Curso:
Confiabilidad estructural de componentes mecánicos con daño
edición 2020
Sesión 7 (parte b): Sorteo de variables aleatorias con distribución arbitraria

docentes del curso: R. Mussini, H. Cancela

dictado semestre 1 - 2020

Generación de variables aleatorias con distribución dada arbitraria

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

Si bien en algunos casos alcanza con generar muestras con distribución uniforme en el intervalo $[0, 1]$, lo que se hace en forma directa utilizando los generadores de números anteriores vistos anterioremente, para la mayor parte de las aplicaciones es necesario generar muestras con otras distribuciones, por ejemplo normal, exponencial, etc.

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.

Discutiremos 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 continuas (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: 2019-05-03)

- 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 http://omega.albany.edu:8008/neal.pdf (último acceso: 2019-05-03).)

- Artículo "The Markov Chain Monte Carlo revolution", Persi Diaconis, disponible en http://statweb.stanford.edu/~cgates/PERSI/papers/MCMCRev.pdf (último acceso: 2019-05-03).

- Libro "Handbook of Markov Chain Monte Carlo", Charles J. Geyer, disponible en http://www.mcmchandbook.net/HandbookChapter1.pdf (último acceso: 2019-05-03).

- 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: 2019-05-03).

- Demostración en línea "Markov Chain Monte Carlo Simulation Using the Metropolis Algorithm", http://demonstrations.wolfram.com/MarkovChainMonteCarloSimulationUsingTheMetropolisAlgorithm/(últim
acceso: 2019-05-03).

Métodos generales para generar una muestra de distribución arbitraria

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.

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 $F^{-1}(u) = -1/\lambda \ln(1 - u)$. Por lo tanto si generamos una uniforme U, podemos calcular 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 $X = -1/\lambda \ln(U)$.

Otro ejemplo de interés para este curso es el caso de una v.a. W de distribución Weibull de parámetros (positivos) (k, λ) (usualmente llamados parámetros de forma y escala). La densidad de distribución de probabilidad de W es $f_W(w) = (k/\lambda)(w/\lambda)^{k-1} e^{-(w/\lambda)^k}$, para $w \geq 0$ (y 0 para $w < 0$). La distribución de probabilidad de W es $F_W(w) = 1 - e^{-(w/\lambda)^k}$, para $w \geq 0$ (y 0 para $w < 0$). . La función inversa de la distribución es

$F_W^{-1}(u) = \lambda(-ln(1-u))^{1/k}$, por lo cual sorteando una v.a. uniforme U y sustituyendo el valor en esta expresión, conseguimos una muestra de W (en muchos lenguajes ya puede estar programada esta inversa).

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. Este es el método más usado para el sorteo de distribuciones discretas cuando el soporte (conjunto de valores posibles de la variable) es finito y de cardinal pequeño.

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

Obtención de variables aleatorias normales - transformada de Box-Muller.

Existen también métodos específicos para distribuciones particulares, que pueden ser más eficientes computacionalmente. En el curso veremos únicamente la transformada de Box-Muller, que permite obtener valores con distribución normal, por ser de mucho interés y utilidad, y en muchos casos ser más eficiente computacionalmente que el método de la transformada inversa en ese caso (por no existir expresión analítica para la inversa de la función de distribución normal).

Sean dos dos valores uniformes U_1 y U_2 aleatorios, independientes.

Sea $Z_1 = \sqrt{-2ln(U_1)}cos(2\pi U_2)$, y $Z_2 = \sqrt{-2ln(U_1)}sen(2\pi U_2)$.
Entonces Z_1 y Z_2 son v.a. normales (0,1) independientes.

No daremos la demostración, puede buscarse por ejemplo en
`https://mathworld.wolfram.com/Box-MullerTransformation.html`
y `https://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo2005/notes/GaussianSampling.pdf`

Transformaciones entre variables aleatorias.

Las siguientes propiedades son útiles para la generación de variables aleatorias:

- Si X e Y son v.a. y a y b constantes numéricas, y $Y = aX + b$, entonces (por linealidad de la esperanza) $E(Y) = aE(X) + b$, y $Var(Y) = a^2 Var(X)$. Para generar una muestra de Y, alcanza con generar una muestra de X y calcular $aX + b$.

- En particular, para generar una muestra de una v.a. normal de media m y desviación estándar s (atención que a veces se da la varianza y no la desviación estándar), alcanza con generar una v.a. X de distribución normal (0,1), y calcular $sX + m$.

- A veces es útil definir una v.a. truncada. Por ejemplo, un fenómeno puede aproximadamente representarse con una v.a. normal truncada en 0 (es decir, solo valores positivos). Sea Y una v.a. de media m y

desviación estándar s. Entonces $F_Y(0) = Prob(Y \leq 0)$ es la probabilidad que Y sea menor que 0. Podemos definir Z una v.a. derivada de la anterior, tal que $F_Z(z) = (F_Y(z) - F_Y(0))/(1 - F_Y(0))$ para $z > 0$ y 0 para $z \leq 0$. La media de Z será mayor que la de Y, y su varianza menor; la diferencia será mayor cuanto mayor sea $F_Y(0)$, si bien dependerá también de la forma de la distribución.

- Una forma de sortear esta v.a. Z es sorteando valores de Y hasta que cumplen la condición requerida:

1. Repetir
   1.1 Sortear Y con distribución F_Y;
   Hasta que $Y > 0$.
2. $Z = Y$.

Este procedimiento es rápido si $F_Y(0)$ es pequeño. Como se comentó, la media y varianza de Z no coincidirán con las de Y, hay que calcularlas aparte (o verificar que el error de aproximación cometido no sea significativo a efectos del estudio en curso).

Bibliotecas en la web

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.

Muchos lenguajes y sistemas de programación ya tienen incluidas sus propias bibliotecas internas con varias funciones de interés.

- CDFLIB: Cumulative Density Functions.
  - Versión FORTRAN: `http://people.sc.fsu.edu/~jburkardt/f_src/cdflib/cdflib.html` (último acceso:2019-05-03)
  - 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:2019-05-03)

- Probability Distributions Java Package: http://www1.fpl.fs.fed.us/distributions.html (último acceso:2019-05-03)

- 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:2019-05-03)

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?

Ejercicio

- Usando una planilla electrónica, generar 100 valores de una v.a. con distribución exponencial de parámetro lambda=0.3 . Graficar. Generar 100 valores de una v.a. con distribución normal de parámetros m=5, s=1. Graficar.

- Usando el lenguaje de programación de su elección, generar 100 valores de una v.a. con distribución exponencial de parámetro lambda=0.3. Calcular la media y varianza de los valores generados. Generar 100 valores de una v.a. con distribución normal de parámetros m=5, s=1. Calcular la media y varianza de los valores generados.