Generación de valores de las variables aleatorias

Juan F. Olivares-Pacheco*

14 de junio de 2007

Resumen

En todo modelo de simulación estocástico, existen una o varias variables aleatorias interactuando. Generalmente, estas variables siguen distribuciones de probabilidad teóricas o empíricas diferentes a la distribución uniforme. Por consiguiente, para simular este tipo de variables, es necesario contar con un generador de números uniformes y una función que a través de un método específico, transforme estos números en valores de la distribución de probabilidad deseada.

1. La generación de estadísticas simuladas

La generación de estadísticas simuladas, o sea de valores de las variables aleatorias, tienen una naturaleza enteramente numérica y deben configurarse mediante la aportación de números aleatorios. Estos números se introducen al proceso o sistema bajo estudio (en donde el sistema se representa por un modelo probabilístico) a fin de obtener ciertas cifras (o valores de las variables aleatorias) de las cuales se obtengan las respuestas. Como regla general, el proceso de simulación estocástica comprende una actividad de reemplazo del universo estadístico de elementos que se emplean en el sistema por su contraparte teórica, un universo descrito por una distribución probabilística supuesta (por ejemplo, una distribución normal), seguido de un muestreo efectuado sobre esta población teórica, con la ayuda de cierto tipo de generador de números aleatorios.

Sin embargo, en algunos casos es posible que sea difícil encontrar una distribución teórica convencional que describa un proceso estocástico particular o alguno de los componentes de dicho proceso.

*Departamento de Matemática, Universidad de Atacama, CHILE. E-mail: jolivares@mat.uda.cl

En estos casos, el proceso estocástico se puede reproducir (o si se quiere simular) tan sólo mediante un muestreo aplicado sobre las distribuciones empíricas en lugar de considerar algunas de las distribuciones teóricas conocidas. Resulta aconsejable el empleo, en primer lugar, de las distribuciones teóricas convencionales y si ninguna de ellas describe adecuadamente el comportamiento del proceso, entonces deberemos, necesariamente, recurrir a distribuciones empíricas.

La primera meta de este capítulo, es proveer un conjunto de técnicas específicas para generar (con una computadora) valores de variables aleatorias a partir de las distribuciones de probabilidad más conocidas, así como también de ciertos métodos generales para generar los citados valores tomados como base cualquier distribución empírica que probablemente se configure al intentar la solución de problemas estocásticos.

Al considerar los procesos estocásticos que involucran variables continuas o discretas, pero siempre de tipo aleatorio, definimos una función $F_X(x)$ llamada *función de distribución acumulada* de x, la cual denota la probabilidad de que una variable aleatoria X tome un valor menor o igual a x. Si la variable aleatoria es discreta, entonces x tendrá valores específicos y $F_X(x)$ será una función escalonada. Si $F_X(x)$ es continua en el dominio de x, entonces esta función es podrá diferenciar, para lo cual se define $f_X(x) = d\ F_X(x)/dx$. La derivada $f_X(x)$ recibe el nombre de función de densidad de probabilidad. La función de distribución acumulada se puede proponer matemáticamente como sigue:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\ dt \tag{1}$$

donde $F_X(x)$ se define en el intervalo $0 \leq F_X(x) \leq 1$ y $f_X(t)$ representa el valor de la función de densidad de probabilidad de la variable aleatoria X cuando $X = t$.

Los valores de tales variables uniformemente distribuidas en el intervalo $(0, 1)$ juegan un papel muy importante en la generación de los valores de variables aleatorias, obtenidas a partir de otros tipos de distribución de probabilidad.

Se tienen tres métodos básicos para generar los valores de variables aleatorias a partir de las distribuciones de probabilidad: *el método de la transformada inversa*, *el de rechazo* y el *de composición*. Estos métodos o alguna de sus variantes respectivas, proporcionan la base general para simular la mayoría de las distribuciones.

2. El método de la transformada inversa

El método de la transformada inversa, es el método más utilizado en la obtención de variables aleatorias para experimentos de simulación. Para aplicar éste método suponga que queremos generar el valor de una variable aleatoria discreta X con función de masa de probabilidad:

$$P\left(X = x_j\right) = p_j,\ j = 0,\ 1,\ \ldots,\ \sum_j p_j = 1 \tag{2}$$

Para esto, generamos un número aleatorio R; es decir, R está distribuido uniformemente en $(0,\ 1)$, y sea:

$$X = \begin{cases} x_0 & \text{, si} \quad R < p_0 \\ x_1 & \text{, si} \quad p_0 \leq R < p_0 + p_1 \\ \vdots & \\ x_j & \text{, si} \quad \sum_{i=1}^{j-1} p_j \leq R < \sum_{i=1}^{j} p_j \\ \vdots & \end{cases} \tag{3}$$

Como $P\left(a \leq U < b\right) = b - a$ para $0 < a < b < 1$, tenemos que:

$$P\left(X = x_j\right) = P\left(\sum_{i=1}^{j-1} p_i \leq R < \sum_{i=1}^{j} p_i\right) = p_j \tag{4}$$

y entonces X tiene la distribución deseada.

Para este método se pueden realizar ciertas observaciones, por ejemplo, podemos escribir lo anterior es forma algorítmica.

Algoritmo 2.1 *Método de la transformada inversa para variables aleatorias discretas.*

1. Generar un número aleatorio R.
2. Si $R < p_0$ hacer $X = x_0$ y terminar.
3. Si $R < p_0 + p_1$ hacer $X = x_1$ y terminar.
4. Si $R < \sum_{i=1}^{j} p_j$ hacer $X = x_j$ y terminar.

Si los x_i, $i \geq 0$, están ordenados de modo que $x_0 < x_1 < x_2 < \ldots$ y si F denota la función de distribución de X, entonces $F_X\left(x_k\right) = \sum_{i=0}^{k} p_i$ y:

X será igual a x_j, si $F_X\left(x_{j-1}\right) \leq R < F_X\left(x_j\right)$

En otras palabras, después de generar un número aleatorio R determinamos el valor de X hallando el intervalo $(F_X(x_{j-1}),\ F_X(x_j))$ en el que está R (o, de forma equivalente hallando la inversa de $F_X(R)$). Es por esta razón que el anterior se llama método de la transformada inversa discreta para generar X.

Ejemplo 2.1 *Si queremos simular una variable aleatoria X tal que*

$$f_X(x) = P(X = x) = \begin{cases} 0{,}20 & , \textit{ si } x = 1 \\ 0{,}15 & , \textit{ si } x = 2 \\ 0{,}25 & , \textit{ si } x = 3 \\ 0{,}40 & , \textit{ si } x = 4 \end{cases}$$

entonces si obtenemos la función de distribución, tenemos:

$$F_X(x) = P(X < x) = \begin{cases} 0{,}20 & , \textit{ si } x \leq 1 \\ 0{,}35 & , \textit{ si } 1 < x \leq 2 \\ 0{,}60 & , \textit{ si } 2 < x \leq 3 \\ 1{,}00 & , \textit{ si } x > 4 \end{cases}$$

Finalmente, las variables aleatorias X se obtienen a través de:

$$X = \begin{cases} 1 & , \textit{ si } R < 0{,}20 \\ 2 & , \textit{ si } 0{,}20 \leq R < 0{,}35 \\ 3 & , \textit{ si } 0{,}35 \leq R < 0{,}60 \\ 4 & , \textit{ si } 0{,}60 \leq R < 1 \end{cases}$$

■

Ahora consideremos si la variable que se desea generar es una variable aleatoria continua con función de distribución F. El método de la transformada inversa para variables continuas, se basa en la siguiente proposición.

Proposición 2.1 *Sea R una variable aleatoria uniforme en $(0,\ 1)$. Para cualquier función de distribución continua F, invertible, la variable aleatoria X definida como:*

$$X = F_X^{-1}(r)$$

tiene distribución F. (F_X^{-1} se define como el valor de x tal que $F_X(x) = r$).

La proposición anterior muestra entonces que para generar una variable aleatoria X a partir de la función de distribución continua F, generamos un número aleatorio R y hacemos entonces $X = F_X^{-1}(R)$.

Ejemplo 2.2 *Distribución triangular. Suponga que queremos generar una variable aleatoria X con función continua de probabilidad:*

$$f_X(x) = \begin{cases} 1+x & , \; si \; -1 \leq x < 0 \\ 1-x & , \; si \; 0 \leq x \leq 1 \end{cases}$$

Si obtenemos $F_X(r)$, tenemos:

$$F_X(x) = P(X < x) = \begin{cases} 0 & , \; si \; x < -1 \\ \frac{(1+x)^2}{2} & , \; si \; -1 \leq x < 0 \\ 1 - \frac{(1-x)^2}{2} & , \; si \; 0 \leq x < 1 \\ 1 & , \; si \; x \geq 1 \end{cases}$$

Ahora calculando la función inversa de la función de distribución acumulada, tenemos que:

$$X = \begin{cases} \sqrt{2r} - 1 & , \; si \; 0 \leq r \leq \frac{1}{2} \\ 1 - \sqrt{2-2r} & , \; si \; \frac{1}{2} < r \leq 1 \end{cases}$$

Por lo tanto, para generar la variable aleatoria X podemos utilizar el siguiente algoritmo.

Algoritmo 2.2 *Generación de variables aleatorias con distribución triangular.*

1. *Obtener un numero aleatorio R uniformemente distribuido en $(0,\ 1)$.*
2. *Si $R \leq \frac{1}{2}$, hacer $X = \sqrt{2r} - 1$. En caso contrario hacer $X = 1 - \sqrt{2-2r}$.*
3. *Repetir los pasos 1 y 2, tantas veces como variables aleatorias se desean generar.*

Utilizando el algoritmo anterior generamos una muestra aleatoria de tamaño $n = 10000$, y obteniendo el histograma con el respectivo ajuste de los datos tenemos (ver figura 1).

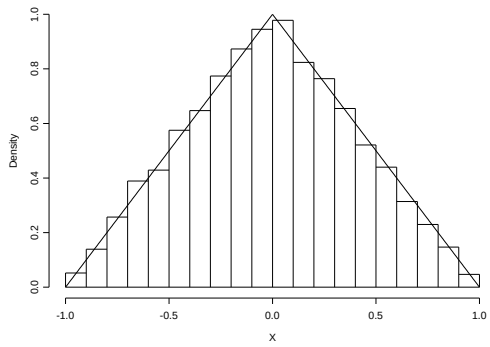

Figura 1: Histograma de los datos generados con el algoritmo para obtener variables aleatoria triangular.

■

Desafortunadamente, para muchas de las distribuciones de probabilidad, resulta imposible o extremadamente difícil expresar a x en terminos de la transformación inversa. Cuando es éste el caso, el único recurso de que se dispone, consiste en obtener una aproximación numérica para la transformación inversa, o bien recurrir a alguno de los siguientes dos métodos.

3. El método de rechazo

Otro procedimiento para generar valores de variables aleatorias de distribuciones de probabilidad no uniformes, es el método de rechazo. Este método consiste primeramente en generar un valor de la variable aleatoria y en seguida probar que dicho valor simulado proviene de la distribución de probabilidad que se está analizando. Para comprender la lógica de este método, suponga que $f(x)$ es una distribución de probabilidad *acotada y con rango finito*, es decir, $a \leq x \leq b$. De acuerdo a esta función de probabilidad, la aplicación del método de rechazo implica el desarrollo del siguiente algoritmo:

Algoritmo 3.1 *Método de rechazo.*

1. Normalizar el rango de f mediante un factor de escala c tal que:

$$cf(x) \leq 1, \quad a \leq x \leq b$$

2. Generar dos números uniformes R_1 y R_2.

3. Determinar el valor de la variable aleatoria x de acuerdo a la siguiente relación líneal de R_1:

$$x = a + (b - a) R_1$$

4. Evaluar la función de probabilidad en $x = a + (b - a) R_1$.

5. Determinar si la siguiente desigualdad se cumple:

$$R_2 \leq cf\left(a + (b - a) R_1\right)$$

   Se utiliza a $x = a + (b - a) R_1$ si la respuesta es afirmativa como un valor simulado de la variable aleatoria. De lo contrario, es necesario regresar al paso 1 tantas veces como sea necesario.

La teoría sobre la que se apoya este método se basa en el hecho de que la probabilidad de que $R_2 \leq cf(x)$ es exactamente $cf(x)$. Por consiguiente , si un número es escogido al azar de acuerdo a $x = a + (b - a) R_1$ y rechazado si $R_2 > cf(x)$, entonces la distribución de probabilidad de las $x's$ aceptadas será exactamente $f(x)$. Por otra parte, conviene señalar que si todas las $x's$ fueran aceptadas, entonces x estaría uniformemente distribuida entre a y b.

Finalmente, es necesario mencionar que algunos autores como Tocher, han demostrado que el número esperado de intentos para que x sea aceptada como una variable aleatoria que sigue una distribución de probabilidad $f(x)$, es $1/c$. Esto significa que este método podría ser un tanto ineficiente para ciertas distribuciones de probabilidad en las cuales la moda sea grande.

Ejemplo 3.1 *Distribución parabólica. Se desea generar valores de variables aleatorias que sigan la siguiente distribución de probabilidad:*

$$f_X(x) = \frac{3}{4}\left(1 - x^2\right), \; si \; -1 \leq x \leq 1$$

Entonces $a = -1$, $b = 1$ *implica que* $x = 2R - 1$, *pero* $f_X(x)$ *esta definida en el intervalo* $0 \leq f_X(x) \leq \frac{3}{4}$, *si se normaliza haciendo* $c = \frac{4}{3}$, *se transformará a* $f_X(x)$ *al intervalo unitario. Ahora, para generar valores de variables aleatoria con distribución de probabilidad* $f_X(x)$, *tenemos que aplicar el siguiente algoritmo:*

Algoritmo 3.2 *Generación de variables aleatorias con distribución parabólica.*

1. *Generar dos números aleatorios* R_1 *y* R_2.

2. *Calcular* $x = 2R_1 - 1$.

3. *Si* $R_2 \leq \frac{4}{3} f(2R_1 - 1)$. *Si la respuesta es afirmativa, entonces* $x = 2R_1 - 1$ *es un valor simulado de la variable aleatoria. De lo contrario, se requiere regresar al paso* 1 *tantas veces como sea necesario.*

Utilizando el algoritmo anterior generamos una muestra aleatoria de tamaño $n = 10000$, *y obteniendo el histograma con el respectivo ajuste de los datos tenemos (ver figura 2).*

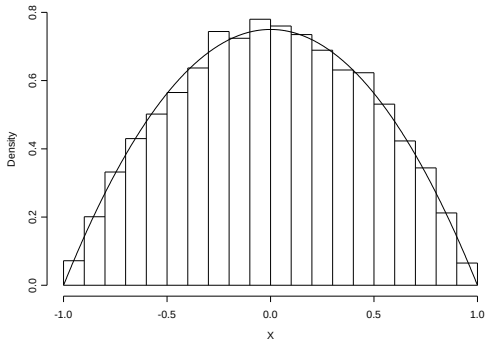

Figura 2: Histograma de los datos generados con el algoritmo para obtener variables aleatoria parabólicas.

■

4. El método de composición

El tercer método existente para la generación de variables aleatorias utilizando computadoras es el llamado método de composición o método de mezclas.

Mediante este método la distribución de probabilidad $f_X(x)$ se expresa como una mezcla de varias distribuciones de probabilidad $f_i(x)$ seleccionadas adecuadamente.

$$f_X(x) = \sum_{i=1}^{n} A_i f_i(x) \text{ y } \sum_{i=1}^{n} A_i = 1 \tag{5}$$

La guía para la selección de los $f_i(x)$ está dada sobre las consideraciones relativas a la bondad de ajuste y el objetivo de minimizar $\sum_{i=1}^{n} T_i A_i$ donde T_i es el tiempo esperado de computación para generar valores de variables aleatorias a partir de $f_i(x)$.

Los pasos requeridos para la aplicación de este método en la simulación de variables aleatorias son los siguientes:

Algoritmo 4.1 *Método de composición.*

1. Dividir la distribución de probabilidad original en subáreas.
2. Definir una distribución de probabilidad para cada subárea.
3. Expresar la distribución de probabilidad original en la forma siguiente:

$$f(x) = A_1 f_1(x) + A_2 f_2(x) + \cdots + A_n f_n(x) = \sum_{i=1}^{n} A_i f_i(x)$$

4. Obtener la función de distribución de las áreas.
5. Generar dos números aleatorios R_1 y R_2.
6. Seleccionar la distribución de probabilidad $f_i(x)$ con la cual se va a simular el valor de X. La selección de esta distribución se obtiene al aplicar el método de la transformación inversa, en la cual el eje de las ordenadas está representado por la distribución acumulada de las áreas, y el eje de las abscisas por las distribuciones $f_i(x)$. Para esta selección se utiliza el número aleatorio R_1.
7. Utilizar el número aleatorio R_2 para simular por el método de la transformada inversa o algún otro procedimiento especial, números al azar que sigan la distribución de probabilidad $f_i(x)$ seleccionada en el paso anterior.

Ejemplo 4.1 *Se desea generar variables aleatorias de la siguiente distribución de probabilidad:*

$$f_X(x) = \begin{cases} 1+x & , \; si \; -1 \leq x \leq 0 \\ 1-x & , \; si \; 0 \leq x \leq 1 \end{cases}$$

Siguiendo los pasos descritos previamente, la generación de variables aleatorias, puede ser resumida en los siguientes pasos.

1. *La distribución de probabilidad original, se va ha dividir en dos áreas, definidas por los límites de la misma, entonces estas áreas son:* $A_1 = \frac{1}{2}$ *que es el área definida en el intervalo* $-1 \leq x \leq 0$, *y* $A_2 = \frac{1}{2}$ *que corresponde al área definida en el intervalo* $0 \leq x \leq 1$.

2. *En seguida se determinan las distribuciones de probabilidad y distribución acumulada de las áreas definidas en el paso anterior.*

$$f_X(x)_1 = 2(1+x) \;\; y \;\; F_X(x)_1 = (x+1)^2$$

$$f_X(x)_2 = 2(1-x) \;\; y \;\; F_X(x)_2 = 2x - x^2$$

3. *La distribución de probabilidad original, se puede expresar como:*

$$\begin{aligned} f_X(x) &= A_1 f_X(x)_1 + A_2 f_X(x)_2 \\ &= \frac{1}{2}(2(1+x)) + \frac{1}{2}(2(1-x)) \end{aligned}$$

4. *Con las áreas y distribuciones* $f_i(x)$ *definidas en los pasos anteriores, la distribución acumulada de las áreas sería:*

$$F_A = \begin{cases} \frac{1}{2} & , \textit{ si } x \textit{ está definido en } f_X(x)_1 \\ 1 & , \textit{ si } x \textit{ está definido en } f_X(x)_2 \end{cases}$$

5. *Generar dos números aleatorios* R_1 *y* R_2.

6. *Si* $R_1 < \frac{1}{2}$, *entonces se simulan valores de la distribución* $f_X(x)_1$*:*

$$(x+1)^2 = R_2$$

$$x = \sqrt{R_2} - 1$$

en caso contrario, se simulan valores de la distribución $f_X(x)_2$*:*

$$2x - x^2 = R_2$$

$$x = 1 - \sqrt{1 - R_2}$$

7. *Repetir los pasos anteriores tantas veces como se desee.*

Con estos pasos, generamos una muestra aleatoria de tamaño $n = 10000$, y obteniendo el histograma con el respectivo ajuste de los datos tenemos (ver figura 3).

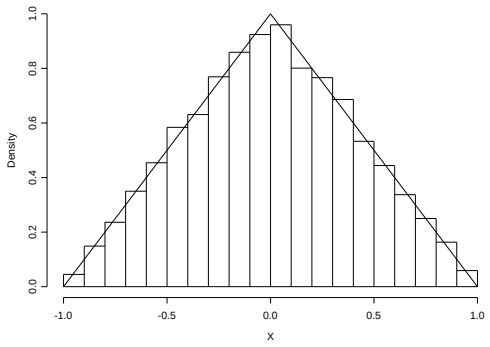

Figura 3: Histograma de los datos generados con el algoritmo para obtener variables aleatoria triangular.

■

5. Distribuciones continuas de probabilidad

5.1. Distribución uniforme

Quizá la función de densidad de probabilidad más simple es aquella que se caracteriza por ser constante, en el intervalo $(a,\ b)$ y cero fuera de él. Esta función de densidad define la distribución conocida como uniforme o rectangular. El valor más sobresaliente que puede tener la distribución uniforme respecto a las técnicas de simulación radica en su simplicidad y en el hecho de que tal distribución se puede emplear para simular variables aleatorias a partir de casi cualquier tipo de distribución de probabilidad.

Matemáticamente, la función de densidad uniforme se define como sigue:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{, si } a < x < b \\ 0 & \text{, si } a > x > b \end{cases} \tag{6}$$

En la ecuación 6, X es una variable aleatoria definida en el intervalo $(a,\ b)$. La función de la distribución acumulada $F_X(x)$, para una variable aleatoria X uniformemente distribuida, se puede

representar por:

$$F_X(x) = \int_a^x \frac{1}{b-a}\, dt = \frac{x-a}{b-a}, \quad 0 \leq F(x) \leq 1 \tag{7}$$

El valor esperado y la varianza de una variable aleatoria uniformemente distribuida están dados por las siguientes expresiones:

$$E[X] = \frac{b+a}{2} \tag{8}$$

$$V(X) = \frac{(b-a)^2}{12} \tag{9}$$

Al efectuar aplicaciones de esta función, los parámetros de la función de densidad uniforme 6 esto es, los valores numéricos de a y de b, no necesariamente deben ser conocidos en forma directa. En casos típicos, aunque esto no sucede en todas las distribuciones uniformes, solamente conocemos la media y la varianza de la estadística que se va a generar. En estos casos, los valores de los parámetros se deben derivar al resolver el sistema que consta de las ecuaciones 8 y 9, para a y para b, pues se supone que $E[X]$ y $V(X)$ son conocidos. Este procedimiento, semejante a una técnica de estimación conocida en la literatura estadística como método de momentos, proporciona las siguientes expresiones:

$$a = E[X] - \sqrt{3V(X)} \tag{10}$$

$$b = 2E[X] - a \tag{11}$$

Para simular una distribución uniforme sobre cierto intervalo conocido $(a,\ b)$ deberemos, en primer lugar, obtener la transformación inversa para la ecuación 7, entonces:

$$x = a + (b-a)R, \quad 0 \leq R \leq 1 \tag{12}$$

En seguida generamos un conjunto de números aleatorios correspondiente al rango de la probabilidades acumulativas, es decir, los valores de variables aleatorias uniformes definidas sobre el rango 0 a 1. Cada número aleatorio R determina, de manera única, un valor de la variable aleatoria x uniformemente distribuida.

En la figura 4 se tiene una muestra de variables aleatorias uniformes con $a = 5$ y $b = 10$, y podemos observar el comportamiento de los valores generados.

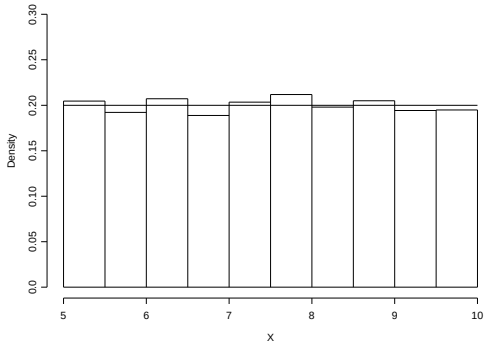

Figura 4: Variables aleatorias uniformes.

5.2. Distribución exponencial

Durante nuestra experiencia diaria, observamos cómo transcurren los intervalos de tiempo definidos entre las ocurrencias de los eventos aleatorios distintos, y sobre la base de un plan de tiempos completamente independientes, recibimos información sobre numerosos eventos que ocurren en nuestro alrededor. Bastaría con citar los nacimiento, defunciones, accidentes, o conflictos mundiales, para mencionar sólo algunos. Si es muy pequeña la probabilidad de que ocurra un evento en un intervalo corto, y si la ocurrencia de tal evento es, estadísticamente independiente respecto a la ocurrencia de otros eventos, entonces el intervalo de tiempo entre ocurrencias de eventos de este tipo estará distribuido en forma exponencial.

Se dice que una variable aleatoria X tiene una distribución exponencial, si se puede definir a su función de densidad como:

$$f_X(x) = \alpha e^{-\alpha x} \tag{13}$$

con $\alpha > 0$ y $x \geq 0$. La función de distribución acumulada de X está dada por:

$$F_X(x) = \int_0^x \alpha e^{-\alpha t}\, dt = 1 - e^{-\alpha x} \tag{14}$$

y la media junto con la varianza de X se puede expresar como:

$$E[X] = \frac{1}{\alpha} \tag{15}$$

$$V(X) = \frac{1}{\alpha^2} \tag{16}$$

Como la distribución exponencial solamente tiene un parámetro α, es posible expresarlo como:

$$\alpha = \frac{1}{E[X]} \tag{17}$$

Existen muchas maneras para lograr la generación de valores de variables aleatorias exponenciales. Puesto que $F_X(x)$ existe explícitamente, la técnica de la transformada inversa nos permite desarrollar métodos directos para dicha generación. Por tanto:

$$R = 1 - e^{-\alpha x} \tag{18}$$

y consecuentemente:

$$x = -\frac{1}{\alpha}\ln(1-R) = -E[X]\ln(1-R) \tag{19}$$

Por consiguiente, para cada valor del número aleatorio R se determina un único valor para x. Los valores de x toman tan sólo magnitudes no negativas, debido a que $\log R \leq 0$ para $0 \leq R \leq 1$, y además se ajustan a la función de densidad exponencial con un valor esperado $E[X]$.

En la figura 5 se tiene una muestra de variables aleatorias exponenciales con $\alpha = 1$, y podemos observar el comportamiento de los valores generados.

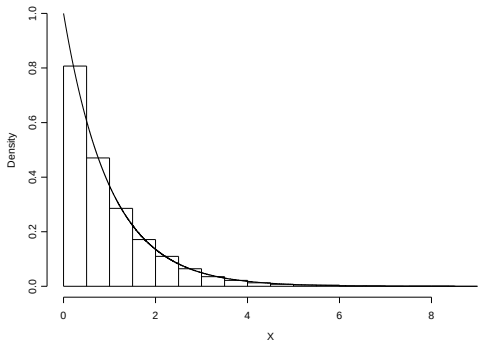

Figura 5: Variables aleatorias exponencial.

5.3. Distribución normal

La más conocida y más ampliamente utilizada distribución de probabilidad es sin duda la distribución normal y su popularidad se debe cuando menos a dos razones que presentan sus propiedades

generales. Las pruebas matemáticas nos señalan, que bajo ciertas condiciones de calidad, resulta justificado que esperamos una distribución normal mientras que la experiencia estadística muestre que, de hecho, muy a menudo las distribuciones se aproximan a la normal.

La distribución normal basa su utilidad en el teorema del límite central. Este teorema postula que, la distribución de probabilidad de la suma de N valores de variables aleatorias x_i independiente pero idénticamente distribuidos, con medias respectivas μ_i y variancias σ_i^2 se aproxima asintóticamente a una distribución normal, a medida que N se hace muy grande, y que dicha distribución tiene como media y varianzas respectivamente, a:

$$\mu = \sum_{i=1}^{N} \mu_i \tag{20}$$

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2 \tag{21}$$

A partir de la distribución normal, se pueden derivar otras muchas de las distribuciones existentes que juegan un papel muy importante en la estadística moderna, por ejemplo la Chi-cuadrada, la t, y la distribución F, las cuales se originan a partir de consideraciones hechas sobre la distribución de probabilidad de la suma de los cuadrados de un número específico de valores de variables aleatorias con una distribución normal estándar.

Si la variable aleatoria X tiene una función de densidad $f_X(x)$ dada como:

$$f_X(x) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}, \quad -\infty < x < \infty \tag{22}$$

con σ_x positiva, entonces se dice que X tiene una distribución normal o Gaussiana, con parámetros μ_x y σ_x.

Si los parámetros de la distribución normal tienen los valores de $\mu_x = 0$ y $\sigma_x = 1$, la función de distribución recibirá el nombre de *distribución normal estándar*, con función de densidad:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}, \quad -\infty < z < \infty \tag{23}$$

Cualquier distribución normal se puede convertir a la forma estándar, mediante la siguiente substitución:

$$z = \frac{x - \mu_x}{\sigma_x} \tag{24}$$

La función de distribución acumulada $F_X(x)$ o $F_Z(z)$ no existe en forma explícita; sin embargo, esta última se encuentra totalmente tabulada en cualquier libro sobre estadística. El valor esperado y la varianza de la distribución normal no estándar están dados por:

$$E[X] = \mu_x \tag{25}$$

$$V(X) = \sigma_x^2 \tag{26}$$

Existen varios métodos para generar en una computadora, valores de variable aleatoria distribuidos en forma normal. Debido a su popularidad sólo se discutirá en detalle el procedimiento llamado del límite central.

A fin de simular una distribución normal con media μ_x y varianza σ_x^2 dadas, se debe proponer la siguiente interpretación matemática del teorema del límite central. Si $r_1,\ r_2,\ \ldots,\ r_N$ representan variables aleatorias independientes, cada una de las cuales posee la misma distribución de probabilidad caracterizada por $E[r_i] = \theta$ y $V(r_i) = \sigma^2$, entonces:

$$\lim_{N\to\infty} P\left(a < \frac{\sum_{i=1}^{N} r_i - N\theta}{\sqrt{N}\sigma} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left\{-\frac{1}{2}z^2\right\}\ dz \tag{27}$$

donde

$$E\left[\sum_{i=1}^{N} r_i\right] = N\theta \tag{28}$$

$$V\left(\sum_{i=1}^{N} r_i\right) = N\sigma^2 \tag{29}$$

$$z = \frac{\sum_{i=1}^{N} r_i - N\theta}{\sigma\sqrt{N}} \tag{30}$$

Tanto de la definición de la distribución normal estándar como de la ecuación 24, se sigue que z es un valor de variable aleatoria con distribución normal estándar.

El procedimiento para simular valores normales utilizando computadoras requiere el uso de la suma de K valores de variable aleatoria distribuidos uniformemente; esto es, la suma de $r_1,\ r_2,\ \ldots,\ r_K$, con cada r_i definida en el intervalo $0 < r_1 < 1$. Así, tenemos:

$$\theta = \frac{1}{2} \tag{31}$$

$$\sigma = \frac{1}{\sqrt{12}} \tag{32}$$

$$z = \frac{\sum_{i=1}^{K} r_i - \frac{K}{2}}{\sqrt{\frac{K}{12}}} \tag{33}$$

Pero, por definición, z es un valor de variable aleatoria con distribución normal estándar que se puede escribir en la forma sugerida por la ecuación 24, donde x es un valor de variable aleatoria distribuido en forma normal que se va a simular, con media μ_x y varianza σ_x^2. Igualando las ecuaciones 24 y 33 obtenemos:

$$\frac{x - \mu_x}{\sigma_x} = \frac{\sum_{i=1}^{K} r_i - \frac{K}{2}}{\sqrt{\frac{K}{12}}} \tag{34}$$

y resolviendo para x, tenemos que:

$$x = \sigma_x \sqrt{\frac{12}{K}} \left(\sum_{i=1}^{K} r_i - \frac{K}{2} \right) + \mu_x \tag{35}$$

Por lo tanto, mediante la ecuación 35 podemos proporcionar una formulación muy simple para generar valores de variable aleatoria normalmente distribuidos, cuya media sea igual a μ_x y varianza σ_x^2. Para generar un solo valor de x (un valor de variable aleatoria con distribución normal) bastará con sumar K números aleatorios definidos en el intervalo de 0 a 1. Substituyendo el valor de esta suma en la ecuación 35, así como también los valores de μ_x y σ_x para la distribución deseada, encontraremos que se ha determinado un valor particular de x. Ciertamente, este procedimiento se puede repetir tantas veces como valores de variable aleatoria normalmente distribuidos se requieran.

En la figura 6 se tiene una muestra de variables aleatorias normales con $\mu = 0$ y $\sigma = 1$, y podemos observar el comportamiento de los valores generados.

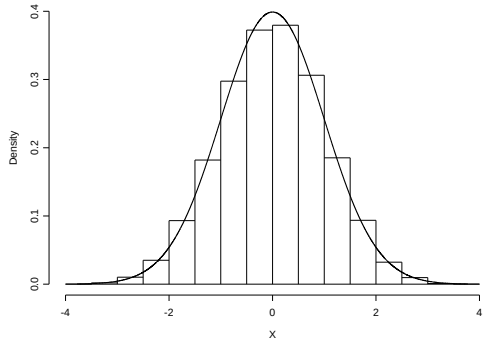

Figura 6: Variables aleatorias normales.

5.4. Distribución logarítmica normal

Si el logaritmo de una variable aleatoria tiene una distribución normal, entonces la variable aleatoria tendrá una distribución continua sesgada positivamente, conocida con el nombre de distribución logarítmica normal. Frecuentemente se hace uso de esta distribución para describir procesos aleatorios que representan al producto de varios eventos pequeños e independientes. Esta propiedad de la distribución logarítmica normal es conocida como *ley de efectos proporcionales*, que establecen una base sobre la que podemos decidir la validez con la cual una distribución logarítmica normal sirva o no para describir una variable aleatoria particular.

Las aplicaciones más importantes de la distriubución logarítmica normal quedan comprendidas en áreas del análisis de rendimientos y ganancias, análisis de ventas y teoría de puntos de ruptura. Esta última proporciona criterios básicos para identificar la distribución de ciertos tamaños de lotes, así como también la distribución del ingreso, la cual se comporta en forma logarítmica normal.

Se dice que X tiene una distribución logarítmica normal si cuando solamente se consideran los valores positivos de x, el logaritmo natural (en base e) de la variable aleatoria X tiene una función de densidad $f(y)$ dada como sigue:

$$f(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\left\{ \left(-\frac{1}{2} \right) \left(\frac{y - \mu_y}{\sigma_y} \right)^2 \right\}, \quad -\infty < y < \infty \tag{36}$$

con $y = \ln(x)$. Los parámetros μ_y y σ_y^2 que aparecen en la expresión, corresponde a la media y varianza de y, respectivamente.

El valor esperado y la varianza de los valores de variable aleatoria, distribuidos en forma logarítmica normal x, están dados por las fórmulas siguientes:

$$E[X] = \exp\left\{\mu_y + \frac{\sigma_y^2}{2}\right\} \tag{37}$$

$$V(X) = (E[X])^2 \left[\exp\left\{\sigma_y^2\right\} - 1\right] \tag{38}$$

La simulación de valores de variable aleatoria logarítmica normal con una media y varianza dadas, requiere necesariamente que μ_y y σ_y^2 estén expresadas en términos de $E[X]$ y de $V(X)$, lo cual se puede lograr con sólo resolver la ecuación 38 para $\exp\left\{\sigma_y^2\right\}$. Tenemos, que:

$$\exp\left\{\sigma_y^2\right\} = \frac{V(X)}{(E[X])^2} + 1 \tag{39}$$

Tomando ahora el logaritmo de ambos miembros de la ecuación 39 obtenemos:

$$\sigma_y^2 = \ln\left[\frac{V(X)}{(E[X])^2} + 1\right] \tag{40}$$

A continuación, tomamos el logaritmo de ambos miembros de la ecuación 37:

$$\ln(E[X]) = \mu_y + \frac{\sigma_y^2}{2} \tag{41}$$

y resolviendo para μ_y:

$$\mu_y = \ln(E[X]) - \frac{1}{2}\ln\left[\frac{V(X)}{(E[X])^2} + 1\right] \tag{42}$$

Ahora que tanto μ_y como σ_y^2 han quedado expresadas en términos de la media y la varianza de x, el valor logarítmico normal de la variable aleatoria que se va a generar o sea el valor de variable aleatoria z con distribución normal estándar, se puede definir como sigue:

$$z = \frac{\ln(x) - \mu_y}{\sigma_y} \tag{43}$$

Resolviendo la ecuación 43 para $\ln(x)$ y tomando el antilogaritmo de ambas partes de la ecuación, obtenemos:

$$\ln(x) = \mu_y + \sigma_y z \tag{44}$$

$$x = \exp\{\mu_y + \sigma_y z\} \tag{45}$$

Substituyendo el valor de z de la ecuación 33 en la ecuación 45, tendremos:

$$x = \exp\left\{\mu_y + \sigma_y \left(\frac{K}{12}\right)^{-1/2} \left(\sum_{i=1}^{K} r_i - \frac{K}{12}\right)\right\} \tag{46}$$

Resumiendo, para generar valores logarítmicos normales de variable aleatoria $x_1,\ x_2,\ \ldots,\ x_n$ con $E[X]$ y $V(X)$ dadas, debemos en primer lugar determinar μ_y y σ_y de las ecuaciones 40 y 42, y después substituir estos valores en la ecuación 46.

En la figura 7 se tiene una muestra de variables aleatorias logarítmicas normales con $\mu = 0$ y $\sigma = 0{,}01$, y podemos observar el comportamiento de los valores generados.

Figura 7: Variables aleatorias logarítmicas normales.

6. Distribuciones discretas de probabilidad

6.1. Distribución geométrica

Entre las primeras y probablemente más simples de las formulaciones matemáticas de procesos estocásticos, se encuentra la llamada *de ensayos* de Bernoulli. Estos ensayos son experimentos independientes al azar, en los que el resultado de cada ensayo queda registrado, ya sea como un éxito o un fracaso. La probabilidad de éxito se denota por p $(0 \leq p \leq 1)$ y se supone que p es constante para cualquier sucesión particular de ensayos. La probabilidad de un fracaso se denota por q, donde:

$$q = 1 - p \tag{47}$$

Una sucesión de ensayos de Bernoulli, combinada con cierto proceso de conteo, viene a construir la base conceptual para una gran familia de distribuciones discretas de probabilidad, incluyendo

la geométrica, binomial negativa, Poisson y otras distribuciones binomiales. Los valores de variable aleatoria que se generan al contar el número de fracasos en una sucesión de ensayos o eventos antes de que ocurra el primer éxito, son valores de variable aleatoria que se ajustan a una distribución geométrica. La distribución geométrica de probabilidad tiene un gran valor y utilidad en el área del control estadístico de calidad, así como también para las distribuciones de rezagos y movimientos en modelos econométricos.

La distribución geométrica queda descrita por la siguiente función de probabilidad:

$$f(x) = pq^x \quad \forall\ x = 0,\ 1,\ 2,\ \ldots \tag{48}$$

y la función de distribución acumulada está definida por:

$$F(x) = \sum_{t=0}^{x} pq^t \tag{49}$$

Puesto que por definición se tiene $F(x) = P(X \leq x)$, y como $P(X = 0) = F(0) = p$, el rango de $F(x)$ es $p \leq F(x) \leq 1$. Por otra parte, $P(X > x) = 1 - F(x)$, lo que implica que $P(X > 0) = q$ y además:

$$1 - F(x) = q^{x+1} \tag{50}$$

El valor esperado y la varianza de la variable geométrica, están dadas por:

$$E[X] = \frac{q}{p} \tag{51}$$

$$V(X) = \frac{q}{p^2} = \frac{E[X]}{p} \tag{52}$$

La distribución geométrica tiene sólo un parámetro p, el cual se puede expresar como una función de la media $E[X]$.

$$p = \frac{1}{1 + E[X]} \tag{53}$$

Para generar en una computadora valores de variable aleatoria con distribución geométrica se emplea la técnica de la transformación inversa y la fórmula que aparece en la ecuación 50. Al observar que el rango de la expresión $[(1 - F(x))/q]$ es unitario, resulta que:

$$r = q^x \tag{54}$$

y consecuentemente:

$$x = \frac{\log(r)}{\log(q)} \tag{55}$$

donde al valor x siempre se le redondea al entero que sea menor. Esto se puede lograr de manera muy simple con sólo retener los números hasta antes del punto decimal, o bien convirtiendo los números en modo de punto flotante al de punto fijo. Para generar un valor de variable aleatoria con distribución geométrica haciendo uso de esta técnica, se requiere únicamente el empleo de un número aleatorio uniforme, tal como lo muestra la ecuación 55.

6.2. Distribución binomial negativa

Cuando los procesos de ensayos de Bernoulli, tal como se han descrito en la sección anterior, se repiten hasta lograr que ocurran k éxitos ($k > 1$), la variable aleatoria que caracteriza al número de fallas tendrá una distribución binomial negativa. Por consiguiente, los valores de variable aleatorias con distribución binomial negativa coincide esencialmente con la suma de k valores de variable aleatoria con distribución geométrica; en este caso, k es un número entero y la distribución recibe el nombre de distribución de Pascal. En consecuencia, la distribución geométrica constituye un caso particular de la distribución de Pascal, especificada para k igual a uno.

La función de distribución de probabilidad para una distribución binomial negativa está dada por:

$$f(x) = \binom{k+x-1}{x} p^k q^x, \ \forall x = 0,\ 1,\ 2,\ \ldots \tag{56}$$

donde k es el números total de éxitos en una sucesión de $k+x$ ensayos, con x el número de fallas que ocurren antes de obtener k éxitos. El valor esperado y la varianza de X se representa con:

$$E[X] = \frac{kq}{p} \tag{57}$$

$$V(X) = \frac{kq}{p^2} \tag{58}$$

Se debe hacer notar que tanto la distribución geométrica como la binomial negativa se caracteriza por una sobredispersión, esto es, $V(X) > E[X]$.

Para una media y una varianza dadas, se pueden determinar los parámetros p y k de la siguiente manera:

$$p = \frac{E[X]}{V(X)} \tag{59}$$

$$k = \frac{E^2[X]}{V(X) - E[X]} \tag{60}$$

Sin embargo, puede suceder que el proceso de simulación se complique considerablemente cuando resulte que en la ecuación 60 el valor que se obtenga al efectuar el cómputo de k no sea entero.

Cuando k es un entero, los valores de la variable aleatoria con distribución de Pascal se pueden generar con sólo considerar la suma de k valores con distribución geométrica. En consecuencia:

$$x = \frac{\sum_{i=1}^{k} \log(r_i)}{\log(q)} = \frac{\log\left(\prod_{i=1}^{k} r_i\right)}{\log(q)} \tag{61}$$

viene a ser un valor de variable aleatoria con distribución de Pascal, una vez que su magnitud se redondea con respecto al menor entero más próximo al valor calculado.

6.3. Distribución binomial

Las variables aleatorias definidas por el número de eventos exitosos en una sucesión de n ensayos independientes de Bernoulli, para los cuales la probabilidad de éxito es p en cada ensayo, siguen una distribución binomial. Este modelo estocástico también se puede aplicar al proceso de muestreo aleatorio con reemplazo, cuando los elementos muestreados tienen sólo dos tipos de atributos (por ejemplo *si* y *no*, o respuestas como *defectuoso* o *aceptable*). El diseño de una muestra aleatoria de n elementos es análoga a n ensayos independientes de Bernoulli, en los que x es un valor binomial que está denotando al número de elementos de una muestra de tamaño n con atributos idénticos. Es ésta la analogía que sitúa la distribución binomial como uno de los modelos más importantes en las áreas del muestreo estadístico y del control de calidad.

La distribución binomial proporciona la probabilidad de que un evento o acontecimiento tenga lugar x veces en un conjunto de n ensayos, donde la probabilidad de éxito está dada por p. La función de probabilidad para la distribución binomial se puede expresar de la manera siguiente:

$$f(x) = \binom{n}{x} p^x q^{n-x} \tag{62}$$

donde x se toma como un entero definido en el intervalo finito $0,\ 1,\ 2,\ \ldots,\ n$, y al que se le asocia el valor $q = (1 - p)$.

El valor esperado y la varianza de la variable binomial X son:

$$E[X] = np \tag{63}$$

$$V(X) = npq \tag{64}$$

Cuando se conoce la media y la varianza, resulta inmediata la determinación de p y de n, las cuales pueden calcularse como sigue:

$$p = \frac{E[X] - V(X)}{E[X]} \tag{65}$$

$$n = \frac{E^2[X]}{E[X] - V(X)} \tag{66}$$

Los valores de variable aleatoria con distribución binomial se pueden generar de muy diversos modos, aunque uno de los métodos más simples, que en el caso de que el valor de n sea moderado resulta uno de los métodos más eficientes, es el basado en la reproducción de ensayos de Bernoulli, siguiendo el método de rechazos. Este método empieza con valores conocidos de p y de n y consiste en generar n números aleatorios después de fijar x_0 igual a cero. Para cada número aleatorio r_i $(1 \leq i \leq n)$ se efectúa una prueba y la variable x_i se incrementa de acuerdo con el siguiente criterio:

$$x_i = x_{i-1} + 1, \text{ si } r_i \leq p \tag{67}$$

$$x_i = x_{i-1}, \text{ si } r_i > p \tag{68}$$

Después de haberse generado n números aleatorios, el valor de x_n será igual al valor de la variable aleatoria con distribución binomial x. Este procedimiento se puede repetir tantas veces como valores binomiales se requieran.

6.4. Distribución de Poisson

Si tomamos una serie de n ensayos independientes de Bernoulli, en cada uno de los cuales se tenga una probabilidad p muy pequeña relativa a la ocurrencia de un cierto evento, a medida que n tiende al infinito, la probabilidad de x ocurrencias está dada por la distribución de Poisson:

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \;\; \forall\; x = 0,\; 1,\; 2,\; \ldots \text{ y } \lambda > 0 \tag{69}$$

siempre y cuando permitamos que p se aproxime a cero de manera que se satisfaga la relación $\lambda = np$ consistentemente. De nuestra discusión previa sabemos que np es el valor esperado de la distribución

binomial y se puede demostrar que λ es el valor esperado para la distribución de Poisson. De hecho, tanto el valor esperado como la varianza de la distribución de Poisson coinciden en el valor λ. También se puede demostrar que si x es una variable de Poisson con parámetro λ, entonces para valores muy grandes de λ ($\lambda > 10$), se puede utilizar la distribución normal con $E[X] = \lambda$ y $V(X) = \lambda$ para aproximar la distribución de x.

Los eventos que se distribuyen e forma poissoniana ocurren frecuentemente en la naturaleza; por ejemplo, el número de aeroplanos que descienden en un aeropuerto en un período de veinticuatro horas puede ser considerablemente grande. Aun así, resulta muy pequeña la probabilidad de que un avión aterrice durante un segundo determinado. Por lo tanto, podemos esperar que en un período determinado, la probabilidad de que desciendan $0,\ 1,\ 2, \ldots$ aviones, obedecerá a las leyes de la distribución de Poisson. Esta distribución es particularmente útil cuando tratamos con problemas en los que se da la ocurrencia de eventos aislados sobre un intervalo continuo de tiempo, o bien cuando resulta posible prescribir el número de veces que ocurre en evento aunque no el número de veces que no ocurre.

Para simular una distribución de Poisson con parámetro λ, nos podemos servir ventajosamente de la relación conocida entre las distribuciones exponenciales y de Poisson. Se puede justificar que si 1) el número total de eventos que ocurren durante un intervalo de tiempo dado es independiente del número de eventos que ya han ocurrido previamente al inicio del intervalo y 2) la probabilidad de que un evento ocurra en el intervalo de t a $t + \Delta t$ es aproximadamente $t\Delta t$ para todos los valores de t, entonces: a) la función de densidad del intervalo t entre las ocurrencias de eventos consecutivos es $f(t) = \lambda e^{-\lambda t}$, y b) la probabilidad de que ocurran x eventos durante el tiempo t es:

$$f(x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!} \tag{70}$$

para toda x y toda t.

En términos matemáticos el valor poissoniano x se determina haciendo uso de las siguientes desigualdad:

$$\sum_{i=0}^{x} t_i \leq \lambda < \sum_{i=0}^{x+1} t_i \quad (x = 0,\ 1,\ 2,\ \ldots) \tag{71}$$

donde los valores de la variable aleatoria t_i se generan por medio de la formula:

$$t_i = -\log(r_i) \tag{72}$$

con una media unitaria.

Referencias

[1] BARCELÓ, J. *"Simulación de sistemas discretos"*. Primera edición. Madrid: Isdefe, 1996. ISBN: 84-89338-12-4.

[2] CASELLA, G and BERGER, R. *"Statistical inference"*. Primera edición. EEUU: Duxbury, 1990. ISBN: 0-534-11958-1.

[3] CHURCHMAN, C. W. *"A analisys of the concept of simulation"*. Simposium on Simulation Model. Editado por Austin C. Hoggatt y Frederich E. Balderston. Página 12.

[4] INTERNATIONAL BUSINESS MACHINE CORPORATION. *"Random number generation and testing"*. Reference manual (C20-8011). Nueva York. 1959.

[5] LEHMER, D. H. *"Matematical methods in large-scale computing units"*. Annal Computer Laboratory Hardvard University. XXVI. 1951. Páginas: 141-146.

[6] NAYLOR, T. et al. *"Técnicas de simulación en computadoras"*. Primera edición. México: Limusa, 1993. ISBN: 968-18-0839-8.

[7] OLIVARES-PACHECO, J. F. *"Simulación de distribuciones Skew con dominio acotado"*. Trabajo de titulación para optar al titulo de Ingeniero Civil en Computación e Informática. Universidad de Atacama. Copiapó–Chile. 2005.

[8] ROSS, S. M. *"Simulación"*. Segunda edición. México: Prentice Hall, 1999. ISBN: 970-17-0259-X.

[9] WALPOLE, R. E. and MYERS, R. H. *"Probabilidad y estadística"*. Cuarta edición. México: McGraw-Hill, 1992. ISBN: 968-422-993-5.

[10] TOCHER, K. D. *"The application of automatic to sampling experiments"*. Journal of the Royal Statistical Society B16. 1954. Página 40.