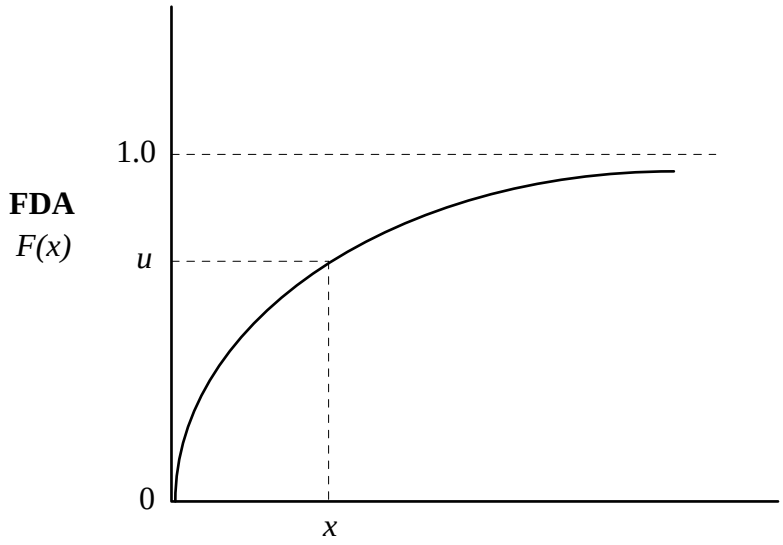
GENERACION DE VARIABLES ALEATORIAS

Hay una variedad de métodos para generar variables aleatorias. Cada método se aplica solo a un subconjunto de distribuciones y para una distribución en particular un método puede ser más eficiente que otro.

I. TRANSFORMACIÓN INVERSA

Si la variable aleatoria X tiene una FDA $F(x)$, entonces la variable $u = F(x)$ está distribuida uniformemente entre 0 y 1. Por lo tanto, X se puede obtener generando números uniformes y calculando $x = F^{-1}(u)$.



Prueba:

Sea $u = g(x)$ tal que $x = g^{-1}(u)$:

$$F_U(u) = P(U \le u) = P(X \le g^{-1}(u)) = F_X(g^{-1}(u))$$

Seleccionemos $g(\ )$ de forma que $g(x) = F_X(x)$, o $u = F_X(x)$, y que u sea una variable aleatoria entre 0 y 1 con distribución dada por

$$F_U(u) = F_X(g^{-1}(u)) = F_X(F_X^{-1}(u)) = u$$

y

$$f(u) = dF/du = 1$$

o sea que u está distribuida uniformemente entre 0 y 1.

Este método nos permite generar variables aleatorias siempre que se pueda determinar $F^{-1}(x)$ analíticamente o empíricamente.

Ejemplo (determinación analítica)**:**

Sea X exponencial con $f(x) = \lambda e^{-\lambda x}$. La FDA es $F(x) = 1 - e^{-\lambda x} = u$ o $x = -\frac{1}{\lambda}\ln(1-u)$. Si u es uniforme entre 0 y 1, entonces 1-u también está distribuida uniformemente entre 0 y 1. Por lo tanto podemos generar variables aleatorias exponenciales generando u y después calculando $x = -\frac{1}{\lambda}\ln(u)$.

Ejemplo (determinación empírica)**:**

El tamaño de los paquetes en una red fueron medidos y encontrados trimodales con las siguientes probabilidades:

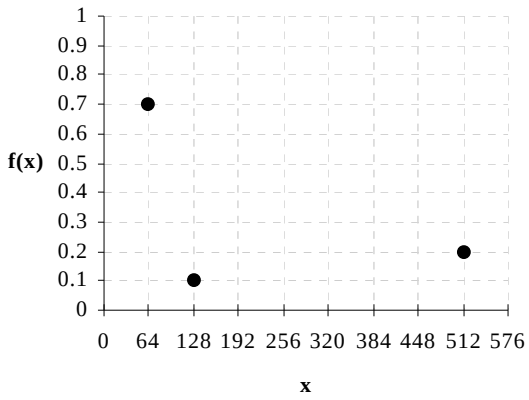
Tamaño (bytes)	Probabilidad
64	0.7
128	0.1
512	0.2

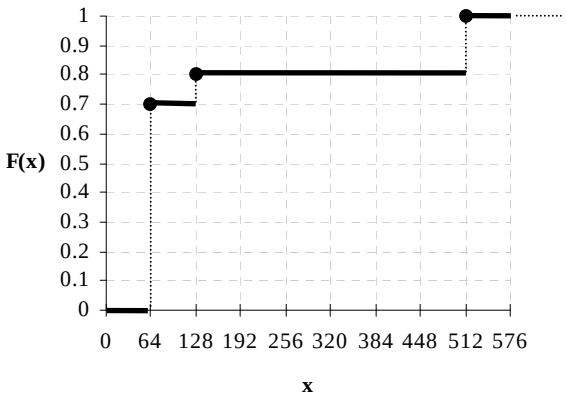
La FDA viene dada por:

$$F(x) = \begin{cases} 0.0 & 0 \le x < 64 \\ 0.7 & 64 \le x < 128 \\ 0.8 & 128 \le x < 512 \\ 1.0 & 512 \le x \end{cases}$$

y la inversa está dada por:

$$F^{-1}(u) = \begin{cases} 64 & 0 < u \le 0.7 \\ 128 & 0.7 < u \le 0.8 \\ 512 & 0.8 < u \le 1 \end{cases}$$





II. MÉTODO DEL RECHAZO

Esta técnica se puede usar si existe otra función de densidad $g(x)$ tal que $cg(x)$ *supera* la función de densidad $f(x)$, es decir, $cg(x) > f(x)$ para todos los valores de x. Si esta función existe, entonces se pueden aplicar los siguientes pasos:

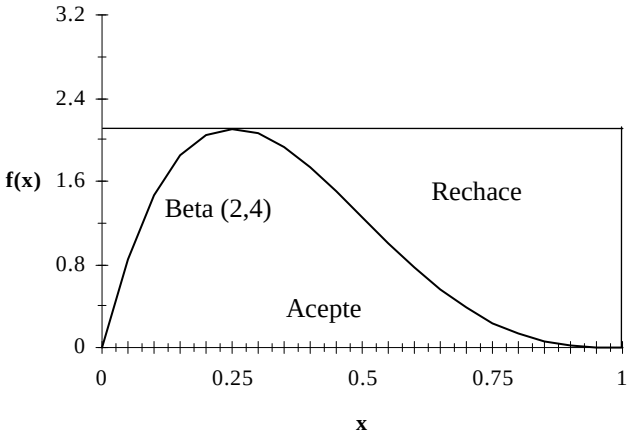
1. Genere x con la densidad $g(x)$.
2. Genere y uniforme en $[0, cg(x)]$.
3. Si $y \leq f(x)$, devuelva x y retorne. De lo contrario repita desde el paso 1.

El algoritmo permanece *rechazando* las variables x y y hasta que la condición $y \leq f(x)$ sea satisfecha.

Ejemplo:

Consideremos la función de densidad beta(2,4):

$$f(x) = 20x(1-x)^3 \qquad 0 \leq x \leq 1$$



Esta función se muestra en la figura y puede ser limitada por el rectángulo de altura 2,11. Por lo tanto podemos usar $c = 2{,}11$ y $g(x) = 1$ para $0 \leq x \leq 1$. La variables beta (2,4) pueden ser generadas como sigue:

1. Genere x uniforme en [0, 1].
2. Genere y uniforme en [0, 2,11].
3. Si $y \le 20x(1-x)^3$, devuelva x y retorne. De lo contrario vuelva al paso 1.

Los pasos 1 y 2 generan un punto (x, y) distribuido uniformemente en el rectángulo en la figura. Si el punto cae sobre la densidad $f(x)$, entonces el paso 3 rechaza x.

La eficiencia del método depende de que tan bien $g(x)$ limita a $f(x)$. Si hay una brecha muy grande entre $cg(x)$ y $f(x)$, entonces un gran número de puntos generados en los pasos 1 y 2 serán rechazados. Similarmente, si la generación de variables aleatorias con $g(x)$ es compleja, entonces el método puede ser ineficiente.

III. COMPOSICIÓN

Este método se puede usar si la FDA $F(x)$ deseada se puede expresar como una suma ponderada de otras n FDA $F_1(x)$, ..., $F_n(x)$:

$$F(x) = \sum_{i=1}^{n} p_i F_i(x) \qquad p_i \ge 0, \quad \text{y} \quad \sum_{i=1}^{n} p_i = 1$$

El número de funciones n puede ser finito o infinito, y las n FDA son compuestas para formar la FDA deseada; de aquí el nombre de la técnica. Esto también se puede ver como que la FDA deseada es descompuesta en otras n FDA; por esto la técnica a veces es llamada *descomposición*.

La técnica también se puede usar si la función de densidad $f(x)$ puede ser descompuesta como una suma ponderada de otras n densidades:

$$f(x) = \sum_{i=1}^{n} p_i f_i(x) \qquad p_i \ge 0, \quad \text{y} \quad \sum_{i=1}^{n} p_i = 1$$

En cualquier caso, los pasos a seguir so:

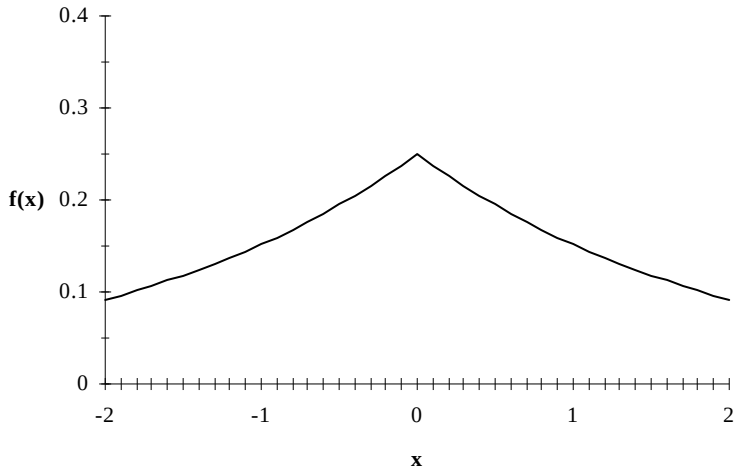
1. Genere un entero aleatorio I tal que $P(I = i) = p_i$. Esto puede ser hecho con el método de transformación inversa.
2. Genere x con la i-esima densidad $f_i(x)$ y retorne.

Ejemplo:

Consideremos la densidad de Laplace dada por

$$f(x) = \frac{1}{2a} e^{-|x|/a} \qquad -\infty < x < \infty$$

La siguiente figura muestra la densidad para $a = 2$.



Esta densidad es una composición de dos densidades exponenciales. La probabilidad de que x sea positiva es 1/2, y de que sea negativa también es 1/2. Usando la técnica de composición podemos generar variables de Laplace de la siguiente forma:

1. Genere $u_1 \sim U(0,1)$, y $u_2 \sim U(0,1)$.
2. Si $u_1 < 0.5$, retorne $x = -a \ln u_2$, de lo contrario retorne $x = a \ln u_2$.

Es de hacer notar que estas variables se pueden generar más eficientemente usando la técnica de transformación inversa.

IV. CONVOLUCIÓN

Esta técnica puede ser usada si la variable aleatoria x puede ser expresada como la suma de n variables aleatorias y_1 , ..., y_n que puedan ser generadas fácilmente:

$$x = y_1 + y_2 + ... + y_n$$

En este caso x se puede generar generando n variables aleatorias y_1 , ..., y_n y sumándolas. Si x es la suma de dos variables aleatorias y_1 y y_2 , entonces la densidad de x puede ser obtenida analíticamente por la convolución de las densidades de y_1 y y_2 ; de aquí el nombre de la técnica a pesar de que la convolución no es necesaria para la generación de números aleatorios.

Nótese la diferencia entre composición y convolución. La primera se usa cuando la densidad o FDA puede ser expresada como la suma de otras densidades o FDA. La segunda se usa cuando la variable misma puede ser expresada como la suma de otras variables.

A continuación se dan unos ejemplos de aplicación de esta técnica:

- Una variable Erlang-k es la suma de k exponenciales.
- Una variable Binomial de parámetros n y p es la suma de n variable Bernulli con probabilidad de éxito p.
- La chi-cuadrado con v grados de libertad es la suma de cuadrados de v normales $N(0,1)$.

- La suma de un gran número de variables de determinada distribución tiene una distribución normal. Este hecho es usado para generar variables normales a partir de la suma de números $U(0,1)$ adecuados.
- Una variable Pascal es la suma de m geométricas.
- La suma de dos uniformes tiene una densidad triangular.

V. CARACTERIZACIÓN

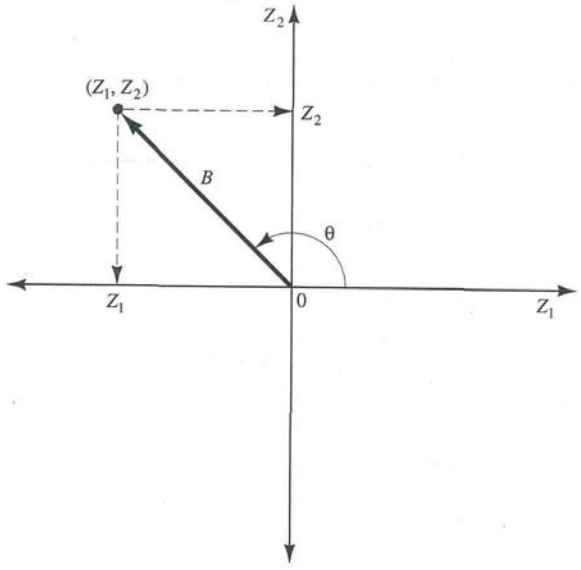
Características especiales de ciertas distribuciones permiten generar sus variables usando algoritmos especialmente ajustados para ellas. Todos estos algoritmos están clasificados bajo una técnica llamada **caracterización**.

Ejemplos de variables generadas usando caracterización son:

- Si los tiempos entre llegadas son exponenciales con media $1/\lambda$, el número de llegadas n en cierto intervalo T es Poisson con parámetro λT. Por lo tanto una Poisson puede ser obtenida generando exponenciales hasta que su suma supere T y devolviendo el número de exponenciales usadas.
- El a-esimo menor número en una secuencia de $a + b + 1$ variables $U(0,1)$ tiene distribución beta(a , b).
- La razón de dos normales estándar en Cauchy(0,1).
- Una chi-cuadrado con un número par de grados de libertad $\chi^2(v)$ es un gamma $\gamma(2,v/2)$.
- Si x_1 y x_2 son dos gammas $\gamma(a , b)$ y $\gamma(a , c)$ respectivamente, la razón $x_1 / (x_1 + x_2)$ es beta(b,c).

Distribución Normal

Considere dos distribuciones normales estándar Z_1 y Z_2 graficadas en el plano tal como se muestra en la figura:



siendo su representación en coordenadas polares la siguiente:

$$Z_1 = B\cos\theta$$
$$Z_2 = Bsen\theta$$

Es sabido que $B^2 = Z_1^2 + Z_2^2$ tiene distribución chi-cuadrado con 2 grados de libertad que es equivalente a una distribución exponencial con media 2:

$$f(x) = \frac{x^{(2-2)/2}e^{-x/2}}{2^{2/2}\Gamma(2/2)} = \frac{1}{2}\ e^{-x/2}\ \ x \geq 0$$

por lo tanto el radio *B* lo podemos generar por transformación inversa usando

$$B^2 = -2\ln(u) \quad \text{o} \quad B = \sqrt{-2\ln(u)}$$

Por simetría de la distribución normal, es razonable suponer que el ángulo θ está distribuido uniformemente entre *0* y 2π. Igualmente se puede considerar que el radio *B* y el ángulo θ son independientes. Con esto podemos generar dos variables aleatorias normales estándar independientes Z_1 y Z_2 a partir de dos números uniformes u_1 y u_2 (nótese que si *u* es uniforme entre 0 y 1, entonces $2\pi u$ es uniforme entre 0 y 2π):

$$Z_1 = \sqrt{-2\ln(u_1)}\cos(2\pi u_2)$$
$$Z_2 = \sqrt{-2\ln(u_1)}sen(2\pi u_2)$$

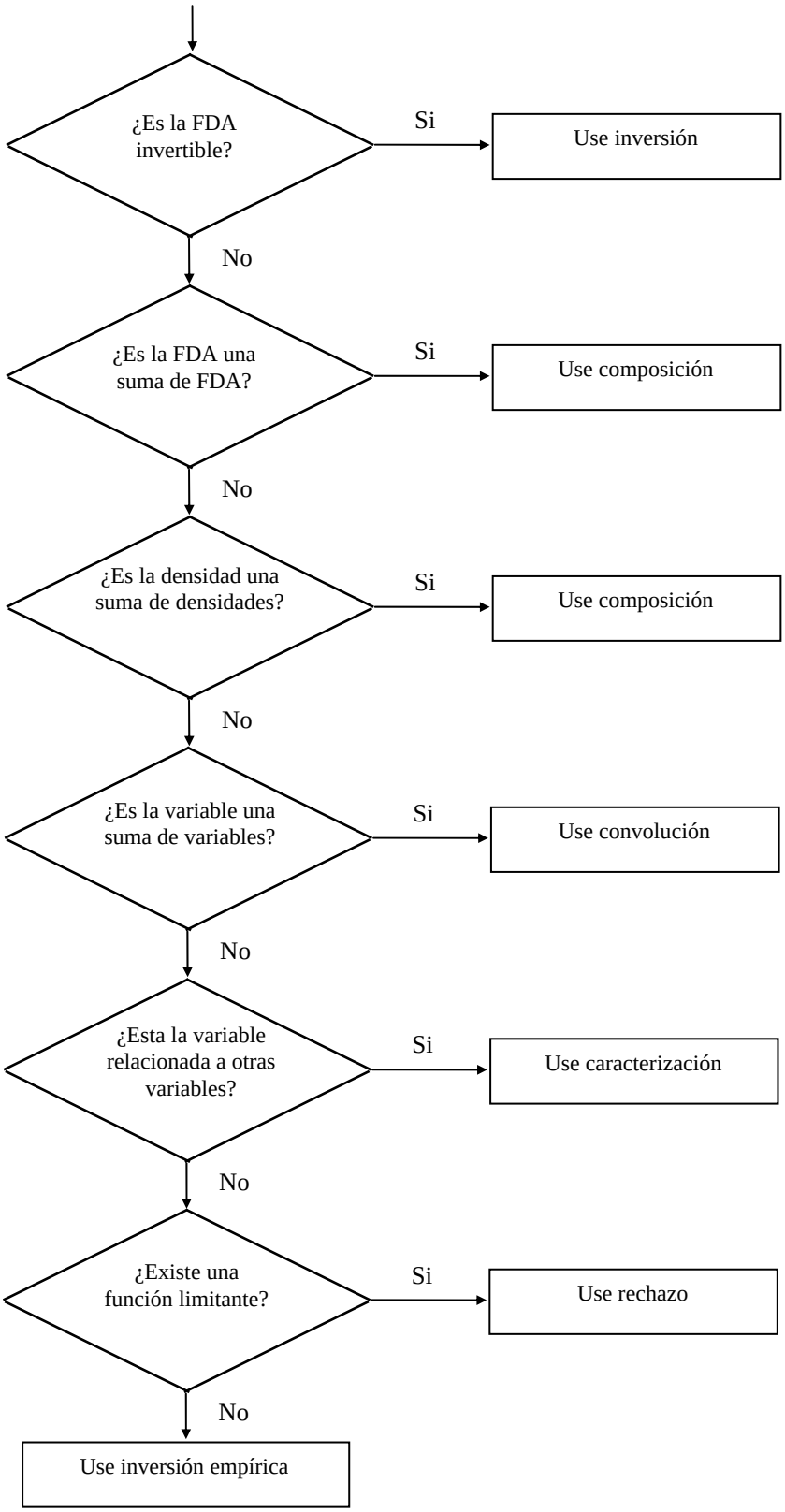
y para obtener normales con parámetros (μ_1, σ_1^2) y (μ_2, σ_2^2) usamos las transformaciones:

$$X_1 = \mu_1 + \sigma_1 Z_1$$
$$X_2 = \mu_2 + \sigma_2 Z_2$$

Por ejemplo, asumamos que queremos generar dos variables aleatorias normales independientes con parámetros (8,4) y que nuestro generador de números aleatorios nos proporciona u_1 = 0.2375 y u_2 = 0.8561. Aplicando las ecuaciones anteriores nos da:

$$Z_1 = \sqrt{-2\ln(0.2375)}\cos(2\pi 0.8561) = 1.0485$$
$$Z_2 = \sqrt{-2\ln(0.2375)}sen(2\pi 0.8561) = -1.3326$$
$$X_1 = 8 + 4(1.0485) = 12.1940$$
$$X_2 = 8 + 4(-1.3326) = 2.6696$$

A continuación se presenta un diagrama de flujo que ayuda a decidir cuál de las técnicas anteriores se debe usar:



VI. BOOTSTRAP

El bootstrap es un método estadístico creado para facilitar los cálculos que no se pueden hacer con fórmulas simples (por medio de las técnicas estadísticas clásicas) teniendo como herramienta importante la ayuda del computador.

Este método consiste básicamente en sustituir la distribución teórica por la muestral y estudiar las propiedades del estimador remuestreando de esa nueva población en las mismas condiciones en que se obtuvo la muestra original. Se basa en el muestreo de muestras.

El método bootstrap trabaja como sigue:

1. Se tiene un conjunto de muestra aleatoria (el cual se trabaja con reemplazo) de tamaño *n*, en donde $x = \{x_1, x_2, ..., x_n\}$ son los valores observados de dicha muestra.
2. Se crea una nueva muestra del mismo tamaño muestreando aleatoriamente n veces con reemplazo de la muestra original $\{x_1, x_2, ..., x_n\}$, donde la probabilidad de escoger cualquier punto de los datos es $1/n$.
3. Luego se calcula el estadístico de interés $\hat{\theta}$ para cada una de las muestras bootstrap, a partir de la remuestra obtenida, dando así $\hat{\theta}_b^*$. (donde $b = 1,2,..., B$).
4. Se repiten los puntos 2 y 3 *B* veces, donde *B* es un número grande que representa la cantidad de remuestras hechas. La magnitud de *B* en la práctica depende de las pruebas que se van aplicar a los datos. En general, *B* debería ser de entre 50 a 200 para estimar el error estándar de $\hat{\theta}$, y al menos de 1000 para estimar intervalos de confianza en un punto o alrededor de $\hat{\theta}$.

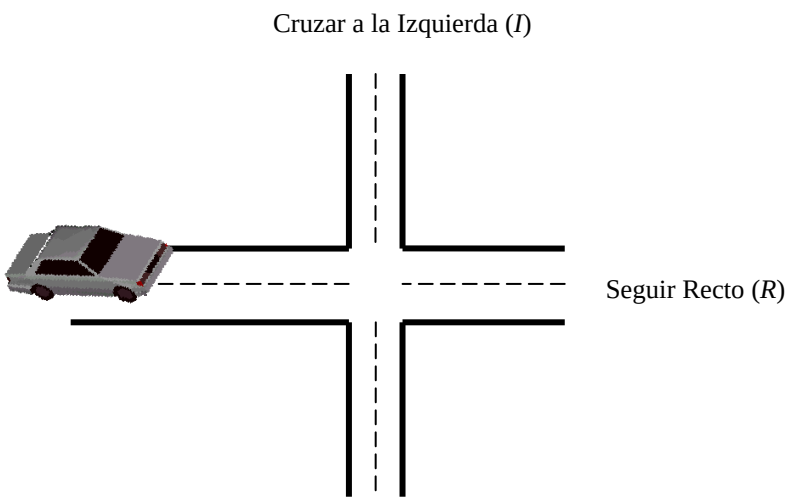
Por ejemplo, supongamos que $\hat{\theta}$ es la media o $\frac{1}{n}\sum_{i=1}^{n} x_i$, que es el estimador de la media poblacional μ. Si remuestreamos *B* veces, cada una es una muestras de tamaño *n* con reemplazo de la muestra original, obtenemos $\hat{\theta}_1^* \ldots \hat{\theta}_B^*$ que son las medias de cada demuestra. La distribución empírica de los $\hat{\theta}_b^*$ estima la distribución de $\hat{\theta}$ y a partir de esta podemos calcular, por ejemplo, intervalos de confianza para la media.

Método bootstrap para Generar Variables Aleatorias

En este caso usamos la muestra $\{x_1, x_2, ..., x_n\}$ para generar una variable aleatoria que siga la distribución de la misma. Consideremos dos caso, cuando la variable es discreta y cuando es continua.

Caso variables aleatorias discretas

Suponga un cruce en donde los carros tienen la opción de cruzar a la derecha (*D*), cruzar a la izquierda (*I*) o seguir recto (*R*), tal como se muestra en el diagrama a continuación:



Supongamos que observamos el cruce y obtenemos una muestra:

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\} = \{I, R, R, I, D, R, D, R, I, R\}$$

Analizando la muestra tenemos 3 *I*, 5 *R*, y 2 *D*. Es decir, si la muestra es representativa, la densidad aproximada de la dirección que puede tomar un carro es:

$$f(x) = \begin{cases} 0.3 & si \quad x = I \\ 0.5 & si \quad x = R \\ 0.2 & si \quad x = D \end{cases}$$

y para generar la dirección que tomará el carro por métodos clásicos podemos usar la inversa empírica:

$$F^{-1}(u) = \begin{cases} I & 0 < u \le 0.3 \\ R & 0.3 < u \le 0.8 \\ D & 0.8 < u \le 1 \end{cases} \qquad \text{con} \quad u \sim unif(0,1)$$

Ahora, si no analizamos la muestra (no determinamos *f(x)*) y simplemente generamos un valor uniforme *i* entre 1 y 10, y devolvemos x_i, la densidad del valor devuelto (x_i) sigue precisamente *f(x)*, es decir, será *I* con probabilidad 0.3, *R* con probabilidad 0.5 y *D* con probabilidad 0.2. Esta es la base del método bootstrap, y en resume el método se reduce a:

1. Genere *i* uniforme entre 1 y *n* (el tamaño de la muestra).
2. Devuelva $X = x_i$ como el valor de la varaible aleatoria requerida.

Debe ser obvio, que si la muestra es representativa, los valores de x_i generados siguen la densidad correcta.

Caso variables aleatorias continuas

Supongamos que observamos la estatura de una población y obtenemos la siguiente muestra:

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\} = \{1.78,\ 1.63,\ 1.82,\ 1.72,\ 1.60,\ 1.67,\ 1.77,\ 1.81,\ 1.73,\ 1.59\}$$

Los pasos que se describieron anteriormente se pueden utilizar cuando lidiamos con distribuciones discretas, pero para el caso de variables aleatorias continuas, este procedimiento no el más adecuado debido a que solo se podrían generar los puntos que se encuentran en la muestra original (para el ejemplo, solo habría 10 valores posibles). Si aplicamos el método anterior para generar estaturas, queda claro que jamás se podrá generar valores intermedios o más allá del mínimo y máximo de la muestra. Es decir, nunca se podrá generar, por ejemplo, un 1.64, o un 1.57, o un 1.87. Esto causa inquietud.

Para generar variables aleatorias continuas usando la técnica bootstrap se agrega un "ruido" al valor muestreado de la muestra:

$$X = x_i + ruido$$

Esto nos permite generar valores fuera de los que están explícitamente en la muestra. La ecuación seria:

$$X = x_i + hK(u)$$

donde *h* se denomina parámetro suavizador, y *K(u)* el kernel, que es la distribución del ruido aleatorio y debe ser simétrico alrededor de cero.

Algoritmo:

Entrada: una muestra aleatoria $\{x_1, x_2, ..., x_n\}$

Salida: una variable aleatoria continua que sigue la distribución de los datos muéstrales

1. Escoger el parámetro suavizador h
   Para calcular el h se plantea la siguiente fórmula: $h = \alpha(k)1.364\frac{\hat{\sigma}}{n^{1/5}}$
   donde la constante $\alpha(k)$ es 0.776 para un kernel Gaussiano (Normal), $\hat{\sigma}$ denota la desviación estándar de la muestra y n su tamaño. La formula queda descrita por: $h = 1.06\frac{\hat{\sigma}}{n^{1/5}}$
2. Generar i (entero) uniforme en [1,*n*]
3. Generar W según la distribución kernel $K(u)$
   En este caso, la elección es una de las funciones kernel más conocidas, la Normal(0,1):
   $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty$$
4. Devolver $X = x_i + hW$ o $X = x_i + 1.06\frac{\hat{\sigma}}{n^{1/5}} Normal(0,1)$

Cuando los datos no siguen una distribución conocida no es posible generarla por métodos clásicos. En este caso bootstrap se vuelve particularmente útil para generar este tipo de variables. Recuerde que boostrap no requiere analizar la muestra para descubrir que distribución se podría atribuir a los datos, usa la muestra directamente.