TP 2.1 - GENERADORES PSEUDOALEATORIOS

Abud Santiago Elias
Legajo 47015
sabudvicco@gmail.com

Buchhamer Ariel
Legajo 46217
arielbuchhamer1@outlook.com

Castellano Marcelo
Legajo 39028
marce.geek22@gmail.com

Dolan Guillermo Patricio
Legajo 46101
guillermo230899@gmail.com

Navarro Franco
Legajo 46387
franconavarro1889@gmail.com

6 de mayo de 2022

ABSTRACT

Los números pseudoaleatorios se generan de manera secuencial con un algoritmo determinístico. Construir un buen algoritmo de números pseudoaleatorios es complicado, por eso hemos hecho un estudio sobre cómo funcionan y de la manera que se comportan.

1. Introducción

Un proceso que parece generar números pseudoaleatorios al azar, pero no lo hace realmente. Una forma de analizar esto y probarlos es con los generadores de números pseudoaleatorios. Uno de estos generadores es el método de los cuadrados medios, que fue generado en 1946 y fue uno de los más populares, otro es el de GCL que en el presente es el más utilizado.

2. Descripción del trabajo

Utilizando en Python 3.7, reproducimos y comparamos algoritmos generadores de números pseudoaleatorios, poniendo a prueba la aleatoriedad de cada método por medio de diferentes pruebas.

Los algoritmos generadores de números pseudoaleatorios utilizados fueron:

1. Método de los cuadrados medios.
2. Generador lineal congruencial (GLC), en sus formas:
   a) Ansi
   b) Randu

Y los "tests.º pruebas realizadas, fueron:

1. Test de Chi Cuadrado
2. Test de póker

3. Marco teórico

Existen diferentes generadores números pseudoaleatorios, uno de ellos es el método de los cuadrados medio que Jon Von Neuman sugirió usar las operaciones aritméticas de una computadora para generar secuencias de número pseudoaleatorios. Con este procedimiento se generan números pseudoaleatorios de 4 dígitos de la siguiente forma:

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.

Este generador cae rápidamente en ciclos cortos, por ejemplo, si aparece un cero se propagará por siempre. A inicios de 1950s se exploró el método y se propusieron mejoras, por ejemplo para evitar caer en cero. Metrópolis logró obtener una secuencia de 750,000 números distintos al usar semillas de 38 bits (usaba sistema binario), además la secuencia de Metrópolis mostraba propiedades deseables. No obstante, el método del valor medio no es considerado un método bueno por lo común de los ciclos cortos.

Otro generador de número pseudoaleatorio es el de GLC que se introdujeron en 1949 por D.H. Lehemer, son muy populares y los más utilizados. Los generadores como rand y randu se denominan generadores congruenciales. Tienen la forma:

$$X_{n+1} = (aX_n + c)mod(m) \tag{1}$$

Donde a es el multiplicador, m el módulo, c el incremento y X la semilla.

Vale la pena notar que un periodo grande no determina que el generador congruencial es bueno, debemos verificar que los números que generan se comportan como si fueran aleatorios.Los GCLs continúan siendo utilizados en muchas aplicaciones porque con una elección cuidadosa de los parámetros (la elección de los parámetros determina la calidad del generador) pueden pasar muchas pruebas de aleatoriedad, son rápidos y requieren poca memoria.

Recordemos que lo que nos interesa para trabajar con un buen generador de números aleatorios es que la distribución de los números obtenidos tiene que ser uniforme, no deben de haber correlaciones entre los términos de la secuencia, el periodo debe ser lo más largo posible, y el algoritmo debe ser de ejecución rápida.

El problema es saber que generador de números es mejor, ya que la razón es que si su generador de números aleatorios es bueno, es igualmente probable que aparezca cada posible secuencia de valores . Esto significa que un buen generador de números aleatorios también producirá secuencias que parecen no aleatorias para el ojo humano y que también fallan en cualquier prueba estadística a la que podamos exponerlo.

Es imposible probar definitivamente la aleatoriedad. Una forma de aproximar esto es tomar muchas secuencias de números aleatorios de un generador dado y someterlos a una batería de pruebas estadísticas. A medida que las secuencias pasan más pruebas, aumenta la confianza en la aleatoriedad de los números y también la confianza en el generador. Sin embargo, debido a que esperamos que algunas secuencias no parezcan aleatorias, debemos esperar que algunas de las secuencias fallen al menos en algunas de las pruebas. Sin embargo, si muchas secuencias fallan en las pruebas, deberíamos sospechar.

Hay varias formas de examinar un generador de números aleatorios, las distintas pruebas estadisticas son: Analisis visual simple, Análisis estadístico de Charmaine Kenny y Análisis estadístico de Louise Foley.

3.1. Pruebas estadísticas

3.1.1. Prueba χ^2 de Pearson

La prueba χ^2 de Pearson se considera una prueba no paramétrica que mide la discrepancia entre una distribución observada y otra teórica (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis.[**eswiki2022pearson**]

Pone a prueba una hipótesis nula que establece que la distribución de frecuencia de ciertos eventos obervados en una muestra es consistente con una distribución teórica en particular. Los eventos considerados deben ser mutuamente excluyentes y tener una probabilidad total igual a 1. Un caso común para esto es donde cada uno de los eventos cubren un resultado de una variable categórica.

La prueba χ^2 de Pearson se usa para realizar distintos tipos de análisis. Para nuestro estudio, como los valores generados son números enteros y deberían estar distribuidos de manera uniforme, llevamos a cabo pruebas de bondad de ajuste a la distribución uniforme discreta.

Una prueba de bondad de ajuste determina si una distribución de frecuencia observada difiere de una distribución teórica dada.

Planteamos el test de hipótesis siguiente, la hipótesis nula (H_0) de que los datos siguen la distribución esperada y la alternativa (H_1), de que los datos no siguen la misma:

$$H_0 : \text{no hay diferencia entre las distribuciones}$$
$$H_1 : \text{hay diferencia entre las distribuciones}$$

El procedimiento de cálculo de la prueba de χ^2 para bondad de ajuste incluye los pasos siguientes:

1. Calcular el estádistico χ^2, que se asemeja a una suma normalizada de los desvíos entre las frecuencias observadas y las teóricas al cuadrado (ecuación **??**).
2. Determinar los grados de libertad, **gl**, del estadístico: para nuestro caso, de bondad de ajuste, gl = $n - m$, donde n es el número de valores distintos de la distribución, y m es el número de parámetros ajustados para hacer que la distribución se ajuste mejor a las observaciones: el número de valores reducidos por el número de parámetros ajustados en la distribución.
3. Seleccionar el nivel deseado de confianza (nivel de significancia, valor p[1] o el nivel alfa correspondiente) para el resultado de la prueba. Usualmente y para nuestro caso seleccionamos un alfa de 0.05, lo que corresponde a un 95 % de confianza.
4. Comparar χ^2 con una distribución χ^2 con gl grados de libertad para obtener el valor p correspondiente y emplear y el nivel de confianza seleccionado (de un solo lado, puesto que la prueba es solamente en una dirección, esto es, ¿es el valor del estadístico de prueba mayor que valor el crítico? o ¿es el valor p menor o igual al alfa?), lo que en muchos casos da una buena aproximación de la distribución χ^2.
5. Rechazar o mantener la hipótesis nula de que la distribución de frecuencias observadas es la misma que la teórica empleando el valor p correspondiente. Si el valor p es menor o igual que el nivel alfa seleccionado, se rechaza la hipótesis nula (H_0) y se acepta la alternativa (H_1), con el nivel de confianza seleccionado. Si en cambio el valor p supera dicho umbral, no se puede llegar a una conclusión clara, y se mantiene la hipótesis nula (no la podemos rechazar), lo que no significa necesariamente que la misma sea aceptada.

3.1.1.1 Prueba de bondad de ajuste - distribución discreta uniforme

En este caso se dividen N observaciones entre n valores. Una aplicación simple es probar la hipótesis de que, en la población general, los distintos valores se producirán con la misma frecuencia. La frecuencia teórica absoluta para cualquier valor (bajo la hipótesis nula de una distribucíón discreta uniforme) se calcula como

$$E_i = \frac{N}{n},$$

y la reducción en los grados de libertad es $p = 1$, dado que las frecuencias observadas O_i deben cumplir con la restricción de sumar N. Esto es, los grados de libertad resultan ser gl = $n - 1$

El valor del estadístico de prueba es

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i} \tag{2}$$

donde

- χ^2: estadístico acumulativo de prueba de Pearson, que se aproxima asintóticamente a una distribución χ^2
- O_i: número de observaciones de tipo i
- $E_i = Np_i$: la cantidad esperada (teórica) de observaciones de tipo i, afirmada por la hipótesis nula de que la fracción de observaciones de tipo i en la población es p_i

Finalmente, el estadístico χ^2 puede entonces emplearse para calcular un valor p comparando el valor del estadístico con una distribución χ^2. Esto es

$$\text{valor p} = 1 - P(X \leq \chi^2) \tag{3}$$

Como ya se ha mencionado, si el valor p resulta menor o igual que el nivel de significancia $\alpha = 0{,}05$, rechazamos la hipótesis nula y concluimos con un 95 % de confianza que los valores generados difieren de manera estadísticamente significativa de una distribución discreta uniforme y por lo tanto no son aleatorios, en caso contrario, concluimos que no se puede rechazar la hipótesis nula y por lo tanto los valores pueden ser aleatorios.

[1] Probabilidad de obtener resultados de prueba al menos tan extremos como los observados, bajo la suposición de que la hipótesis nula es correcta[**enwiki2022pvalue**]

Implementación En nuestro trabajo empleamos la función `scipy.stats.chisquare`, que se encarga tanto de calcular el estadístico como el valor p correspondiente y ya tiene en cuenta por defecto la reducción en los grados de libertad $p = 1$.

4. Análisis de resultados

5. Conclusiones