Acerca de la simulación de observaciones

Andrés Gutiérrez

2009

1. Método de la transformación uniforme

Al momento de la simulación estocástica de observaciones provenientes de alguna distribución de interés, la distribución uniforme es quizás la más usada y la más importante. El siguiente resultado, adaptado de Robert y Casella (1999), así lo confirma.

Resultado 1.1 (Transformación integral de probabilidad)**.** *Si U es una variable aleatoria con distribución uniforme en el intervalo $(0, 1)$, entonces la la variable aleatoria $F^{-1}(U)$ tiene distribución F.*

Aunque la función F no necesariamente es una función uno a uno, por lo menos no lo es en el caso discreto, sí se puede verificar que $F^{-1}(U)$ es única con probabilidad uno. Una definición general, que encaja en el caso continuo o discreto, de la función F inversa es la siguiente

Definición 1.1. *Para cualquier función F definida sobre $\mathbb{R}$, se define la función inversa generalizada de F como*

$$F^{-1}(u) = \inf\{x|F(x) \geq u\} \tag{1}$$

Ejemplo 1.1. Suponga que X es una variable aleatoria con distribución exponencial dada por (B.2.4). De esta forma, su función de densidad acumulativa viene dada por

$$F(x) = 1 - \exp\{-\theta x\}$$

Del anterior resultado se tiene que si u es una realización de una variable $U \sim Uniforme(0, 1)$, entonces $F^{-1}(u)$ es una realización de una variable con distribución exponencial. Como $x = F^{-1}(u)$, entonces $F(x) = u$ y despejando x, se llega a que la siguiente expresión

$$F^{-1}(u) = -\frac{\ln(1-u)}{\theta}$$

entrega una forma diáfana para la simulación de una observación con distribución exponencial. Para simular una muestra de n observaciones, simplemente se repite el anterior procedimiento n veces. En R, el código necesario para la simulación de una muestra de tamaño 1000 proveniente de una distribución exponencial con parámetro $\theta = 5$ es

```
> theta <- 5
> u <- runif(1000)
> rexpo <- log(1-u)/(-theta)
> 1/mean(rexpo)
[1] 5.076471

> hist(rexpo,breaks=100)
> lines(density(rexpo),col=2)
```

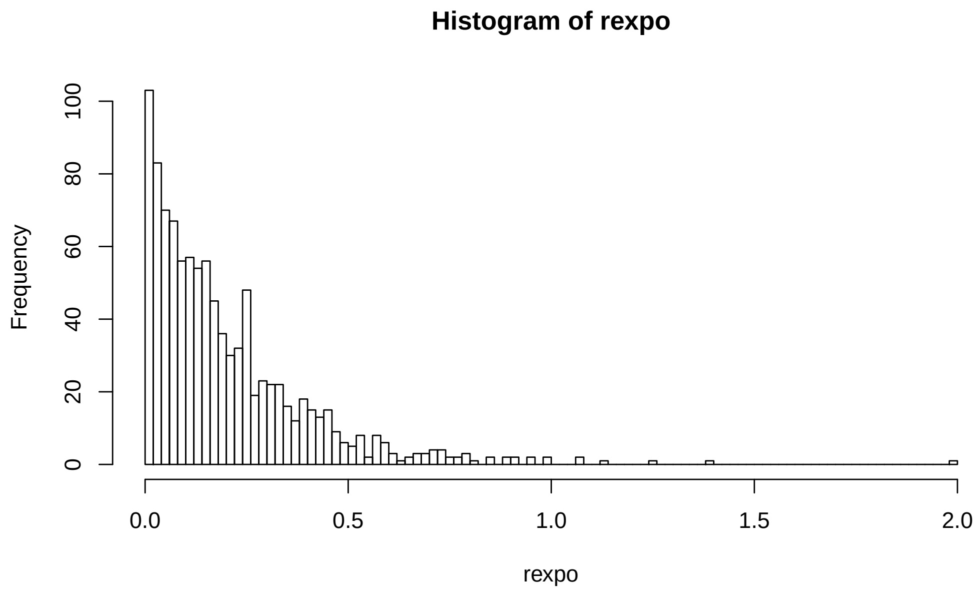

Figura 1: *Histograma de* $n = 1000$ *observaciones con distribución exponencial*

1.1. Método de la grilla

Existen distribuciones de probabilidad cuya forma estructural es muy compleja. Mas aún, existen distribuciones de probabilidad conocidas para las cuales la inversa de la función de de densidad acumulativa es difícil de solucionar analíticamente. En los anteriores casos, el método analítico dado por el teorema de la transformación integral de probabilidad no siempre resulta efectivo. Sin embargo, es posible realizar una variante, manteniendo el espíritu de la anterior técnica.

El presente método utiliza una distribución discreta para aproximar cualquier tipo de distribución (discreta o continua) sin importar su nivel de complejidad. El algoritmo que enmarca este método se da a continuación:

1. Escribir la densidad de interés como $f(\cdot)$ y establecer el rango de la variable aleatoria de interés.

2. Fijar un conjunto de n valores $x_1 < \cdots < x_n$ equiespaciados que cubran una gran parte del rango de la variable aleatoria.

3. Para x_k $(k = 1, \ldots, n)$ calcular $f(x_k)$ que equivale al valor de la densidad en el punto x_k. Nótese que si $f(\cdot)$ es una función de densidad continua, entonces $f(x_k)$ no corresponde a una probabilidad;

4. Calcular la probabilidad asociada al punto x_k definida por la aproximación discreta a $f(\cdot)$ y dada por

$$p(x_k) = \frac{f(x_k)}{\sum_{k=1}^{n} f(x_k)}$$

5. Calcular la función de densidad acumulativa aproximada definida como

$$F(x) = \begin{cases} 0, & \text{si } x < x_1 \\ \sum_{l=1}^{k} p(x_l), & \text{si } x_k \leq x < x_{k+1} \\ 1, & \text{si } x > x_n \end{cases}$$

6. Simular una observación u proveniente de una distribución uniforme continua en el intervalo $(0, 1)$.

7. Si $F(x_k) < u \leq F(x_{k+1})$, entonces $F^{-1}(u) = x_{k+1}$ y por consiguiente el valor x_{k+1} es una pseudo-observación proveniente de la densidad de interés.

Nótese que en el anterior proceso, la unidad x_{k+1} es seleccionada con probabilidad p_{k+1}; puesto que

$$\begin{aligned} P(F(x_k) < U \leq F(x_{k+1})) &= F(x_{k+1}) - F(x_k) \\ &= \sum_{l=1}^{k+1} p(x_l) - \sum_{l=1}^{k} p(x_l) = p_{k+1} \end{aligned}$$

Si se quiere extraer una muestra aleatoria de N observaciones provenientes de la distribución de interés, entonces basta con repetir el anterior proceso N veces. Por supuesto, como se trata de una muestra aleatoria cada selección se debe realizar con repetición; de esta manera no importa si $N > n$. Suponiendo que el conjunto $x_1, \ldots, x_n$ conforma una grilla de puntos lo suficientemente cercanos y que no sucede nada importante entre cada uno de ellos, entonces esta técnica debe tener un buen funcionamiento.

```
> x.grid<-seq(0,100,by=0.01)
> p.x<-expo(x.grid,5)
> rexpo<-sample(x.grid,1000,prob=p.x,replace=T)
> 1/mean(rexpo)
[1] 5.022349
> hist(rexpo,breaks=100)
> lines(density(rexpo),col=2)
```

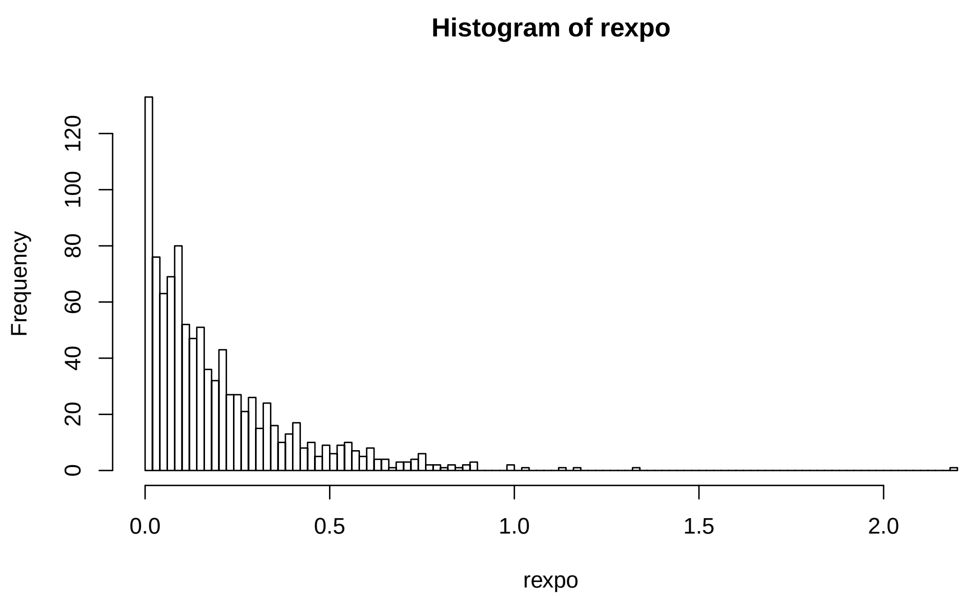

Figura 2: *Histograma de $n = 1000$ pseudo-observaciones con distribución exponencial*