Análisis de Datos - Final- Agosto 2018

Leer antes de empezar: en cada ejercicio definir apropiadamente los eventos o variables aleatorias a utilizar (y cuando corresponda sus distribuciones). Si utiliza un Teorema o Propiedad enunciarlo y verificar que se cumplen las hipótesis. En todos los ejercicios dar explícitamente la respuesta al problema planteado.

1. En una determinada ciudad se ha cometido un asesinato. De la investigación se encarga un detective, que tiene 5 sospechosos entre los que se encuentra el asesino. Se sabe que el detective trabaja con un pequeño margen de error, de forma que la probabilidad de creer inocente al verdadero asesino es de 0,05 y la probabilidad de creer culpable a una persona inocente es de 0,08.

   i. Si el detective cree que una persona es culpable, ¿cuál es la probabilidad de que esa persona sea el asesino?

   ii. Si la probabilidad de que una persona vaya presa es 0,5 y la de que la persona no sea el verdadero asesino y vaya presa es 0,01, ¿hay independencia entre ir preso y ser culpable?

2. El número de vehículos que llegan a una estación de servicio por minuto es un proceso de Poisson con media igual a 2.

   i. ¿Cuál es la probabilidad de que en un minuto lleguen 5 vehículos?

   ii. ¿y la de que en 5 minutos no llegue ninguno?

   iii. Suponiendo que acaba de llegar un vehículo, calcule la probabilidad de que transcurran más de 3 minutos hasta que aparezca el siguiente.

3. El peso de las naranjas de un determinado calibre, fluctúa normalmente con media 150 gr. Y desviación típica 30 gr. Si una bolsa se llena con 15 naranjas seleccionadas al azar. ¿Cuál es la probabilidad de que el peso total de la bolsa sea inferior a 2 kilos?

4. Para estudiar la efectividad de un medicamento contra la diabetes se mide la cantidad de glucemia en sangre antes y después de la administración de dicho medicamento, obteniéndose los resultados siguientes:
   Antes 7.2, 7.3, 6.5, 4.2, 3.1, 5.3, 5.6
   Después 5.2, 5.4, 5.3, 4.7, 4.1, 5.4, 4.9.
   Se supone que la distribución de la diferencia de las mediciones es normal.

   i. Estimar la reducción en la glucemia media producida por el medicamento mediante un intervalo de confianza de nivel 0,95 % (dar la función pivote y su distribución).

   ii ¿Puede afirmar con nivel de significación 0,05 que la reducción en la glucemia media es distinta de 0? Justifique.

5. **i.** Test de hipótesis: definir nivel de significación y p-valor.

   ii. Se estudiaron 100 muestras de aceite crudo de determinado proveedor con el fin de detectar la presencia del níquel mediante una prueba que nunca da un resultado erróneo. Si en 5 de dichas muestras se observo la presencia de níquel ¿podemos creerle al proveedor cuando asegura que menos del 10 % de las muestras contienen níquel? Plantear las hipótesis adecuadas. Indicar el estadístico de prueba y su distribución bajo H_0. Indicar la zona de rechazo para un nivel $\alpha = 0,05$. Expresar su conclusión en términos del problema. Dar una idea del p-valor.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad f(k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad f(x) = \frac{1}{b-a} \qquad f(x) = \lambda e^{-\lambda x}$$

$$\frac{(\overline{X}-\mu)}{S/\sqrt{n}} \qquad \frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \qquad \frac{(n-1)S^2}{\sigma^2} \qquad \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}}$$

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \qquad \frac{\widehat{\beta}_1-\beta_1}{s_r/\sqrt{S_{xx}}} \qquad \frac{\widehat{\beta}_0-\beta_0}{s_r\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}}} \qquad S_{rr} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$\sqrt{\left[h_x(\mu_X,\mu_Y)\sigma_X\right]^2 + \left[h_Y(\mu_X,\mu_Y)\sigma_Y\right]^2} \qquad \frac{\widehat{y}_0-(\beta_0+\beta_1 x_0)}{s_r\sqrt{\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}}$$

Análisis de Datos - Final- Agosto 2018 - Bosquejo de Resolución

1. i) Para la resolución del problema se definen los siguientes eventos:

$$\begin{aligned} A &= \{\text{ser asesino}\} \\ I &= \{\text{ser enjuiciado inocente}\} \\ C &= \{\text{ser enjuiciado culpable}\}. \end{aligned}$$

De esta forma se tiene que:

a) Hay un asesino de 5 sospechosos, por tanto, la probabilidad de que una persona elegida al azar sea el asesino es:
$$P(A) = \frac{1}{5} \qquad P(A^c) = \frac{4}{5}$$

b) La probabilidad de creer inocente al verdadero asesino, es la probabilidad de ser enjuiciado inocente condicionada a que es el asesino, es decir, $P(I|A) = 0,05$. Además, se sabe que $P(C|A) = 1 - P(I|A) = 0,95$.

c) La probabilidad de creer culpable a una persona inocente, es la probabilidad de ser enjuiciado culpable condicionada a que no es el asesino, es decir, $P(C|A^c) = 0,08$.

En el problema se pide la probabilidad de que una persona asesina haya sido enjuiciada culpable, es decir, la probabilidad de que una persona sea asesina condicionada a que ha sido enjuiciada culpable. Por tanto, la probabilidad requerida es $P(A|C)$, para calcular dicha probabilidad se recurre al teorema de Bayes (debe ser enunciado y chequeadas las hipótesis),

$$\begin{aligned} P(A|C) &= \frac{P(C|A)P(A)}{P(C|A)P(A) + P(C|A^c)P(A^c)} \\ &= \frac{0,95 \cdot 0,2}{0,95 \cdot 0,2 + 0,08 \cdot 0,8} \\ &= \frac{0,19}{0,254} = 0,748. \end{aligned}$$

i) Para la resolución del problema necesitamos definir el evento

$$B = \{\text{ser enviado a prisión}\}.$$

Luego $P(B) = 0{,}5$, $P(A^c) = 4/5$ y $P(A^c \cap B) = 0{,}01$, como $P(A^c) \cdot P(B) = 2/5$, A^c y B no son independientes luego A y B tampoco lo son.

2. i) Se trata de una distribución de Poisson, para la que se conoce el número medio, que es igual al parámetro λ, por lo tanto $X =$ número de vehículos que llegan a una estación de servicio por minuto y $X \sim P(2)$. La primera probabilidad pedida es:

$$P(X = 5) = \frac{2^5}{5!}e^{-2} = 0,036.$$

ii) Hay que hacer un cambio de parámetro pues la longitud del intervalo cambia.

Si en un minuto el número medio de vehículos es 2, en cinco minutos será 10. Por lo tanto se tiene que si $Y =$ número de vehículos que llegan a una estación de servicio en 5 minutos, $Y \sim P(10)$.

$$P(Y = 0) = \frac{10^0}{0!}e^{-10} = 0,000045.$$

iii) Sea T = tiempo que transcurre desde que pasa un vehículo hasta el siguiente, sigue una distribución exponencial de parámetro igual a 2. La probabilidad pedida es:

$$\begin{aligned} P(T > 3) &= 1 - P(T \leq 3) = 1 - (1 - e^{-2\cdot 3}) \\ &= 0{,}00248. \end{aligned}$$

3. Sean X_i = peso de la i-esima naranja en gr, con $i = 1, \ldots, 15$. Cada $X_i \sim N(150, 30^2)$ y son independientes. Luego $Y = \sum_{i=1}^{15} X_i \sim N(15 \cdot 150, 15 \cdot 30^2)$ y

$$P(Y < 2000) = P(Z < ((2000 - 2250)/(116, 189)) = P(Z < -2, 15) = (tabla) = 1 - 0, 9842 = 0, 0158.$$

4. i) Sean X_{Ai} =valor obtenido en la medición anterior a la administración del medicamento en la i-ésima persona y X_{Pi} =valor obtenido en la medición posterior a la administración del medicamento en la i-ésima persona, es decir, sobre cada individuo de la población se obtiene una observación de la variable y posteriormente se repite la observación una vez que el individuo ha tomado el medicamento siendo esta última observación dependiente de la primera. En este caso, la diferencia de ambas mediciones sigue una distribución normal de media la diferencia de ambas ($\mu_D = \mu_A - \mu_P$) y desviación típica desconocida, $N(\mu_D, \sigma_D^2)$, por lo que un intervalo de confianza de nivel $1 - \alpha = 0{,}95$, viene dado por:

$$IC_{0{,}95}(\mu_D) = (\bar{D} - t_{0{,}025{,}6} S_D/\sqrt{7}, \bar{D} + t_{0{,}025{,}6} S_D/\sqrt{7})$$

donde $\bar{D} = \sum_{i=1}^{7}(X_{Ai} - X_{Pi})/7 = 0{,}6$ y $S_D = 1{,}17473$ y utilizando la tabla de la distribución de Student obtenemos $t_{0{,}025{,}6} = 2{,}447$. Luego el intervalo de confianza para la reducción de glucemia por el medicamento de nivel 0,95 es $(-0{,}486; 1{,}686)$. **ii)** Queremos contrastar $H_0 : \mu_D = 0$ contra $H_A : \mu_D \neq 0$ con nivel $\alpha = 0{,}05$. Como $0 \in IC_{0{,}95}(\mu_D) = (-0{,}486; 1{,}686)$ no podemos afirmar con nivel de significación 0,05 que la reducción en la glucemia media es distinta de 0.

5. Llamemos X a la cantidad de muestras que contienen níquel entre las 100 y p a la proporción de muestras que contienen níquel. $X \sim B(100, p)$. Contrastamos la hipótesis nula

$$H_0 : p = 0, 1$$

.

Frente a la alternativa

$$H_A : p < 0, 1.$$

Definimos $p_0 = 0{,}1$. Si la prueba nunca da un resultado erróneo la variable $\hat{p} = X/100$, que representa la proporción de pruebas positivas al analizar 100 muestras satisface que

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/100}}$$

tiene aproximadamente distribución normal típica (por TCL ya que $100p(1 - p) > 5$, debe enunciar teorema) bajo H_0. Al tratarse de un test unilateral, con la región de rechazo a izquierda, esta corresponde a valores de la distribución del estadístico de prueba inferiores a $-z_{0{,}95} = -1, 645$, si consideramos un nivel de significación $\alpha =$ 0,05

En nuestro caso $\hat{p} = 5/100 = 0, 05$ de modo que $z = \frac{0{,}05 - 0{,}1}{\sqrt{0{,}1(1-0{,}1)/100}} = -1, 667 \leq -1, 645$. En conclusión, podemos afirmar con un nivel de significación 0,05 que menos del 10 % de las muestras contienen níquel, es decir podemos creerle al proveedor.