2022年12月19日

“Stochastic Modelling for System Biology” Chapter 4

入江　海地（Irie Kaichi）
聖林G　京都大学経済学部2回生

目次

目次

本章のトピック

- 確率変数のシミュレーション
- 全ての動きが決定的であるコンピューターで、どのようにランダムな数（乱数）を作り出すか、についての話。
- シミュレーションするだけなら、既に便利な関数が用意されているが、その裏側にある数理に迫る。
- ※確率論・統計学の知識は最低限必要かもしれません

シミュレーションの基本は一様分布

STEP 1

(0,1)の一様分布を生成

- 一様分布とは？
  - $U(a,b)$なら、開区間(a,b)上のすべての値の出方が等しい分布。
  - 例①：時計を見た時の、コンマ以下の値$\sim U(0,1)$（連続一様分布）
  - 例②：サイコロ$\sim U\{1,2,\dots,6\}$（離散一様分布）
- どうやって生成するかが問題
  - 線形合同法

STEP 2

一様分布を応用

例えば、$U \sim U(0,1)$なら、

$$(2U+1) \sim U(1,3)$$

$$\left(\sum_{i=1}^{12} U_i - 6\right) \sim N(0,1) \quad ※近似的$$

他にももっとたくさん応用できる。

- Transformation methods
- Lookup methods
- Rejection samplers

今後の流れ

- 一様分布の生成→**線形合同法**
- 一様分布の応用
  - 連続確率変数をシミュレーションしたい場合
    - CDFの逆関数が存在する場合→ **Transformation methods**
    - CDFの逆関数が存在しない場合→**Rejection samplers**
  - 離散確率変数をシミュレーションしたい場合

    → **Lookup methods**

目次

線形合同法（Linear congruential generators）

a, b, Nをあらかじめ定めておいて、seed値$x_0 (\neq 0)$を与え、

$$x_{n+1} = \left(ax_n + b \bmod 2^N\right) \ (n \geq 0)$$

のようにして、 $x_1, x_2 \ldots, x_n$を生成すると、

$$x_1, x_2, \ldots x_n \sim U\{0,1, \ldots, 2^N - 1\}$$

とみなせる。$\{x_n\}_{n\geq 0}$は2^N以下の周期を持つが、実用上困らないほどに長い周期を持つようにa, b, Nを定めればよい。

よく用いられるのは、 $a = 13^{13}, b = 0, N = 59$

線形合同法（続き）

実際使われているのは、更に難解で効率的な手法だが、本質は同じ。

$$x_1, x_2, \ldots x_n \sim U\{0,1,\ldots,M\}$$ ：離散一様分布

が得られたとすると、

$$\frac{x_1}{M}, \frac{x_2}{M}, \ldots \frac{x_n}{M} \sim U(0,1)$$ ：連続一様分布

が近似的に成り立つから、 $U(0,1)$に従うn個の確率変数が生成できた。

目次

用語確認

連続確率変数

Xを連続確率変数とする。

①

$$F_X(x) = P(X \leq x)$$

を満たす$F_X(x)$をXの**CDF(cumulative density function：累積密度関数)**という。

②

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

を満たす$f_X(x)$をXの**PDF(probability density function：確率密度関数)**という。

離散確率変数

Xを離散確率変数とする。

①

$$F_X(k) = P(X \leq k)$$

を満たす$F_X(k)$をXの~~**CMF(cumulative density function：累積密度関数)**~~という。

正）CMF(cumulative mass function：累積質量関数)

②

$$P(X = k) = f_X(k)$$

$$\left(\Leftrightarrow P(X \leq k) = \sum_{i \in \{k, k-1, \ldots\}} f_X(i) \right)$$

を満たす$f_X(x)$をXの~~**PMF(probability density function：確率密度関数)**~~という。

正）PMF(probability mass function：確率質量関数)

目次

Transformation methods

CDFとして$F(x)$を持つXをシミュレーションしたいとする。$F(x)$が逆関数を持つとする。

命題：$U \sim U(0,1), X = F^{-1}(U)$として$X$を定めると、

$$P(X \leq x) = F(x)$$

となる。つまり、XがCDFとして$F(x)$を持つ。

証明：$F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) = P\left(U \leq F(x)\right) = F_U\left(F(x)\right) = F(x)$

ただし、$F(x)$の単調増加性と$F_U(x) = x$を用いた。

例：指数分布$\mathrm{Exp}(\lambda)$のシミュレーション

$X \sim \mathrm{Exp}(\lambda)$のCDFは、$F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}\ (x \geq 0)$なので、

$$F^{-1}(y) = -\frac{1}{\lambda}\log(1-y)\ (0 \leq y \leq 1)$$

$$U \sim U(0,1), X = -\frac{1}{\lambda}\log(1-U)$$

としてXを定めると、

$$X \sim \mathrm{Exp}(\lambda)$$

※$U \sim (1-U)$なので、$X = -\frac{1}{\lambda}\log U$でもよい

Transformation methodsのその他の例

$U \sim U(0,1)$とする。

- 一様分布：$V = (b-a)U + b\ (a < b)$とすれば、$V \sim U(a,b)$
- 指数分布：$X = -\frac{1}{\lambda}\log(U)$とすれば、$X \sim \mathrm{Exp}(\lambda)$
- ガンマ分布：$\forall i \in \{1, 2, ..., n\}$: $Y_i \sim \mathrm{Exp}(\lambda)$ならば$\sum_{i=i}^{n} Y_i \sim \mathrm{Ga}(n, \lambda)$
- ポアソン分布：時間$[0,T]$で平均λ回の割合で起こる事象について、時間$[0,T]$で事象がX回起こるとすると、$X \sim \mathrm{Po}(\lambda)$である。このとき、$(i-1)$回目に事象が起こってから、もう一度事象が起こるまでの時間を$t_i (i \geq 1)$とすると、$t_i \sim \mathrm{Exp}(\lambda)$であることを利用する。$X_k = X_{k-1} + t_k (k \geq 0), X_0 = 0$として、初めて$X_{k+1} > T > X_k$を満たす$k$を$X$とおくと、$X \sim \mathrm{Po}(\lambda)$である。
- 正規分布：$\Theta \sim U(0, 2\pi), R^2 \sim Exp(1/2)$として、$X = R\cos\Theta, Y = R\sin\Theta$とすれば、$X \sim N(0,1), Y \sim N(0,1)$ で、X, Yは独立。

Rejection samplers

CDFを計算できない、一般の連続分布に従う確率変数のシミュレーション。
PDFはわかっているとする。

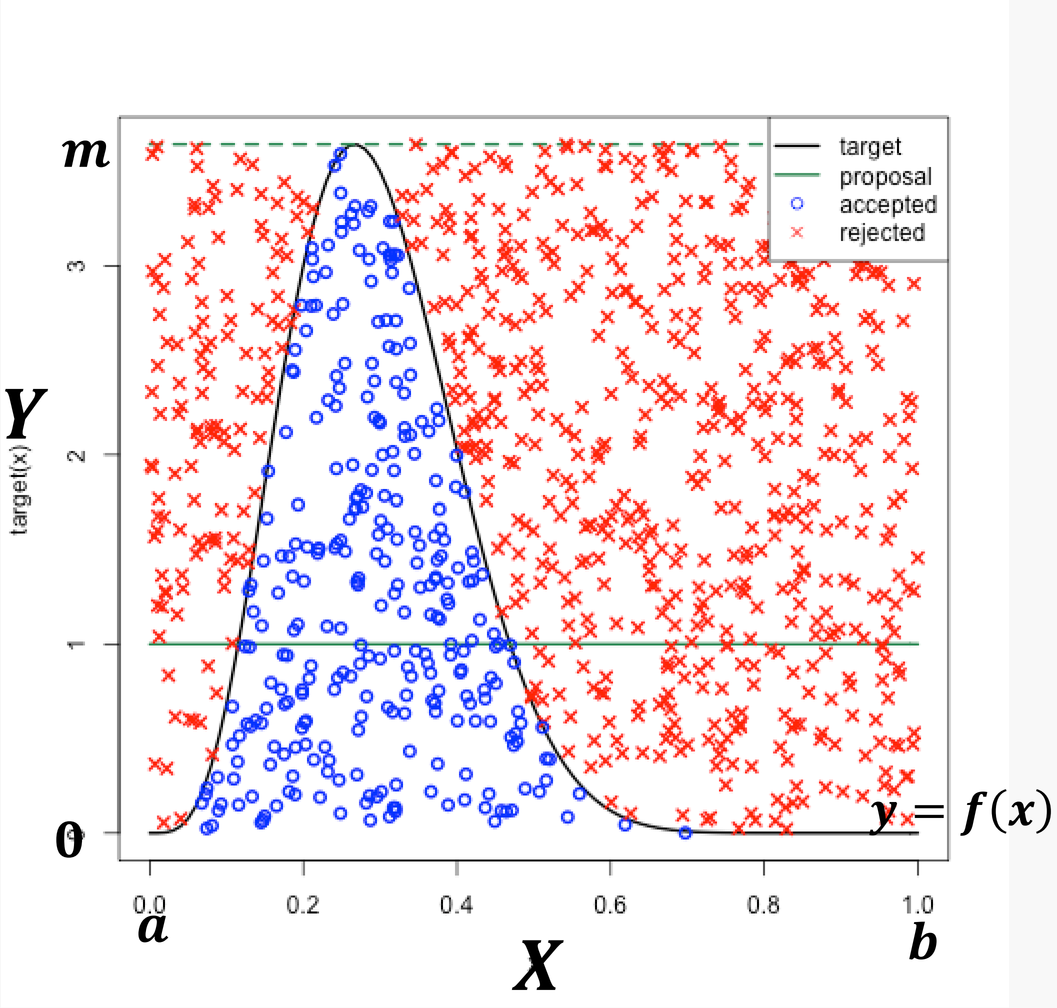
命題：

$f(x)$をPDFに持つXを区間$[a,b]$に限ってシミュレーションしたいとする。また、$\forall x \in [a,b]$: $f(x) \leq m$とする。このとき、$X \sim U(a,b), Y \sim U(0,m)$として$X,Y$を生成し、そのうち$Y < f(X)$を満たす$X$を$X_{\text{Accept}}$とすると、$X_{\text{Accept}}$はPDF$f(x)$を持つ。

注意点：$P(Accept) = \frac{1}{m(b-a)}$であり、もし$P(Accept)$がとても低ければ、生成する$X,Y$のほとんどが使えないので、とても効率が悪い。それをある程度解決した、envelop methodsというのもある。

Rejection samplersの定性的解釈

- $X \sim U(a,b), Y \sim U(0,m)$なる、$(X,Y)$をとるということは、領域$\{(x,y) | a \le x \le b, 0 \le y \le m\}$にランダムに一つ点を打つことに対応している。
- 領域$\{(x,y) | a \le x \le b, 0 \le y \le f(x)\}$に点が入れば、Acceptする。
- 左の図で言うと、$x = 0.3$近傍では、たくさんの点がAcceptされるので、X_{Accept}の値は0.3近傍の値に集中することになる。結果X_{Accept}は、PDF として$f(x)$を持つ。



（jarad.meより）

Lookup methods

任意の離散確率分布を持つ（$P(X=k)=p_k$なる）Xをシミュレーションする方法。 ※わかりやすさのために、$k \geq 0$とする。

任意の離散確率分布を持つ（$P(X=k)=p_k$なる）Xをシミュレーションする方法。

命題：

$U \sim U(0,1)$とする。$q_k := \sum_{i=0}^{k} p_i$,$X = \min\{k|q_k \geq U\}$とすれば、

$P(X=k)=p_k,\ q_k = P(X \leq k)$

証明：

$P(X=k) = P(q_{k-1} < U \leq q_k) = q_k - q_{k-1} = p_k$

Lookup methodsの解説

右図のような分布に従う離散確率変数をシミュレーションしたいとする。つまり、

$p_0 = 0.2, p_2 = 0.1, p_5 = 0.3, p_7 = 0.4$

p_k はPMF、q_k はCMFを表す。

$$\min\{k|q_k \geq U\}$$

を具体的に追っていけば、何をしているのかがわかる。

実はとても単純。

任意の離散確率分布を持つ（$P(X=k)=p_k$なる）Xをシミュレーションする方法。

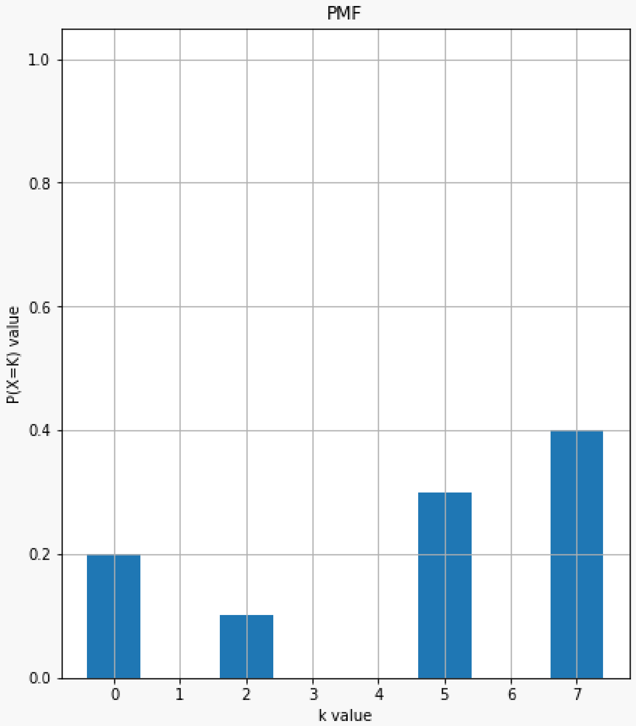
命題：

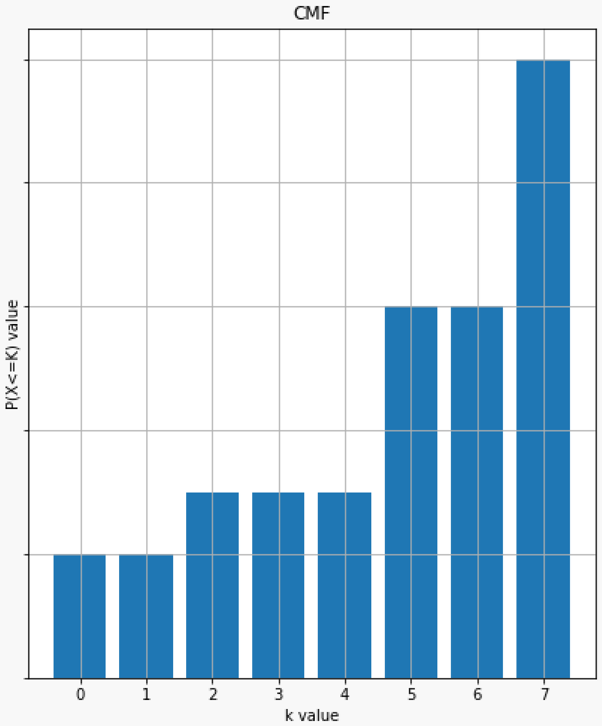
$U \sim U(0,1)$とする。$q_k := \sum_{i=0}^{k} p_i$,$X = \min\{k|q_k \geq U\}$とすれば、

$P(X=k) = p_k$，$q_k = P(X \leq k)$

証明：

$P(X=k) = P(q_{k-1} < U \leq q_k) = q_k - q_{k-1} = p_k$





目次

- **確率変数のシミュレーションの応用として、モンテカルロ法がある。**
- **これは積分値を数値的に推定する手法**

$X \overset{d}{=} f(x)dx$なるXをシミュレーションし、$\{x_1, ..., x_n\}$が得られたとすれば、大数の法則より

$$E(g(X)) = \int_X g(x)f(x)dx \approx \frac{1}{n}\sum_i g(x_i)$$

更に、$X \overset{d}{=} f(x)dx$なるXをシミュレーションできずとも、$Y \overset{d}{=} h(y)dy$なるY（ただし、$h(y)$はある条件を満たす必要あり。）をシミュレーションでき、$\{y_1, ..., y_n\}$が得られたとすれば、

$$E(g(X)) = \int_X g(x)f(x)dx = \int_Y \frac{g(y)f(y)}{h(y)}dy \approx \frac{1}{n}\sum_i \frac{g(y_i)f(y_i)}{h(y_i)}$$

参考文献

- **Darren J. Wilkinson, “Stochastic Modelling for System Biology”. CRC Press, 2011年.**
- **jarad.me. “Rejection sampling”. https://www.jarad.me/teaching/2013/10/03/rejection-sampling , (2022年12月09日)**