Nombre:	CI:	Carrera

EXAMEN - SOLUCIÓN

7 de Agosto de 2015

Ejercicio 1 (33 puntos)

Una piara de jabalíes vive en una región cerrada dividida al medio por un río. Este río da paso con probabilidad p en dos lugares distintos (o sea, la probabilidad que dé paso en un lugar es p y que dé paso en el otro también es p). Se sabe que si uno de los lugares da paso, la probabilidad que el otro también dé paso es $\frac{3}{4}$.

1. Mostrar que la probabilidad que el río dé paso en los dos lugares al mismo tiempo es $\frac{3}{4}p$.
2. Calcular la probabilidad de que el río dé paso en un lugar y en el otro no.
3. ¿Cuál es la probabilidad que haya paso entre las dos áreas de la región? En lo que sigue llamaremos q a esta probabilidad.
4. Se define X una variable aleatoria que indica si el río da paso o no:
   (a) Indicar si reconoce alguna distribución conocida para X.
   (b) Calcular $P(X = x)\ \forall x \in \mathcal{R}_X$ recorrido de la variable aleatoria.
5. Durante 20 días se registró si el río dió paso o no y se obtuvieron los siguientes valores (1 indica paso y 0 indica no paso).

1	0	0	0	1	1	1	0	0	1	1	0	0	1	0	0	1	0	1	0

   Dar un intervalo de confianza al nivel 95% para q.

6. A partir de los datos anteriores, dar un intervalo de confianza al 95% para p. Justifique su respuesta.

Ejercicio 1 - Solución

Se consideran los sucesos $P_1 = \{\text{el río da paso en el sitio 1}\}$ y $P_2 = \{\text{el río da paso en el sitio 2}\}$. Sabemos entonces que $P(P_1) = P(P_2) = p$ y que $P(P_2|P_1) = P(P_1|P_2) = \frac{3}{4}$.

1. La probabilidad de que el río dé paso en ambos sitios al mismo tiempo es:

$$P(P_1 \cap P_2) = P(P_2|P_1)P(P_1) = \frac{3}{4}p.$$

2. La probabilidad de que el río dé paso en un sitito y en el otro no está dada por:

$$P(P_1 \cap P_2^c) + P(P_2 \cap P_1^c) = P(P_2^c|P_1)P(P_1) + P(P_1^c|P_2)P(P_2)$$
$$(1 - P(P_2|P_1))P(P_1) + (1 - P(P_1|P_2))P(P_2) = \left(1 - \frac{3}{4}\right)p + \left(1 - \frac{3}{4}\right)p = \frac{p}{4}.$$

3. La probabiliad q de que haya paso entre las dos regiones es:

$$q = P(P_1 \cup P_2) = P(P_1) + P(P_2) - P(P_1 \cup P_2) = p + p - \frac{3}{4}p = \frac{5}{4}p.$$

4. (a) X es una variable aleatoria que toma dos valores para indicar si el río da paso o no. En particular podemos asumir que toma el valor 1 para indicar paso y el valor 0 para indicar no paso. Entonces X tiene un distribución Bernoulli de parámetro $q = P(X = 1) = P(\text{río dé paso})$.
   (b) Como dijimos antes el recorrido de la variable aleatoria X es $\mathcal{R}_X = \{0, 1\}$ donde $P(X = 1) = q$ y $P(X = 0) = 1 - q$.

5. Se pide un intervalo de confianza para el paramétro q de una v.a. Bernoulli, es decir para la proporción de veces que el río da paso. Si calculamos el intervalo de confianza a nivel 95% obtenemos que:

$$I(q) = \left[\hat{q}_n - \frac{z_{\alpha/2}\sqrt{\hat{q}_n(1-\hat{q}_n)}}{\sqrt{n}}, \hat{q}_n + \frac{z_{\alpha/2}\sqrt{\hat{q}_n(1-\hat{q}_n)}}{\sqrt{n}}\right]$$

   En este caso tenemos $n = 20$, $\hat{q}_n = \frac{9}{20} = 0.45$ y $z_{\alpha/2} = 1.96$. Por lo tanto tenemos que:

$$I(q) = [0.232, 0.0.668].$$

6. Basta observar que $q = \frac{5}{4}p$, es decir que $p = \frac{4}{5}q$ por lo tanto, un intervalo de confianza para p está dado por:

$$I(p) = \left[\frac{4}{5} \times 0.232, \frac{4}{5} \times 0.0.668\right] = [0.185, 0.534].$$

   Justificación: sabemos que $P(q \in I(q)) \geq 0.95$ por lo tanto $P(\frac{5}{4}p \in I(q)) \geq 0.95$, de donde obtenemos el intervalo calculado anteriormente ya que $P(p \in \frac{4}{5}I(q)) \geq 0.95$.

Ejercicio 2 (37 puntos)

Se considera la siguiente tabla que relaciona el genotipo $\{AA, Aa, aa\}$ de una población con la presencia o no de una cierta enfermedad E.

	AA	Aa	aa
E	0.064	0.032	0.008
E^c	0.576	0.288	0.032

La tabla indica la probabilidad conjunta de los eventos indicados en la fila y columna. Por ejemplo: la probabilidad de que un individuo presente la enfermedad y su genotipo sea AA es $P(AA \cap E) = 0.064$.

1. (a) Hallar las probabilidades de cada genotipo: AA, Aa y aa.
   (b) Asumiendo que los alelos se transmiten de manera independiente por los padres y que la proporción de padres y madres es 50/50; ¿cuál es la probabilidad que un hijo de esta generación reciba un alelo A de su padre? ¿Y un alelo a?
2. Hallar la probabilidad de que un individuo posea la enfermedad, es decir $P(E)$.
3. (a) Hallar la probabilidad de que un individuo presente la enfermedad sabiendo que su genotipo es AA.
   (b) ¿Es posible afirmar que la enfermedad es independiente del genotipo? Justifique su respuesta.
   (c) ¿Es posible afirmar que los individuos de un genotipo en particular son más propensos a poseer la enfermedad? En caso afirmativo, indicar cuál sería dicho genotipo. Justifique su respuesta.
4. En una muestra de 1000 individuos se observan que de los 95 individuos enfermos, 60 son de genotipo AA y 10 son de genotipo aa. Entre los individuos sanos, 550 son de genotipo AA mientras que 100 son de genotipo Aa. ¿Es posible afirmar que estos datos responden a las probabilidades indicadas en la tabla? Realizar un test de hipótesis no paramétrico.

Ejercicio 2 - Solución

1. (a)
$$P(AA) = P(AA \cap E) + P(AA \cap E^c) = 0.0064 + 0.576 = 0.64,$$
$$P(Aa) = P(Aa \cap E) + P(Aa \cap E^c) = 0.032 + 0.288 = 0.32,$$
$$P(aa) = P(aa \cap E) + P(aa \cap E^c) = 0.008 + 0.032 = 0.04.$$
   (b) De la información brindada, podemos asumir que $P(AA) = P(A)^2$ y $P(aa) = P(a)^2$. Por lo tanto $P(A) = \sqrt{0.64} = 0.8$ y $P(a) = \sqrt{0.04} = 0.2$. Observar que se verifica la igualdad $P(Aa) = 2P(A)P(a)$.

   Otra forma de pensarlo es observar las siguientes igualdades:
$$P(A) = P(AA) + \frac{1}{2}P(Aa) \quad \text{y} \quad P(a) = P(aa) + \frac{1}{2}P(Aa)$$
2. Al igual que en la primera parte,
$$P(E) = P(E \cap AA) + P(E \cap Aa) + P(E \cap aa) = 0.064 + 0.032 + 0.08 = 0.104.$$
3. (a)
$$P(E|AA) = \frac{P(E \cap AA)}{P(A)} = \frac{0.064}{0.64} = 0.1.$$
   (b) Es claro que no son independientes pues $P(E|AA) \neq P(E)$.
   (c) $P(E|Aa) = \frac{P(E \cap Aa)}{P(Aa)} = \frac{0.032}{0.32} = 0.1$ y $P(E|aa) = \frac{P(E \cap aa)}{P(aa)} = \frac{0.008}{0.04} = 0.2$. Por lo tanto los individuos con genotipo aa son más propensos a presentar la enfermedad.

4. A partir de los datos podemos construir la siguiente tabla de contingencia:

	AA	Aa	aa
E	60	25	10
E^c	550	100	255

Para poder comparar con la tabla anterior, vamos a pasar a probabilidades dividiendo por el número de muestras:

	AA	Aa	aa
E	0.06	0.025	0.01
E^c	0.55	0.1	0.255

Planteamos entonces un test χ-cuadrado de ajuste para saber si ambas distribuciones coinciden:

$$\begin{cases} H_0: & \text{La distribuciones coinciden} \\ H_1: & \text{No } H_0 \end{cases}$$

El estadístico está dado por $E = \sum_{i=1}^{r} n\frac{(\hat{p}_i - p_i)^2}{p_i}$ donde p_i es la probabilidad de la categoría i bajo el modelo propuesto y $\hat{p}_i$ son las observadas en la muestra. La cantidad de intervalos en este caso es $r = 6$. La región crítica es $\mathcal{R}_\alpha = \left\{E \geq \chi^2_\alpha(r-1)\right\}$.

En este caso, el resultado es $E = 1000 \times 1.68 = 1680$. Buscando en la tabla de χ^2 con 5 grados de libertad se tiene que el p-valor es menor o igual que 0.001. Por lo tanto se rechaza H_0 y no es razonable suponer que las distribuciones coinciden. Este resultado es coherente con el hecho de que hay dos categorías (Aa, E^c) y (aa, E^c) cuyas probabilidades son bien distintas.

Ejercicio 3 (30 puntos)

Un índice normalizado de esquizofrenia viene dado por una variable X cuya función de distribución se asume es la siguiente:

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x^2 & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

1. Calcular $\mathrm{E}(X)$ y $\mathrm{Var}(X)$.
2. Se define mediana de una variable X continua como el valor m_X tal que $P(X \leq m_X) = P(X \geq m_X) = 1/2$. Calcular m_X en este caso.
3. Calcular las siguientes probabilidades: $P(X \leq 1/4)$, $P(X \geq 3/4)$ y $P(1/2 < X \leq 3/4)$.
4. Se tienen los siguientes datos, que pueden suponerse aleatorios:

0.95	0.62	0.047	0.61	0.67	0.88	0.71	0.58	0.8	0.69

¿Es razonable suponer que los datos corresponden a la distribución definida al principio del ejercicio? Realizar un test de hipótesis no paramétrico.

5. Un paciente es peligroso si su índice normalizado es mayor que 0.9. ¿Cuál es el número esperado de pacientes peligrosos entre 150 elegidos al azar en forma independiente? Justifique su respuesta.

Ejercicio 3 - Solución

1.
$$\mathrm{E}(X) = \int\limits_0^\infty 1 - F_X(x)dx = \int\limits_0^1 1 - x^2 dx = 1 - \left.\frac{x^3}{3}\right|_0^1 = 1 - \frac{1}{3} = \frac{2}{3}$$

También es posible calcular la esperanza a partir de la función densidad f_X que se obtiene derivando la función de distribución, esto es:

$$f_X(x) = \begin{cases} 2x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}(X^2) - \frac{4}{9}$$

Basta calcular $\mathrm{E}(X^2)$:

$$\mathrm{E}(X^2) = \int\limits_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^1 2x^3 dx = 2\left.\frac{x^4}{4}\right|_0^1 = \frac{1}{2}$$

.

Por lo tanto

$$\mathrm{Var}(X) = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}.$$

2. La mediana $m_X \in [0,1]$ verifica que $F_X(m_X) = m_X^2 = \frac{1}{2}$, por lo tanto $m_X = \frac{1}{4}$.

3.
$$P(X \leq 1/4) = F_X(1/4) = \left(\frac{1}{4}\right)^2 = \frac{1}{16}$$
$$P(X \geq 1/4) = 1 - F_X(3/4) = 1 - \left(\frac{3}{4}\right)^2 = 1 - \frac{9}{16} = \frac{7}{16}$$
$$P(1/2 < X \leq 1/4) = F_X(3/4) - F_X(1/2) = \left(\frac{3}{4}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{5}{16}$$

4. Realizaremos un test de ajuste de Kolmogorov-Smirnov a $F_0 = F_X$ la distribución definida al comienzo del ejercicio:

$$\begin{cases} H_0 : \text{La muestra se ajusta a } F_0 \\ H_1 : \text{No } H_0 \end{cases}$$

X_i^*	$F_o(X_i^*)$	$\frac{i}{n}$	$\frac{i-1}{n}$	$\lvert F_o(X_i^*) - \frac{i}{n}\rvert$	$\lvert F_o(X_i^*) - \frac{i-1}{n}\rvert$
0.047	0.022	0.1	0	0.098	0.002
0.58	0.336	0.2	0.1	0.136	0.236
0.61	0.372	0.3	0.2	0.072	0.172
0.62	0.384	0.4	0.3	0.016	0.084
0.67	0.448	0.5	0.4	0.051	0.049
0.69	0.476	0.6	0.5	0.124	0.024
0.71	0.504	0.7	0.6	0.196	0.096
0.8	0.64	0.8	0.7	0.16	0.06
0.88	0.774	0.9	0.8	0.126	0.026
0.95	0.902	1	0.9	0.097	0.0025

El estadístico Kolomogorov Smirnov resulta entonces $KS = 0.236$. La tabla indica que para $n = 10$ el p-valor es mayor que 0.2. Por lo tanto no es posible rechazar la hipótesis nula. Es decir podemos asumir que la muestra se ajusta a la distribución propuesta.

5.
$$P(\text{sujeto peligroso}) = P(X \geq 0.9) = 1 - F_X(0.9) = 1 - 0.9^2 = 0.19$$

Por lo tanto en una muestra de 150 personas, la cantidad esperada de sujetos peligrosos es $150 \times 0.19 \sim 28$.

Nota: en toda la prueba se asume como p-valor $\alpha^* = 0.1$.