Tarea 5
Simulación estocástica

Joel Chacón Castillo
Cómputo Científico

9 de octubre de 2019

1. CDF Inversa generalizada

Definir la cdf inversa generalizada F_x^- y demostrar que en el caso de variables aleatorias continuas esta coincide con la inversa usual. Demostrar además que en general para simular X podemos simular $u \sim U(0,1)$ y $F_X^-(u)$ se distribuye como X.

Comentarios

El método de la transformada inversa es un método básico para el muestreo de números pseudo-aleatorios, es decir para generar muestras de números aleatorios de cualquier distribución (que cumpla las propiedades que se describen a continuación) de probabilidad dada su función de distribución acumulativa. Éste método toma muestras uniformes de un número u que se encuentran en $[0,1]$, y regresa el número más largo x del dominio de la distribución $P(x)$ tal que $P(-\infty < X < x) \leq u$. Sea $F(x)$, $x \in \Re$ una función de distribución acumulativa (cdf) (contínua o no). Recordar que $F : \Re \to [0,1]$ es entonces una función no negativa y no decresiente (monótona) que es continua de la derecha y tiene límites por la izquierda, con valores en $[0,1]$; mas aún $F(\infty) = 1$ y $F(-\infty) = 0$. El propósito de éste método es generar o simular variables aleatorias X distribuidas como F; esto es que se quiere simular una variable aleatoris X tal que $P(X \leq x) = F(x), x \in \Re$. Así la inversa generalizada de F, $F^- : [0,1] \to \Re$, $x \in \Re$ se define de la sigueinte forma:

$$F^-(y) = min\{x : F(x) \geq y\}, \quad y \in [0,1] \tag{1}$$

Así si D es continua, entonces F es invertible (dado que es continua y estrictamente creciente), en tal caso la función ordinaria inversa es $F^-(y) = min\{x : F(x) = y\}$, y por lo tanto $F(F^-(y)) = y$ y $F^-(F(x)) = x$. En caso general (funciones discretas y continuas) se mantiene que $F(F^-(y)) \geq y$ y $F^-(F(x)) \leq x$. Además $F^-(y)$ es una función no decreciente (monótona) en y.

El método de la tranformada inversa (Definición)

Dado $F(x)$, $X \in \Re$, que denota cualquier función de distribución acumulativa (cdf) (continua o no). Sea $F^{-}(y), y \in [0,1]$ denote la función inversa definida en la ecuación (1). Definiendo $X = F^{-}(U)$, donde U tiene una distribución uniforme continua en el intervalo $[0,1]$. Entonces X está distribuída como F, esto es, $P(X \leq x) = F(x), x \in \Re$.

Demostración

Lo que se debe demostrar es que $P(X \leq x) = P(F^{-}(U) \leq x) = F(x), x \in \Re$. Primero se supone que F es continua (como se muestra en la figura 1). Entonce se muestra bajo igualdad de eventos que $\{F^{-}(U) \leq x\} = \{U \leq F(x)\}$, por lo tanto tomando probabilidades (y dejando que $a = F(x)$) en $P(U \leq a)$ genera el siguiente resultado: $P(F^{-}(U) \leq x) = P(U \leq F(x)) = F(x)$. Hasta este punto sólo se han considerado funciones continuas $F(F^{-}(y) = y)$ y por lo tanto (por monotonicidad de F) si $F^{-}(U) \leq x$, entonces $U = F(F^{-}(U)) \leq F(x)$, o $U \leq F(x)$. Similarmente $F^{-}(F(x)) = x$ y por lo tanto $U \leq F(x)$, entonces $F^{-}(U) \leq x$. Finalmente se concluye por la igualdad de los dos eventos (mencionados previamente). En general sea el caso continuo o no, se demuestra que

$$\{U < F(x)\} \subseteq \{F^{-}(U) \leq x\} \subseteq \{U \leq F(x)\}, \tag{2}$$

alcanza el mismo resultado después de tomar probabilidades dado que $P(U = F(x))$ desde que U es una variable aleatoria continua.

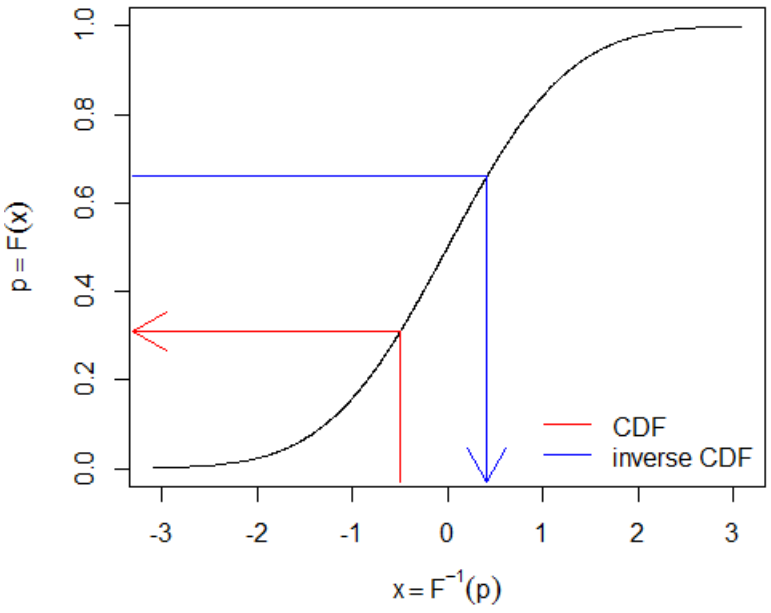

Figura 1: TDiagrama que explica la equivalencia que existe entre la función $F(x)$ y $F^{-}(x)$ en el caso continuo..

2. Simulación de variables aleatoria uniformes

Implementar el siguiente algortimo para simular variables aleatorias uniformes

$$x_i = (107374182x_{i-1} + 104420x_{i-5})mod(2^{32} - 1) \tag{3}$$

regresa a x_i, y recorrer el estado, esto $x_{j-1} - x_j$; $j = 1, 2, 3, 4, 5$. ¿Tienen distribución similar a $U(0, 1)$?

Comentarios

Una forma de caracterizar el rendimiento de los generadores de enteros pseudoaleatorios es a través de la noción de periodo.

Definición *El petiodo, T_0, de un generador es el entero más pequeño T tal que $u_{i+T} = u_i$ para cada i, esto es, tal que D^T es igual a la funcióon identidad.*

El periodo es un parámetro muy importante, teniendo un impacto directo en la eficacia de un generador aleatorio. Si el número de generaciones requeridas excede el periodo de un generador, entonces podrían existir variaciones no controladas en la secuencia (fenómenos cíclicos, ordenaciones fales, entre otros). Desafortunadamente, un generador de la form $X_{n+1} = f(X_n)$ tiene un periodo menor que $M + 1$.

Definición *Un generador congruencial en $\{0, 1, ..., M\}$ es definido por la función*

$$D(x) = (ax + b)mod(M + 1) \tag{4}$$

.

El periodo, y el rendimiento de los generadores congruenciales dependen mucho de la elección de (a, b). Con la elección de a racional, un generador congruencial producirá pares $(X_n, D(x_n))$ que caen en línea paralelas. Por lo tanto es importante seleccionar a tal que maximice el número de segmentos paralelos in $[0, 1]^2$.

Una forma para probar un generador de número aleatorios es por medio de la prueba de hipótesis de Kolmogorov-Smirnov en comparación a una muestra que proviene de una distribución uniforme. Es decir si se tiene una muestra que se desea probar y si se tiene la muestra de una distribución deseada, es posible aplicar la prueba de Kolmogorov-Smirnov, ya que es una prueba no paramétrica. Se generaron distintas muestras con distintos tamaños, a las cuales se les aplicaron pruebas de Kolmogorov. La prueba *k-test* regresa un estadístico y un p-valor correspondiente de la estadística D. La estadística D es el máxima distancia absoluta (sup) entre los CDFs de dos muestras. Entre más cercano sea el estadístico a cero, es mas probable que dos muestras provengan de la misma distribución. El p-valor tiene el mismo significado que otros p-valores. Se rechaza la hipótesis nula de que dos muestras son generadas de la misma distribución si el p-valor es menor que el nivel de significancia. En la tabla 1 se presentan los resultados de aplicar el estadístico, se observa que conforme se aumenta el tamaño de la muestra el valor del estadístico D tiende a ser menor, es decir que el generador propuesto sí tiene distribución similar a la uniforma. Otra forma de realizar esto es tomando la entropía de una muestra, sabiendo que una distribución uniforme tiene entropía máxima. Otra alternativa es analizar visualmente el CDF de la muestra generada.

Tamaño muestra	Statistic -D	P-value
1	1.000	0.000
10	0.373	0.029
100	0.096	0.280
1000	0.017	0.945
10000	0.005	0.965
100000	0.003	0.194

Cuadro 1: Resultados de realizar pruebas de Kolmogorov-Smirnov (no paramétricas) partiendo del generador de números aleatorios

.

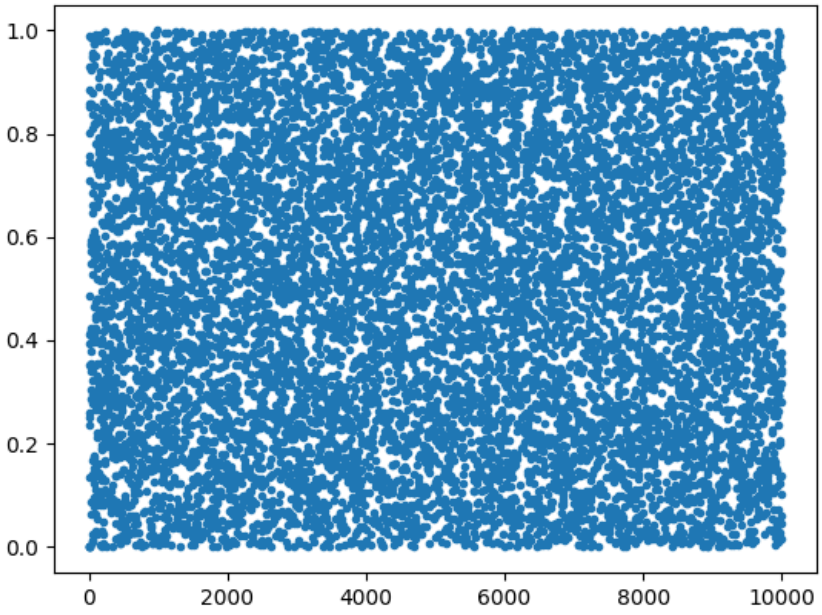

Figura 2: Muestra de tamaño 10000 del generador propesto.

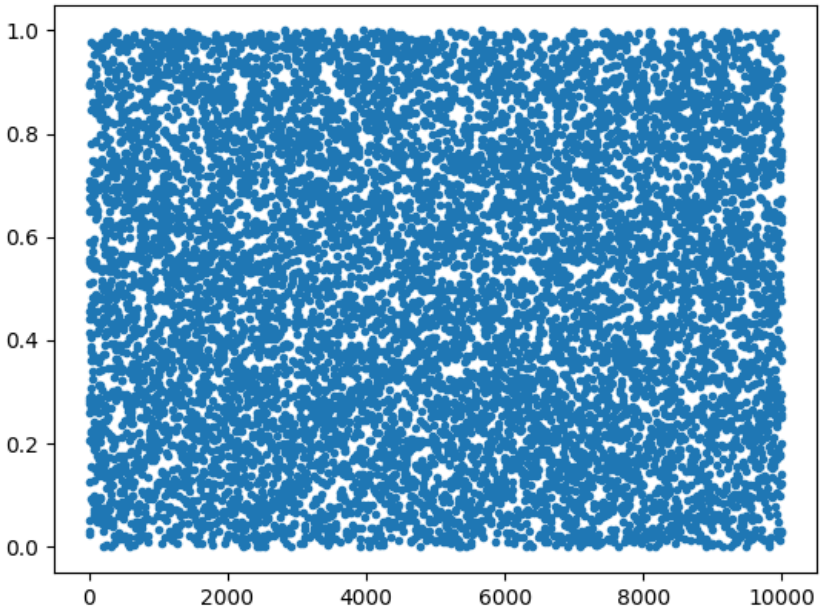

Figura 3: Muestra de tamaño 10000 de de la librería *numpy.random.uniform*.

3. Generador de números aleatorios

¿Cuál es el algoritmo que usa *scipy.stats.uniform* para generar números aleatorios?. ¿Cómo se pone la semilla? ¿Y en R?

Comentarios

La librería de *scipy.stats.uniform* utiliza *numpy.random* para generar sus números aleatorios, por lo tanto para asignar una semilla basta con poner al inicio o antes de utilizar el generador pseudo-aleatorio *np.random.seed(seed=1)* donde se da la semilla como 1.

Por otra parte en R la semilla se pone con *set.seed(42)* y el generador aleatorio por defecto es *Mersenne-Twister* como se muestra en este enlace [1].

4. Simulación de variables aleatorias discretas en *scipy*

Considerando la librería de *scipy*, ¿Qué funciones hay para simular una variable aleatoria genérica discreta? ¿Tienen pre-proceso?

[1] https://stat.ethz.ch/R-manual/R-devel/library/base/html/Random.html

Comentarios

Para simular una variable aleatoria está *stats.rv_discrete* que es una clase base para construir clases de distribución específicas e instancia de variables aleatorias discretas. Adeás puede ser utilizado para construir distribuciones arbitrarias definidas por una lista de puntos de soporte y probabilidades correspondientes. Esta clase base principalmente maneja varios métodos útiles como son:

- rvs: Random variables
- pfm: Probability mass function.
- cdf: Cumulative distribution function of the given random variable.

La librería *stats.rv_discrete* requiere como argumento una tabla de probabilidades, se podría ver esta tabla como un pre-procesamiento, es decir, de forma eficiente se puede mantener una tabla de probabilidades la cual puede ser actualizada eficientemente (esta parte depende del usuario).

5. Adaptive Rejection Sampling (ARS)

Implemementar el algoritmo Adaptive Rejection Sampling y simular de una *Gamma(2,1)* 10,000 muestras.

Comentarios

Se implementó el método explicado en el libro *Monte Carlo Statistical Methods* de Christian P. Robert y Gerorge Casella Pag. 57, además el modelo que se considera en cada intervalo se tomó el considerado en la pag. 71. Es importante mencionar que la implementación se puede hacer de forma más eficiente, en lugar de actualizar todos los intervalos (como se hace en el código), sólo es necesario actualizar el intervalo en el que se van agregando los nuevos puntos, aún más para ubicar el intervalo al que pertenece el nuevo punto se podría hacer un búsqueda binaria. En la figura 4 se puede observar que la diferencia entre la implementación y la librería no son muy distintos.

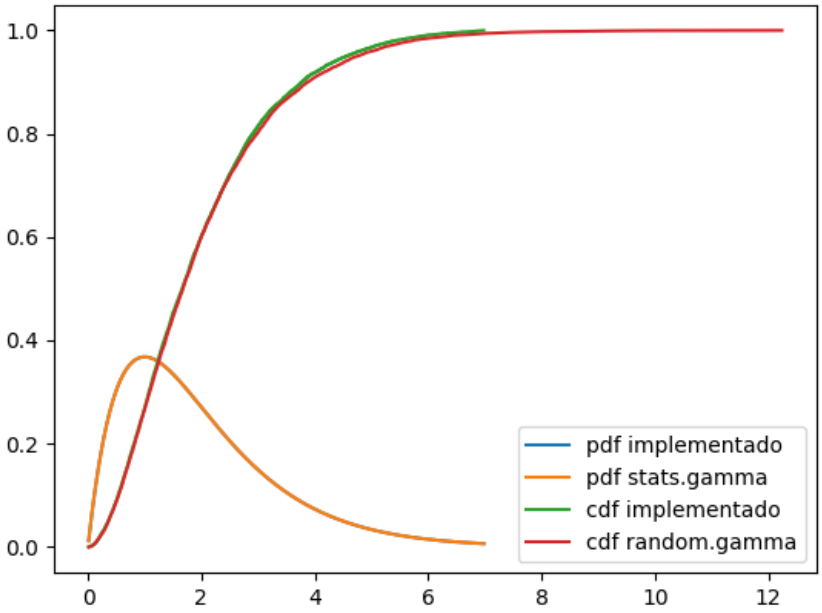

Figura 4: PDF y CDF de una función $Gamma(2{,}0, 1{,}0)$ considerando 10,000 puntos.

.

¿Cuándo es conveniente dejar de adaptar la envolvente?

En distintas configuraciones, la distribución asociada con la densidad f es difícil de simular dado que la complejidad de la misma función f, la cual podría requerir un tiempo significativo en cada evaluación. Se sugiere dejar de adaptar la envolvente en las siguientes situaciones:

- Cuando la razón entre el envolvente y la función a simular se aproxima a la unidad. Es decir, si se tiene una función de densidad a simular $f(x)$ y una función envolvente $g(x)$ se tiene que $f(x) \leq Mg(x)$, entonces se conviene dejar de simular cuando $f(x)/(Mg(x)) \approx 1$.

- La probabilidad de rechazar una solución es menor de forma significativa, esto se puede hacer empíricamente.

- Se podría aplicar la divergencia de Kullback, cuando sea suficientemente cercano a cero.

- La diferencia entre el CDF de la distribución objetivo $f(x)$ y el CDF de la *mezcla* $g(x)$ sea aproximadamente cero, esto es que el CDF de $Mg(x) \approx 1$.

Además no sería necesario aplicar este método cuando se desea simular una función que no es computacionalmente costoso de evaluar.