TP2.1 - GENERADORES PSEUDOALEATORIOS

Carlucci, Gino
Universidad Tecnológica Nacional
Rosario, Santa fe
ginocarlucci@hotmail.com

Docampo, Juan Manuel
Universidad Tecnológica Nacional
Rosario, Santa fe
docampojuan@gmail.com

Menegozzi, Milton
Universidad Tecnológica Nacional
Rosario, Santa fe
miltonmenegozzi@gmail.com

1 de julio de 2020

ABSTRACT

El presente trabajo práctico se basa en el estudio de generadores de números pseudoaleatorios. El objetivo es crear un Generador congruencial lineal y aplicarle al mismo diversas pruebas para determinar la calidad del generador.

1. Introducción

Un número se lo llama pseudoaleatorio ya que el mismo no es obtenido al azar, si no que fue generado por un algoritmo completamente determinista, en el que las mismas condiciones iniciales producen siempre el mismo resultado.
Para que un generador sea aceptado, debe cumplir especificamente dos propiedades:

1. Presentar una distribución uniforme.
2. Presentar independencia entre los números generados.

Por lo tanto, nuestro objetivo es generar sucesiones de numeros independientes que no muestren ningún patrón o regularidad desde el punto de vista estadístico. Más precisamente, lo que queremos lograr son sucesiones de números independientes que se puedan considerar como observaciones de una distribución uniforme en el intervalo (0, 1). Para ello, se programaron dos Generadores:

- Generador parte media de los cuadrados.
- Generador congruencial lineal.

El inconveniente al evaluar estos generadores, es que resulta muy difícil para el ojo humano poder detectar pratones en las sucesiones generadas. Es por eso que para determinar la calidad de los mismos se realizan pruebas estadísticas en lugar de realizar un análisis visual. Las pruebas existentes se pueden clasificar en dos grandes ramas:

- Empiricas: Evalúan estadísticas de sucesiones de números.
- Teoricas: Se establecen las características de las sucesiones usando métodos de teoría de números con base en la regla de recurrencia que generó la sucesión..

Para nuestro generador congruencial lineal (GCL), se aplicarán las siguientes pruebas empíricas:

- Prueba de Bondad de ajuste (Chi Cuadrado)
- Prueba de Paridad
- Prueba de Kolmogorov-Smirnov
- Prueba de corrida

2. Descripción de Generadores:

2.1. Generador media de los cuadrados

Generador creado por Jon Von Neuman para generar secuencias de número pseudoaleatorios de 4 dígitos.

El metodo consiste en 3 simples pasos descriptos a continuación:

1. Se inicia con una semilla de 4 dígitos
2. La semilla se eleva al cuadrado, produciendo un número de 8 dígitos (si el resultado tiene menos de 8 dígitos se añaden ceros al inicio)
3. Los 4 números del centro serán el siguiente número en la secuencia, y se devuelven como resultado

Para nuestro caso, se inicia con una semilla = **1391** y se muestran los resultados de las primeras **10** iteraciones.

Iteración	Semilla	Valor
1	7287	3728761
2	1003	53100369
3	60	1006009
4	36	3600
5	12	1296
6	1	144
7	0	1
8	0	0
9	0	0
10	0	0

Cuadro 1: Generador media de los cuadrados

Como podemos visualizar en el Cuadro 1, el principal problema con este tipo de generador es que el mismo esta condicionado por la semilla que genera la sucesión, ya que al estar en la presencia de ceros, los mismos se propagan y provocan que las suceciones tengan ciclos con una longitud pequeña. Por tal motivo, el mismo no será de análisis en este estudio y las pruebas serán aplicadas al generador GLC que veremos a continuación.

2.2. Generador congruencial lineal

Un generador lineal congruencial es un algoritmo que permite obtener una secuencia de números pseudoaleatorios calculados con una función lineal definida a trozos discontinua. Es uno de los métodos más antiguos y conocidos para la generación de números pseudoaleatorios.
En los generadores congruenciales lineales se considera una combinación lineal de los últimos k enteros generados y se calcula su resto al dividir por un entero fijo m. En el método congruencial simple (de orden $k = 1$), partiendo de una semilla inicial x_0, el algoritmo secuencial es el siguiente:

$$x_{n+1} = (ax_n + c)mod(m) \tag{1}$$

Delimitado por:

- Módulo $m > 0$
- Multiplicador $\leq a < m$
- incremento $c \leq m$
- Módulo $0 \leq x_0 < m$

donde a, c y m son parámetros enteros del generador fijados de antemano.

3. Pruebas

3.1. Descripciones

El fin de realizar las pruebas, es determinar si nuestro generador congruencial lineal puede ser aceptado como un generador de números aleatorios. Como bien mencionamos anteriormente, el mismo debe cumplir con dos propiedades, por lo que las siguientes pruebas serán para determinar la uniformidad de los datos y su aleatoriedad.
Para la sucesión del GCL se generaron 1000 números y se utilizaron los siguientes parámetros:

semilla = 1234
a = 134775813
c = 1
m = 2^{32}

3.1.1. Prueba de Bondad de ajuste Chi Cuadrado

En general un test de bondad de ajuste se utiliza para discriminar si una colección de datos o muestra se ajusta a una distribución teórica de una determinada población. En otras palabras, nos dice si la muestra disponible representa (o ajusta) razonablemente los datos que uno esperaría encontrar en la población.

Hipotesis:

H_0 Los datos son muestra de la distribución uniforme
H_1 Los datos no son muestra de la distribución uniforme

Prueba:

1. Partir la distribución en n celdas que son exhautivas y mutuamente excluyentes.

2. Contar el número de observaciones O_i encontrados en cada celda

3. Calcular el valor esperado en cada celda

$$e_i = np_i \tag{2}$$

4. Calcular la sumatoria de las estadística de prueba

$$x^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \tag{3}$$

5. $X^2 < X_{p,q}$, se acepta la hipotesis H_0

   p: Representa el intervalo de confianza
   q: Representa los grados de libertad

Resultados:

Se obtuvieron los siguentes valores adjuntados en la Cuadro 2

Intervalo	$\frac{(O_i - E_i)^2}{E_i}$
1	0.64
2	0.16
3	0.04
4	0.04
5	0.16
6	0.16
7	1.44
8	0.04
9	1.96
10	0.16
Total	$\sum_{i=1}^{10} \frac{(o_i - e_i)^2}{e_i} = 4{,}8$

Cuadro 2: Sumatoria estadísticas de prueba

Para un intervalo de confianza del 95 % y 9 grados de libertad, se obtuvo el valor de la tabla Chi Cuadrado = 19.9189

Como $X^2 < X_{0,05,9}$, se acepta la hipotesis H_0 y podemos afirmar que no existe diferencia entre la distribución de números generada y la distribución uniforme.

3.1.2. Prueba de Kolmogorov-Smirnov

La prueba de Kolmogórov-Smirnov es una prueba perteneciente a la estadística, concretamente a la estadística inferencial. La estadística inferencial pretende extraer información sobre las poblaciones. Al igual que la prueba Chi Cuadrado, es un procedimiento de bondad de ajuste, es decir, permite la medición del grado de concordancia existente entre la distribución de un conjunto y una distribución teorica específica. Por lo tanto, nuestro objetivo es señalar y determinar si la secuencia de numeros generada por el generador GCL provienen de una población que tiene una distribución uniforme.

A diferencia de la prueba anterior, esta es más fácil de calcular y usar, ya que no requiere agrupación de los datos y el estadístico es independiente de la distribución de frecuencias esperada, solo depende del tamaño de la muestra.

Hipotesis:

H_0 Los datos pertenecen a una distribución uniforme
H_1 Los datos no pertenecen a una distribución uniforme

Prueba:

1. Ordenar la serie de menor a mayor $x_i, ..., x_n$
2. Calcular las desviaciones maximas

$$D^+ = max_{1 \leq i \leq N} \left[\frac{i}{N} - R_i \right] \tag{4}$$

$$D^- = max_{1 \leq i \leq N} \left[\frac{i}{N} - R_i \right] \tag{5}$$

3. Obtener el valor máximo de las desviaciones

$$D_{max} = max(D^+, D^-) \tag{6}$$

4. Obtener el valor crítico Dα de la tabla Kolmogorov-Smirnov
5. D <Dα Se acepta la hipótesis H_0

Resultados:

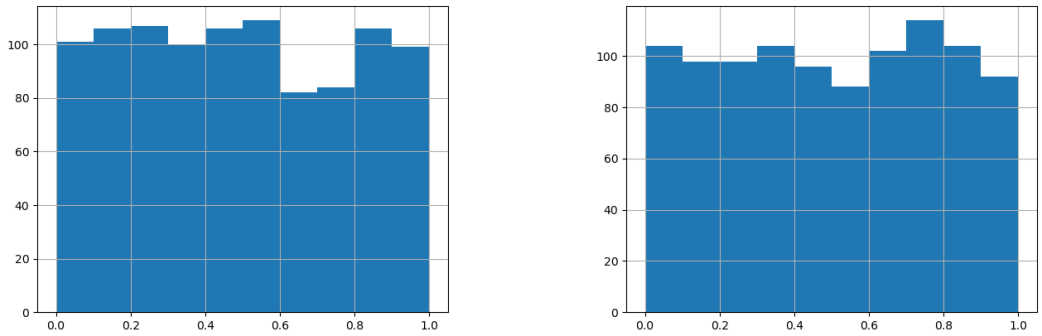
D^+	D^-	D_{max}	Dα
0.02369243192672732	-0.021692431926727318	0.02369243192672732	0.04277658868221006

Cuadro 3: Prueba de Kolmogorov-Smirnov

Al observar los resultados, podemos ver que el valor crítico obtenido de la tabla KS para un nivel de significancia $0{,}05$ y tamaño 1000 es menor al valor máximo D obtenido. Por lo tanto, podemos aceptar la Hipótesis H_0 y afirmar que la sucesión de números generada pertenecen a una distribución uniforme

3.1.3. Analisis Visual

Para complementar las pruebas de bondad de ajuste, graficaremos un histograma para ver la distribución de valores en el intervalo $(0, 1)$
Si se reparte el intervalo en subintervalos de igual longitud, se esperará encontrar la misma cantidada de datos en cada subintervalo.



(a) Numpy (b) GCL

Figura 1: Histogramas de las distribuciones generadas

Como podemos observar en la Figura 1, y con las pruebas previas realizadas, ambos generadores siguen una distribución Uniforme. A continuación, procederemos a aplicar una prueba de paridad y otra de aleatoriedad para determinar la calidad del generador.

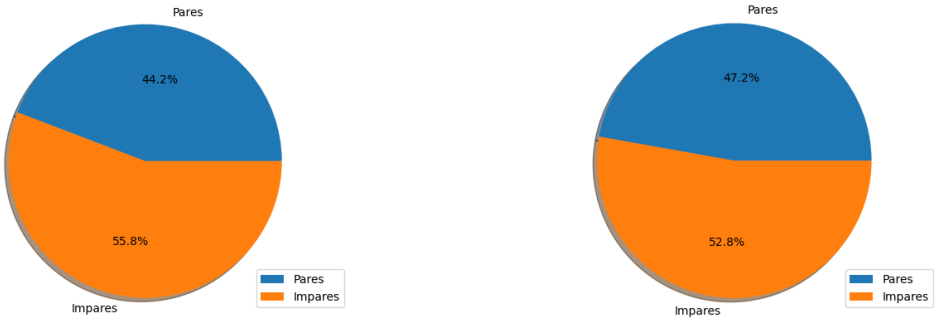
3.1.4. Prueba de Paridad

Prueba:

La prueba de paridad se basa en el supuesto de que, al generar n cantidad de numeros pseudialatorios con una distribucion uniforme, deberia suceder que cerca del 50 % de ellos pertenecen a los pares y el otro 50 % restante pertenecen a los impares. Para complementar el estudio, se incluirá el generador de la biblioteca científica de Python (NumPy) para comparar ambos resultados.

Para llevar adelante la prueba, se analizaran los porcentajes de paridad mediante un gráfico pastel con sus respectivos porcentajes.

Resultados:



(a) Numpy (b) GCL

Figura 2: Porcentajes de paridad obtenida de los generadores

Como podemos observar para ambos gráficos, los porcentajes obtenidos son similares a los esperado, por lo que podemos afirmar que ambos generadores pasan la prueba de paridad.
A medida que se incrementa la cantidad de números generados, es de esperarse que los porcentajes tiendan a acercarse al 50 %

3.1.5. Prueba de corrida ascendente

A diferencia de la Prueba de Kolmogorov-Smirnov y Chi cuadrado, donde se buscar validar que la sucesión generada tenga una distribución uniforme, lo que buscaremos con la prueba de corrida es verificar la independencia de los números generados.

Un test de Corridas es un método que nos ayuda a evaluar el carácter de aleatoriedad de una secuencia de números estadísticamente independientes y números uniformemente distribuidos. Es decir dado una serie de números determinar si son o no aleatorios. Los números pueden estar uniformemente distribuidos y aún no ser independientes uno de otros. Por tal motivo, la siguiente prueba será evaluar el caracter de aleatoriedad.

Hipótesis:

H_0 La secuencia de números es independiente, y por lo tanto, aleatoria.
H_1 La secuencia de números no es independiente, y por lo tanto, no es aleatoria.

Prueba:

1. Asignar a cada número de la secuencia un signo $+$ o $-$ hasta $(N-1)$, según el siguiente criterio:

   Si $X_i < X_{i+1}$ asignar un $+$
   Si $X_i > X_{i+1}$ asignar un $-$
   (N representa el tamaño de la muestra).
2. Calcular el total de corridas que resultan de la suma de suma de corridas ascendente con la descendente, es decir, calcular el total de cambios de signos dentro de la sucesión generada aplicando el criterio anterior.
3. Calcular la media y la varianza:
$$\mu = \frac{2N-1}{3} \tag{7}$$
$$\sigma^2 = \frac{16N-29}{90} \tag{8}$$
4. Calcular la estadística de prueba:
$$z = \frac{a1-\mu}{\sigma} \tag{9}$$

5. Obtener el valor $Z_{1-\alpha/2}$ de una distribución normal.
6. Si $Z < Z_{1-\alpha/2}$ se acepta la hipótesis H_0

Resultados:

Para la sucesión de números generados por el generador GLC y el generador de la biblioteca Numpy, ambos con un $N = 1000$ y $\alpha = 0{,}05$, se obtuvieron los resultados expresados en la siguiente Tabla:

Como el tamaño para ambos generados fue el mismo, se obtuvo:

$$\mu = 666{,}33$$
$$\sigma^2 = 177{,}45$$

Generador	a	Z	$Z_{1-\alpha/2}$	Prueba corrida
GLC	490	13,23	1,9559	False
NumPy	662	0,3252	1,9559	True

Cuadro 4: Resultados prueba de corridas ascendente

Al observar los resultados para el generador GLC, podemos ver que $Z > Z_{1-0{,}05/2}$, por lo que aceptamos la hipótesis H_1 y concluimos que los números no provienen de un generador Aleatorio.
Por otro lado, para el caso del generador de NumPy, $Z < Z_{1-0{,}05/2}$, por lo que podemos aceptar la hipótesis H_0 y concluir que los números generados son independientes.

4. Conclusiones

Como se ha estudiado, para que un generador de números sea útil de ser usado, debe cumplir con dos propiedades, la uniformidad en el intervalo (0,1) y la independencia entre términos generados. De acuerdo a las pruebas realizadas y la evidencia presentada, podemos afirmar que el generador GLC presenta una distribución Uniforme, al haber pasado las pruebas de Bondad de ajuste, Por otro lado, el mismo no genera una secuencia de números aleatorios, (los números generados no son independientes entre sí) ya que no pasó la prueba de las corridas.
Se concluye que el generador GLC, con los parametros (semilla = 1234, $a = 134775813$, $c = 1$, $m = 2^{32}$, $N = 1000$) no es apto para ser usado como un generador de números aleatrorios.

Referencias

[1] Generador GCL. In *https://www.lawebdefisica.com/apuntsmat/num$_a leatorios//$*.

[2] Numeros pseudoaleatorios y pruebas. In *https://tereom.github.io/est-computacional-2018/numeros-pseudoaleatorios.html//*.

[3] Pruebas de bondad de ajuste. In *http://www.juntadeandalucia.es/averroes/centros-tic/14002996/helvia/aula/archivos/repositorio/250/295/html/estadistica/bondad.htm*.

[4] Pruebas de corrida. In *https://campusvirtual.univalle.edu.co/moodle/pluginfile.php/1174508/mod$_r esource/content/1/SC1.pdf$*.