Nombre:	**CI:**	**Carrera**

SOLUCIÓN DEL EXAMEN

12 de Febrero de 2015

Ejercicio 1 (25 puntos)

Se considera un gen bi-alélico con dos alelos A y a. Se asume que el alelo a es recesivo y es causante de una cierta enfermedad (sólo personas con genotipo aa tienen la enfermedad). Se definen los siguientes eventos de interés:

- $E = \{\text{enfermedad}\} = \{aa\}$
- $S = \{\text{no enfermedad}\} = \{Aa, AA\}$
- $HM = \{\text{homocigoto}\} = \{AA, aa\}$
- $HT = \{\text{heterocigoto}\} = \{Aa\}$

Se tienen los siguientes datos: $P(AA) = 0.49$, $P(Aa) = 0.42$ y $P(aa) = 0.09$.

1. ¿Cuál es la probabilidad de cada uno de los 4 eventos antes definidos?
2. ¿Cuál es la probabilidad de ser heterocigoto y no estar enfermo? ¿y cuál la de ser heterocigoto y estar enfermo?
3. Calcular las probabilidades $P(HM \cup S)$ y $P(HM \cup E)$.
4. Sabiendo que el gen es homocigoto ¿cuál es la probabilidad de estar enfermo? ¿y cuál la de no estar enfermo?
5. ¿Los eventos HM y E son independientes? Justifique su respuesta.

Solución Ejercicio 1

1. $P(E) = P(aa) = 0.09$

   $P(S) = P(Aa) + P(AA) = 0.91$ o bien $P(S) = 1 - P(E) = 0.91$

   $P(HM) = P(HM \cap S) + P(HM \cap E) = P(AA) + P(E) = 0.58$

   $P(HT) = P(Aa) = 0.42$

2. La probabilidad de ser heterocigoto y no estar enfermo es $P(HT \cap S) = P(S) = 0.42$.

   La probabilidad de ser heterocigoto y estar enfermo es $P(HT \cap E) = P(\emptyset) = 0$.

3. $$P(HM \cup S) = P(HM) + P(S) - P(HM \cap S) = 0.58 + 0.91 - 0.49 = 1$$
   $$P(HM \cup E) = P(HM) + P(E) - P(HM \cap E) = 0.58 + 0.09 - 0.09 = 0.58$$

4. Si se sabe que el gen es homocigoto la probabilidad de estar enfermo es:

$$P(E|HM) = \frac{P(E \cap HM)}{P(HM)} = \frac{0.09}{0.58} = 0.15517.$$

Por otra parte, si se sabe que el gen es homocigoto la probabilidad de no estar enfermo es:

$$P(S|HM) = \frac{P(S \cap HM)}{P(HM)} = \frac{0.49}{0.58} = 0.8448.$$

5. Los eventos HM y E *no* son independientes ya que $P(E \cap HM) \neq P(E)P(HM)$.

Ejercicio 2 (35 puntos)

Se considera un bolillero contiene 100 bolillas que pueden ser rojas, blancas o azules. Se sabe que hay 30 bolillas rojas y 40 blancas. El resto son azules. Se selecciona al azar con **reposición** una muestra con 10 bolillas.

Se definen las siguientes variables aleatorias:

- $X = \{$número de bolillas rojas observadas en la muestra$\}$
- $Y = \{$número de bolillas blancas observadas en la muestra$\}$

1. Identificar la distribución de X e Y e indicar sus funciones de probabilidad.
2. Calcular E(X) y Var(X).
3. (a) Calcular E($X + Y$).
   (b) Sabiendo que la covarianza entre X e Y es Cov$(X, Y) = 1.2$, calcular Var$(X + Y)$.
   (c) ¿ Las variables X e Y son independientes? Justifique su respuesta.
4. Calcular $P(X + Y = k)$ para los valores $k = 0, 1, 2$.
5. Si se repite 2500 veces de manera independiente la experiencia de anterior y se obtiene una cantidad promedio de bolillas rojas de 3.2 y de 4.6 de bolillas blancas, ¿le parece creíble que en el bolillero haya más de 35 bolillas azules? Justifique la respuesta.

Solución Ejercicio 2

1. Dado que X es la variable aleatoria que cuenta el número de bolillas rojas observadas en la muestra de tamaño 10 y el muestreo es sin reposición se tiene que:

$$X \sim Bin\left(3, \frac{3}{10}\right)$$

es decir X es Binomial con parámetros $n = 10$ y $p_r = \frac{3}{10}$ (probabilidad de que salga roja). Análogamente:

$$Y \sim Bin\left(3, \frac{4}{10}\right)$$

es decir Y es Binomial con parámetros $n = 10$ y $p_b = \frac{4}{10}$ (probabilidad de que salga blanca)

2. Como $X \sim Bin(n, p_r)$ se tiene que:

$$E(X) = np_r = 10\frac{3}{10} = 3$$

$$\text{Var}(X) = np_r(1-p_r) = 10\frac{3}{10}\frac{7}{10} = 2,1$$

3. (a) $E(X+Y) = E(X) + E(Y) = 3 + 10\frac{4}{10} = 7$
   (b) $\text{Var}(X+Y) = \text{Var(X)} + \text{Var(Y)} + 2\text{Cov(X, Y)} = 2,1 + 2,4 + 2 \times 1,2 = 6,9$
   (c) X e Y no son independientes, pues $\text{Cov}(X, Y) \neq 0$

4. Sea Z la variable aleatoria que cuenta el número de bolillas azules observadas en la muestra de tamaño 10, al igual que para X e Y, se tiene que:

$$Z \sim Bin(n, p_a)$$

   con $n = 10$ y $p_a = \frac{3}{10}$ (probabilidad de que salga azul).
   Se tiene que:

$$P(X+Y=0) = P(Z=10) = C_{10}^{10} p_a^{10}(1-p_a)^0 = \left(\frac{3}{10}\right)^{10} = \left(\frac{3}{10}\right)^{10} = 5,9049 \times 10^{-6}$$

$$P(X+Y=1) = P(Z=9) = C_9^{10} p_a^9(1-p_a)^1 = 10\left(\frac{3}{10}\right)^9 \frac{7}{10} = 1,37781 \times 10^{-4}$$

$$P(X+Y=2) = P(Z=8) = C_8^{10} p_a^8(1-p_a)^2 = 10 \times 9 \times \left(\frac{3}{10}\right)^8 \left(\frac{7}{10}\right)^2 = 0,0014467$$

   También es posible resolver el ejercicio calculando las probabilidades pedidas como casos favorables sobre casos posibles.

5. Como en promedio la cantidad de bolillas rojas es de 3,2 y de blancas es 4,6 se tiene que la cantidad promedio de azules es 2,2. Como el promedio es un estimador del valor esperado, se espera que entre todas las bolillas tengamos 22 azules. Por lo cual es poco creíble que en el bolillero haya más de 35 bolillas azules.

Ejercicio 3 (40 puntos)

En un laboratorio de análisis clínicos se decide estudiar el flujo de pacientes que llegan a hacerse el análisis **A**. Como los análisis llevan aproximadamente 20 minutos en hacerse, si llegan en promedio 4 o más pacientes por hora, se debe incrementar el equipo que realiza dicho análisis. La recepcionista estima que llegan en promedio unos 3 pacientes por hora, por lo que no sería necesario agregar personal.

1. Para verificar la suposición de la recepcionaista, se le pide que registre el número de pacientes que llegan durante 1 hora, en 5 horas elegidas al azar. Los datos obtenidos son:

X_1	X_2	X_3	X_4	X_5
6	4	0	2	3

¿Es razonable suponer que estos datos representan una muestra independiente e idénticamente distribuida? Realice dos test de hipótesis.

2. Si el laboratorio quisiera testear la hipótesis nula de que los datos obtenidos corresponden a una distribución de Poisson de parámetro $\lambda = 3$ sobre la base de los datos anteriores, ¿cuál sería su conclusión? Realice un test de hipótesis.

3. El laboratorio acepta que es una distribución de Poisson, pero quiere tener más certeza sobre el valor del parámetro. Con los datos de la parte 1, el laboratorio decide testear si tiene evidencia suficiente para rechazar la hipótesis que en promedio llegan 4 pacientes por hora (hipótesis nula) frente a la alternativa que llegan menos. Plantee un test de hipótesis (hipótesis nula, hipótesis alternativa y región crítica) para este problema.

4. La dirección decide testear si el equipo que realiza un segundo análisis **B** tiene más trabajo que el equipo anterior (análisis **A**). Como a la recepcionista le encanta registrar todo, ya tenía registrado la cantidad de pacientes promedio por hora que llegaron durante otras 5 horas para realizarse el análisis **B**. Los datos obtenidos son:

Y_1	Y_2	Y_3	Y_4	Y_5
5	2	1	9	3

¿Qué conclusión saca la dirección del laboratorio de la comparación de los datos obtenidos? Realice un test de hipótesis. Los datos pueden suponerse independientes.

Para el ejercicio 3, se asume como p-valor $\alpha^* = 0.1$.

Solución Ejercicio 3

1. Test de Rachas: El estadístico que se obtiene es $R = 2 < \frac{2n-1}{3} = \frac{9}{3}$.
   Si se plantea el test a una cola:

$$\begin{cases} H_0: \text{ La muestra es aleatoria} \\ H_1: \text{ La muestra presenta pocas rachas} \end{cases}$$

   Entonces la tabla da directamente el p-valor $\alpha^* = 0.25 > 0.1$. Por lo tanto, no es posible rechazar la hipótesis nula.

   La conclusión es la misma si se plantea un test a dos colas, es decir:

$$\begin{cases} H_0: \text{La muestra es aleatoria} \\ H_1: \text{La muestra no es aleatoria} \end{cases}$$

   Test de Spearman: el vector posición (o vector de rangos) es $R_i = (5, 4, 1, 2, 3)$. Por lo tanto el estadístico es:

$$R_S = 1-6\frac{\sum_{i=1}^{5}(R_i-i)^2}{N\times(N^2-1)} = 1-6\frac{\sum_{i=1}^{5}(R_i-i)^2}{6\times 35} = 1-\frac{6}{5*24}(16+4+4+4+4) = 1-\frac{6*32}{5*24} = -0.6$$

Si se plantea el test a una cola:

$$\begin{cases} H_0: \text{ La muestra es aleatoria} \\ H_1: \text{ Hay dependencia negativa} \end{cases}$$

Entonces la tabla da directamente el p-valor $\alpha^* = 0.175 > 0.1$. Por lo tanto, no es posible rechazar la hipótesis nula.

La conclusión es la misma si se plantea un test a dos colas, es decir:

$$\begin{cases} H_0: \text{La muestra es aleatoria} \\ H_1: \text{La muestra no es aleatoria} \end{cases}$$

Visto los resultados del test de Rachas y el test de Correlación de Rangos de Spearman, podemos suponer que la muestra es aleatoria (i.i.d).

2. Vamos a realizar el test de ajuste de Kolmogorov-Smirnov a la distribución propuesta F_0 Poisson de parámetro $\lambda = 3$:

$$F_0(k) = e^{-\lambda}\sum_{i=0}^{k}\frac{\lambda^i}{i!}$$

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

X_i^*	$F_o(X_i^*)$	$\frac{i-1}{n}$	$\frac{i}{n}$	$\lvert F_o(X_i^*) - \frac{i}{n-1}\rvert$	$\lvert F_o(X_i^*) - \frac{i}{n}\rvert$
0	0.05	0.00	0.20	0.05	0.15
2	0.42	0.20	0.40	0.22	0.02
3	0.65	0.40	0.60	0.25	0.05
4	0.82	0.60	0.80	0.22	0.02
6	0.97	0.80	1.00	0.17	0.03

El estadístico Kolomogorov Smirnov resulta entonces $KS = 0.25$. La tabla indica que 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.

3. Para testear la hipótesis el arribo de pacientes se realiza

$$\begin{cases} H_0: \lambda = 4 \\ H_1: \lambda < 4 \end{cases}$$

Es razonable suponer que $\lambda = 4$ si el promedio muestral $\frac{1}{n}\sum_{i=1}^{n} X_i$ es cercano a 4. Cómo la hipótesis alternativa es $\lambda < 4$, la región crítica debe ser:

$$\mathcal{R}_\alpha = \left\{ \frac{1}{n}\sum_{i=1}^{n} X_i \leq C_\alpha \right\}$$

donde C_α es tal que $P_{H_0}(\mathcal{R}_\alpha) = 1 - \alpha$.

En este caso como los datos son pocos ($n = 5$), no es posible utilizar una aproximación gaussiana para el estadístico. Sin embargo, utilizando la sugerencia se tiene que $\sum_{i=1}^{5} X_i$ tiene también distribución Poisson de parámetro $\hat{\lambda} = 20$ (ya que bajo H_0 cada X_i es Poisson de parámetro $\lambda = 4$.

4. Para testear si ambas muestras tienen la misma distribución, vamos a realizar un test de Kolmogorov-Smirnov de comparación de dos muestras.

$$\begin{cases} H_0 : X \sim Y \\ H_1 : \text{No } \ H_0 \end{cases}$$

Z_i^*	Muestra	$F_m^*(Z^*)$	$F_n^*(Z^*)$	$\lvert F_m^*(Z^*) - F_n^*(Z^*)\rvert$
0	X	0.20	0.00	0.20
1	Y	0.20	0.20	0
2	X/Y	0.40	0.40	0
3	X/Y	0.60	0.60	0
4	X	0.80	0.60	0.20
5	Y	0.80	0.80	0
6	X	1.00	0.80	0.20
9	Y	1.00	1.00	0

El estadístico de Kolmogorov Smirnov es $mnD_{mn} = 5 \times 5 \times \frac{1}{5} = 5$. La tabla indica que el p-valor es mayor que 0.1. Por lo tanto no es posible rechazar la hipótesis nula. Es decir, podemos asumir que la distribución de trabajo en el análisis A coincide con la del análisis B.