Bioestadística

- Soluciones Final Marzo 2016 -

Ejercicio 1. Un detector de mentiras se usa con frecuencia en distintas aplicaciones. Sea D^+ el evento que el detector de una lectura positiva, indicando que el sujeto está mintiendo. D^- el evento que el detector da una lectura negativa, indicando que el sujeto no está mintiendo. Sea T el evento que el sujeto dice la verdad, y L el evento que el sujeto está mintiendo. Se sabe que

$$\mathbb{P}(D^+|L) = 0.9, \text{ y que } \mathbb{P}(D^-|T) = 0.88.$$

O sea, si una persona está mintiendo la probabilidad de que sea detectada por el polígrafo es 0.9, y si está diciendo la verdad el polígrafo indica que esta diciendo la verdad con probabilidad 0.88.

Supongamos que se comienza con una pregunta rutinaria para la cual la mayoria de las personas no tienen motivos para mentir, de modo que $P(T) = 0.99$. Supongamos que queremos someter a un individuo al detector.

a) ¿Cuál es la probabilidad $\mathbb{P}(D^+)$ de que el detector de una lectura positiva?

b) ¿Cuál es la probabilidad $\mathbb{P}(T|D^+)$ de que el detector se haya equivocado cuando esté diciendo la verdad?

c) ¿Qué diferencia observa en la probabilidad $\mathbb{P}(T|D^+)$ de que el detector se equivoque cuando esté diciendo la verdad, si ahora cambiamos la pregunta y $\mathbb{P}(T) = 0.3$?

d) Si ahora se somete a 10 personas al detector y todas dieron lecturas positivas, ¿cuál es la probabilidad de que el detector se haya equivocado en al menos una lectura?

e) Supongamos que una persona que siempre miente es sometida al detector hasta que este de 10 lecturas positivas. ¿Cuál es el número esperado de intentos?

Solución:

a) La probabilidad de que la lectura sea positiva es $\mathbb{P}(D^+) = \mathbb{P}(D^+ \cap L) + \mathbb{P}(D^+ \cap T)$

Tenemos que

$$\mathbb{P}(D^+ \cap L) = \mathbb{P}(D^+ \mid L)\mathbb{P}(L) = 0.009$$

y

$$\mathbb{P}(D^+ \cap T) = \mathbb{P}(T) - \mathbb{P}(D^- \cap T) = \mathbb{P}(T) - \mathbb{P}(D^- \mid T)\mathbb{P}(T) = 0.99(1-0.88) = 0.1188$$

Concluimos así que $\mathbb{P}(D^+) = 0.1278$

b)
$$\mathbb{P}(T \mid D^+) = \frac{\mathbb{P}(D^+ \cap T)}{\mathbb{P}(D^+)} = \frac{0.1188}{0.1278} \cong 0.93$$

c) Repitiendo las cuentas de las partes a y b para $\mathbb{P}(T) = 0.3$

$$\mathbb{P}(T \mid D^+) = \frac{\mathbb{P}(D^+ \cap T)}{\mathbb{P}(D^+)} = \frac{0.036}{0.666} \cong 0.054$$

d) Si el detector dio siempre positivo, para que haya al menos un error, al menos uno debio decir la verdad, o lo que es lo mismo

$$1 - \mathbb{P}(L \mid D^+)^{10} = 1 - \left(\mathbb{P}(D^+ \mid L)\left(\frac{\mathbb{P}(L)}{\mathbb{P}(D^+)}\right)\right)^{10} \cong 1 - 0.574 = 0.426$$

e) El proceso de repetir un experimento hasta encontrar 10 exitos esta dado por la distribución binomial negativa con parametros , $\theta = 0.9$ y $r = 10$ de donde su esperanza es $\mu = \frac{r}{\theta} = 11$

Ejercicio 2. Se tienen dos muestras de tamaños 12 y 10 respectivamente de variables aleatorias independientes, normales ambas con varianza conocida e igual a 1.

Muestra 1	-0.44	0.32	0.76	0.90	0.97	1.00
	1.01	1.02	1.18	1.32	2.25	2.68
Muestra 2	-0.90	-0.84	-0.51	-0.37	0.01	0.31
	0.69	0.71	1.39	1.83		

a) Calcular la media de cada muestra.

b) Testear al nivel $\alpha = 0.05$ si las dos muestras tienen la misma media.

c) Hallar un intervalo de confianza para la diferencia de medias de nivel 0.9.

Solución:

a) Llamamos $X_1, \ldots, X_{12}$ a los datos de la primera muestra e $Y_1, \ldots, Y_{10}$ a los de la segunda. Entonces tenemos que $\overline{X} \approx 1,081$ y $\overline{Y} = 0,232$.

b) Realizamos un test paramétrico de comparación de medias de muestras normales con varianza conocida. El mismo tiene el siguiente estadístico:

$$E = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}}} \approx \frac{1,081 - 0,232}{\sqrt{\dfrac{1}{12} + \dfrac{1}{10}}} \approx 1,98$$

Como la región crítica es $\mathbb{RC} = \left\{|E| \geq z_{\alpha/2}\right\} = \{|E| \geq 1,96\}$ concluimos que se rechaza la hipótesis nula, por lo que asumimos que las medias son distintas.

c) El intervalo de confianza para la diferencia de medias de variables normales con varianza conocida es el siguiente:

$$I = \left[\overline{X}_n - \overline{Y}_m \ \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right] \approx [0,85 \ \pm 0,7] = [0,15; 1,55]$$

Ejercicio 3. Un laboratorio decide investigar una colonia de bacterias para estudiar la efectividad de la acción de un medicamento. Luego de haber observado varias colonias, un investigador propone modelar la variable aleatoria X que mide el tiempo de vida de una bacteria en minutos luego de ser expuesta al fármaco con la siguiente función de densidad:

$$f_X(x) = \begin{cases} Kx & \text{si } x \in [0,1] \\ 1 - \frac{1}{2}K & \text{si } x \in [1,2] \\ 0 & \text{en otro caso} \end{cases}$$

donde K es un parámetro desconocido.

1. (a) Probar que si $K \in [0,2]$ entonces f_X es efectivamente una densidad.

   (b) Probar que la siguiente función F_X es la distribución asociada a la densidad anterior:

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ Kx^2 & \text{si } x \in [0,1] \\ \frac{K}{2} + (1 - \frac{1}{2}K)(x-1) & \text{si } x \in [1,2] \\ 1 & \text{si } x > 2 \end{cases}$$

   (c) Hallar la esperanza de X en función de K.

   (d) Proponer un estimador para K dada una muestra $X_1 \dots X_n$ de X.

2. El investigador que propuso esta distribución decide medir los tiempos de vida en una nueva colonia de 10 bacterias y obtiene los siguientes datos (en minutos):

0.51	0.12	1.27	0.91	1.69	0.75	0.89	1.34	1.84	1.37

   (a) Estimar el valor de K a partir de la muestra.

   (b) Realizar un test de hipótesis para verificar si es razonable que estos datos corresponden a la distribución propuesta por el investigador.

Solución:

1. (a) Primero, notar que $f_X \geq 0 \ \forall K \in [0,2]$, ademas,

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^2 f(x)dx = \int_0^1 Kxdx + \int_1^2 1 - \frac{K}{2}dx =$$

$$= \frac{K}{2}x^2 \mid_{x=0}^{x=1} + 1 - \frac{K}{2} \mid_{x=1}^{x=2} = 1$$

de donde podemos concluir que f_X es una funcion de densidad.

(b) Se deduce de $F_X(x) = \int_{-\infty}^{x} f_x(t)dt$ y las primitivas de la parte anterior.

(c) La esperanza de una variable aleatoria continua es

$$\int_{-\infty}^{+\infty} f_X(x)xdx = \int_0^1 Kx^2dx + \int_1^2 \left(1 - \frac{1}{2}K\right) xdx =$$

$$= \frac{K}{3}x^3 \mid_{x=0}^{x=1} + \left(1 - \frac{K}{2}\right) \frac{x^2}{2} \mid_{x=1}^{x=2} = \frac{-5K}{12} + \frac{3}{2}$$

(d) Como $\bar{X}_n$ es im estimador para $\mathbb{E}(X) = \frac{-5K}{12} + \frac{3}{2}$, despejando K tenemos que

$$K_n = \left(\frac{3}{2} - \bar{X}_n\right)\left(\frac{12}{5}\right)$$

es un estimador para K

2. (a) Para la muestra se tiene que $\bar{X}_n = 1.069$, por tanto $K_n = 1.0344$

(b) Realizamos un test de ajuste de Kolmogorov-Smirnov a $F_0 = F_X$ como en la parte 1, para $K = K_n$.

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

X_i^*	$F_0(X_i^*)$	$\frac{i}{n}$	$\frac{i-1}{n}$	$\mid F_0(X_i^*) - \frac{i}{n} \mid$	$\mid F_0(X_i^*) - \frac{i-1}{n} \mid$
0.12	0.007	0.1	0	0.093	0.007
0.51	0.134	0.2	0.1	0.065	0.035
0.75	0.29	0.3	0.2	0.009	0.091
0.89	0.41	0.4	0.3	0.01	0.110
0.91	0.428	0.5	0.4	0.071	0.028
1.27	0.65	0.6	0.5	0.048	0.148
1.34	0.681	0.7	0.6	0.019	0.081
1.37	0.696	0.8	0.7	0.104	0.004
1.69	0.850	0.9	0.8	0.05	0.050
1.84	0.922	1	0.9	0.077	0.023

El estadistico KS es $KS = 0.148$ de donde el p-valor es mayor que 0.2. Por lo tanto no es posible rechazar la hipotesis nula

Ejercicio 4. En un estudio de la relación del tipo de sangre con una enfermedad se obtuvieron los siguientes resultados en las ciudades de Londres y Manchester:

Datos de Londres

	Control	*Ulcera*
Grupo A	4219	579
Grupo O	4578	911

Datos de Manchester

	Control	*Ulcera*
Grupo A	3775	246
Grupo O	4532	361

a) Considere los datos de Londres. ¿Encuentra en ellos una relación entre el tipo de sangre y la propención a una úlcera péptica? Use 0.05 como nivel.

b) Considere los datos de Manchester. ¿Encuentra en ellos una relación entre el tipo de sangre y la propención a una úlcera péptica? Use 0.05 como nivel.

c) Interprete: ¿son los datos de Londres y Manchester comparables?

Solución:

a) Para estudiar si hay una relación entre el tipo de sangre y la propención a una úlcera péptica, vamos a realizar un test de independencia. Si rechazamos la hipótesis nula, entonces tendremos suficiente evidencia para suponer que ambas variables están correlacionadas.

De los datos podemos estimar las probabilidades de cada variable como $p_A \approx 0,47$, $p_O \approx 0,53$ y $p_{control} \approx 0,86$, $p_{ulcera} \approx 0,14$. La tabla de valores esperados para Londres sería aproximadamente:

	Control	*Ulcera*
Grupo A	4103	695
Grupo O	4694	795

por lo que el valor del estadístico sería $\chi^2 \approx 42,4$. Como los grados de libertad son $(2-1)\times(2-1) = 1$, mirando la tabla chi cuadrado llegamos a que el p-valor es aproximadamente 0 por lo que rechazamos la hipótesis de independencia al nivel 0,05.

b) Usando el mismo razonamiento que en la parte anterior, estimamos las probabilidades en $p_A \approx 0,45$, $p_O \approx 0,55$ y $p_{control} \approx 0,93$, $p_{ulcera} \approx 0,07$. La tabla de valores esperados para Manchester sería aproximadamente:

	Control	*Ulcera*
Grupo A	3731	281
Grupo O	4560	343

por lo que el valor del estadístico sería $\chi^2 \approx 5,5$. Mirando la tabla chi cuadrado llegamos a que el p-valor cumple $0.05 > p - valor > 0.01$ por lo que rechazamos la hipótesis de independencia al nivel 0,05.

c) (Para esta pregunta no hay una única posible respuesta, pero damos una a modo de ejemplo):

A pesar de que en ambas ciudades tenemos evidencia para afirmar una relación entre el tipo de sangre y la úlcera péptica, podemos ver que la proporción de la población que sufre de la enfermedad es significativamente mayor en Londres que en Manchester, aún cuando las proporciones de tipo de sangre se mantienen casi iguales. Esto nos dice que la distribución de la enfermedad hace que los datos no sean comparables.