Universidad Nacional Santiago Antúnez de Mayolo
Facultad de Ciencias
Escuela Profesional de Estadística e Informática

COMPUTACIÓN ESTADÍSTICA

M.Sc. Emerson D. Norabuena Figueroa

https://orcid.org/0000-0003-2909-7080

PLANNING

TASK CHECKLIST

- ❑ Algoritmos para generar muestras simuladas
- ❑ Compara métodos de generación

OBJECTIVES

- Desarrollar, implementar y comparar algoritmos para la generación de muestras simuladas, con el fin de evaluar el comportamiento de estimadores, validar supuestos estadísticos y optimizar procesos de inferencia en contextos aplicados.

INTRODUCTION

La generación de muestras simuladas constituye una herramienta fundamental en estadística computacional, permitiendo evaluar el comportamiento de estimadores, validar supuestos teóricos y optimizar procesos de inferencia en entornos aplicados. A través de algoritmos como bootstrap, Monte Carlo y simulaciones paramétricas, es posible replicar escenarios controlados, analizar la robustez de modelos y fortalecer la toma de decisiones basada en evidencia. Esta práctica resulta especialmente valiosa en disciplinas como banca, seguros y ciencia de datos, donde la precisión y la validación técnica son esenciales.

MARCO TEÓRICO

Algunos métodos de generación:
Existen muchos métodos para la generación de números aleatorios entre 0 y 1, la importancia radica en que los números deben cumplir ciertas características.

Clases de generadores:
- Generadores congruenciales.
- Generadores de registro de desplazamiento.
- Generadores de Fibonacci retardados.
- Generadores no lineales.
- Combinación de generadores.
- Generadores paralelos de números aleatorios.
- Generadores comerciales.

MARCO TEÓRICO

La mayoría de los RNG están basados en congruenciales lineales de la forma:

$$\boldsymbol{x_n = (a_1 x_{n-1} + \cdots + a_k x_{n-k}) \bmod m}$$

donde m se llama *módulo,* a_1*,...,* a_k son enteros entre –m + 1 y m – 1 llamados multiplicadores con $\boldsymbol{(a_k \neq 0)}$ y k es el orden de la recurrencia. Se puede definir el estado de la recurrencia en el paso n como $\boldsymbol{S_n = (x_n, \ldots, x_{n+k-1})}$. La longitud del periodo máximo es $\boldsymbol{m^k - 1}$.

- **Método de congruencia lineal**

Esta técnica fue introducida por Lehmer en 1951. Una secuencia de números enteros Z_1, Z_2,...está definida por la fórmula recursiva:

$$\boldsymbol{Z_i = (aZ_{i-1} + c) \bmod m}$$

a, es el multiplicador $(0 \leq a < m)$
m, el módulo $(m > 0)$
c, el incremento $(0 \leq c < m)$
X_0, la semilla $(0 \leq X_0 < m)$

MARCO TEÓRICO

Dada la forma de la expresión de $\boldsymbol{Z_i}$ es inevitable el comportamiento como un bucle, es decir, que en el momento que se repita un $\boldsymbol{Z_i}$ todos los siguientes serán iguales a los obtenidos hasta ese momento.
La longitud de cada uno de esos ciclos se conoce como el periodo del generador y se representa por p. Como $\boldsymbol{Z_{i-1}}$ sólo depende de $\boldsymbol{Z_{i-1}}$ y se verifica que $0 \leq \boldsymbol{Z_i} \leq m-1$ entonces se cumple que $p \leq m$.
Si p = m el generador se llama de periodo total.

Case

Sea m = 16, a = 5, c = 3 y $Z_0 = 7$, la secuencia de los Z_i obtenidos será: $Z_i = (5\ Z_{i-1} + 3) \bmod 16$

i	Z_i	U_i	i	Z_i	U_i	i	Z_i	U_i	i	Z_i	U_i
0	7	-	5	10	0.625	10	9	0.563	15	4	0.250
1	6	0.375	6	5	0.313	11	0	0.000	16	7	0.438
2	1	0.063	7	12	0.750	12	3	0.188	17	6	0.375
3	8	0.500	8	15	0.938	13	2	0.125	18	1	0.063
4	11	0.688	9	14	0.875	14	13	0.813	19	8	0.500

En simulación se necesitan generadores con periodos largos, fijando un m grande sería conveniente conseguir que tuviera periodo total.

La mayoría de los programas estadísticos usan LCG con $a = 7^5 = 16807$, c = 0 y $m = 2^{31} - 1 = 2147483647$ para generar números aleatorios. Estas constantes fueron propuestas por Lewis, Godman y Miller en 1969.

MARCO TEÓRICO

Teorema:

El generador definido de la forma $\boldsymbol{Z_i = (aZ_{i-1} + c) \bmod m}$ tiene periodo total si y sólo si se cumplen las siguientes condiciones:

- c y m son primos entre sí.
- Si q es un número primo que divide a m entonces q divide a (a -1).
- Si 4 divide a m entonces 4 divide a (a-1).

Dependiendo del valor de c el comportamiento de los generadores puede ser distinto por lo que se distinguen generadores mixtos (c>0) y generadores multiplicativos (c = 0).

MARCO TEÓRICO

- **Generador de cuadrados medios**

Fue introducido por Von Neumann. El procedimiento para obtener números pseudoaleatorios con este tipo de generador es el siguiente:

- Generar una semilla X_0 de 2n cifras, se recomiendan 4 cifras.
- Elevar al cuadrado la semilla para obtener un número de hasta 4n cifras, de obtener un número con menos cifras aumentar ceros a la izquierda hasta lograr 4n cifras.
- Tomar las 2n cifras centrales del número.
- El número pseudoaleatorio se forma aumentando un punto decimal delante de las 2n cifras.
- Ahora este número obtenido se eleva al cuadrado y el proceso continua hasta formar el total de números pseudoaleatorios.

Case

Generar números pseudoaleatorios de cuatro dígitos con semilla inicial 3708:

$x0 = 3708 \Rightarrow x02 = 13|7492|64 \Rightarrow x1 = 7492 \Rightarrow u1 = 0.7492$
$x1 = 7492 \Rightarrow x12 = 56|1300|64 \Rightarrow x2 = 1300 \Rightarrow u2 = 0.1300$
$x2 = 1300 \Rightarrow x22 = 1|6900|00 \Rightarrow x3 = 6900 \Rightarrow u3 = 0.6900$
$x3 = 6900 \Rightarrow x32 = 47|6100|00 \Rightarrow x4 = 6100 \Rightarrow u4 = 0.6100$
$x4 = 6100 \Rightarrow x42 = 37|2100|00 \Rightarrow x5 = 2100 \Rightarrow u5 = 0.2100$
$x5 = 2100 \Rightarrow x52 = 4|4100|00 \Rightarrow x6 = 4100 \Rightarrow u6 = 0.4100$
$x6 = 4100 \Rightarrow x62 = 16|8100|00 \Rightarrow x7 = 8100 \Rightarrow u7 = 0.8100$
$x7 = 8100 \Rightarrow x72 = 65|6100|00 \Rightarrow x8 = 6100 \Rightarrow u8 = 0.6100$

MARCO TEÓRICO

- **Pruebas de hipótesis para los números pseudoaleatorios**

Una vez generados los números pseudoaleatorios, se debe verificar si cumplen con las siguientes características:
- Uniformemente distribuidos en (0, 1).
- Independientemente distribuidos.
- Con media igual a 0.5.
- Con varianza igual a 1/12.
- Su periodo o ciclo debe ser largo.

- Prueba de forma

Se utiliza la prueba de bondad de ajuste Chi-cuadrado o Kolmogorov - Smirnov.

Hipótesis:

H_0: $\{ u_i \} \sim U(0, 1)$

H_1: $\{ u_i \}$ no tienen distribución U(0, 1).

Estadístico de prueba: Chi-cuadrado

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

k = número de categorías de la variable. Criterio de decisión: Para un nivel de significancia α, rechazamos la hipótesis nula si $\chi^2 > \chi^0_{(k-r-1,1-\alpha)}$ donde r es el número de parámetros estimados.

MARCO TEÓRICO

- Prueba de independencia

H_0: { u_i } son independientes (aleatorios)

H_1: { u_i } no son independientes.

Clasificar cada número como:

Si $u_i \leq u_{i-1}$ entonces asignar un signo − .

Si $u_i > u_{i-1}$ entonces asignar un signo +.

Calcular el número de rachas observadas (r).

$$E(r) = \frac{2n-1}{3}, V(r) = \frac{16n-29}{90}$$

Estadística de prueba:

$$Z = \frac{r - E(r)}{\sqrt{V(r)}} \rightarrow N(0,1)$$

Regla de decisión: Para un nivel de significancia α, se rechaza la hipótesis nula si $|Z| > Z_{1-\alpha/2}$.

MARCO TEÓRICO

- Prueba para la media

H_0: $\mu = 0.5$

H_1: $\mu \neq 0.5$

Estadístico de prueba:

$$Z = \frac{\bar{X} - 0.5}{\sqrt{1/12n}}$$

donde $\bar{x}$ es la media aritmética de los números aleatorios generados.

Regla de decisión:

Para un nivel de significancia α, se rechaza la hipótesis nula si $|Z| > Z_{1-\alpha/2}$.

MARCO TEÓRICO

- Prueba para la varianza

H_0: $\sigma^2 = 1/12$

H_1: $\sigma^2 \neq 1/12$

Estadístico de prueba:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

donde S^2 es la varianza de los números aleatorios generados.

Regla de decisión:

Para un nivel de significancia α, se rechaza la hipótesis nula si

$\chi^2 > \chi^2_{(n-1,1-\alpha/2)}$ o $\chi^2 < \chi^2_{(n,-1\,\alpha/2)}$

- **Prueba de la distancia**

Este contraste analiza la tendencia de los datos a concentrarse dentro de cierto subintervalo del [0,1).

- Si los números pseudoaleatorios se consideran como números reales:

Entonces, para realizar esta prueba es necesario seleccionar un intervalo (α, β), el cual debe estar contenido en el intervalo (0, 1), es decir $0 \leq \alpha \leq \beta \leq 1$. Luego para cada número generado se pregunta si es o no elemento del intervalo (α, β). Si U_j (número uniforme generado) es elemento de (α, β), U_{j+1} hasta U_{j+i} no son elementos de dicho intervalo y U_{j+i+1} vuelve a ser elemento del intervalo (α, β), entonces tiene un hueco de tamaño i.
O sea llamaremos hueco a cualquier serie de números de la secuencia de partida, que no están dentro del intervalo (α, β) y que están comprendidos entre dos números que sí lo están. A la cantidad de números en ese hueco que no pertenecen a (α, β) la llamaremos longitud del hueco. Por ejemplo, si α = 0.6 y β = 0.9 y los números pseudoaleatorios generados son como sigue: 0.7, 0.3, 0.1, 0.2, 0.8, 0.75, entonces los números : 0.7, 0.3, 0.1, 0.2, 0.8, forman un hueco de tamaño 3, y los números 0.8 y 0.75 forman un hueco de longitud 0.

MARCO TEÓRICO

La distribución de probabilidad del tamaño del hueco es como sigue:

$$P_i = \theta\,(1-\theta)^i, \text{ para } i = 0, 1, 2, \ldots$$

Donde $\theta = \beta - \alpha$ representa la probabilidad de caer en el intervalo (α, β). Los valores de α y β no influyen en la bondad de la prueba.

Es conveniente agrupar las probabilidades para valores de $i \geq n$. El valor de n debe ser tal que todas las frecuencias esperadas deben ser mayor o igual a 5.

Se calculan las frecuencias esperadas multiplicando las probabilidades por el total de las frecuencias observadas.

Estadístico de prueba:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

k = número de categorías de la variable número de huecos

Regla de decisión:

Para un nivel de significancia α, si

$\chi^2 < \chi^2_{(k-1,1-\alpha)}$ concluimos que los números generados pasan la prueba de la distancia.

MARCO TEÓRICO

- Prueba de series

Esta prueba se utiliza para comprobar el grado de aleatoriedad entre números sucesivos. Usualmente esta prueba consiste en formar parejas de números, las cuales son consideradas como coordenadas en un cuadrado unitario dividido en n2 celdas. Esta idea se puede extender a las ternas de números pseudoaleatorios que representan puntos en un cubo unitario. Sin embargo, en este caso sólo veremos para dos dimensiones.

La prueba de series consiste en generar N números pseudoaleatorios de los cuales se forman parejas aleatorias entre Ui y Ui+1, es decir, si se generan 10 números, entonces las parejas aleatorias que se podrían formar serían:

$$(U_1, U_2), (U_2, U_3), (U_3, U_4), (U_4, U_5), (U_5, U_6), (U_6, U_7),$$
$$(U_l, U_8), (U_8, U_9), (U_9, U_{10})$$

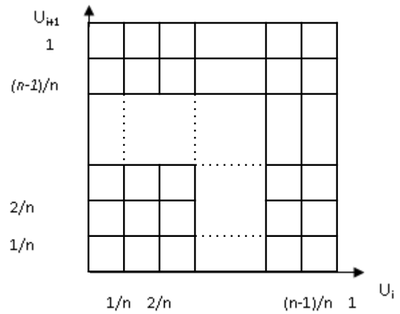
En seguida, se determina la celda a la que pertenece cada pareja ordenada, con lo cual se obtiene la frecuencia observada de cada celda. La frecuencia esperada de cada celda se obtiene al dividir el total de parejas coordenadas (N-1) por el total de celdas (n2). Finalmente, se calcula el estadístico de prueba:

$$\chi_0^2 = \frac{n^2}{N-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(O_{ij} - (N-1)/n^2\right)^2$$

MARCO TEÓRICO

- Prueba de series

Donde Oij es la frecuencia observada en la celda ij. Si $\chi 2 < \chi 2(g.l.;1-\alpha)$, (g.l. = n2-1) entonces no se puede rechazar la hipótesis de que los números provienen de una distribución uniforme.



- Prueba del Póker

Esta prueba examina en forma individual los dígitos del número pseudoaleatorio generado. La forma como esta prueba se realiza es tomando 5 dígitos a la vez y clasificándolos como: par, dos pares, tercia, póker, quintilla, full, y todos diferentes. Lo anterior significa que los números pseudoaleatorios generados son de 5 dígitos cada uno, o bien, en el caso de que el número tenga más de 5 dígitos, solamente se consideran los 5 primeros. Las probabilidades de cada una de las manos de póker se muestran en seguida:

Mano de póker	Probabilidad
Todos diferentes	$\frac{10 \times 9 \times 8 \times 7 \times 6}{10^5} = 0.3024$
Un par	$\binom{5}{2}\frac{10 \times 1 \times 9 \times 8 \times 7}{10^5} = 0.5040$
Dos pares	$\binom{5}{2}\binom{3}{2}\frac{1}{2}\frac{10x1x9x1x8}{10^5} = 0.1080$
Tercia	$\binom{5}{3}\frac{10 \times 1 \times 1 \times 9 \times 8}{10^5} = 0.0720$
Full	$\binom{5}{3}\binom{2}{2}\frac{10x1x1x9x1}{10^5} = 0.0090$
Póker	$\binom{5}{4}\frac{10x1x1x1x9}{10^5} = 0.0045$
Quintilla	$\binom{5}{5}\frac{10x1x1x1x1}{10^5} = 0.0001$

MARCO TEÓRICO

- Prueba del Póker

Con las probabilidades anteriores y con el número de números pseudoaleatorios generados, se puede calcular la frecuencia esperada de cada posible resultado.

Estadístico de prueba:

$$\chi_0^2 = \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i}$$

Regla de decisión:

Si $\chi_0^2 < \chi^2_{(6,1-\alpha)}$ entonces no se puede rechazar la hipótesis nula de que los números pseudoaleatorios provienen de una distribución uniforme.

Case

Para los siguientes números pseudoaleatorios, aplicar todas las pruebas explicadas. Los números se leen de izquierda a derecha y de arriba hacia abajo.

0.03991	0.10461	0.93716	0.16894	0.98953
0.38555	0.95554	0.32886	0.59780	0.09958
0.17546	0.73704	0.92052	0.46215	0.15917
0.32643	0.52861	0.95819	0.06831	0.19640
0.69572	0.68777	0.39510	0.35905	0.85244
0.24122	0.66591	0.27699	0.06494	0.03152
0.61196	0.30231	0.92962	0.61773	0.22109
0.30532	0.21704	0.10274	0.12202	0.94205
0.03788	0.97599	0.75867	0.20717	0.82037
0.48228	0.63379	0.85783	0.47619	0.87481
0.88618	0.19161	0.41290	0.63312	0.71857
0.71299	0.23853	0.05870	0.01119	0.92784
0.27954	0.58909	0.82444	0.99005	0.04921
0.80863	0.00514	0.20247	0.81759	0.45197
0.33564	0.60780	0.48460	0.85558	0.15191
0.90899	0.75754	0.60833	0.25983	0.01291
0.78038	0.70267	0.43529	0.06318	0.38384
0.55986	0.86485	0.88722	0.56736	0.66164
0.87539	0.08823	0.94813	0.31900	0.54155
0.16818	0.60311	0.74457	0.90561	0.72848

REFERENCIAS BIBLIOGRÁFICAS

- Méndez Arias, L. (2022). Métodos de generación de variables aleatorias [Trabajo de fin de grado, Universidad de Zaragoza]. Repositorio Zaguán.

OTROS RECURSOS DE INTERÉS

https://www.kdnuggets.com/

http://archive.ics.uci.edu/ml/datasets.php

https://www.cs.ubc.ca/labs/beta/Projects/autoweka/datasets/

https://explodat.cl/Analytics/business-intelligence/la-metodologia-kimball-para-data-warehouses-y-bi-exitosos/

Dive straight in!