)}} \leq \frac{\sqrt{81}\left(\frac{100}{81} - 2p\right)}{\sqrt{2p(1-p)}} \right) \\
&\simeq \Phi\left(\frac{\sqrt{81}\left(\frac{100}{81} - 2p\right)}{\sqrt{2p(1-p)}} \right) = \frac{1}{2}
\end{aligned}$$

Usando que $\Phi(0) = \frac{1}{2}$, estimamos p por el valor que cumple

$$\frac{100}{81} - 2p = 0,$$

esto es:

$$p = \frac{100}{2 \times 81} = 0.617$$

Problema 2

a) Dado que f es una función positiva, para que f sea una densidad de probabilidad debe cumplir $\int_{-\infty}^{\infty} f(x)dx = 1$.

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^1 axdx + \int_1^b adx = \frac{a}{2} + a(b-1) = -\frac{a}{2} + ab = 1$$

de donde

$$b = \frac{1}{a} + \frac{1}{2}$$

b)

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 ax^2dx + \int_1^b axdx = \frac{a}{3} + \frac{a}{2}(b^2-1).$$

Por ser $a > 1$ se tiene que

$$\int_0^1 axdx = \frac{a}{2} > \frac{1}{2}.$$

Por lo tanto, si m es la mediana de X, entonces debe cumplirse $m < 1$; así que calculamos m a partir de la condición:

$$\frac{1}{2} = \int_0^m axdx = \frac{am^2}{2}.$$

De modo que:

$$m = \frac{1}{\sqrt{a}}.$$

c) La mediana empírica $\widehat{m}_n$ es un estimador consistente de m, de manera que

$$\widehat{a}_n = \frac{1}{\widehat{m}_n^2}$$

es un estimador consistente del parámetro a.

Otra forma posible consiste en observar que $\widehat{b}_n = \max\{X_1, \cdots, X_n\}$ es un estimador consistente de b. Entonces, usando la parte (a), se tiene que

$$\widehat{a}_n = \frac{2}{2\widehat{b}_n - 1}$$

es un estimador consistente del parámetro a.

Problema 3

a) *Test de rachas*

El estadístico del test es el número de rachas R.

$$U_i = \begin{cases} 1 & \text{si} \quad \mathrm{X_i} \leq \mathrm{X_{i+1}} \\ 0 & \text{en otro caso} \end{cases}$$

X_i	U_i
0.98	1
2.25	0
1.19	1
1.30	1
1.73	0
1.27	0
1.20	0
0.86	1
1.38	0
0.63	1
1.57	1
1.99	

De manera que $R = 7$. Como $R = 7 < \frac{2(12)-1}{3} = 7.6667$, planteamos

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

Usando la tabla tenemos que el p-valor es $\alpha^* = 0.4453 > 0.1$. Conclusión: No se rechaza $\mathbf{H}_0$.

Test de Spearmann

El estadístico del test es:

$$R_s = 1 - 6 \frac{\sum\limits_{j=1}^{j=12} \left(R\left(X_j\right) - j\right)^2}{n\left(n^2 - 1\right)}$$

Índice$=j$	Rango$= R\left(X_j\right)$	$\left(R\left(X_j\right) - j\right)^2$
1	3	4
2	12	100
3	4	1
4	7	9
5	10	25
6	6	0
7	5	4
8	2	36
9	8	1
10	1	81
11	9	4
12	11	1

Entonces,

$$\sum_{j=1}^{j=12} \left(R\left(X_j\right) - j\right)^2 = 266$$

de donde

$$\mathbf{R_s = 0.06993}$$

Como el estadístico de Spearman dio positivo planteamos

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, ..., X_n$ son *iid*	hay tendencia creciente

Usando la tabla tenemos que $\alpha^* > 0.1$. Conclusión: No se rechaza $\mathbf{H}_0$.

No podemos rechazar que los datos sean *iid*.

b) Para testear la hipótesis de normalidad:

Test de D'Agostino

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, ..., X_n$ son normales	$X_1, ..., X_n$ no son normales

El estadístico del test es $DA_n = \dfrac{\sum_{i=1}^n \left(i - \frac{n+1}{2}\right) X_i^*}{n^2 \sigma_n}$

i	X_i^*	$\left(i - \frac{n+1}{2}\right) X_i^*$
1	0.63	-3.465
2	0.86	-3.87
3	0.98	-3.43
4	1.19	-2.975
5	1.20	-1.8
6	1.27	-0.635
7	1.30	0.65
8	1.38	2.07
9	1.57	3.925
10	1.73	6.055
11	1.99	8.955
12	2.25	12.375

$\sum_{i=1}^n \left(i - \frac{n+1}{2}\right) X_i^* = 17.855$ y es dato de la letra que $\sigma_n = \sqrt{0.197} = 0.44$. Por lo tanto

$$DA_n = 0.2818$$

y de la tabla se obtiene que el p-valor es $\alpha^* > 0.2$. Conclusión: No se rechaza $\mathbf{H}_0$

Test de Shapiro-Wilk

$\mathbf{H}_0$	$\mathbf{H}_1$
$X_1, ..., X_n$ son normales	$X_1, ..., X_n$ no son normales

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	$a_i(X_{n-i+1}^* - X_i^*)$
2.25	0.63	0.5475	0.887
1.99	0.86	0.3325	0.376
1.73	0.98	0.2347	0.176
1.57	1.19	0.1586	0.061
1.38	1.20	0.0922	0.0166
1.30	1.27	0.0303	0.0000089

Por lo tanto $\mathbf{W_n} = \mathbf{0.9638}$ y de la tabla se obtiene que el p-valor es $\alpha^* > 0.5$. Conclusión: No se rechaza $\mathbf{H}_0$

No podemos rechazar que los datos sean normales. Para este examen basta con realizar una sola prueba de hipótesis.

c) Asumiendo que los datos son normales

i) *Test sobre la varianza*
Como $s_n = 0.463$, planteamos el test:

$\mathbf{H}_0$	$\mathbf{H}_1$
$\sigma = 0.46$	$\sigma > 0.46$

El estadístico del test es:

$$E_n = \frac{(n-1)s_n^2}{\sigma^2} = 11.193$$

y de la tabla de la distribución χ^2, se obtiene que el p-valor es $0.1 < \alpha^* < 0.5$. Conclusión: No se rechaza $\mathbf{H}_0$.

ii) *Test para la media*
Realizamos un test para la media de muestras gaussianas con varianza desconocida.
Como $\overline{X}_n = 1.3625 > 1.25$ hacemos el test

$\mathbf{H}_0$	$\mathbf{H}_1$
$\mu = 1.25$	$\mu > 1.25$

El estadístico del test es:

$$E_n = \frac{\sqrt{n}(\overline{X}_n - 1.25)}{s_n}$$

con $s_n = 0.4636$, entonces $\mathbf{E_n} = \mathbf{0.8417}$ y de la tabla de la distribución t tenemos que el p-valor es $0.1 < \alpha^* < 0.25$. Conclusión: No se rechaza $\mathbf{H}_0$.

2) Podemos también resolver el problema en forma no paramétrica y hacer tests para la mediana, pues al asumir que los datos son gaussianos, media y mediana coinciden.

Test de signos

El estadístico es

$$E_n = \sum_{i=1}^{n} 1_{\{X_i \geqslant m\}} = 7$$

Entonces testeamos

$\mathbf{H}_0$	$\mathbf{H}_1$
$m = 1.25$	$m > 1.25$

y de la tabla de la distribución $\mathrm{Bin}(12, 0.5)$ el p.valor es

$$\alpha^* = 1 - \mathrm{Bin}(12, 0.5)(6) = 1 - 0.6128 = 0.3872$$

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

Test de Wilcoxon

Nuevamente planteamos el test:

$\mathbf{H}_0$	$\mathbf{H}_1$
$m = 1.25$	$m > 1.25$

El estadístico del test es:

$$E_n = \sum_{i=1}^{n} R_i^+ 1_{\{X_i - m_0 \geqslant 0\}}$$

Datos	$\|X_i - m_0\|$	$R_i^+ 1_{\{X_i - m_0 > 0\}}$
0.98	0.27	0
2.25	1	12
1.19	0.06	0
1.30	0.05	2.5
1.73	0.48	9
1.27	0.02	1
1.20	0.05	0
0.86	0.39	0
1.38	0.13	5
0.63	0.62	0
1.57	0.32	7
1.99	0.74	11

Entonces $\mathbf{E_n} = \mathbf{47.5}$. De la tabla se obtiene que el p-valor es $0.259 < \alpha^* < 0.285$. Conclusión: No se rechaza $\mathbf{H}_0$.

Tanto por un camino paramétrico (test de media para datos gaussianos) como por un camino no paramétrico (test de signos, test de Wilcoxon) concluimos que no se puede rechazar que la media de los datos es $\mu = 1.25$.