DOI:10.13860/j.cnki.sltj.2006.06.021
2006年 11月 数理统计与管理 Nov., 2006
第25卷 第6期 Application of Statistics and Management Vol.25 No.6

文章编号:1002-1566(2006)06-0750-07

非均匀随机数产生

杨振海 程维虎

(北京工业大学应用数理学院,北京,100022)

摘要:本文详细介绍了产生非均匀随机数的一般方法,常用连续分布(正态分布,Gamma分布,Beta分布,χ^2分布和F分布)的抽样法及利用R软件产生随机数的方法。

关键词:随机数生成;离散分布;连续分布;R软件

中图分类号:O212 文献标识码:A

The Common Method of Generating Random Number for the Non-uniform Distribution

YANG Zhen-hai, CHENG Wei-hu

(College of Applied Sciences, Beijing University of Technology, Beijing, 100022, China)

Abstract: In the section, we summarize the common method of generating random number for the non-uniform distribution and some often used continuous univariate distributions (Normal distribution, Gamma distribution, Beta distribution, Chi-square distribution and F distribution. And then, introduce the method of generating arndom number by R software.

Key words: generating random number; discrete distributions; continuous univariate distribution; R software.

8 非均匀随机数产生

前面我们介绍过一些产生非均匀随机数的方法,但还不够全面。本文将全面、系统地介绍产生非均匀随机数的一般方法;给出常用连续分布:正态分布,Gamma分布,Beta分布,χ^2分布及F分布随机数产生的常用方法。

8.1 产生随机数的一般方法

产生随机数的一般方法有:直接抽样法,变换抽样法,舍选抽样法,复合抽样法(合成法)及近似抽样法等。其中直接抽样法,舍选抽样法与复合抽样法等我们均做过介绍[1,2,3]。

8.1.1 直接抽样法

直接抽样法包括:连续分布直接抽样法(又称反函数法)与离散分布直接抽样法两部分。前者在文献[1]中有详细介绍。下面介绍后者:

设X是离散型随机变量,所有可能取的值为$a_1, a_2, \cdots$,概率分布为

$$P\{X=a_i\}=p_i,\ i=1,2,\cdots. \tag{8.1}$$

收稿日期:2005年2月
修改日期:2006年2月
基金项目:北京市自然科学基金资助,项目编号:1062001。

令 $p^{(0)}=0$, $p^{(i)}=\sum_{j=1}^{i} p_j$, $i=1,2,\cdots$。用$\{p^{(i)}\}$作区间$[0,1)$的分位点。若 $R\sim U(0,1)$,当且仅当 $p^{(i-1)}<R\leqslant p^{(i)}$时,取 $X=a_i$.则

$$P\{p^{(i-1)}<R\leqslant p^{(i)}\}=P\{X=a_i\}=p_i,\ i=1,2,\cdots.$$

故,产生概率分布由(8.1)式给出的离散分布的随机数 $x_1, x_2, \cdots, x_n$ 的方法是:

(1) 产生均匀随机数 $r_1, r_2, \cdots, r_n$;

(2) 对 $i=1,2,\cdots,n$,若存在 $j\in\{1,2,\cdots\}$,使得 $p^{(j-1)}<r_i\leqslant p^{(j)}$,令 $x_i=a_j$。

8.1.2 变换抽样法

变换抽样法包括:一维变换抽样法与多维变换抽样法(包括值序抽样法)。

1. 关于一维变换抽样法有如下结果:

设 $r_1, r_2, \cdots, r_n$ 为均匀随机数,若 $X=g(R)$为常见分布,则 X 的随机数 $x_1, x_2, \cdots, x_n$ 可由公式 $x_i=G(r_i)$, $i=1,2,\cdots,n$ 直接给出(实例见表 8.1)。

表 8.1 均匀分布 R 常见变换的随机数生成

变换公式 $x_i=G(r_i)$	X 的概率密度 $f(x)$	X 分布类
$x_i=ar_i+b(a>0)$	$1/a$, $x\in[b, a+b)$	均匀分布 $U[b, b+a]$
$x_i=r_i^k(k>1$ 为整数)	$k^{-1}x^{1/k-1}$, $x\in[0,1]$	Beta 分布 $BE(1/k, 1)$
$x_i=r_i^{1/k}(k>1$ 为整数)	kx^{k-1}, $x\in[0,1]$	Beta 分布 $BE(k, 1)$
$x_i=-\lambda^{-1}\ln r_i(\lambda>0)$	$\lambda e^{-\lambda x}$, $x>0$	指数分布 $E(0,\lambda)$
$x_i=\tan[(r_i-0.5)\pi]$	$\dfrac{1}{\pi(1+x^2)}$	标准 Cauchy 分布
$x_i=\ln[r_i/(1-r_i)]$	$\dfrac{e^{-x}}{(1+e^{-x})^2}$	标准 Logistic 分布
$x_i=\sqrt{-2\ln r_i}$	$f(x)=x\exp(-x^2/2)$, $x\geqslant 0$	标准 Rayleigh 分布
$x_i=(-a\ln r_i)^{1/m}(a>0, m>0)$	$\dfrac{m}{a}x^{m-1}e^{-x^m/a}$, $x>0$	Weibull 分布 $W(m, a)$

2. 关于二维变换抽样法有如下结果:

设 $r_1, r_2, \cdots, r_{2n}$为均匀随机数,对 $i=1,2,\cdots,n$,

若取 $x_{2i-1}=\sqrt{-2\ln r_{2i-1}}\cos(2\pi r_{2i})$, $x_{2i}=\sqrt{-2\ln r_{2i-1}}\sin(2\pi r_{2i})$,,则 $x_1, x_2, \cdots, x_{2n}$为标准正态分布 $N(0,1)$随机数;

若取 $x_i=r_{2i-1}+r_{2i}$,则 $x_1, x_2, \cdots, x_n$ 为三角分布随机数,三角分布的概率密度函数为

$$f(x)=\begin{cases} x, & 0<x\leqslant 1,\\ 2-x, & 1<x\leqslant 2,\\ 0, & \text{其他}. \end{cases}$$

3. 关于多维变换抽样法有如下结果:

设 $r_1, r_2, \cdots, r_{mn}$为均匀随机数 $m\geqslant 2$ 为整数,对 $i=1,2,\cdots,n$,

若取 $x_i=-\sum_{j=(i-1)m+1}^{im}\ln r_j$,则 $x_1, x_2, \cdots, x_n$ 为 Gamma 分布 $G(m,1)$的随机数;

若取 $x_i=\#\{r_j\geqslant p, j=(i-1)m+1, (i-1)m+2, \cdots, mi\}$,其中 $0<p<1$ 为常数,则 $x_1, x_2, \cdots, x_n$ 为二项分布 $B(m, p)$的随机数;

若取 x_i 为 $r_{(i-1)m+1}$, $r_{(i-1)m+2}$, …, r_{im} 的第 k 个次序样本, $1\leqslant k\leqslant m$, 则 x_1, x_2, …, x_n 为 Beta 分布 $BE(k, m+1-k)$ 的随机数。

8.1.3　舍选抽样法(详见文献[1, 2])

8.1.4　复合抽样法

复合抽样法是 1961 年 Marsaglia 提出的。其方法是: 首先将欲抽取分布的分布函数 $F(x)$ 分解成若干容易生成分布分布函数的加权和, $F(x)=\sum_j p_j F_j(x)$, 其中 $p_j>0$, $\sum_j p_j=1$。

若 X 是连续型随机变量, 有概率密度函数 $f(x)$, 则将 $f(x)$ 分解成 $f(x)=\sum_j p_j f_j(x)$, 其中 $f_j(x)$ 是容易生成分布的概率密度函数。

$F(x)$(或 $f(x)$)随机数生成过程如下:

(1) 产生一个正的随机整数 J, 使得: $P\{J=j\}=p_j$, $j=1, 2, \cdots$;

(2) 产生分布函数为 $F_J(x)$ 随机数 X。

重复(1)和(2), 可产生分布函数为 $F(x)$ 的随机数列。

例 8.1(混合正态分布抽样)随机变量 X 有概率密度函数 $f(x)=\alpha f_1(x)+(1-\alpha)f_2(x)$, 其中 $\alpha\in(0, 1)$ 为常数; $f_j(x)$ 为正态分布 $N(\mu_j, \sigma_j^2)$ 的概率密度函数, $j=1, 2$, 。试产生随机变量 X 的随机数 x_1, x_2, …, x_n。

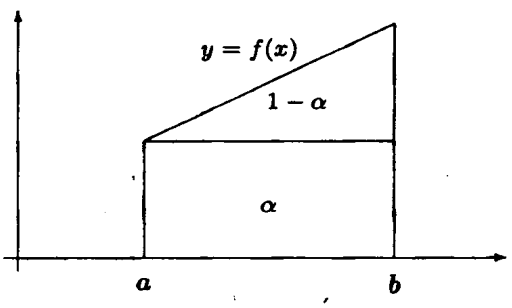
解　对 $i=1, 2, \cdots, n$:

(1) 产生独立的均匀随机数 U_i, V_i, $W_i\sim U(0, 1)$;

(2) 若 $U_i\leqslant\alpha$, 取 $x_i=\mu_1+\sigma_1\sqrt{-2\ln V_i}\cos(2\pi W_i)$; 否则, 取 $x_i=\mu_2+\sigma_2\sqrt{-2\ln V_i}\cos(2\pi W_i)$。

例 8.2(一般梯形分布抽样)若随机变量 X 以 $[a, b]$ 区间上直线 $y=f(x)$ 为概率密度

设下边矩形面积为 α, 上边三角形面积为 $1-\alpha$, 则矩形高度 $\alpha/(b-a)$, 三角形高度为 $2(1-\alpha)/(b-a)$。



当 $f(a)<f(b)$ 时(如上图), $f(x)=\alpha f_1(x)+(1-\alpha)f_2(x)$;

当 $f(a)\geqslant f(b)$ 时, $f(x)=\alpha f_1(x)+(1-\alpha)f_3(x)$。

其中

$$f_1(x)=\frac{1}{b-a}I_{[a, b]}(x),$$

$$f_2(x)=\frac{2(x-a)}{(b-a)^2}I_{[a, b]}(x),$$

$$f_3(x)=\frac{2(b-x)}{(b-a)^2}I_{[a, b]}(x).$$

易见: $f_1(x)$ 为 $[a, b]$ 区间上均匀分布, $f_2(x)$ 为 $[a, b]$ 区间上右三角形分布, $f_3(x)$ 为 $[a, b]$ 区间上左三角形分布。

可以证明: 若 R_1, R_2 为独立的均匀随机数, 则

$$\zeta=a+(b-a)\max\{R_1, R_2\}\sim f_2(x),\ \eta=a+(b-a)\min\{R_1, R_2\}\sim f_3(x).$$

8.1.5　近似抽样法

近拟抽样法的思想是: 为得到分布函数为 $F(x)$的随机数, 产生与 $F(x)$很接近的分布函数 $G(x)$的随机数, 但有一定的系统误差, 该方法可以使用的前提条件是: $G(x)$的随机数容易产生, 且适当控制系统误差后可使系统误差与模拟的随机误差相比可以被忽略。近似抽样法主要包括: 利用中心极限定理抽样法; 分段线性密度近似抽样法(已在文献[3] 中详细介绍); 经验分布抽样法和频数观测数据抽样法等。下面介绍经验分布抽样法和频数观测数据抽样法。

经验分布抽样法

在实际问题中, 数据的分布函数 $F(x)$往往未知, 此时, 利用观测数据 $x_1, x_2, \cdots, x_n$ 可建立观测数据(样本)的经验分布函数 $F_n(x)$, 利用 n 充分大时"$F_n(x) \approx F(x)$"的事实, 产生 $F_n(x)$的随机数, 并将其近似地看成 $F(x)$的随机数。产生随机数过程如下:

(1) 产生 $r \sim U(0, 1)$, 记 $Y=(n-1)r$, $I=[Y]+1$;

(2) 取 $X=x_{(I)}+(I-Y)(x_{(I+1)}-x_{(I)})$。

则 X 近似为$F(x)$的随机数。

注 1: $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$为观测数据 $x_1, x_2, \cdots, x_n$ 的次序样本;

注 2: 这样产生随机数方法简单, 但产生随机数的取值范围是$[x_{(1)}, x_{(n)}]$。

频数观测数据抽样法

若仅知 n 个观测数据 $x_1, x_2, \cdots, x_n$ 在 m 个连续小区间$[a_0, a_1), [a_1, a_2), \cdots, [a_{m-1}, a_m)$内观测频数分别为 $n_1, n_2, \cdots, n_m$, $n_1+n_2+\cdots+n_m=n$。利用这些观测数据, 也可建立观测数据的经验分布函数 $F_n(x)$, 产生 $F_n(x)$的随机数, 并将其近似地看成 $F(x)$的随机数。产生随机数过程如下:

(1) 产生 $r \sim U(0, 1)$, 若存在 k 使得$F_n(a_{k-1}) < r \leqslant F_n(a_k)$, 则记 $J=k$;

(2) 取 $X=a_{k-1}+\dfrac{r-F_n(a_{k-1})}{F_n(a_k)-F_n(a_{k-1})}(a_k-a_{k-1})$,

则 X 近似地为$F(x)$的随机数。

8.2 常用连续分布抽样法

指数分布抽样法已在文献[1] 中详细介绍, 而 Weibull 分布又可用变换抽样法得到(见本文 §8.1)。故本文只介绍其他常用连续分布抽样法。

8.2.1 正态分布抽样法

基于中心极限定理的近拟抽样法

设 $r_1, r_2, \cdots, r_{12}$为独立的均匀随机数, 则 $X=\sum_{i=1}^{6}(r_{2i}-r_{2i-1}) \sim N(0, 1)$;

Box-Muller(1958)变换抽样法

设 r_1, r_2 为独立的均匀随机数, 则 $X_1=\sqrt{-2\ln r_1}\cos(2\pi r_2)$, $X_2=\sqrt{-2\ln r_1}\sin(2\pi r_2)$为独立的 $N(0, 1)$随机数;

Hasting 近似直接抽样法

利用 $R \sim U(0, 1)$时, $\Phi^{-1}(R) \sim N(0, 1)$, 及 Hasting 有理逼近法得到标准正态分布分位点的近似计算式:

(1) 产生 $r \sim U(0, 1)$;

(2) 若 $r \leqslant 0.5$, 取 $a=r$; 否则, 取 $a=1-r$;

(3) 计算 $y=\sqrt{-2\ln a}$, 令 $X=sign(r-0.5)\left(y-\dfrac{c_0+c_1y+c_2y^2}{1+d_1y+d_2y^2+d_3y^3}\right)$,

其中

$$c_0=2.515517,\ c_1=0.802853,\ c_2=0.010328,$$

$$d_0=1.432788,\ d_2=0.189269,\ d_3=0.001308,$$

则 X 为 $N(0,1)$随机数。

8.2.2 Gamma 分布

a 为整数

利用 $\Gamma(n,1)$分布为 Erlang 分布,即 $X_1, X_2, \cdots, X_n$ 为独立的标准指数分布时,$X=X_1+X_2+\cdots+X_n \sim \Gamma(n,1)$的结果;

a>1 且 a 不为整数

•Naylor(1966)近似复合抽样法

(1) 以概率 $p_1=[a]+1-a$ 产生 $\Gamma([a],1)$;

(2) 以概率 $p_2=a-[a]$ 产生 $\Gamma([a]+1,1)$。则 $X \sim \Gamma(a,1)$。

•Fishman(1976)舍选变换抽样法

(1) 产生独立的 U, $V \sim U(0,1)$, 令 $Y=-\ln V$;

(2) 若 $U \leqslant (Y/e^{Y+1})^{a-1}$, 取 $X=aY$;否则,转(1)。则 $X \sim \Gamma(a,1)$。

0<a<1

记 $c=1+a/e$, Ahrens-Dieter(1974)给出了舍选法:

(1) 产生独立的 R, $U \sim U(0,1)$, 令 $Y=cR$;

(2) 当 $Y \leqslant 1$ 时,令 $X=Y^{a-1}$,若 $U>e^{-X}$, 转(1);。

(3) 令 $X=\ln[a/(c-Y)]$, 若 $U>X^{a-1}$, 转(1);

(4) 输出 X。

则 $X \sim \Gamma(a,1)$。

对 $b \neq 1$ 情形

利用 $X \sim \Gamma(a,1)$时, $Y=X/b \sim \Gamma(a,b)$的结果,不难导出 $\Gamma(a,b)$随机数的产生过程。

8.2.3 Beta 分布

利用 Beta 分布与 Gamma 分布关系

若 $X_1 \sim \Gamma(a,1)$, $X_2 \sim \Gamma(b,1)$, 且二者独立,则 $X=X_1/(X_1+X_2) \sim BE(a,b)$;

a=1 或 b=1 时用直接抽样法

产生 $r \sim U(0,1)$, 令 $X=\begin{cases} r^{1/a}, & b=1, \\ 1-r^{1/b}, & a=1, \end{cases}$ $X \sim BE(a,b)$;

a, b 均为整数时用值序抽样法

产生 $n=a+b-1$ 个独立的均匀随机数 $r_1, r_2, \cdots, r_n$, 并排成次序统计量 $r_{(1)}, r_{(2)}, \cdots, r_{(n)}$, 则 $X=r_{(a)} \sim BE(a,b)$;

其他情形

舍选法(参见文献[3])。

8.2.4 χ^2 分布抽样法

由正态分布产生

利用 χ^2_n 分布与标准正态分布关系,得如下抽样法:

(1) 产生独立的 $X_1, X_2, \cdots, X_n \sim N(0,1)$;

(2) 令 $X=X_1^2+X_2^2+\cdots+X_n^2$。则 $X\sim\chi_n^2$。

由Gamma分布产生

利用 χ_n^2 分布与 $\Gamma(n/2,1)$分布关系:若 $Y\sim\Gamma(n/2,1)$,则 $X=2Y\sim\chi_n^2$,特别当 n 为偶数时,抽样法更为简单。具体算法如下:

(1) 产生独立的 $r_1,r_2,\cdots,r_k\sim U(0,1)$,其中

$$k=\begin{cases} n/2, & n\text{ 为偶数},\\ (n-1)/2, & n\text{ 为奇数};\end{cases}$$

(2) 计算 $Y=-\ln(r_1,r_2,\cdots,r_k)$;

(3) 当 n 为偶数时,令 $X=2Y$;否则,产生 $Z\sim N(0,1)$与 Y 独立,令 $X=2Y+Z^2$。

则 $X\sim\chi_n^2$。

8.2.5 $F_{m,n}$分布抽样法

$F_{m,n}$分布

由 χ^2 分布产生

(1) 产生独立的 $Y_1\sim\chi_m^2$, $Y_2\sim\chi_n^2$,且二者独立;

(2) 令 $X=\dfrac{Y_1/m}{Y_2/n}$。则 $X\sim F_{m,n}$。

由Beta分布产生

利用 F 分布与Beta分布关系;若 $Y\sim BE(m/2,n/2)$,则 $X=\dfrac{nY}{m(1-Y)}\sim F_{m,n}$,得如下算法:

(1) $Y\sim BE(m/2,n/2)$;

(2) 则 $X=\dfrac{nY}{m(1-Y)}$。

则 $X\sim F_{m,n}$。

表8.2 R软件包中概率分布

分布	R中的名称	附加参数
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df1, ncp
gamma	gamma	shape, scale
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student's t	t	df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale

8.3 利用R软件产生随机数法

利用R软件,可方便地求各种常见概率分布的分布函数,分位点及生成各种常见分布的

随机数等。表 8.2 列出了 R 软件包中各种常用的概率分布的名称及参数名称。

在各种分布名称中加上不同的前缀表示不同的意义如：p－求分布函数，q－求分位点，r－产生随机数等。

例如：自由度等于 13，自变量取值为－2.43 的 t 分布的分布函数在 R 中的表示为：

＞pt(－2.43，df＝13)

自由度为(2，7)的 F 分布 $F_{2,7}$ 的左侧(也称下侧)99%分位点在 R 中的表示为：

＞qf(0.99，2，7)

产生 100 个期望值为 0，标准差为 2 的正态分布的随机数在 R 中的表示为：

＞rnorm(100，mean＝0，sd＝2)

[参考文献]

[1] 杨振海，程维虎. 统计模拟[J]. 数理统计与管理，2006，25(1)：117－126.

[2] 杨振海，张国志. 随机数生成[J]. 数理统计与管理，2006，25(2)：244－252.

[3] 程维虎，杨振海. 舍选法几何解释及曲边梯形概率密度随机数生产算法[J]. 数理统计与管理，2006，25(4).

[4] 高惠璇，统计计算[M]. 北京：北京大学出版社，1995.

[5] Von Neumann，J. Various technique used in connection with random digits[J]. U.S. Nat. Bur. Stand. Apll. Math. Ser.，1951，12：36－38.

[6] Fang K. T.，Yang Z. H. and Kotz，S.. Generation of Maltivariate Distribution by Vertical Desity Representation [J]. Statics，2001，35：281－293.

[7] Troutt，M. D. A Theorem on the Density of the Density Ordinate and an Alternative Interpretation of the Box-Xuller method[J]. Statistics，1991，22(3)：463－466.

[8] Troutt，M. D.. Vertical density Representation and a Further Remark on the Box-Muller Method[J]. Statistics，1993，24：81－83.

[9] Cheng，R. C. H.. Generating Beta Variates with Nonintegral Shape Parameters[J]. Management Science/Operation Research，1978，21(4)：317－322.

[10] Pang，W. K.，Yang，Z. H.，Hou，S. H. and Troutt，M. D.. Non-uniform random Variate generation by vertical strip method with giveb density[J]. Euroupean Joural of Operation research，1978，35：463－477.