Curso:
Métodos de Monte Carlo
Unidad 4, Sesión 11: Sorteo de variables aleatorias con distribución arbitraria

Departamento de Investigación Operativa
Instituto de Computación, Facultad de Ingeniería
Universidad de la República, Montevideo, Uruguay

dictado semestre 1 - 2024

Generación de variables aleatorias con distribución dada arbitraria

La generación de muestras de variables aleatorias constituye un elemento central paralos métodos de Monte Carlo.

Como ya hemos visto, si deseamos estimar la integral $\zeta = \int_{\mathcal{R}} \phi(\mathbf{x}) d\mathbf{x}$ para una región $\mathcal{R} \subseteq \mathcal{J}^m$ en el hipercubo $\mathcal{J}^m = [0,1]m$, entonces alcanza con generar puntos con distribución uniforme en el hipercubo, lo que puede hacerse fácilmente sorteando cada dimensión de manera independiente empleando los generadores de números aleatorios que hemos visto en las sesiones anteriores.

En muchos casos, y tal como se mencionó en la sesión 6, es más conveniente plantearse calcular la integral $\zeta = \int_{R^m} \kappa(\mathbf{z}) dF(\mathbf{z})$, donde $F()$ es una función de distribución de un vector aleatorio en R^m, y $\kappa()$ una función arbitraria en ese dominio.

Cuando podemos expresar la medida de interés bajo esta forma, es posible utilizar el método de Monte Carlo realizando un muestreo de valores $\mathbf{Z}^{(i)}$

de distribución $F()$, y calculando el estimador siguiente:

$$\ddot{\zeta}_n = n^{-1} \sum_{i=1}^{n} \kappa(\mathbf{Z}^{(i)})$$

que es un estimador insesgado de ζ.

Aunque generar los valores $\mathbf{Z}^{(i)}$ es más costoso que generar valores $\mathbf{X}^{(i)}$ con distribución uniforme, la evaluación de $\kappa()$ puede significar beneficios computacionales frente a la evaluación de $\phi()$.

Además, el utilizar la fórmula más genérica $\zeta = \int_{R^m} \kappa(\mathbf{z}) dF(\mathbf{z})$ crea oportunidades adicionales para mejorar la performance de un método de Monte Carlo. Sea F^* una segunda función de distribución, tal que $dF(\mathbf{z})/dF^*(\mathbf{z})$ exista para todo $\mathbf{z} \in R^m$. Definamos $\kappa^*(\mathbf{z}) = \kappa(\mathbf{z}) dF(\mathbf{z})/dF^*(\mathbf{z})$, $\mathbf{z} \in R^m$. Entonces es posible muestrear

valores $\mathbf{Y}^{(i)}$ de distribución $F^*()$, y calcular el estimador siguiente:

$$n^{-1} \sum_{i=1}^{n} \kappa^*(\mathbf{Y}^{(i)})$$

que es también un estimador insesgado de ζ. Como veremos en el próximo capítulo del curso (en la sesión dedicada a métodos para aumentar la eficiencia computacional de Monte Carlo, conocidos habitualmente como métodos de reducción de la varianza), una elección sensata de F^* puede llevar a una menor varianza del estimador y a un mejor desempeño del método.

Es por lo tanto importante ser capaz de obtener muestras de una distribución arbitraria, tanto continua como discreta.

Este es un tema clásico en el área, que ha sido objeto de cientos de artículos científicos discutiendo tanto métodos generales como métodos aplicables a distribuciones específicas. Por lo tanto, no pretendemos aquí

cubrir en forma exhaustiva ni siquiera un panorama del estado del arte al respecto.

Nos limitaremos a discutir algunas consideraciones generales sobre la importancia de ver si la función de distribución F puede o no expresarse mediante producto de funciones de distribución unidimensionales independientes, y luego presentar algunos métodos generales para distribuciones unidimensionales (que no son siempre los más eficientes para una distribución específica, pero sí los de mayor utilidad dado lo amplio de su espectro de aplicación).

Dependencia versus independencia

La función de distribución F es una función definida en R^m, por lo tanto en varias variables (es la distribución de un vector aleatorio), cumpliéndose (por definición de función de distribución) que

$$F(\mathbf{z}) = \text{Prob}\left(Z_1 \leq z_1, \ldots, Z_m \leq z_m\right).$$

Cuando las distintas componentes son independientes, tenemos que

$$F(\mathbf{z}) = \text{Prob}\left(Z_1 \leq z_1, \ldots, Z_m \leq z_m\right) = \prod_{j=1}^{m} \text{Prob}\left(Z_j \leq z_j\right) = \prod_{j=1}^{m} F_j(z_j),$$

donde los $F_j()$ son funciones de distribución en R, y decimos que F tiene una representación multiplicativa separable.

En este caso, alcanza con sortear de manera independiente cada uno de los componentes Z_j del vector aleatorio $\mathbf{Z}$ de acuerdo a las distribuciones F_j,

aplicándose directamente los métodos para distribuciones unidimensionales que veremos luego.

El panorama es distinto cuando las componentes no son independientes, pues entonces no es posible proceder de manera tan sencilla. En este segundo caso, no tendremos una representación separable, sino que deberemos recurrir a la función de densidad de probabilidad (o en el caso discreto, desarrollable de manera análoga, a las probabilidades de cada valor posible). Sea $f(\mathbf{z})$ la función de densidad de probabilidad correspondiente a F (supuesta continua). Podemos representar a f con la siguiente forma producto (no separable):

$$f(\mathbf{z}) = f_1(z_1) \prod_{j=2}^{m} f_i(z_i|z_1, \ldots, z_{i-1}),$$

donde los f_i son funciones de densidad de probabilidad condicionales.

Si conocemos una representación de esta forma, podemos entonces generar un vector $(\mathbf{Z})$ a través del algoritmo siguiente:

Generar Z_1 de función de densidad de distribución $f_1(z)$.
Para $j = 2$ hasta m
Generar Z_j de función de densidad de distribución $f_i(z|Z_1, \ldots, Z_{j-1})$.
Devolver $Z = (Z_1, \ldots, Z_m)$.

Este método es de alcance general. Es preciso sin embargo hacer algunas observaciones.

En primer término, dado que la formulación previa depende del orden en el cual se tomen las distintas componentes de Z, para una misma distribución F existen $m!$ formas distintas de descomponerla, que si bien de un punto de vista conceptual son equivalentes, pueden tener propiedades muy distintas cuando se trata de implementarlas, o aún de obtener su formulación analítica.

En segundo término, que si bien puede parecer entonces que alcanza con conseguir una cualquiera de estas formulaciones, puede no ser sencillo en la práctica encontrar ni siquiera una, ya que en ciertos casos no poseemos

toda la información o las herramientas necesarias para hacer esta derivación y obtener formulas aplicables. En ciertos casos, es posible conocer los cocientes $f(\mathbf{z})/f(\mathbf{y})$ o las probabilidades condicionales $f_i(z_i|z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_m)$; existe un conjunto de métodos, conocidos bajo los nombres de *muestreo de Metropolis* y *muestreo de Gibbs* (dentro de la familia de métodos de Cadenas de Markov en Monte Carlo) que permiten generar (con una cierta aproximación) las distribuciones correspondientes. Este tema no será tratado en el curso, pero lo mencionamos por completitud. Por más información, ver por ejemplo las fuentes siguientes:

- un breve artículo en la Wikipedia `http://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo`(último acceso: 2023-03-02)

- Reporte técnico "Probabilistic Inference Using Markov Chain Monte Carlo Methods", Radford M. Neal, Technical Report CRG-TR-93-1, Department of Computer Science University of Toronto, 25 September

1993, disponible en https://omega0.xyz/omega8008/neal.pdf (último acceso: 2023-03-02).)

- Artículo "The Markov Chain Monte Carlo revolution", Persi Diaconis, disponible en http://statweb.stanford.edu/~cgates/PERSI/papers/MCMCRev.pdf (último acceso: 2023-03-02).

- Libro "Handbook of Markov Chain Monte Carlo", Charles J. Geyer, disponible en http://www.mcmchandbook.net/HandbookChapter1.pdf (último acceso: 2023-03-02).

- Tutorial: "Markov chain Monte Carlo". Iain Murray, Machine Learning Summer School 2009, disponible en http://homepages.inf.ed.ac.uk/imurray2/teaching/09mlss/slides.pdf (último acceso: 2023-03-02).

- Demostración en línea "Markov Chain Monte Carlo Simulation Using the Metropolis Algorithm", http://demonstrations.wolfram.com/MarkovChainMonteCarloSimulationUsingTheMetropolisAlgorithm/(últim
acceso: 2023-03-02).

Métodos generales para generar una muestra de distribución arbitraria (en una variable)

Los siguientes cuatro métodos se encuentran entre los más empleados y robustos para transformar números aleatorios (es decir, muestras de variables aleatorias uniformes) en muestras de variables de una distribución específica:

- Método de la transformada inversa.
- Método de composición.
- Método de aceptación-rechazo.
- Método de cociente de uniformes.

Veremos en las siguientes transparencias un breve esquema del primero de ellos, que es el más empleado.

Como ya hemos dicho, existen también métodos específicos para distribuciones particulares (a veces de gran interés, como por ejemplo la distribución uniforme), pero no entraremos en ellos en este curso

Método de la transformada inversa.

Este es el método más directo (y muchas veces el más eficiente) para generar una muestra de una variable aleatoria arbitraria.

Se basa en el siguiente teorema:

Teorema 1. *Sea $F(z)$, $a \leq z \leq b$ una función de distribución, y su inversa F^{-1} definida por*

$$F^{-1}(u) = \inf\{z \in [a,b] : F(z) \geq u, 0 \leq u \leq 1\}.$$

Sea U una variable aleatoria de distribución uniforme $U(0,1)$. Entonces $Z = F^{-1}(U)$ es una variable aleatoria de distribución F.

Prueba. La prueba es directa, si se observa que (por la monotonía de toda función de distribución F y la uniformidad de U),
$\text{Prob}\,(Z \leq z) = \text{Prob}\left(F^{-1}(U) \leq z\right) = \text{Prob}\,(U \leq F(z)) = F(z)$. .

A partir de este resultado, si F es una distribución continua y conocemos una expresión analítica (o una numérica aproximada) para su inversa F^{-1}, resulta inmediato generar una v.a. Z de distribución F, simplemente generando una v.a. uniforme $(0, 1)$ y calculando $Z = F^{-1}(U)$.

Para dar un ejemplo sencillo, supongamos que queremos generar una v.a. X de distribución exponencial y parámetro λ, cuya distribución de probabilidad es $F_X(x) = 1 - e^{-\lambda x}$. Calculando llegamos a que la inversa es $F^{-1}(u) = -1/\lambda \ln(1 - u)$, y por lo tanto podemos generar una uniforme U y una muestra de $X = -1/\lambda \ln(1 - U)$. Observando que si U es uniforme $1 - U$ también lo será, podemos simplificar la expresión a $X = -1/\lambda \ln(U)$.

Este método tiene varias ventajas, una de ellas es que necesita un único número aleatorio uniforme para generar una muestra de distribución arbitraria; otra que por la monotonía de F, si tenemos dos valores uniformes U_1 y U_2 correlacionados, se mantendrá el mismo signo de correlación en las muestras correspondientes, esto es importante para ciertos esquemas de reducción de la varianza. También es fácilmente

aplicable para generar muestras condicionales del tipo $F(z|a' \leq Z \leq b')$.

Las dificultades de aplicación surgen cuando no es posible calcular de manera analítica F^{-1}. Es posible utilizar algun esquema numérico, con el consiguiente error de aproximación, y con el costo computacional asociado (es por ejemplo el caso de la distribución normal, para la que existen otras maneras más eficientes de generación).

Si tenemos una distribución discreta, el método de transformada inversa también es aplicable, y admite la siguiente formulación sencilla.

Sea una v.a. Z tal que Prob $(Z = a_k) = p_k$, $1 \leq k \leq L$, k entero, con $\sum_{k=1}^{L} p_k = 1$ y con $a_k < a_{k+1}$ para todo k. Entonces $F(z) = \text{Prob}\,(Z \leq z) = \sum_{k/a_k \leq z} p_k$.

Si notamos $q_k = \sum_{l \leq k} p_l$, entonces para sortear una v.a. con distribución Z, sorteamos una v.a. U uniforme $(0, 1)$, y buscamos el valor k tal que $q_{k-1} < U \leq q_k$ y asignamos a Z el valor a_k correspondiente. La v.a. Z tendrá entonces la distribución deseada.

Con este método, la probabilidad de obtener el valor a_k es igual a $q_k - q_{k-1} = p_k$, tal como deseábamos.

Material adicional

Damos a continuación referencias a bibliotecas para distribuciones de probabilidad y para generación de variables aleatorias. Todas las bibliotecas que poseen implementaciones robustas para calcular los inversos de funciones de distribución pueden ser empleados con el método de transformada inversa para generar muestras de variables con esas distribuciones.

- CDFLIB: Cumulative Density Functions.
  - Versión FORTRAN: `http://people.sc.fsu.edu/~jburkardt/f_src/cdflib/cdflib.html` (último acceso:2023-03-02)
  - Versión C++: `http://people.sc.fsu.edu/~jburkardt/cpp_src/cdflib/cdflib.html`(último acceso:2019-04-03)
  - Versión C: `http://people.sc.fsu.edu/~jburkardt/c_src/cdflib/cdflib.html`(último acceso:2023-03-02)

- Java Libraries for MC simulation (de Pierre L'Ecuyer): http://www.iro.umontreal.ca/~simardr/ssj/doc/html/umontreal/iro/lecuyer/randvar/package-summary.html (último acceso:2023-03-02)

- Statistical functions in Python: https://docs.scipy.org/doc/scipy/reference/stats.html (último acceso:2023-03-02)

Preguntas para auto-estudio

- ¿Por qué es importante contar con métodos para generar muestras de una distribución arbitraria?
- ¿Qué métodos generales conoce para obtener muestras de una variable aleatoria que sigue una distribución específica?
- ¿Cómo funciona el método de transformada inversa?

Entrega 6

Ejercicio 11.1: (individual)

Para generar un punto aleatorio (X_1, X_2) en un círculo de centro $(0, 0)$ y radio 1, es posible hacerlo de la forma siguiente (derivación disponible en las páginas 234 y 235 del libro de referencia del curso, "Monte Carlo: concepts, algorithms and applications", Fishman 1996):

- se genera un valor aleatorio r, de distribución $F_r(x) = x^2$ para $0 \leq x \leq 1$, y 0 para cualquier otro x;
- se generan dos v.a. independientes Z_1 y Z_2 de distribución normal $(0, 1)$;
- se calcula $X_1 = rZ_1/\sqrt{(Z_1^2 + Z_2^2)}$ y $X_2 = rZ_2/\sqrt{(Z_1^2 + Z_2^2)}$.

Utilizar esta propiedad para volver a resolver el Ejercicio 6.1 parte a, pero

generando únicamente valores de puntos dentro del círculo de base de la montaña:

Problema: se idealiza una montaña como un cono inscrito en una región cuadrada de lado 1 km. La base de la montaña es circular, con centro en $(0.5, 0.5)$ y radio $r = 0.4$km, y la altura es $H = 8$km. La altura de cada punto (x, y) de la montaña está dada por la función $f(x, y) = H - H/r \times \sqrt{(x - 0.5)^2 + (y - 0.5)^2}$, en la zona definida por el círculo, y 0 fuera del círculo. El volumen total de la montaña (en km cúbicos) puede verse como la integral de la función altura en la región.
Parte a: escribir un programa para calcular el volúmen por Monte Carlo. Realizar 10^6 replicaciones y estimar el valor de ζ y el error cometido (con nivel de confianza 0.95), utilizando como criterio la aproximación normal.

Comparar la precisión obtenida con la alcanzada en el ejercicio 6.1.

Sugerencia: tener en cuenta que al estar generando puntos dentro del círculo, estamos calculando una integral de Lebesgue-Stieltjes, por lo que es necesario ajustar el integrando de manera que quede explícita la integral

en la forma $\int k(z)dF(z)$, con z un vector. En particular, por ser un sorteo uniforme dentro del círculo, la densidad de probabilidad en el círculo es $1/$(área del círculo), y 0 afuera del mismo.

Fecha entrega: Ver cronograma del curso.