Tarea 3 - Análisis de datos
Giovanni Gamaliel López Padilla

Problema 3

Sea X una v.a. con la siguiente distribución. Calcula Var(X^2) y E(X^2 | X>1)

	1	2	3	4
P(X=x)	0.2	0.1	0.4	0.3

Se tiene que:

$$Var(X^2) = E(X^4) - (E(X^2))^2 \tag{1}$$

Calculando $E(X^2)$, se tiene que:

$$\begin{aligned} E(X^2) &= \sum_i x_i^2 P(x_i) \\ &= 1^2(0.2) + 2^2(0.1) + 3^2(0.4) + 4^2(0.3) \\ &= 9 \end{aligned}$$

Calculando $E(X^4)$, se tiene que:

$$\begin{aligned} E(X^4) &= \sum_i x_i^2 P(x_i) \\ &= 1^4(0.2) + 2^4(0.1) + 3^4(0.4) + 4^4(0.3) \\ &= 111 \end{aligned}$$

Entonces:

$$\begin{aligned} Var(X^2) &= 111 - 9^2 \\ &= 111 - 81 \\ &= 30 \end{aligned}$$

por lo tanto, para la distribución dada se tiene:

$$Var(X^2) = 30$$

Definimos al conjunto A como el siguiente:

$$A = \{x : x > 1\} \tag{2}$$

Usando la distribución dada, entonces el conjunto A esta constituido de la siguiente manera:

$$A = \{2, 3, 4\}$$

Calculando $E(X^2|X > 1)$, se obtiene lo siguiente:

$$E(X^2|X > 1) = \sum_i x_i^2 P(X = x_i|X > 1) \tag{3}$$

Si $x_i \in A$, entonces se obtiene lo siguiente:

$$\begin{aligned} P(X = x_i|X > 1) &= \frac{P(x_i, X > 1)}{P(X > 1)} \\ &= \frac{P(x_i)}{P(X > 1)} \end{aligned}$$

Esto es porque la interseccion de x_i con A es el mismo x_i. En el caso contrario $x_i \notin A$, entonces $P(X = x_i|X > 1) = 0$, ya que no existira intersección entre x_i y A. Por lo tanto:

$$P(X = x_i|X > 1) = \begin{cases} 0 & \text{para} \quad x_i \notin A \\ \frac{P(X=x_i)}{P(A)} & \text{para} \quad x_i \in A \end{cases} \tag{4}$$

Calculando $P(X > 1)$, se obtiene lo siguiente:

$$\begin{aligned} P(X > 1) &= P(X = 2) + P(X = 3) + P(X = 4) \\ &= 0.1 + 0.4 + 0.3 \\ P(X > 1) &= 0.8 \end{aligned} \tag{5}$$

Usando las ecuaciones 4 y 5 en la expansión de la ecuación 3 se obtiene lo siguiente:

$$\begin{aligned} E(X^2|X > 1) &= \sum_i x_i^2 P(X = x_i|X > 1) \\ &= 1^2 P(X = 1|X > 1) + 2^2 P(X = 2|X > 1) + 3^2 P(X = 3|X > 1) + 4^2 P(X = 1|X > 1) \\ &= 4\left(\frac{P(X = 2)}{P(A)}\right) + 9\left(\frac{P(X = 3)}{P(A)}\right) + \left(16\frac{P(X = 4)}{P(A)}\right) \\ &= \frac{1}{P(A)}\left(4P(X = 2) + 9P(X = 3) + 16P(X = 4)\right) \\ &= \frac{1}{0.8}\left(4(0.1) + 9(0.4) + 16(0.3)\right) \\ &= \frac{1}{0.8}\left(8.8\right) \\ E(X^2|X > 1) &= 11 \end{aligned}$$

Por lo tanto:

$$E(X^2|X > 1) = 11$$

Problema 4

Verifica que si X y Y son independientes:

$$H(X, Y) = H(X) + H(Y)$$

Partiendo de la definición de entropia se tiene que:

$$H(X, Y) = -\sum_i \sum_j P(x_i, y_j) \log_2 \left(P(x_i, y_j)\right) \tag{6}$$

Calculando $P(x_i, y_j)$, se tiene que:

$$P(x_i, y_j) = P(y_j)P(x_i)$$

Entonces, la ecuación 6 se puede escribir como:

$$\begin{aligned} H(X,Y) &= -\sum_i \sum_j P(x_i, y_j) \log_2 \left(P(x_i, y_j)\right) \\ &= -\sum_i \sum_j P(x_i)P(y_j) \log_2 \left(P(y_j)P(x_i)\right) \\ &= -\sum_i \sum_j P(x_i)P(y_j) \left[\log_2 P(x_i) + \log_2 P(y_j)\right] \\ &= -\sum_i \sum_j P(x_i)P(y_j) \log_2 P(x_i) - \sum_i \sum_j P(x_i)P(y_j) \log_2 P(y_j) \end{aligned}$$

Como X y Y son independientes, entonces los contadores i,j puedes intercambiarse, entonces:

$$\begin{aligned} H(X,Y) &= -\sum_i \sum_j P(x_i)P(y_j) \log_2 P(x_i) - \sum_i \sum_j P(x_i)P(y_j) \log_2 P(y_j) \\ &= -\sum_i \sum_j P(x_i)P(y_j) \log_2 P(x_i) - \sum_j \sum_i P(x_i)P(y_j) \log_2 P(y_j) \\ &= -\sum_i P(x_i) \log_2 P(x_i) \sum_j P(y_j) - \sum_j P(y_j) \log_2 P(y_j) \sum_i P(x_i) \end{aligned}$$

donde

$$\sum_i P(x_i) = 1 \qquad \sum_j P(y_j) = 1$$

por lo tanto

$$\begin{aligned} H(X,Y) &= -\sum_i P(x_i) \log_2 P(x_i) \sum_j P(y_j) - \sum_j P(y_j) \log_2 P(y_j) \sum_i P(x_i) \\ &= -\sum_i P(x_i) \log_2 P(x_i) - \sum_j P(y_j) \log_2 P(y_j) \\ &= H(X) + H(Y) \end{aligned}$$

Problema 5

Considera una secuencia de lanzamientos independientes de una moneda. Calcula la probabilidad que en el veintésimo lanzamiento se obtiene por cuarta vez aguila. En promedio ¿cuántas veces se va a tener que lanzar la moneda para obtener por cuarta vez aguila?

Como cada lanzamiento es independiente y la probabilidad se mantiene en cada uno de ellos, entonces llamaremos a A, la probabilidad de obtener aguila, entonces, su probabilidad de obtener un evento l en A es la siguiente:

$$P(A) = \frac{1}{2}$$

El evento de interés, que llamaremos B, es obtener cuatro aguilas en veinte lanzamientos, en el cual, el ultimo lanzamiento obtenemos una aguila, entonces podemos realizar el calculo de obtener tres aguilar en 19 lanzamientos y la probabilidad de obtener un aguila, ya que al ser independientes los eventos de cada lanzamiento este no afectara al ultimo lanzamiento. La probabilidad de obtener tres aguilas en 19 lanzamientos podemos calcularla usando la distribucion binomial. Entonces, se tiene que:

$$\begin{aligned} P(X=3) &= \binom{19}{3} (P(A))^3 (1-P(A))^{19-3} \\ &= \binom{19}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{16} \\ P(X=3) &= \binom{19}{3} \left(\frac{1}{2}\right)^{19} \end{aligned}$$

Entonces, la probabilidad que se presente el evento B es:

$$\begin{aligned} P(B,A) &= P(B)P(A) \\ &= \binom{19}{3} \left(\frac{1}{2}\right)^{19} \left(\frac{1}{2}\right) \\ &= \binom{19}{3} \left(\frac{1}{2}\right)^{20} \\ &= \frac{969}{2^{20}} \\ &= \frac{969}{1048576} \\ P(B,A) &= 0.00092411 \end{aligned}$$

Sea N el numero de lanzamientos, entonces calculando $E[N]$, se obtiene que:

$$\begin{aligned} E[N] &= E\left[\sum_i N_i\right] \\ &= \sum_i E[N_i] \end{aligned}$$

Separando el evento en cuatro diferences, lanzaremos cada moneda hasta que se obtenga un aguila, entonces este lo podemos llevar a una distribución geométrica, por ende su valor esperado es

$$E[N_i] = \frac{1}{p}$$

entonces, el promedio de numero de lanzamientos es:

$$\begin{aligned} E[N] &= \sum_{i=1}^{4} E[N_i] \\ &= \sum_{i=4}^{4} \frac{1}{p} \\ &= \frac{4}{p} \\ &= \frac{4}{\frac{1}{2}} \\ E[N] &= 8 \end{aligned}$$

Problema 6

Verifica que para cualquier v.a. X y Y con una misma distribución:

$$E(X - Y|X + Y) = 0$$

Como es una esperanza condicional, entonces se tiene lo siguiente:

$$E(X - Y|X + Y) = E(X|X + Y) - E(Y|X + Y)$$

donde el segundo termino puede intercambiarse X por Y y viceversa, esto porque X y Y tienen una misma distribución. Entonces:

$$\begin{aligned} E(X - Y|X + Y) &= E(X|X + Y) - E(Y|X + Y) \\ &= E(X|X + Y) - E(X|Y + X) \\ &= E(X|X + Y) - E(X|X + Y) \end{aligned}$$

Por lo tanto:

$$E(X - Y|X + Y) = 0$$

Problema 7

Se debe sujetar N personas a una prueba de sangre para detectar la posible presencia de una cierta enfermedad. Con ese fin se dividen las personas al azar en subgrupos de tamaño k (puedes suponer que N es un múltiple de k). Se toma una muestra de sangre de cada persona y se mezclan las que pertenecen a personas de un mismo subgrupo; se aplica la prueba a estas k mezclas. Si el resultado es positivo, se sujeta cada persona del subgrupo correspondiente a una prueba separada. Suponiendo que la probabilidad de tener la enfermedad es 0.01, y que la presencia de la enfermedad entre las personas ocurre de manera independiente: calcula el promedio del número de pruebas que se va a tener que aplicar.

Como se tienen grupos de k personas, entonces podemos decir que se formaran m grupos, ya que $N = mk$, entonces tenemos inicialmente que se aplicarán N/k pruebas. La presencia de la enfermedad es independiente entre personas, por lo que podemos suponer que se trara de una distribucion binomial con probabilidad $p = 0.01$. Entonces si llamamos a el total de pruebas como P, se tiene lo siguiente:

$$\begin{aligned} E(P) &= E\left(\sum_m P_m\right) \\ &= \sum_m E(P_m) \end{aligned}$$

Al tratarse de una distribucion binomial, entonces:

$$\begin{aligned} E(P) &= \sum_m E(P_m) \\ &= \sum_m kp \\ &= mkp \\ &= \left(\frac{N}{k}\right) kp \\ &= Np \end{aligned}$$

Como se aplicaran pruebas a cada integrante del grupo si se detecta positiva la prueba grupal, entonces se tiene que el promedio del número de pruebas que se aplicaran es:

$$\bar{P} = \frac{N}{k} + Npk$$

Problema 8

Problema 8a

Elige al azar sin remplazo n números de {1, ..., n}. Da un argumento porque la probabilidad que el último obtenido sea el k-ésimo mayor es igual a $\frac{1}{n}(1 \leq k \leq n)$. Sea

$$\Omega = \{\Omega_i : 1, 2, \ldots, n\}$$

Y sea $A \subset \Omega$ tal que $\omega_k \notin A$ donde $\omega_k > \omega_i$. Entonces en el primer intento se tiene que la probabilidad de obtener un número contenido en A es:

$$P(A_1) = \frac{n-1}{n}$$

para el segundo intento es

$$P(A_2) = \frac{n-2}{n-1}$$

si realizamos esta sucesion para obtener todos los números contenidos en A se tiene que:

$$\begin{aligned} P(A_1, A_2, \ldots, A_{n-1}) &= P(A_1)P(A_2)\ldots P(A_{n-1}) \\ &= \left(\frac{n-1}{n}\right)\left(\frac{n-2}{n-1}\right)\ldots\left(\frac{1}{2}\right) \\ &= \frac{1}{n} \end{aligned}$$

Como el único número que no ha sido obtenido es Ω_k, entonces, la probabilidad de obtenerlo es $\frac{1}{n}$.

Problema 8b

Considera el siguiente código para encontrar el máximo en un arreglo de n números enteros positivos (todos diferentes).

```
1     max=-1
2     for (i in 1:n)
3         {
4         if (A[i] > max)
5             max = A[i]
6         }
```

Supongamos que el orden de los elementos es totalmente al azar. Define X el número de veces que se actualiza la variable max. Calcula EX (usa el inciso anterior).
Sea A el conjunto de números tal que

$$A = \{A_i \in \mathbb{N} : A_i \neq A_j \forall i, j\}$$

donde

$$A_{max} = max(A)$$

La variable max se actualizará hasta que esta obtenga el valor de A_{max}, ya que es el número máximo del conjunto A. La probabilidad de obtener a A_{max} en el intento i es:

$$P(A_i) = \prod_{j=1}^{i} \frac{n-j}{n-j+1} = \frac{n-i}{n}$$

Entonces, el valor esperado de las veces en que A_{max} es obtenido es:

$$\begin{aligned}
E(A) &= \sum_{i}^{n} iP(A_i) \\
&= \sum_{i}^{n} i \left[\frac{n-i}{n}\right] \\
&= \sum_{i}^{n} i - \frac{1}{n}\sum_{i}^{n} i^2 \\
&= \frac{n(n+1)}{2} - \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6}\right] \\
&= \left(\frac{n+1}{2}\right)\left(n - \frac{2n+1}{3}\right) \\
&= \left(\frac{n+1}{2}\right)\left(\frac{3n-2n-1}{3}\right) \\
&= \frac{(n+1)(n-1)}{6} \\
E(A) &= \frac{n^2-1}{6}
\end{aligned}$$