Trabajo Práctico Generador Números Aleatorios

Bengoechea, Guadalupe María[a], Corsetti,Ornela Milagros[a], Garcia, Narela Rosana[a], Golzman,Gabriel[a]

[a] *Universidad Tecnológica Nacional Regional Rosario*

Enunciado

Este trabajo consiste en la generación de 10000 números aleatorios mediante la aplicación del método del cuadrado de los medios, un Generador Congruencial mixto y mediante el uso de la página web *Random.org*.
A estos 3 métodos, comprobamos su uniformidad con la utilización de un test de Chi-Cuadrado y luego aplicamos el test de corridas con 4000 numeros aleatorios

Palabras Clave:
Generador,python,gráficas,excel,chi-cuadrado

1. Marco Teórico

1.1. Introducción

Los números aleatorios son la base esencial de la simulación. Usualmente, toda la aleatoriedad involucrada en el modelo se obtiene a partir de un generador de numeros aleatorios que produce una sucesión de valores que supuestamente son realizaciones de una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) $U(0, 1)$

Posteriormente estos números aleatorios se transforman convenientemente para simular las diferentes distribuciones de probabilidad que se requieran en el modelo. En general, la validez de los métodos de transformación dependen fuertemente de la hipótesis de que los valores de partida son realizaciones de variables aleatorias iid $U(0, 1)$, pero esta suposición realmente no se cumple, puesto que los generadores de números aleatorios son simplemente programas determinısticos que intentan reproducir una sucesión de valores que parezca aleatoria.

1.2. Generador de números pseudo-aleatorios

El método mas conveniente y mas fiable de generar números aleatorios es utilizar algoritmos determinısticos que posean alguna base matemática solida. Estos algoritmos producen una sucesión de números que se asemeja a la de una sucesión de realizaciones de variables aleatorias iid $U(0, 1)$, aunque realmente no lo sea. Es por ello que este tipo de números se denominan pseudo-aleatorios y el algoritmo que los produce se llama generador de números pseudo-aleatorios.

Un generador de números (pseudo)aleatorios es una estructura $G = (X, x_0, T, U, g)$, donde X es un conjunto finito de estados, $x_0 \in X$ es el estado inicial(semilla), la aplicación $T : X \to X$ es la función de transición, U es el conjunto finito de posibles observaciones, y $G : X \to U$ es la función de salida.

1.3. Propiedades de un buen generador de números aleatorios

- Los valores generados están distribuidos uniformes en el intervalo $(0, 1)$.
- Los valores generados son valores correlacionados es decir independientes en términos estadísticos.
- Es posible regenerar una secuencia de números aleatorios.Esto es necesario cuando se quiere comparar dos o más alternativas y se desea someterla a la misma situación de eventos generados.
- Debe ser computacionalmente eficiente.

1.4. Distribución de χ^2

La distribución de chi-cuadrada es una distribución continua que se especifica por los grados de libertad y el parámetro de no centralidad. La distribución es positivamente asimétrica, pero la asimetría disminuye al aumentar los grados de libertad.

La distribución de χ^2 es utilizada en pruebas de significancia estadística para:

- Comprobar qué tan bien se ajusta una muestra a una distribución teórica. Por ejemplo, puede utilizar una prueba de bondad de ajuste de chi-cuadrada para determinar si los datos de la muestra se ajustan a una distribución de Poisson.
- Comprobar la independencia de las variables categóricas.

Cuando los grados de libertad son 30 o más, la distribución de chi-cuadrada puede aproximarse razonablemente con una distribución normal, como se ilustra en las siguientes gráficas:

Correos Electrnicos: `guadabengoechea@gmail.com` (Bengoechea, Guadalupe María), `ocorsetti@frro.utn.edu.ar` (Corsetti,Ornela Milagros), `nargarcia@frro.utn.edu.ar` (Garcia, Narela Rosana), `ggolzman@frro.utn.edu.ar` (Golzman,Gabriel)

Figura 1: Distribución de chi-cuadrada con 20 grados de libertad

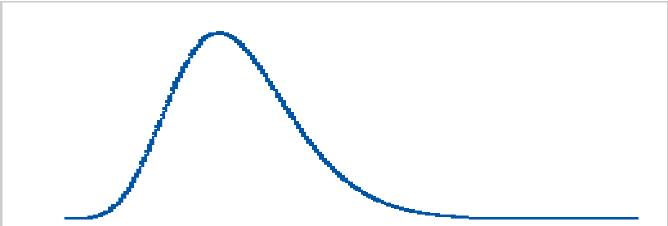
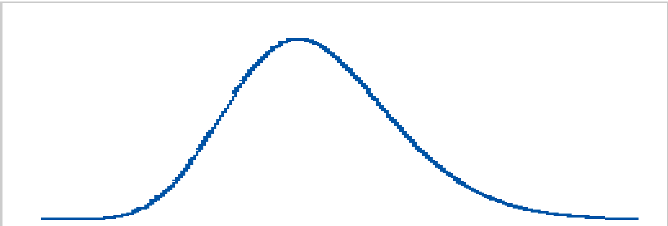


Figura 2: Distribución de chi-cuadrada con 40 grados de libertad



1.5. Generador Congruencial lineal

Un generador lineal congruencial (GLC) 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.

La teoría que sustenta el proceso es relativamente fácil de entender, el algoritmo en si es de fácil implementación y su ejecución es rápida, especialmente cuando el hardware del ordenador puede soportar aritmética modular al truncar el bit de almacenamiento correspondiente.

$$X_{(n+1)} = (\alpha X_n + c) \mod m$$

donde X es la secuencia de valores pseudoaleatorios, y

$$m, \ 0 < m$$

es el ”módulo”

$$\alpha, \ 0 < \alpha < m$$

es el ”multiplicador”

$$c, \ 0 \leq c < m$$

es el ïncremento”

$$X_0, \ 0 \leq X_0 < m$$

es la ”semilla.º ”valor inicial”

2. Metodología

2.1. Método cuadrado de los medios

El método del cuadrado de los medios para la generación de números aleatorios consiste en los siguientes pasos:

- Obtener un entero de 4 cifras como la semilla del generador (Z_0).
- Elevar cada Z_i al cuadrado.
- Completar con ceros a la izquierda si el número obtenido no tiene 8 cifras.
- Tomar las 4 cifras centrales como el próximo Z_i.
- Calcular $U_i = Z_i/10000$.
- Iterar.

2.2. Generador Congruencial lineal mixto($c > 0$)

Utilizando la fórmula:

$$Z_i = (\alpha * Z_{(i-1)} + c) \quad \text{mód } m$$

siendo
a el multiplicador y entero > 0
c el incremento y entero > 0
m es el módulo
Z_i es entero > 0
Z_0 es la semilla a partir de la cuál se puede regenerar una secuencia.

Además debe ser:
$m > 0$
$m > a$
$m > c$
$m > Z_0$

La metodología para este generador,Una vez determinados los parámetros enunciados anteriormente, consiste en:

- Obtener la semilla Z_i.
- Realizar el cálculo de U_i, de manera que:

$$U_i = \frac{Z_i}{m}$$

- Iterar.

Posteriormente, se debe comprobar la validez los números generados con dichos parámetros. Esto se realiza con el siguiente teorema.

- El único entero que divide a m y c en forma simultanea (primos entre si) es 1.
- Si q es un numero primo que divide a m, entonces q divide a $a - 1$.
- Si 4 divide a m, entonces 4 divide a $a - 1$.

Para verificar la validez de los números generados se deben cumplir todos y cada uno de los ítems enumerados anteriormente. De pasarlo, se dice que el generador es de *período completo*, y por ende es aceptable para simulación.

2.3. Test de Chi-Cuadrado

La prueba de Chi-Cuadrado se realiza calculando el valor de Chi-Cuadrado con la muestra de valores aleatorios obtenidos de los distintos generadores y posteriormente comparándolo con el valor esperado fijo de Chi-Cuadrado según la tabla. Para calcular el valor que toma la muestra se utiliza el siguiente método.

- Se divide el intervalo (0,1) en k subintervalos de igual longitud con $k \geq 100$.
- Generar U_i con $i = 1, ..., n$ con n determinado de manera que $\frac{n}{k} \geq 5$.
- Calcular f_i donde f_i es la frecuencia absoluta de la cantidad de U_i que cayeron en el intervalo j.
- Calcular la variable de Chi-Cuadrado $\chi^2 > \frac{k}{n} \sum_{j=1}^{k} (f_i - \frac{n}{k})^2$.
- Comparar con el valor de la tabla de un Chi-cuadrado con k-1 grados de libertad, donde $1 - \alpha$ nivel de confianza $\chi^2 > {\chi^2}_{k-1,1-\alpha}$
- Si la desigualdad se comprueba entonces se rechaza la hipótesis nula H_0 de que los números aleatorios se comportan uniformemente.

2.4. Test empírico de Corridas

Se utiliza para determinar secuencias, patrones sistemáticos, cambios secuenciales, etc., que denoten tendencias predicibles.
Ejemplos de estas secuencias son los cambios en los indices de la bolsa de valores, las decisiones de compra de los consumidores en una presentación, los niveles de consumo después de una campaña de publicidad, los votos hacia un candidato después de un debate televisado, etc.

Para el tema de estudio, se asigna un signo + para un .exito”, y un signo - para un ”fracaso”. De esta forma se tiene una cadena de + y de -. Por ejemplo, para la siguiente secuencia de + y de -:

+ + - - + + + - - - + - - + + + - - +

Cada grupo de + y de - se identifica como una corrida. En la secuencia mostrada existen 19 observaciones, 9 corridas, 10 .exitos”(signos +), 9 ”fracasos”(signos -)

El número de corridas en una secuencia se usa como un estadístico para determinar si existe aleatoriedad o no, en una secuencia de observaciones.

Los pasos que utilizamos para realizar este test son:

- Generar U_i con $i = 1, n$ con $n \leq 4000$
- Examinar los U_i generados identificando las subsecuencias crecientes y continuas de U_i de longitud máxima.
- Calcular los r_i donde

$$r_i = \left\{ \frac{\text{cantidad de subsecuencias de longitud i para i = 1,2,3,4,5}}{cantidaddesubsecuenciasdelongitud \geq 6 \quad para \quad i = 6} \right\}$$

- Calcular la variable Chi-cuadrado R $= \frac{1}{n} \sum_{i=1}^{6} \sum_{j=1}^{6} a_{ij}(r_i - nb_i)(r_j - nb_j)$
- Comparar $R > \chi^2_{6,1-\alpha}$ donde $\chi^2_{6,1-\alpha}$ es el valor de tabla de una Chi-cuadrado con 6 grados de libertad y $1 - \alpha$ de nivel de confianza.

Si la desigualdad se comprueba entonces se rechaza la H_0 hipótesis nula de que los números aleatorios son independientes(no están correlacionados)

La matriz a_{ij}

$$\begin{bmatrix} 4{,}529.4 & 9{,}044.9 & 13{,}568 & 18.091 & 22{,}615 & 27{,}892 \\ 9{,}044.9 & 18{,}097 & 27{,}139 & 36{,}187 & 45{,}234 & 55{,}789 \\ 13{,}568 & 27{,}139 & 40{,}721 & 54{,}281 & 67{,}852 & 83{,}685 \\ 18{,}091 & 36{,}187 & 54{,}281 & 72{,}414 & 90{,}470 & 111{,}580 \\ 22{,}615 & 45{,}234 & 67{,}852 & 90{,}470 & 113{,}262 & 139{,}476 \\ 27{,}892 & 55{,}789 & 83{,}685 & 111{,}580 & 139{,}476 & 172{,}860 \end{bmatrix}$$

El vector b_i o b_j

$$(b_1, b_2, \ldots, b_6) = (\tfrac{1}{6}, \tfrac{5}{24}, \tfrac{11}{120}, \tfrac{19}{720}, \tfrac{29}{5040}, \tfrac{1}{840})$$

2.5. Hipótesis

2.5.1. Método cuadrado de los medios.

Nuestra hipótesis consiste en que los valores generados con este método no son utilizables para simulación, ya que estos no cumplen con la propiedad de uniformidad en el intervalo (0,1). Y tampoco con la de independencia, ya que, todo U_i está determinado por $U_[i-1]$. Esto se debe principalmente a que los valores tienden a cero muy rápidamente y luego no vuelven a valores cercanos a 1.

2.5.2. Método Generador Congruencial lineal mixto($c > 0$)

Este método de generación de números aleatorios puede ser o no aceptable para usar en simulación dependiendo de los parámetros Z_0, a, cym que sean utilizados. En nuestro caso, utilizamos diferentes valores para probar el generador, algunos de estos deberían pasar el teorema 2.2, y por lo tanto, puede ser un buen generador de números pseudo aleatorios. Las combinaciones de valores que decidimos para los parámetros (Z_0, a, c, m) respectivamente es la siguiente:

- $(7, 5, 3, 16)$: Debería superar el teorema.
- $(7, 4, 3, 16)$: No debería superar el teorema.
- $(7, 5, 4, 16)$: No debería superar el teorema.
- $(7, 4, 4, 16)$: No debería superar el teorema.

3. Casos de estudio

3.1. Método cuadrado de los medios

Con un $Z_0 = 1009$ y una cantidad de corridas igual a quince, realizamos el método de la sección 2.1 y los resultados los fuimos guardando en una *lista*, la cual luego mediante la librería *xlwt* pudimos plasmar en un libro de *Excel*.

i	Zi	Ui	Zi2
0	1009		1018081
1	180	0,18	32400
2	324	0,324	104976
3	1049	1,049	1100401
4	1004	1,004	1008016
5	80	0,08	6400
6	64	0,064	4096
7	40	0,04	1600
8	16	0,016	256
9	2	0,002	4
10	0	0	0
11	0	0	0
12	0	0	0
13	0	0	0
14	0	0	0

3.2. Generador Congruencial lineal mixto($c > 0$)

Primero, definimos los valores iniciales

- $m = 16$
- $c = 3$
- $a = 5$
- $Z_0 = 7$

Generaremos 20 números aleatorios que guardaremos en una lista mediante el método explicado y posterior comprobación enunciada del teorema en 2.2. Para este primer set de parámetros, se comprueba el teorema que verifica la validez de dichos parámetros. Luego, probamos el generador con los parámetros $(a = 16, c = 3, a = 4, Z_0 = 7)$,$(a = 16, c = 4, a = 5, Z_0 = 7)$,$(a = 16, c = 4, a = 4, Z_0 = 7)$.

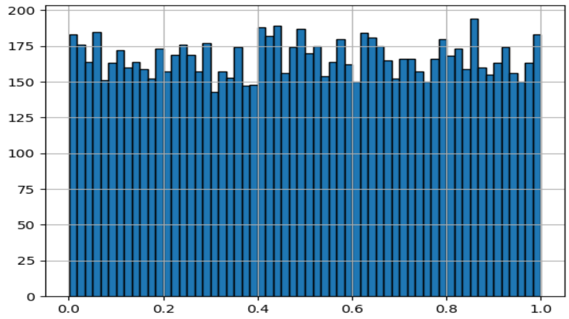
En ninguno de estos caso, los parámetros dados resultan validos para formar un GCL aceptable.

3.3. Random() de Python-Test de Chi-Cuadrado

Declaramos en primera instancia una lista vacía donde se guardaran los números aleatorios generados, definimos la cantidad de corridas (en este caso, 10000), y luego establecemos un $k = 100$, que seran la cantidad de subintervalos de igual longitud tal que $k \geq 100$.

Con el llamado a la función *random* de *python*, fuimos llenando en un bucle la lista con los valores aleatorios, a los cuales luego aplicamos el calculo de sus frecuencias absolutas de la cantidad de U_i que cayeron en el intervalo j como se especifica en la sección 2.3.

Figura 3: Distribución de las frecuencias absolutas obtenidos dentro de los subintervalos



Luego, efectuamos el test de Chi-Cuadrado tal como lo específica la sección 2.3 utilizando la librería científica de *Python*, *scipy* con la función *stats.chi2.ppf* que toma como parámetros un nivel de confianza elegido (en nuestro caso, 0,95) y $k - 1$ grados de libertad (en nuestro caso, 99 grados de libertad). Esto da como resultado:

$$\chi^2{}_{k-1,1-\alpha} = 123{,}22$$

La variable χ^2 calculada posee el valor:

$$\chi^2 = 95{,}68$$

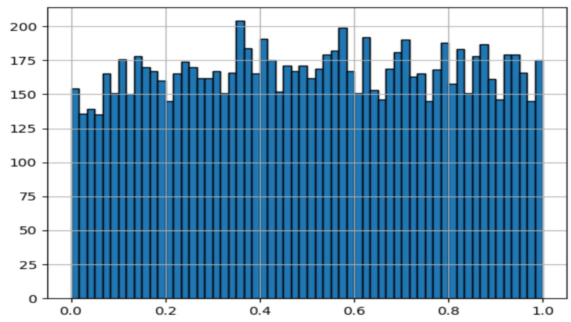
Se acepta entonces la hipótesis nula(H_0) porque no se comprueba la desigualdad $\chi^2 > \chi^2{}_{k-1,1-\alpha}$, y los valores se comportan uniformemente.

3.4. GCL-Test de Chi-Cuadrado

Utilizamos el mismo código de la sección 2.2 para generar los 10000 números aleatorios pero con distintos parámetros para la generación, en este caso utilizamos el modelo propuesto por Coveyou, con parámetros:

- $m = 2^{35}$
- $a = 5^{15}$
- $c = 3$
- $Z_0 = 7$

Prosiguiendo de forma similar que en la sección 3.3 , calculamos las frecuencias absolutas de la cantidad de U_i que cayeron en el intervalo j.



Realizamos el test de Chi-Cuadrado con la misma metodolgía y función estadística que en al sección 3.3 para obtener la variable χ^2 calculada:

$$\chi^2 = 128{,}5$$

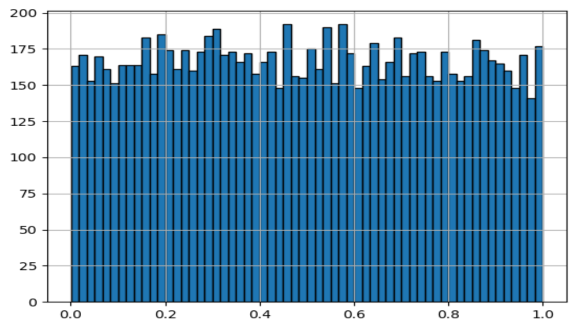
Recordemos que:

$$\chi^2{}_{k-1,1-\alpha} = 123{,}22$$

Al comprobarse la desigualdad $\chi^2 > \chi^2{}_{k-1,1-\alpha}$, se rechaza la hipótesis nula(H_0), lo que significa que los valores no se van a comportar uniformemente.

3.5. Random.org-Test de Chi-Cuadrado

Esta vez, los 10000 valores aleatorios los conseguimos utilizando la función *requests* para poder solicitar los valores al sitio web, dichos valores los guardamos en una lista para proceder de forma similar que en los dos casos anteriores.



Luego de aplicar el test, nuestra variable χ^2 calculada resulta:

$$\chi^2 = 107{,}46$$

Recordemos que:

$$\chi^2{}_{k-1,1-\alpha} = 123{,}22$$

No se comprueba la desigualdad($\chi^2 > \chi^2{}_{k-1,1-\alpha}$), se rechaza la hipótesis nula(H_0) y se confirma que los valores se comportan uniformemente.

3.6. Test de Corridas

Habiendo declarado la matriz a y el vector b (con los valores enunciados en 2.4), establecemos una constante $n = 4000$ con el numero de números aleatorios que vamos a generar, y declaramos la lista r que vamos a llenar con las frecuencias absolutas de las subsecuencias con longitud correspondiente a la posición de la lista donde se va a guardar.

Aplicamos la metodología para el calculo de R, especificado en la sección 2.4. En este caso:

$$R = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} a_{ij}(r_i - nb_i)(r_j - nb_j) = 5{,}65$$

Tomando una variable Chi-Cuadrado (obteniéndola con la función *stats.chi2.ppf*) con 0,9 de nivel de confianza y 6 grados de libertad, tal que:

$$\chi^2_{6,1-\alpha} = 10{,}64$$

Entonces, se comprueba la desigualdad($\chi^2_{6,1-\alpha} >= R$) y no se rechaza la H_0, por ende los valores se comportan de manera independiente.