Simulación Digital - Evaluacion 2

Mauricio Alejandro Romejo Jaimes

Daniel Torres

Rubén Darío Rodríguez Moreno

Punto 1. Solución en el notebook.

Punto 2. Código en el notebook.

a.

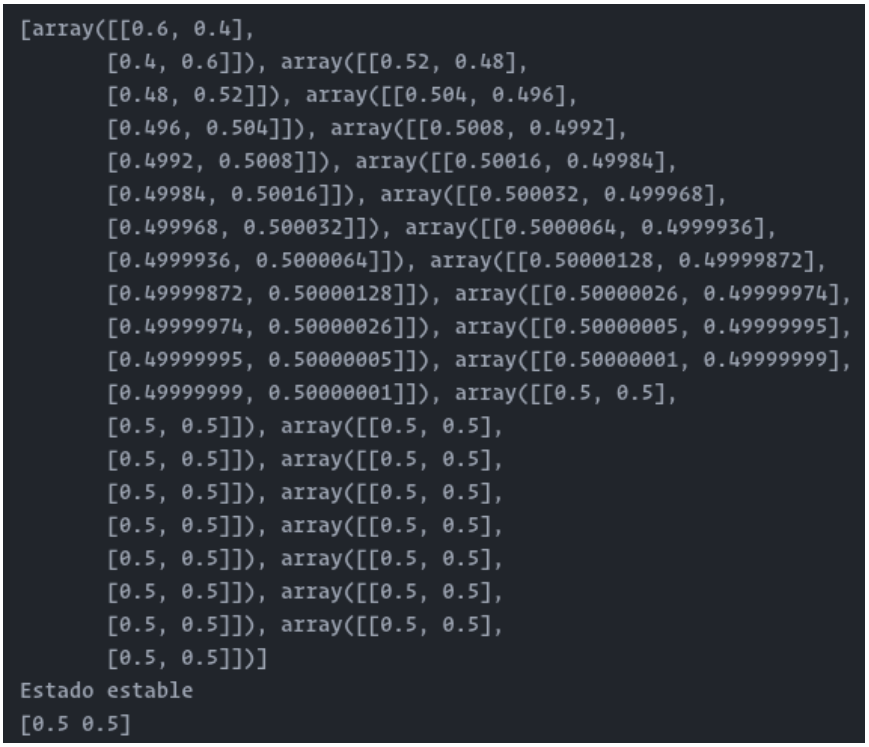
$$P_{11} = P(D_{n+1} = 0) + P(D_{n+1} = 2) + P(D_{n+1} = 4) = 3/5$$
$$P_{12} = P(D_{n+1} = 1) + P(D_{n+1} = 3) = 2/5$$
$$P_{21} = P(D_{n+1} = 1) + P(D_{n+1} = 3) = 2/5$$
$$P_{22} = P(D_{n+1} = 0) + P(D_{n+1} = 2) + P(D_{n+1} = 4) = 3/5$$

P =

$$\begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{3}{5} \end{bmatrix}$$

b.

$$\pi = \pi P$$
$$\pi_1 + \pi_2 = 1$$
$$\pi_1 = \pi_2 = \frac{1}{2}$$



Punto 3.

¿Qué mecanismos podría utilizar para validar los promedios obtenidos de las variables de interés?

Validación simple: Este es un método el cual se basa principalmente en dividir de manera aleatoria y con un tamaño comparable, un conjunto de observaciones en dos conjuntos, el de entrenamiento y el de validación.

Se considera un modelo ajustado a la hora de ejecutar el test de entrenamiento, el cual se usa para predecir las nuevas observaciones del test de prueba que permiten obtener una estimación del error (Mean Squared Error).

El proceso de este método de validación consiste en:

- Se toma el conjunto de datos base que tenemos y se le hace la separación de los datos en función de una variable.
- Se generan dos grupos separando los datos de la muestra aleatoria con una proporción cercana al 50% - 50%.
- Se verifica que la distribución sea proporcional entre el conjunto de entrenamiento y el de prueba.
- Una vez verificado se ajusta el modelo de entrenamiento con los datos asignados.
- Después de tener el modelo, se calculan las predicciones que se pueden obtener con el mismo a partir de la probabilidad a posteriori de direction.

- Finalmente obtenemos el porcentaje de la estimación del test error rate del modelo, con lo que podremos conocer el porcentaje de asertividad que tienen las predicciones en cada uno de los casos.

Para que el modelo predictivo pueda considerarse útil, debe acertar un porcentaje de predicciones superior a lo obtenido por azar o respecto al obtenido de la asignación a la clase mayoritaria.

bootstrapping: Este tipo de muestras Bootstrap son muestras obtenidas a partir e la muestra original por muestreo aleatorio con reposición del mismo tamaño que la muestra original, como resultado de este tipo de muestreo algunas observaciones apareceran múltiples veces en la muestra Bootstrap y otras ninguna. Utilizamos el bootstrap, específicamente el remuestreo de casos, para derivar la distribución de x, pero primero se remuestrean los datos para obtener una muestra bootstrap.

Para realizar el proceso de Bootstrapping se debe:

- Obtener una nueva muestra del mismo tamaño que la muestra original mediante muestreo aleatorio con reposición.
- Ajustar el modelo empleando la nueva muestra generada anteriormente.
- Calcular el error del modelo empleando aquellas observaciones de la muestra original que no se han incluido en la nueva muestra, a este error se le conoce como error de validación.
- Repetir el proceso *n* veces y calcular la media de los *n* errores de validación.
- Finalmente, y tras las *n* repeticiones, se ajusta el modelo final empleando todas las observaciones de entrenamiento originales.

La naturaleza del proceso de *bootstrapping* genera cierto bias en las estimaciones que puede ser problemático cuando el conjunto de entrenamiento es pequeño, pero existen ciertas modificaciones del algoritmo original para corregir este problema.

Un ejemplo de la primer remuestreo podría ser el siguiente X1* = x2, x1, x10, x10, x3, x4, x6, x7, x1, x9. Hay algunos duplicados, ya que una remuestra bootstrap procede de un muestreo con reemplazo de los datos. Además, el número de puntos de datos en una remuestra bootstrap es igual al número de puntos de datos en nuestras observaciones originales.

A continuación, calculamos la media de esta remuestra y obtenemos la primera media bootstrap: μ1*. Repetimos este proceso para obtener la segunda remuestra X2* y calculamos la segunda media bootstrap μ2*.

Si repetimos esto 100 veces, entonces tenemos μ1*, μ2*, ..., μ100*. Esto representa una distribución empírica bootstrap de la media de la muestra. A partir de esta distribución empírica, se puede derivar un intervalo de confianza bootstrap con el fin de probar la hipótesis.

¿Qué mecanismos podría usar para validar un conjunto de datos aleatorios continuos que se comportan como una distribución dada?

Chi cuadrado X^2 :

Uno de los mecanismos para realizar esta validación es el Chi Cuadrado, el cual se encuentra dentro de las pruebas pertenecientes a la estadística descriptiva que se centra en extraer información sobre la muestra, enfocado a el estudio de dos variables y comúnmente utilizada para analizar variables nominales o cualitativas, buscando determinar la existencia o no de independencia entre estas dos. La independencia demuestra si no tienen relación, y por tanto una no depende de la otra, ni viceversa.

Para estimar la independencia entre las variables, se calculan los valores que indicarían la independencia absoluta, lo que se denomina “frecuencias esperadas”, comparándolos con las frecuencias de la muestra.

Descrito de manera general, los pasos a realizar están determinados por:

- Primero realizar una tabla de frecuencia observados (fo) con los datos que tenemos donde se van a medir las variables de interés
- Realizamos la formulación de hipótesis, como es conocido:

  la hipótesis nula (H0) → los parámetros son independientes
  hipótesis alternativa (H1) → los parámetros tienen algún grado de asociación o relación.

- Realizaremos una tabla de frecuencias esperadas (fe)

- Como se utilizara el método estadístico chi cuadrado, a casa de que este permite la comparación entre frecuencias observadas con las esperadas

  Se emplea la siguiente fórmula para calcular el chi cuadrado real

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe}$$

- Se calcula el Chi cuadrado teórico, para esto necesitamos hallar los grados de libertad y nivel de significancia. Al saber la información anterior, buscamos en la tabla Chi cuadrado de acuerdo a los resultados para obtener el Chi teórico.

- Por último se contrastan las hipótesis, si al comparar los resultados, Chi cuadrado real es mayor que Chi cuadrado teórico rechazamos la hipótesis nula y aceptamos la hipótesis alternativa. Si no es así y el chi real es menor al chi teórico aceptamos la hipótesis nula.

En resumen de lo anterior, el proceso de la prueba Chi-cuadrado consiste en:

Utilizar una aproximación a la distribución chi cuadrado para evaluar la probabilidad de una discrepancia igual o mayor que la que existe entre los datos y las frecuencias esperadas según la hipótesis nula. La exactitud de dicha evaluación dependerá de que los valores esperados no sean muy pequeños, y en menor medida de que el contraste entre ellos no sea muy elevado.

Los contrastes de hipótesis se enfocan en la comparación de las frecuencias observadas (frecuencias empíricas) en la muestra, con aquellos resultados a esperar (frecuencias teóricas o esperadas) si la hipótesis nula fuera cierta. Así, la hipótesis nula se rechaza si existe una diferencia significativa entre las frecuencias observadas y las esperadas.

En el proceso funcional, una vez están definidas las hipótesis, se debe realizar el contraste, y para ello disponemos de los datos en una tabla de frecuencias. Se indica una frecuencia absoluta observada o empírica para cada valor. Luego, suponiendo que la hipótesis nula es cierta, para cada valor se calcula la frecuencia absoluta que sería nuestra frecuencia a esperar.

Si el estadístico chi-cuadrado toma un valor grande, denota una discrepancia entre las frecuencias, por ende, se deberá rechazar la hipótesis nula. Mientras, si es un valor igual a 0, significa que hay concordancia entre las frecuencias observadas y las esperadas.

Recapitulando información y juntándose a la teoría del libro de Sheldon que tiene un enfoque más matemático, pero apoya la explicación sencilla del apartado anterior :

Cierta información, ecuaciones y ejercicio de ejemplo tomadas del libro de Sheldon [2].

Supongamos que vamos a observar n variables aleatorias independientes Y1, …, Yn, cada una puede tomar los valores 1,2,..., k y que estamos interesados en verificar la hipótesis de que {pi, i= 1, … , k} es la función de masa de probabilidad de estas variables aleatorias.
Es decir, Y representa cualquiera de las Yj, y sería nuestra hipótesis a verificar, conociéndola como la hipótesis nula (H0):

$$H_0: P\{Y=i\} = p_i, \quad i = 1, \ldots, k$$

Para verificar la hipótesis anterior, sea Ni, para cualquier i = 1,...,k el número de Yj que son iguales a i. Como cada Yj, de manera independiente, es igual a i con probabilidad P{Y=i}, se tiene que según la hipótesis nula, Ni es binomial con parámetros n y pi, entonces, si H0 es cierta:

$$E[N_i] = np_i$$

de modo que $(N_i - np_i)^2$ es un indicador de la probabilidad de que pi sea igual a la probabilidad de que Y = i. Cuando este resultado sea grande, indicará que la hipótesis nula(H0) es incorrecta. A partir del siguiente cálculo podremos considerar el valor final y rechazar la hipótesis nula cuando T sea grande. Para su uso, valores pequeños en el

resultado de T son evidencia en a favor de la hipótesis H0, los valores grandes indican su falsedad.

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

Por último, para analizar el valor p, se supondrá que la hipótesis H0 es correcta, esto ayudará a señalar si la probabilidad de que ocurra un valor grande T, como el observado, corrobora el valor de verdad de la hipótesis nula. Es usual **rechazar la hipótesis** nula diciendo que no es congruente, si se obtiene un valor p pequeño (**valores menores que 0.05, o más conservador, 0.01, considerado como crítico**). Para casos contrarios se acepta la hipótesis nula, afirmando la congruencia de la información:

$$\text{valor } p = P_{H_0}\{T \geq t\}$$

$$\text{valor } p \approx P\{X^2_{k-1} \geq t\}$$

Quitar ejemplo lo más probable porque se deben es simular, pa solucionarlos
Ejemplo: Considera una cantidad aleatoria que toma los valores 1,2,3,4,5. Y se quiere verificar la hipótesis de que los valores presentan la misma probabilidad.

$$H_0: p_i = 0.2 \qquad i = 1,2,3,4,5$$

Una muestra de tamaño 50 arrojó estos valores para N_i

$$12,5,19,7,7$$

Se obtiene el estadístico de prueba T, que da el resultado siguiente

$$T = \frac{4 + 25 + 81 + 9 + 9}{10} = 12.8$$

Buscando el valor de probabilidad, da el resultado de

$$valor\ p \approx P\{X_4^2 > 12.8\} = 0.0122$$

entonces, como se explicó teóricamente, al ser un valor menor de 0.05 la hipótesis nula será rechazada.

Kolmogorov Smirnov

Este procedimiento en una muestra sirve para comparar su función de distribución acumulada observada de una variable con una distribución teórica determinada. El resultado de la prueba se representa mediante la letra (Z), la cual se calcula a partir de la diferencia mayor (en valor absoluto) entre las funciones de distribución acumuladas teórica y observada (empírica). Esta prueba de bondad de ajuste contrasta si las observaciones podrían razonablemente proceder de la distribución especificada.

La prueba de Kolmogorov-Smirnov es un tipo de prueba no paramétrica. También conocidas, como de distribución libre, utilizadas frecuentemente en estadística inferencial, para este caso, útil como prueba de bondad de ajuste y estimador de independencia.

Consideremos la situación en la que Y1, Y2, …, Yn, son variables aleatorias independientes, estamos interesados en verificar la hipótesis nula H0 de que todas tienen una función de distribución común F, donde F es una función de distribución continua dada. Entonces, para poder verificar que las Yj provienen de la función de distribución continua F. Observarmos Y1,Y2,..,Yn y consideramos la función de distribución empirica (Fe) definida por:

$$F_e(x) = \frac{\#\{i \mid Y_i \leq x\}}{n}.$$

Fe (x) → Es la proporción de valores observados menores o iguales a x.

Con base a esto, se tiene que si es correcta la hipótesis nula de que F es la distribución subyacente, deberá encontrarse cerca de F(x).
Por ende la hipótesis se define como:

Hipótesis nula (H0) → Fe(x) es “cercana” a F(x)

Como esto es así para toda x, la cantidad natural sobre la cual basar una prueba para H0, es utilizando el estadístico de Kolmogorov-Smirnov (D):

$$D \equiv \max_x |F_e(x) - F(x)|, \qquad -\infty < x < \infty.$$

Para el proceso de cálculo del valor de D, dado un conjunto de datos observados Yi = Yj, sabiendo que j= 1,....,n. Los valores de Yj se deben ir en orden creciente:

$$y_{(j)} = j - \text{ésimo valor más pequeño}$$

$$y_{(1)} < y_{(2)} < \cdots < y_{(n)}.$$

$$y_1 = 3,\ y_2 = 5,\ y_3 = 1 \text{ y } n = 3,$$

Por ejemplo, si

entonces → $y_{(1)} = 1,\ y_{(2)} = 3,\ y_{(3)} = 5.$

La distribución empírica (Fe) se puede escribir de la siguiente forma:

$$F_e(x) = \begin{cases} 0 & x < y_{(1)} \\ \frac{1}{n} & y_{(1)} \le x < y_{(2)} \\ \vdots & \\ \frac{j}{n} & y_{(j)} \le x < y_{(j+1)} \\ \vdots & \\ 1 & y_{(n)} \le x \end{cases}$$

Como las diferencias (Fe(x)-F(x) y F(x)-Fe(x)) y el análisis de sus valores máximos supremos son no negativos y cumplen ciertas condiciones. Y, después de ciertos procesos con los máximos y las relaciones descritas anteriormente se llega a la conclusión de que la ecuación a trabajar para el estadístico de K-S es:

$$\boxed{D = \max_{1 \le j \le n} \left\{ \frac{j}{n} - F(y_{(j)}),\ F(y_{(j)}) - \frac{j-1}{n} \right\}} \leftarrow \text{Estadístico de K-S}$$

Esta nos permitirá calcular el valor de **D**.

Con esta explicación, el proceso se sintetiza en

- Elegir un grado de significación (nivel de rechazo) α.
- Definición de hipótesis.
- Tomar la muestra y ordenar los datos observador
- Calcular el estadístico D para los datos observados (usar ecuacion Estadistico K-S)
- Valor observado: D = d
- Calcular el valor p = Pf(D≥ d) para saber si se rechaza o no la hipótesis nula
  valor p < α → se rechaza H0
  valor p > α. → no se rechaza H0

El proceso para hallar la estimación del valor p es:
El valor p se puede aproximar mediante una simulación de las siguiente fórmula:

$$\text{valor } p = P_F(D \ge d)$$

Pf es la probabilidad suponiendo que la hipótesis H0 es correcta, este valor p no depende de la distribución F, lo que nos permite usar cualquier distribución continua; entonces podemos trabajar una distribución uniformemente distribuida en (0,1).

Teniendo en cuenta lo anterior, el estadístico D depende de n observaciones, Y tiene una distribución F, además F es una función creciente. Por lo cual la ecuación pasaría de esto:

$$D = \sup_x |F_e(x) - F(x)| = \sup_x \left| \frac{\#\{i \mid Y_i \leq x\}}{n} - F(x) \right|$$

A la siguiente ecuación, la cual ya trabaja la distribución uniformemente distribuida, y F(x) por Y ∈ [0, 1].

$$D = \sup_{0 \leq y \leq 1} \left| \frac{\#\{i \mid U_i \leq y\}}{n} - y \right|$$

Por lo tanto, se estima el valor p como la proporción de veces que se cumple la desigualdad D≥ d, mediante una simulación:

- Generar un conjunto de números aleatorios U1, … , Un
- Evaluar D y comparar con el valor observado d de la muestra original, ecuación principal:

$$\sup_{0 \leq y \leq 1} \left| \frac{\#\{i \mid U_i \leq y\}}{n} - y \right| \geq d$$

- Repetir procedimiento r veces.

Cierta información, ecuaciones y ejercicio de ejemplo tomadas del libro de Sheldon [3].

A continuación, un ejemplo extraído del libro de Sheldon y explicado en páginas de internet, muestra cómo funciona este procedimiento:

- Si $n = 3$ y $U_1 = 0.7$, $U_2 = 0.6$, $U_3 = 0.4$, entonces

$$U_{(1)} = 0.4, \qquad U_{(2)} = 0.6, \qquad U_{(3)} = 0.7,$$

y el valor D para este conjunto de datos es

$$D = \max\left\{\frac{1}{3} - 0.4,\ \frac{2}{3} - 0.6,\ 1 - 0.7, 0.4, 0.6 - \frac{1}{3}, 0.7 - \frac{2}{3}\right\} = 0.4$$

j	valores	$F(j/n)$	$\frac{j}{n} - F\left(\frac{j}{n}\right)$	$\frac{j-1}{n} - F\left(\frac{j}{n}\right)$
1	66	0,48	-0,38	0,48
2	72	0,51	-0,31	0,41
3	81	0,56	-0,26	0,36
4	94	0,61	-0,21	0,31
5	112	0,67	-0,17	0,27
6	116	0,69	-0,09	0,19
7	124	0,71	-0,01	0,11
8	140	0,75	0,05	0,05
9	145	0,77	0,13	-0,03
10	155	0,79	0,21	-0,11
			$d = 0,48315$	

- Calcular el valor p mediante simulaciones.
- Si el p valor es 0.012, se rechaza la hipótesis nula.

Punto 4. Código en el notebook.

¿Qué otra forma existe para generar muestras de una distribución que consista en la multiplicación de dos exponenciales con una tasa λ?

Una forma de generar muestras utilizando exponenciales con una tasa λ, es usando la distribución Erlang.

La distribución Erlang es un caso particular de la distribución Gamma:
pues si $X \sim \Gamma(\alpha, \lambda)$ con $\alpha = n \in \mathbb{N}$ entonces $X \sim \text{Erlang}(n, \lambda)$

Por ende, podremos utilizar los métodos específicos de generación de aleatorios de esta distribución.

Si $X \sim \text{Erlang}(1, \lambda)$ entonces $X \sim \text{Exponencial}(\lambda)$

Erlang es una distribución de probabilidad continua con soporte $x \in [0, \infty)$ y contiene dos parámetros:
$k \rightarrow$ factor de “forma” de la distribución
$\lambda \rightarrow$ factor de “proporción o tasa” de la distribución

Es la distribución de una suma de k variables exponenciales independientes con una media $1/\lambda$ cada uno.

Disponiendo de toda esta información, la función de distribución acumulada se puede expresar de la siguiente manera:

$$F(x; k, \lambda) = 1 - \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\lambda x} (\lambda x)^n.$$

Pero, en el caso de la simulación, podemos generar variables aleatorias distribuidas por Erlang a partir de números aleatorios distribuidos uniformemente $U \in [0, 1]$ utilizando la siguiente ecuación:

Ecuación Principal (EC Final) → La base para el programa

$$E(k, \lambda) = -\frac{1}{\lambda} \ln \prod_{i=1}^{k} U_i = -\frac{1}{\lambda} \sum_{i=1}^{k} \ln U_i$$

Al momento de la simulación por código, es sencillo formular la ecuación de distribución de erlang para producir variables aleatorias. Se solicita un valor de lambda, se define la cantidad de exponenciales que serán partícipes de la sumatoria de la función de distribución acumulado (k) y se define la cantidad de randoms a generar usando esta distribución Erlang (c).

Por medio de un bucle se genera la cantidad de valores aleatorios solicitados (c) que se irán guardando en una lista, donde, también usando otro bucle, se crean aleatorios (Ui), uno nuevo por cada iteración, los cuales irán trabajando con la fórmula principal para la generación de variables aleatorias (EC Final) y la cantidad de veces a iterar en la sumatoria (k). Obteniendo al final de este bucle un valor(random) como resultado acumulado de la sumatoria. Por último se imprime la lista de todos estos randoms originados por Erlang.

Referencias.

1. Laura Ruiz Mitjana. Prueba de chi-cuadrado. Qué es y cómo se usa. Tomado de: https://psicologiaymente.com/miscelanea/prueba-chi-cuadrado

2. [2] Ross, Sheldon M. Simulation - 5th Edition. Cap 11.1. Goodness of Fit Test. Pág 247-250. Tomado de: Statistical Validation Techniques - ScienceDirect

3. [3] Ross, Sheldon M. Simulation - 5th Edition. Cap 11.1. Goodness of Fit Test. Pág 250-254. Tomado de: Statistical Validation Techniques - ScienceDirect