Examen 3

1. Se han tomado datos del número de averías anuales (variable Y) de los equipos informáticos de un departamento administrativo. Los resultados se recogen en la siguiente tabla:

Averías	n_i
1	8
2	6
3	2
4	1
5	3
7	1

   a) Calcula la media muestral, y la varianza y desviación típica muestrales de la variable Y.
   b) ¿Cuál es la moda? Razona la respuesta.
   c) Calcula la mediana.
   d) Determina el valor de los tres cuartiles Q_1, Q_2, Q_3.
   e) Dibuja un boxplot que represente los datos de la variable Y usando los cálculos de los incisos anteriores.
   f) ¿Hay atípicos en el boxplot?
   g) Se quiere relacionar el número de averías (la variable Y) con otra variable que mide el número de años de uso del equipo (variable X). Teniendo en cuenta que la media muestral de esta nueva variable es $\bar{x} = 2.35$, su desviación típica es $s_X = 1.4$ y la covarianza entre las dos variables es $s_{XY} = 1.9$. ¿Qué valor tiene la correlación lineal entre las dos variables? Interprétalo.
   h) Escribe la ecuación del modelo de regresión que explica el número de averías en función de los años de uso del equipo.
   i) Con el modelo anterior, ¿cuántas averías anuales esperaríamos en un equipo que tiene 10 años de uso?

SOLUCIÓN:

   a) Primero, veamos que el tamaño muestral n es igual a la suma de los n_i, por tanto $n = 21$.

   Media muestral de Y:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i * n_i = \frac{1}{21}(1*8 + 2*6 + 3*2 + 4*1 + 5*3 + 7*1)$$

$$= \frac{1}{21}(8 + 12 + 6 + 4 + 15 + 7) = \frac{52}{21} = 2.48$$

   Varianza muestral de Y:

$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 * n_i - \bar{y}^2$$
$$= \frac{1}{21}(1^2 * 8 + 2^2 * 6 + 3^2 * 2 + 4^2 * 1 + 5^2 * 3 + 7^2 * 1) - 2.48^2$$
$$= 2.8976$$

Desviación típica muestral de Y:

$$s_Y = \sqrt{s_Y^2} = \sqrt{2.8976} = 1.7$$

b) La moda es el valor que más se repite en la muestra de datos, por tanto, el que mayor frecuencia absoluta (n_i) tenga, en nuestro caso es **el valor 1** de la variable Y, ya que tiene una frecuencia absoluta de 8.

c) Para calcular la mediana, podemos hacer uso de las frecuencias relativas y sus acumuladas completando la tabla:
y_i: Valores
n_i: Frecuencia absoluta
f_i: Frecuencia relativa
F_i: Frecuencia relativa acumulada

y_i	n_i	f_i	F_i en porcentajes
1	8	8/21=0.38	38%
2	6	6/21=0.29	67%
3	2	2/21=0.09	76%
4	1	1/21=0.05	81%
5	3	3/21=0.14	95%
7	1	1/21=0.05	100%

En la tabla, la primera frecuencia relativa acumulada que supera el 50% es la coloreada en verde, que corresponde al valor 2 de la variable. Entonces la mediana es igual a 2.

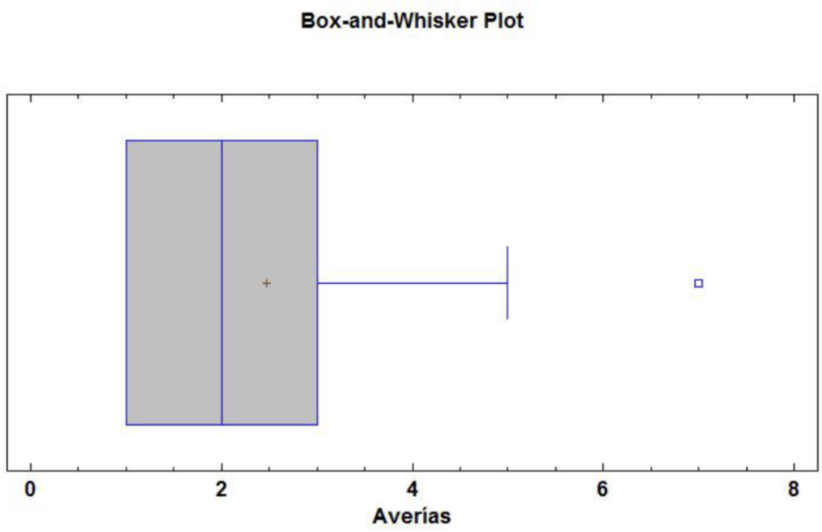
d) El cuartil Q_2 es la mediana, así que $Q_2 = 2$.
Para los otros dos cuartiles Q_1 y Q_3 buscaremos en la tabla las primeras frecuencias relativas acumuladas que superan el 25% y el 75% respectivamente. Son las que están coloreadas en amarillo, que corresponden a los valores 1 y 3 de nuestra variable. Entonces: $Q_1 = 1$ y $Q_3 = 3$.

e) Para dibujar el boxplot, además de los cuartiles necesitamos calcular el rango intercuartílico $RI = Q_3 - Q_1 = 3 - 1 = 2$, y las barreras:

$$B_1 = Q_1 - 1.5 * RI = 1 - 1.5 * 2 = 1 - 3 = -2$$
$$B_2 = Q_3 + 1.5 * RI = 3 + 1.5 * 2 = 6$$

Por el lado izquierdo, como la Barrera 1 está por fuera del valor mínimo de la variable (-2<1), no causa problema. Pero por el lado derecho, como la Barrera 2 sí está antes del valor máximo de la variable (6<7), hay un problema. Entonces el bigote del boxplot, por el lado derecho no puede ser el valor máximo (7) sino que será el máximo valor de la variable que caiga antes del valor de la Barrera

2, y ese es 5. En resumen, los bigotes del boxplot estarán en 1 y 5, mientras que habrá un valor atípico en 7 (ya que está por fuera de la Barrera 2).
La figura quedaría así:



f) Como explicamos en el inciso anterior, el valor 7 es un atípico ya que está por fuera de la barrera. En el gráfico se representa con un símbolo de cuadrado pequeño.

g) En resumen tenemos todos estos datos:

$$\bar{x} = 2.35$$
$$\bar{y} = 2.48$$
$$s_X = 1.4$$
$$s_Y = 1.7$$
$$s_X^2 = 1.96$$
$$s_Y^2 = 2.9$$
$$s_{XY} = 1.9$$

El coeficiente de correlación lineal de Pearson se puede calcular de la siguiente forma:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{1.9}{1.4 * 1.7} = 0.8$$

Recordemos que el coeficiente de correlación siempre está entre -1 y 1, por tanto, este es un valor alto, positivo, y cercano a 1, lo cual indica que la correlación entre las dos variables es alta y positiva (o directa).

h) Nos piden una regresión que explique Y en función de X:

$$\hat{y} = a + bx$$

Donde los coeficientes se pueden calcular de la siguiente forma:

$$b = \frac{s_{XY}}{s_X^2} = \frac{1.9}{1.96} = 0.97$$
$$a = \bar{y} - b\bar{x} = 2.48 - 0.97 * 2.35 = 0.2$$

Entonces la ecuación queda:

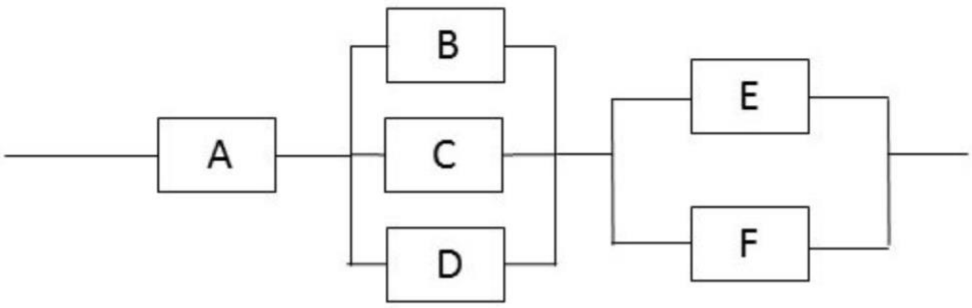
$$\hat{y} = 0.2 + 0.97x$$

i) Para estimar con ese modelo cuántas averías anuales tendría un equipo que tiene 10 años de uso, tenemos que sustituir $x = 10$ y calcular el valor predicho de la $\hat{y}$:

$$\hat{y} = 0.2 + 0.97 * 10 = 9.9 \approx 10$$

Es decir, para un equipo que tiene 10 años de uso esperaríamos que tenga aproximadamente 10 averías anuales.

2. Sea el siguiente sistema, formado por 6 componentes iguales e independientes entre sí:



a) Si la probabilidad de que un componente funcione es 0.95, hallar la probabilidad de que el circuito funcione.

SOLUCIÓN:

Denotemos como $A, B, \ldots$ = "El componente $A, B, \ldots$ Funciona"
Denotemos como $A^C, B^C, \ldots$ = "El componente $A, B, \ldots$ NO Funciona"
Denotemos como S el circuito Funciona.

Para que el circuito funcione, tiene que funcionar el componente A, y tiene que funcionar el sistema formado por los componentes B, C y D, y también tiene que funcionar el sistema formado por los componentes E y F, porque esas tres cosas están conectadas en serie, entonces si falla alguna el circuito no funcionará, por tanto, tienen que funcionar las tres. Además dentro de los dos últimos sistemas, los componentes están conectados en paralelo, lo que significa que cada sistema pequeño (el de B-C-D y el de E-F) funciona si al menos 1 componente funciona.

Para el funcionamiento de sistemas “en serie” se utiliza la “intersección” y para los que se conectan “en paralelo” se utiliza la “unión”:

$$S = A \cap (B \cup C \cup D) \cap (E \cup F)$$

Como todos los componentes son independientes, las probabilidades de las intersecciones se pueden transformar en productos de probabilidades. Entonces la probabilidad de que el sistema funcione es:

$$P(S) = P(A) \cdot P(B \cup C \cup D) \cdot P(E \cup F)$$

Vamos a hallar por separado las dos últimas probabilidades y luego las añadimos a la fórmula.

$$P(B \cup C \cup D) = P(B) + P(C) + P(D) - P(B \cap C) - P(B \cap D) - P(C \cap D) + P(B \cap C \cap D)$$

$$P(B \cup C \cup D) = 0.95 * 3 - 0.95^2 * 3 + 0.95^3 = 0.9999$$

Otra forma de hallar la probabilidad de arriba es haciendo uso de los complementos, porque la probabilidad de que “Funcione” ese sistema B-C-D conectado “en paralelo” es el complemento de que “No Funcione”, que es más fácil de calcular porque sería igual a las intersecciones de que no funcionen ninguno de los componentes B, C, D:

$$P(B \cup C \cup D) = 1 - P(B^C \cap C^C \cap D^C) = 1 - 0.05^3 = 0.9999$$

Ahora solo nos falta hallar esta última probabilidad:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = 0.95 * 2 - 0.95^2 = 0.9975$$

Y sustituyendo en la que nos interesa, que es que el sistema funcione:

$$P(S) = P(A) \cdot P(B \cup C \cup D) \cdot P(E \cup F) = 0.95 * 0.9999 * 0.9975 = 0.9475$$

3. La duración en años (T) de cierto tipo de bombillas sigue una distribución exponencial con $\lambda = 0.67$.
   a) El fabricante quiere determinar la duración k, de modo que la probabilidad de que una bombilla dure menos de k es 0.5 (es decir: $P(T < k) = 0.5$). ¿Cuál es el valor de k?
   b) ¿Cuál es la probabilidad de que una bombilla dure menos de 1 año?
   c) Si consideramos 10 bombillas del tipo anterior, ¿cuál es la probabilidad de que al menos una de ellas dure menos de 1 año?

SOLUCIÓN:

a) Denotemos $T = "Duración\ en\ años"$ con $T \sim Exp(\lambda = 0.67)$
   Sabemos que la función de distribución de una v.a. exponencial $Exp(\lambda)$ tiene esta fórmula: $F_T(k) = 1 - e^{-\lambda k}$ y que a su vez, por definición de distribución $F_T(k) = P(T \leq k)$. En el caso continuo, el “=” da igual ponerlo o no.

   Nos piden averiguar quién es k, sabiendo que $P(T < k) = 0.5$, entonces:

$$0.5 = P(T < k) = F_T(k) = 1 - e^{-0.67k}$$
$$0.5 = 1 - e^{-0.67k}$$
$$e^{-0.67k} = 1 - 0.5 = 0.5$$
$$e^{-0.67k} = 0.5$$

Aplicando logaritmo a ambos lados:

$$\ln(0.5) = \ln(e^{-0.67k}) = -0.67k$$
$$-0.69 = -0.67k$$
$$k = \frac{0.69}{0.67} = 1.0345$$

b) Nos piden la probabilidad:

$$P(T < 1) = F(1) = 1 - e^{-0.67*1} = 0.4883$$

c) Denotemos una nueva variable:

$$Y = \text{"}Número\ de\ bombillas\ que\ duran\ menos\ de\ 1\ año, en\ una\ muestra\ de\ 10\text{"}$$

Esta nueva variable sigue una distribución Binomial con parámetros $n = 10$ que es el total de la muestra (10 bombillas) y con probabilidad $p = 0.4883$ que es la calculada en el inciso anterior (la probabilidad de que 1 bombilla dure menos de 1 año en general):

$$Y \sim Binomial(n = 10, p = 0.4883)$$

Ahora nos piden calcular cuál es la probabilidad de que al menos una de las 10 bombillas dure menos de 1 año, eso es:

$$P(Y \geq 1) = 1 - P(Y < 1)$$

Como esta Binomial tomaría valores 0, 1, 2, …, hasta 10, la probabilidad de que Y sea menor estricto que 1 solo incluye al valor 0, entonces es lo mismo que poner la probabilidad de que Y sea igual a cero, y ahí ya podemos aplicar la función de probabilidad de la Binomial:

$$P(Y < 1) = P(Y = 0) = \binom{10}{0}(0.4883)^{\mathbf{0}}(1 - 0.4883)^{10-0}$$
$$= \frac{10!}{0!\,(10-0)!} * 1 * (1 - 0.4883)^{10} = (1 - 0.4883)^{10} = 0.0012$$

Entonces la probabilidad que buscábamos es:

$$P(Y \geq 1) = 1 - P(Y < 1) = 1 - 0.0012 = 0.9987$$

4. Selecciona la respuesta correcta

Sean $X_1, X_2, \dots, X_{100}$ v.a. i.i.d. de tipo Poisson con parámetro $\lambda = 1/100$, entonces $X_1 + \cdots + X_{100}$ es:

a. $N(0,1)$

b. $Poisson\left(\lambda = \frac{1}{100}\right)$

c. $Poisson\ (\lambda = 1)$

SOLUCIÓN:

El inciso c) es el correcto. Porque la suma de variables aleatorias Poisson independientes, se distribuye como una Poisson con parámetro lambda igual a la suma de los parámetros de las otras variables. En general, si tenemos N v.a. independientes $X_i \sim Poisson(\lambda_i)$, $i = 1, \ldots, N$. Entonces

$$\sum_{i=1}^{N} X_i \sim Poisson\left(\sum_{i=1}^{N} \lambda_i\right)$$

En nuestro caso $\lambda_i = 1/100$ para todas las $i = 1, \ldots, 100$. Por lo cual, como $\sum_{i=1}^{100} 1/100 = 1$, se tiene que:

$$X_1 + \cdots + X_{100} \sim Poisson(1)$$