PRÁCTICA 7

TÉCNICAS DE VALIDACIÓN ESTADÍSTICA

Para estos ejercicios, estudiar

`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2.html`

Otra bibliografia interesante es

`https://relopezbriega.github.io/blog/2016/06/29/distribuciones-de-probabilidad-con-python/`

No utilice implementaciones personales de densidades a menos que el ejercicio se lo pida exactamente.

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.

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.

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.

Ejercicio 4. Calcular una aproximación del $p-$valor de la hipótesis: "Los siguientes 13 valores provienen de una distribución exponencial con media 50":
86, 133, 75, 22, 11, 144, 78, 122, 8, 146, 33, 41, 99.

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.

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.g. un auto 0Km) son mejores que los premios con número bajo (e.g. 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.

Ejercicio 7. Generar los valores correspondientes a 10 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.

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.

(a) 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

(b) Determine cuántas simulaciones son necesarias para asegurar con una confianza del 95% que la media se encuentra a menos de 0.01 del valor poblacional.

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))
```

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 población exponencial.

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.

Ejercicio 11. Un experimento diseñado para comparar dos tratamientos contra la corrosión arrojó los siguientes datos (los cuales representan la máxima profundidad de los agujeros en unidades de milésima de pulgada) en pedazos de alambre sujetos a cada uno de los tratamientos por separado:

Tratamiento 1:	65.2	67.1	69.4	78.4	74.0	80.3
Tratamiento 2:	59.4	72.1	68.0	66.2	58.5	

a) Calcular el $p-$valor exacto de este conjunto de datos, correspondiente a la hipótesis de que ambos tratamientos tienen resultados idénticos.

b) Calcular el $p-$valor aproximado en base a una aproximación normal,

c) Calcular el $p-$valor aproximado en base a una simulación.

Ejercicio 12. Catorce ciudades, aproximadamente del mismo tamaño, se eligen para un estudio de seguridad vial. Siete de ellas se eligen al azar y durante un mes aparecen en los periódicos locales artículos relativos a la seguridad vial. Los números de accidentes de tránsito del mes posterior a la campaña son los siguientes:

Grupo de tratamiento:	19	31	39	45	47	66	75
Grupo de control:	28	36	44	49	52	72	72

a) Calcular el $p-$valor exacto de este conjunto de datos, correspondiente a la hipótesis de que en ambos grupos se tienen resultados idénticos (es decir, los artículos no tuvieron ningún efecto).

b) Calcular el $p-$valor aproximado en base a una aproximación normal,

c) Calcular el $p-$valor aproximado en base a una simulación.

Ejercicio 13. Se requiere averiguar si dos fertilizantes, A y B presentan diferencias significativas en cuanto a sus efectos sobre el aumento de cosecha. Con este propósito se eligieron al azar 15 parcelas a las que se fertilizó aleatoriamente con cada uno de los fertilizantes en cuestión. Los aumentos de cosecha obtenidos fueron los siguientes

Fertilizante	Aumento de cosecha
A	$39, 40, 38.9, 35, 32, 33, 22.8, 36$
B	$36.5, 33.1, 35.2, 30, 29, 26, 35.1$

A la vista de estos datos, estimar el p-valor del test mediante una simulación y una aproximación normal. ¿Puede inferirse que existen diferencias significativas entre los dos fertilizantes a nivel $\alpha = 0.05$?.

Ejercicio 14. El origen de la civilización Etrusca sigue siendo todavía un misterio para los antropólogos. En concreto, una cuestión que se plantea es la de si fueron originarios de la península italiana o si inmigraron a ella procedentes de algún otro lugar. Se pensó que una forma de contestar a esta pregunta sería comparar a los actuales italianos con los restos arqueológicos etruscos mediante un estudio antropométrico. Para ello, se midió, en milímetros, la máxima anchura, X, de 8 cráneos de restos de varones etruscos y la máxima anchura, Y, de la cabeza de 10 varones italianos, todos ellos elegidos al azar. Los resultados obtenidos fueron los siguientes:

Etruscos :	141	132	154	142	143	150	134	140		
Italianos :	133	138	136	125	135	130	127	131	116	128

En base a los datos obtenidos y utilizando un test de Rango, ¿se puede concluir que las diferencias son significativas entre las dos poblaciones a nivel $\alpha = 0,05$?