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

EXAMEN SOLUCIÓN - 21 de Julio de 2015

Ejercicio 1 (33 puntos)

Se sabe que en cierto país europeo la probabilidad de que a un hombre con más de 50 años le sea diagnósticado cáncer de próstata (CP) es de 0.00549.

1. (a) Para una muestra aleatoria de 1000 hombres mayores a 50 años de edad, ¿cuál es la probabilidad de que a ningún hombre le sea diagnósticado CP?
   (b) ¿Y de que a exactamente un hombre le sea diagnósticado CP?
   (c) ¿Cuántos hombres diagnósticados con CP esperaría encontrar en esta muestra?

   Sea X el número de hombres de la muestra diagnósticados con CP. Entonces, X sigue una distribución Binomial con paramétros $n = 1000$ y $p = 0.00549$. Entonces, la función de probabilidad de X está dada por:

$$P(X = k) = C_k^n p^k (1-p)^{n-k} \quad \forall\, k = 0, \ldots, n$$

   (a) $P(X = 0) = (1-p)^{1000} = 0.00406$.
   (b) $P(X = 1) = 1000 \times p \times (1-p)^{999} = 0.0224$
   (c) El número esperado de hombres diagnósticados con CP es $\mathrm{E}(X) = np = 1000 \times 0.00549 = 5,49$

2. Recientemente se encontró que la presencia de una cierta proteína podría ser un indicador para un diagnóstico no invasivo del cáncer de próstata. Basados en un cierto número de estudios, se estima que dicha proteína está presente en 3.4% del total de hombres mayores a 50 años de edad y que además:

   - la proteína está presente en 83% de los hombres diagnosticados con CP
   - la proteína no está presente en 97% de los hombres que no han presentado signos de CP

   ¿Le parece razonable el uso de este test de forma masiva, por ejemplo midiendo una vez al año la presencia de esta proteína en todos los hombres mayores que 50 años? Para contestar la pregunta, calcular

   (a) la probabilidad de ser diagnósticado con CP dado que la proteína está presente
   (b) la probabilidad de no ser diagnósticado con CP dado que la proteína no fue encontrada

   Se definen los siguientes eventos:

$$\begin{aligned} A &= \{ \text{ la proteína está presente en hombres mayores a 50 años } \} \\ B &= \{\text{la proteína está presente en hombres diagnósticados con CP}\} \\ C &= \{\text{la proteína no está presente en hombres no diagnósticados con CP}\} \end{aligned}$$

   Los datos nos dicen que: $P(A) = 0.034$, $P(B) = 0.83$ y $P(C) = 0.97$.

(a) Se define el evento $DCP = \{\text{hombres diagnósticados con CP}\}$, entonces nos piden hallar

$$P(DCP|A) = \frac{P(A|DCP)P(DCP)}{P(A)} = \frac{P(B)P(DCP)}{P(A)} = \frac{0.83 \times 0.00549}{0.034} = 0.134.$$

(b) Nos piden hallar

$$P(A^c|DCP^c) = \frac{P(DCP^c|A^c)P(DCP^c)}{P(A^c} = \frac{P(C)(1 - P(DCP))}{1 - P(A)}$$
$$= \frac{0.97 \times (1 - 0.00549)}{1 - 0.0034} = 0.999.$$

Considerando la respuesta a la pregunta 1, tenemos que la probabilidad de ser dignósticado con CP dado que la proteína está presente es de apenas 0.134. Por lo tanto tendremos alrededor de 13% de casos que resulten en un diagnóstico positivo. No resulta entonces recomendable la aplicación masiva de este test.

3. En un país no europeo, se realiza el mismo estudio sobre 900 hombres mayores que 50 años y se encuentran 4 hombres diagnósticados con cáncer de próstata. ¿Es posible afirmar que la probabilidad de ser diagnósticado CP es la misma que en el país europeo? Utilizar un nivel de confianza del 95%.

   Es necesario realizar un test de hipótesis o un intervalo de confianza para proporciones. Si calculamos el intervalo de confianza a nivel 95% obtenemos que:

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

   En este caso tenemos $n = 900$, $\hat{p}_n = \frac{4}{900} = 0.044$ y $z_{\alpha/2} = 1.96$. Por lo tanto tenemos que:

$$I(p) = [0.000076, 0.0087].$$

   Podemos afirmar entonces que la referida probabilidad es la misma en ambos países.

Ejercicio 2 (37 puntos)

Se considera el caudal de un río medido en metros cúbicos por segundo. Si se asume que este caudal es aleatorio, es posible modelarlo a través de una variable aleatoria X con distribución uniforme. Más aún, normalizando los paramétros es posible asumir que $X \sim \mathcal{U}[0, 1]$. Por lo tanto si medimos Y el caudal de un río con dos afluentes (el caudal resultante es la suma de los caudales de sus dos afluentes, que se asumen independientes), entonces la distribución resultante está dada por la siguiente función f_Y correspondiente a la densidad de la suma de dos variables aleatorias uniformes independientes:

$$f_Y(x) = \begin{cases} x & \text{si} \quad 0 \leq x \leq 1, \\ 2 - x & \text{si} \quad 1 \leq x \leq 2 \\ 0 & \text{en otro caso.} \end{cases}$$

1. Verificar que f_Y es una función de densidad.
2. Calcular $\mathrm{E}(Y)$ valor esperado y $\mathrm{Var}(Y)$ varianza de Y.
3. Hallar F_Y función de distribución de dicha variable aleatoria Y.
4. Hallar la probabilidad de que Y sea mayor o igual que $1/2$.

   (a) Dado que la función es positiva, basta verificar que la integral de la función es igual a 1. El modo más simple es observar que la densidad es triangular y que por lo tanto la integral es el área de un triángulo de base 2 y altura 1.

   (b) Buscamos $F_Y(y) = P(Y \leq y), \forall\, y \in \mathbb{R}$. Es claro que $F_Y(y) = 0, \forall\, y \leq 0$ y que $F_Y(y) = 1, \forall\, y \geq 2$

   Para $y \in [0,1]$: $F_Y(y) = \int_0^y x dx = \frac{y^2}{2}$.

   Para $y \in [1,2]$: $F_Y(y) = \int_0^1 x dx + \int_1^y 2 - x dx = -\frac{y^2}{2} + 2x - 1$.

5. Se considera la siguiente muestra del caudal de un río en las condiciones anteriores (suma de dos caudales uniformes independientes):

0.78	1.45	1.01	0.98	1.39	0.91	0.65	0.74	0.4	1.03

   (a) Verificar por medio de dos tests de hipótesis que la muestra puede considerarse aleatoria.

   (b) Realizar un test de ajuste de los datos a la distribución hallada en la parte 2. ¿Confirma la hipótesis asumida sobre la distribución de la suma de caudales independientes?

   (a) Realizaremos los test de Rachas y de Correlación de Rangos de Spearman.
   Test de Rachas: El estadístico que se obtiene es $R = 7 > \frac{2n-1}{3} = \frac{11}{3}$.
   Si se plantea el test a una cola:

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

   Entonces la tabla da directamente el p-valor $\alpha^* = 0.4524 > 0.1$. 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 $R_i = (4, 10, 7, 6, 9, 5, 2, 3, 1, 8)$.
   Por lo tanto el estadístico es:

$$R_S = 1 - 6\frac{\sum_{i=1}^{6}(R_i - i)^2}{N \times (N^2 - 1)} = 1 - 6\frac{\sum_{i=1}^{6}(R_i - i)^2}{990} = 1 - \frac{6}{990}228 = -0.38$$

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 $\alpha^* = 0.139 > 0.1$. 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:

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

(b) Realizaremos un test de ajuste de Kolmogorov-Smirnov a F_0 la distribución hallada en la parte (2):

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

X_i^*	$F_o(X_i^*)$	$\frac{i}{n}$	$\frac{i-1}{n}$	$\lvert F_o(X_i^*) - \frac{i}{n}\rvert$	$\lvert F_o(X_i^*) - \frac{i-1}{n}\rvert$
0.4	0.08	0.1	0.0	0.02	0.08
0.65	0.21	0.2	0.1	0.01	0.11
0.74	0.27	0.3	0.2	0.03	0.07
0.78	0.30	0.4	0.3	0.1	0
0.91	0.41	0.5	0.4	0.09	0.01
0.98	0.48	0.6	0.5	0.12	0.02
1.01	0.51	0.7	0.6	0.19	0.09
1.03	0.53	0.8	0.7	0.27	0.17
1.39	0.81	0.9	0.8	0.09	0.01
1.45	0.85	1.0	0.9	0.15	0.05

El estadístico Kolomogorov Smirnov resulta entonces $KS = 0.27$. 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.

Ejercicio 3 (30 puntos)

La malnutrición es un rasgo común en pacientes con enfermedades crónicas de riñon. Un indicador de malnutrición es el nivel bajo de albúminas en sangre. Médicos interesados en este fenómeno, registraron las siguientes medidas de este indicador en 12 pacientes de un centro de diálisis.

15.5	12.9	9.6	10.5	8.5	9.0	7.1	7.5	9.7	7.6

1. Calcular la media y la varianza muestrales de los niveles de albúminas.

2. Dar un intervalo de confianza al 95% para la media del nivel de albúminas en los siguientes casos:

   (a) asumiendo que el tamaño de la muestra es grande,

   (b) asumiendo que la distribución de los datos es gaussiana y los datos son pocos.

   (a) $\bar{X_n} = 9.79$ y $\sigma_n^2 = 6.24$.

   (b) Intervalos de confianza aproximado:

      i.

$$I(\mu) = \left[\bar{X_n} - \frac{z_{\alpha/2}\sigma_n}{\sqrt{n}}, \bar{X_n} + \frac{z_{\alpha/2}\sigma_n}{\sqrt{n}}\right] = [8.38, 11.2]$$

      ii. Intervalo de confianza exacta para muestras gaussianas:

$$I(\mu) = \left[\bar{X_n} - \frac{t_{\alpha/2}(n-1)S_n}{\sqrt{n}}, \bar{X_n} + \frac{t_{\alpha/2}(n-1)S_n}{\sqrt{n}}\right] = [8.44, 11.14]$$

3. Se propone un tratamiendo basado en una dieta rica en hierro. Los niveles medidos luego de tres meses de aplicado el tratamiento son los siguientes:

9.2	10.3	8.0	7.5	9.3	13	10

   ¿Es razonable suponer que el tratamiento es efectivo? Realizar un test de hipótesis no paramétrico.

1. Sea A la muestra antes del tratamiento y D después del tratamiento. Para testear si ambas muestras tienen la misma distribución, vamos a realizar un test de Kolmogorov-Smirnov de comparación de dos muestras.

$$\begin{cases} H_0 : F_A \sim F_D \\ H_1 : \text{No } \ H_0 \end{cases}$$

Z_i^*	Muestra	$F_m^*(Z^*)$	$F_n^*(Z^*)$	$\lvert F_m^*(Z^*) - F_n^*(Z^*)\rvert$
7.1	X	0.083	0.0	0.08
7.5	X	0.16	0.0	0.17
7.5	Y	0.16	0.14	0.03
7.6	X	0.25	0.14	0.11
8	Y	0.25	0.28	0.04
8.5	X	0.33	0.28	0.04
9	X	0.42	0.28	0.13
9.2	Y	0.42	0.43	0.01
9.3	Y	0.42	0.57	0.15
9.6	X	0.5	0.57	0.07
9.7	X	0.58	0.57	0.01
10	Y	0.58	0.71	0.13
10.3	Y	0.58	0.86	0.28
10.5	X	0.66	0.86	0.19
12.9	X	0.75	0.86	0.11
13	Y	0.75	1.0	0.25
15.5	X	1.0	1.0	0.00

El estadístico de Kolmogorov Smirnov es $mnD_{mn} = 7 \times 10 \times 0.28 = 19,6$. Dado que $m = 7$ y $= 10$, las tablas nos indican un intervalo que contiene al p-valor dado por los p-valores hallado asumiendo $m = n = 7$ y $m = n = 10$. En el primer caso se tiene que $\alpha* > 0.212$, mientras que en el segundo se obtiene que $\alpha^* > 0.2$. Por lo tanto podemos concluir que el p-valor es mayor que 0.1. Al no poder rechazar la hipótesis nula, asumimos que la distribución no cambia luego del tratamiento y por lo tanto el mismo no es efectivo.

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

Fórmula que puede ser de utilidad

1. $\int_a^b x^\alpha = \frac{x^{\alpha+1}}{\alpha+1}\Big|_a^b$,