Práctico 7 - Técnicas de validación estadística

Ejercicio 1. De acuerdo con la teoría genética de Mendel, cierta planta de guisantes debe producir flores blancas, rosas o rojas con probabilidad $1/4$, $1/2$ y $1/4$, respectivamente. Para verificar experimentalmente la teoría, se estudió una muestra de 564 guisantes, donde se encontró que 141 produjeron flores blancas, 291 flores rosas y 132 flores rojas. Aproximar el p-valor de esta muestra:

a) utilizando la prueba de Pearson con aproximación chi-cuadrada,

b) realizando una simulación.

Llamemos p_1 a la probabilidad de que las flores sean blancas
p_2 " " " " " rosas
p_3 " " " " " rojas

H_0: La producción de flores de cierta planta de guisantes tiene una distribución con $p_1 = 1/4$, $p_2 = 1/2$, $p_3 = 1/4$

H_a: La producción de flores no tiene dicha distribución

a) $N_1 = 141$
$N_2 = 291$
$N_3 = 132$

$$T = \sum_{i=1}^{3} \frac{(N_i - n\cdot p_i)^2}{n\cdot p_i} = \frac{(141 - 564\cdot 1/4)^2}{564\cdot 1/4} + \frac{(291 - 564\cdot 1/2)^2}{564\cdot 1/2} + \frac{(132 - 564\cdot 1/4)^2}{564\cdot 1/4}$$

$$= 0 + \frac{81}{282} + \frac{81}{141}$$

$$= \frac{11421 + 22842}{39762}$$

$$= 0.8617$$

$$p\text{-valor} = P_{H_0}(T > 0.8617) = 1 - P_{H_0}(T \leq 0.8617) = 1 - P_{H_0}(\chi^2_2 \leq 0.8617) \cong 0{,}64996$$

Ejercicio 2. Para verificar que cierto dado no estaba trucado, se registraron 1000 lanzamientos, resultando que el número de veces que el dado arrojó el valor i ($i = 1,2,3,4,5,6$) fue, respectivamente, 158, 172, 164, 181, 160, 165. Aproximar el $p-$valor de la prueba: "el dado es honesto"

a) utilizando la prueba de Pearson con aproximación chi-cuadrada,

b) realizando una simulación.

Para que un dado no este trucado los valores recibidos en un lanzamiento deben tener una distribución uniforme 1/6

H_0: Cada valor del dado tiene una probabilidad 1/6 de aparecer (Distribución uniforme)

H_{alt}: Los lanzamiento del dado no tienen una distribución uniforme.

$N_1 = 158$ $\quad N_2 = 172$ $\quad N_3 = 164$

$N_4 = 181$ $\quad N_5 = 160$ $\quad N_5 = 165$

(a) $$T = \sum_{i=1}^{6} \frac{(N_i - n \cdot 1/6)^2}{1000 \cdot 1/6} = \frac{1}{1000/6} \cdot \sum_{i=1}^{6} (N_i - 1000/6)^2$$

$$= \frac{1}{1000/6} \cdot \left[(158 - 1000/6)^2 + (172 - 1000/6)^2 + (164 - 1000/6)^2 + (181 - 1000/6)^2 + (160 - 1000/6)^2 + (165 - 1000/6)^2 \right]$$

$$= 2.18$$

$$\text{p-valor} = P_{H_0}(T > 2{,}18) = 1 - P_{H_0}(T \leq 2.18) = 1 - P_{H_0}(\chi^2_5 \leq 2{,}18) = 0.82372$$

Ejercicio 3. Calcular una aproximación del $p-$valor de la hipótesis: "Los siguientes 10 números son aleatorios":

0.12, 0.18, 0.06, 0.33, 0.72, 0.83, 0.36, 0.27, 0.77, 0.74.

H_0: Los 10 números son aleatorios uniformes en el intervalo $(0,1)$

H_{alt}: $!H_0$

Recordemos que la función de distribución acumulada para $X \sim U(0,1)$

$$F(x) = \frac{x-a}{b-a} = x$$

$$D = \max_{1 \le j \le 10} \left\{ \frac{j}{10} - F(Y_{(j)}) , F(Y_{(j)}) - \frac{j-1}{10} \right\}$$

Ordenando los valores de menor a mayor

Muestra = 0.06, 0.12, 0.18, 0.27, 0.33, 0.36, 0.72, 0.74, 0.77, 0.83

Calculamos D usando Python

$D = 0.24$

Estimamos el p-valor con Python

Estimación de p-valor = 0.5316

Ejercicio 4. Calcular una aproximación del $p-$valor de la hipótesis: "Los siguientes 13 valores provienen de una distribución exponencial con media 50.0":

86.0 , 133.0 , 75.0 , 22.0 , 11.0 , 144.0 , 78.0 , 122.0 , 8.0 , 146.0 , 33.0 , 41.0 , 99.0 .

Recordemos la función de distribución acumulada de una exponencial

$$F(x) = 1 - e^{-\lambda x}$$

Calculemos el estadístico D con Python

$D \cong 0,3923$

p-valor $\cong$ 0.0249

Ejercicio 5. Calcular una aproximación del $p-$valor de la prueba de que los siguientes datos corresponden a una distribución binomial con parámetros $(n = 8, p)$, donde p no se conoce:

6, 7, 3, 4, 7, 3, 7, 2, 6, 3, 7, 8, 2, 1, 3, 5, 8, 7.

Podemos estimar p usando la media muestral, es decir

Bajo H_0 $\quad E(X) = n \cdot p \quad$ y $\quad \widehat{E(X)} = \overline{X}(n) \Rightarrow p = \dfrac{\overline{X}(n)}{n}$ con $n=8$

Utilizamos Python para esto

$\hat{p}(n) \simeq 0{,}6181 \quad N_0 = 0 \quad N_1 = 1 \quad N_2 = 2 \quad N_3 = 4 \quad N_4 = 1 \quad N_5 = 1 \quad N_6 = 2 \quad N_7 = 5 \quad N_8 = 2$

H_0: La muestra tiene la misma distribución de una X donde

$$P(X=i) = \binom{8}{i} \cdot (0{,}6181)^{i} \cdot (1-0{,}6181)^{8-i}$$

$$t = \sum_{i=0}^{8} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i} \simeq 31{,}499$$

p valor $= P_{H_0}(T > t) = 1 - P_{H_0}(T \leq t) = 1 - P_{H_0}(\chi^2_7 \leq t) \simeq 0{,}0000.5$

Valor estimado de p-valor estimando frecuencias $\simeq 0{,}011$

Ejercicio 6. Un escribano debe validar un juego en cierto programa de televisión. El mismo consiste en hacer girar una rueda y obtener un premio según el sector de la rueda que coincida con una aguja. Hay 10 premios posibles, y las áreas de la rueda para los distintos premios, numerados del 1 al 10, son respectivamente:

31%, 22%, 12%, 10%, 8%, 6%, 4%, 4%, 2% y 1%.

Los premios con número alto (e.j. un auto 0Km) son mejores que los premios con número bajo (e.j. 2x1 para entradas en el cine). El escribano hace girar la rueda hasta que se cansa, y anota cuántas veces sale cada sector. Los resultados, para los premios del 1 al 10, respectivamente, son:

188, 138, 87, 65, 48, 32, 30, 34, 13 y 2.

(a) Construya una tabla con los datos disponibles

(b) Diseñe una prueba de hipótesis para determinar si la rueda es justa

(c) Defina el p-valor a partir de la hipótesis nula

(d) Calcule el p-valor bajo la hipótesis de que la rueda es justa, usando la aproximación chi cuadrado

(e) Calcule el p-valor bajo la hipótesis de que la rueda es justa, usando una simulación.

(a)

Premio	Frecuencia esperada	Frecuencia observada
1	197,47	188
2	140,14	138
3	76,44	87
4	63,7	65
5	50,96	48
6	38,22	32
7	25,48	30
8	25,48	34
9	12,74	13
10	6,37	2

(b) Si la rueda es justa, entonces:

H_0: Los giros de la rueda siguen una distribución donde

$P_1 = 0,31$ $\quad P_2 = 0,22$ $\quad P_3 = 0,12$ $\quad P_4 = 0,10$ $\quad P_5 = 0,8$

$P_6 = 0,6$ $\quad P_7 = 0,4$ $\quad P_8 = 0,4$ $\quad P_9 = 0,2$ $\quad P_{10} = 0,1$

· (c) Bajo la hipótesis nula

$$t = \sum_{i=1}^{10} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i} \quad \text{y} \quad \text{p-valor} = 1 - P_{H_0}(\chi_9^2 \geq t)$$

(d) Usando python para el cálculo tenemos que

$t \cong 9,8104$

$\text{p-valor} = 1 - P(\chi_9^2 \leq 9,8104) \cong 0,3661$

(e) p-valor estimado $\cong$ 0,328

Ejercicio 7. Generar los valores correspondientes a 30 variables aleatorias exponenciales independientes, cada una con media 1. Luego, en base al estadístico de prueba de Kolmogorov-Smirnov, aproxime el $p-$valor de la prueba de que los datos realmente provienen de una distribución exponencial con media 1.

Obs: El pvalor depende de la muestra generada, por ende siempre varía.

Ejercicio 8. Se sortean elementos de un conjunto de datos que tiene una distribución t-student de 11 grados de libertad. El investigador, que no conoce la forma verdadera de la distribución, asume que la misma es normal.

Analice la validez de esta suposición para muestras de tamaños 10, 20, 100 y 1000 elementos Para ello realice simulaciones numéricas e implemente el test de Kolmogorov-Smirnov a los datos simulados, asumiendo una distribución N(0,1). Presente los resultados en una tabla que contenga el número de elementos de la simulación, el valor del estadístico D y el $p-$valor

Ayuda: Función de probabilidad normal: Para obtener la función de probabilidad normal, se puede usar la función `math.erf`. Por ejemplo, la cantidad `math.erf(x/math.sqrt(2.))/2.+0.5` equivale a

$$\int_{-\infty}^{x} N(0,1)(t)\,dt \tag{1}$$

Ayuda: Generación de números aleatorios con una distribución t-student: Utilice el siguiente código para generar números aleatorios que siguen una distribución T-student:

```
import math
import random

def rt(df): # df grados de libertad
    x = random.gauss(0.0, 1.0)
    y = 2.0*random.gammavariate(0.5*df, 2.0)
    return x / (math.sqrt(y/df))
```

n	D	p-valor
10	0,3262	0,181
20	0,1543	0,6803
100	0,1156	0,1242
1000	0,0809	0

Ejercicio 9. En un estudio de vibraciones, una muestra aleatoria de 15 componentes del avión fueron sometidos a fuertes vibraciones hasta que se evidenciaron fallas estructurales. Los datos proporcionados son los minutos transcurridos hasta que se evidenciaron dichas fallas.

1.6 10.3 3.5 13.5 18.4 7.7 24.3 10.7 8.4 4.9 7.9 12 16.2 6.8 14.7

Pruebe la hipótesis nula de que estas observaciones pueden ser consideradas como una muestra de la distribución exponencial.

Ho: La muestra tiene una distribución $\mathcal{E}_x(\lambda)$

Estimamos λ bajo h_0

$$E(x) = \frac{1}{\lambda} \quad y \quad \hat{E}(x) = \overline{x}(n) \Rightarrow \lambda = \frac{1}{\overline{x}(n)}$$

Utilizando Kolmogorov-Smirnov con uniformes y exponenciales tenemos en ambos casos que p-valor > 0.01
Para un nivel de rechazo del 1% no se rechaza Ho

Ejercicio 10. Decidir si los siguientes datos corresponden a una distribución Normal:
91.9 97.8 111.4 122.3 105.4 95.0 103.8 99.6 96.6 119.3 104.8 101.7
Calcular una aproximación del $p-$valor.

Ho: Los datos siguen una distribución $N(\mu, \sigma)$
Halt: Los datos no siguen una distribución $N(\mu, \sigma)$

Bajo Ho estimamos μ y σ

$$\hat{\mu} = \overline{X}(n) \quad y \quad \hat{\sigma} = \sqrt{S^2_{n-1}(n)}$$

Utilizando Kolmogorov-Smirnov con uniformes y exponenciales en ambos casos pvalor > 0.05
Para un nivel de rechazo del 5% no se rechaza Ho.