doi:10.3969/j.issn.1672 - 4348.2014.06.018

基于抽样数据与核估计的随机数生成法

袁晓建[1],吕书龙[2]

(1 福州外语外贸学院,福建 福州 350202; 2 福州大学数学与计算机科学学院,福建 福州 350116)

摘要: 对任意连续总体的一个独立同分布(i.i.d.)样本,在非参数核密度估计的基础上,定义总体分布函数核估计的多种形式,给出基于非参数核估计的随机数产生方法及统计检验。

关键词: 非参数核密度估计; 随机数; 统计检验

中图分类号: O212 **文献标志码**: A **文章编号**: 1672 - 4348(2014)06 - 0595 - 04

Random number generation methods based on data sampling and kernel estimation

Yuan Xiaojian[1], Lü Shulong[2]

(1. Fuzhou Foreign Languages and Foreign Trade College, Fuzhou 350202, China;
2. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

Abstract: For any independently iso-distributed sample (i.i.d.) of a continuous distribution, the multiform of overall distributed function kernel estimation was defined on the basis of non-parametric kernel density estimation. Random number generation method based on non-parametric kernel density estimation and the statistical tests of the method were presented.

Keywords: non-parametric kernel density estimation; random number; statistical test

随机数大量应用于抽样调查、数值计算、信息安全、系统仿真、Monte Carlo 模拟、金融投资、环境重构等多个领域。随着数字时代和虚拟现实的纵深发展,信息的安全性、仿真系统、大数据模拟以及环境重构等日益重要。在这些领域中,统计性能优良、高速稳定的随机发生器和随机数的作用至关重要。关于随机数及其应用的研究经久不衰,可参看文献[1 - 5]。

通常意义下伪随机数的产生都是基于具体的随机分布,而且是在分布已知的前提下来讨论的。最基本的做法是构造均匀分布的随机数发生器,然后应用多种方法如反函数法、变换法或近似法[3-4,6-8],得到其他分布的随机数。在很多实际问题中,抽样数据 $x_1, x_2, \cdots, x_n$ 所服从的分布往往是未知的,在未知分布的前提下,很多基于分布的随机数发生器难以凑效。关键问题在于如何仅利用抽样数据来产生该样本所在总体的随机数,显然这应通过非参数方法加以解决。文献[6 - 7]对这类问题提出了近拟抽样法,包括经验分布抽样法和频数观测数据抽样法。从非参数角度讲,有了样本数据,就可以模拟密度函数和分布函数,进而模拟随机数的产生。

本文利用核估计思想由原始数据估计密度函数和分布函数,并构造基于非参数核估计的随机数发生器,并通过了模拟实验和分布检验。

1 经验分布抽样法

基于经验分布函数 $F_n(x)$ 的随机数产生方法的理论依据是 Glivenko 定理:

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\right) = 1$$

收稿日期: 2014 - 10 - 17
基金项目: 福州大学科技发展基金资助项目(2013 - XY - 20);福州大学研究生重点课程建设项目(Y2013KC03)
第一作者简介: 袁晓建(1981 -),男(汉),河南漯河人,讲师,硕士,研究方向:数理统计。

对于观测数据 $x_1,x_2,\cdots,x_n$，构造其经验分布函数 $F_n(x)$，然后产生 $F_n(x)$ 的随机数，由上述定理可将该随机数近似成 $F(x)$ 的随机数。应用该理论产生随机数的过程如下：

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

(2) 取 $X=x_{(J)}+\dfrac{r-F_n(x_{J-1})}{F_n(x_{(J)})-F_n(x_{J-1})}(x_{(J)}-x_{(J-1)})$，则 X 近似为 $F(x)$ 的随机数。其中 $x_{(1)},x_{(2)},\cdots,x_{(n)}$ 是 $x_1,x_2,\cdots,x_n$ 的次序统计量。该方法仅能产生的随机数范围介于 $[x_{(1)},x_{(n)}]$。注意：若 $x_1,x_2,\cdots,x_n$ 中有重复值，则 J 值出现跳跃。

该方法在产生随机数时，引入线性插值求近似随机数，方法简洁，可操作性强。通常情况下，线性插值并不满足分布曲线的特征，因此产生的随机数还不够精细，有改进余地。

2 核估计随机数发生器的基本理论

非参数方法在密度估计方面的理论[9-13]相当丰富，而基于非参数密度估计的最直接应用在于概率的估计[9]，还可以用于数据挖掘领域[14]等，但是将非参数密度估计应用到随机数抽样中还不多见。下面阐述如何使用非参数核估计法产生近似随机数并对其进行统计性质的检验。

定义1 设 $K(u)$ 为定义在 $(-\infty,+\infty)$ 上的一个 Borel 可测函数，而 $h_n>0$ 为常数。$x_1,x_2,\cdots,x_n$ 是总体 X 的一个 iid 样本，则称

$$f_n(x)=\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right) \tag{1}$$

为总体密度函数 $f(x)$ 的一个核估计。其中 $K(u)$ 称为核函数，一般取对称型的密度函数，h_n 称为窗宽。

文献[9-10]对核估计的大样本性质作了细致的讨论和证明，此处仅给出5个定理结论。

定理1 核密度估计具有渐进无偏性

$$\lim_{n\to\infty}Ef_n(x)=f(x) \tag{2}$$

定理2 若 $\lim\limits_{n\to\infty}h_n=0$，$f(x)$ 在全空间上一致连续，且满足(2)式，则有

$$\lim_{n\to\infty}\{\sup_x|Ef_n(x)-f(x)|\}=0 \tag{3}$$

定理3 核密度估计的均方相合性

$$\lim_{n\to\infty}E[f_n(x)-f(x)]^2=0 \tag{4}$$

定理4 核密度估计的依概率一致收敛性

$$\lim_{n\to\infty}\sup_x|f_n(x)-f(x)|=0 \tag{5}$$

定理5 核密度估计的强相合性

$$\lim_{n\to\infty}f_n(x)=f(x),\text{a. s.}$$

$$\lim_{n\to\infty}\{\sup_x|f_n(x)-f(x)|\}=0,\text{a. s.} \tag{6}$$

基于定义1和上述5个定理，可以从核估计角度来刻画总体分布函数，有了分布函数的估计，从理论上讲得到总体的近似随机数就比较容易了，这也是本文讨论的重点。

定义2 设 $f_n(x)$ 为总体密度函数 $f(x)$ 的一个核估计，则称

$$\tilde{F}_n(x)=\int_{-\infty}^{x}f_n(t)\,dt \tag{7}$$

为总体分布函数 $F(x)$ 的一个核估计。此处采用记号 $\tilde{F}_n(x)$ 以区别于经验分布函数 $F_n(x)$。

对于定义2中的 ，代入定义1的 ，可得到更一般的表示形式，具体如下：

$$\tilde{F}_n(x)=\int_{-\infty}^{x}f_n(t)\,dt=\int_{-\infty}^{x}\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right)dt=$$

$$\sum_{i=1}^{n}\int_{-\infty}^{x}\frac{1}{nh_n}K\left(\frac{x-X_i}{h_n}\right)dt\xleftarrow{y=\frac{x-X_i}{h_n}}=$$

$$\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-X_i}{h_n}}\frac{1}{nh_n}K(y)\cdot h_n\cdot dy=$$

$$\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-X_i}{h_n}}\frac{1}{nh_n}K(y)\,dy \tag{8}$$

下面推导关于具体核函数的 $\tilde{F}_n(x)$ 表示(以高斯核和均匀核为例)，设 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 为样本的次序统计量，当 $x\in[X_{(1)},X_{(n)}]$ 时，一定存在 p,q 使得 $X_{(p)}$ 和 $X_{(q)}\in[x-h_n,x+h_n]$。但 $X_{(p-1)}$ 和 $X_{(q+1)}\notin[x-h_n,x+h_b]$，$p\leqslant q$。且有 $p\leqslant k\leqslant q$，使得 $X_{(k)}\leqslant x,X_{(k+1)}>x$。

1)若 $K(u)$ 为高斯核，即 $K(u)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$，则(8)式可变成

$$\tilde{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x-X_i}{h_n}\right) \tag{9}$$

2) 若 $K(u)$ 为均匀核，即 $K(u)=1/2$，$|u|<1$。p,q 满足上述条件，则(8)式可变成

$$\tilde{F}_n(x)=\frac{1}{2n}\left[(q-p+1)+\sum_{i=p}^{q}\frac{x-X_{(i)}}{h_n}\right] \tag{10}$$

由于 $\tilde{F}_n(x)$ 是分布函数 $F(x)$ 的一个优良估

计,所以构造近似方程组如下:

$$\begin{cases}\tilde{F}_n(\tilde{x}_i) \cong F_{(x_i)} \\ F(x_i) = r_i\end{cases} \Rightarrow \begin{cases}\tilde{F}_n(\tilde{x}_i) \cong r_i \\ i = 1,2,\cdots,N\end{cases} \quad (11)$$

定义 3　对于 $r.v.r \sim U(0,1)$ 的随机序列 $r_1,r_2,\cdots,r_N$, 称 $\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_N$ 为总体分布函数 $F(x)$ 的一个近似随机数序列,其中 $\tilde{x}_i$ 满足(11)式,即 $\tilde{F}_n(\tilde{x}_i) = r_i, i = 1,2,\cdots,N$。

由定义 3 推出基于核估计的随机数产生算法:

步骤 1　由观测值 $x_1,x_2,\cdots,x_n$,以合适的窗宽 h_n 和核函数 $K(u)$, 构造分布函数的核估计 $\tilde{F}_n(x)$。

步骤 2　产生均匀分布随机数 r,即 $r \sim U(0,1)$。

步骤 3　由 $\tilde{F}_n(\tilde{x})$,计算 $\tilde{x}$,则 $\tilde{x}$ 近似为 $F(x)$ 的随机数。

3　实验分析

3.1.3　模拟实验 1

生成 N(0.1) 随机数 600 个 $x_1,x_2,\cdots,x_{600}$ 取前 300 个作为原随机数,利用上述方法和步骤生成密度和分布函数的核估计,再生成 300 个随机数(新随机数),最后分布与样本中前 300 个和后 300 个随机数(预留随机数)比较并检验。下面通过多种检验手段说明核估计产生的随机数符合随机数发生器的要求。

(1)同分布检验 p 值。

表 1　三类同分布检验的 p 值

Tab. 1　P value of three iso-distributed tests

次数	新随机数作正态性检验	与原随机数同分布检验	与预留随机数作同分布检验
1	0.491 108 9	0.970 040 9	0.721 162 5
2	0.943 685 2	0.940 842 3	0.395 298 3
3	0.647 202 7	0.847 488 5	0.787 044 3
4	0.828 594 0	0.940 842 3	0.787 044 3
5	0.514 038 0	0.721 162 5	0.987 698 3
平均	0.684 925 8	0.884 075 3	0.735 649 5

从上述 3 类检验值可知,核估计产生的随机数服从原分布,且与经典方法产生的随机数序列来自同一分布。

(2)原随机数和生成随机数作核密度图比较。

从图 1 可知,原随机数和本文方法产生的随机数的正态密度曲线吻合度比较高。

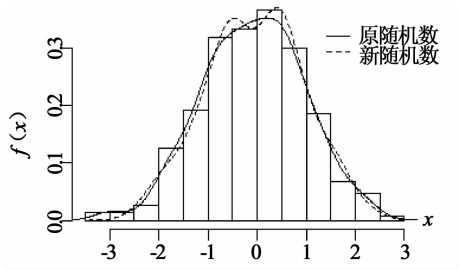

图 1　两种随机数的密度曲线对比

Fig. 1　Comparison between the density curves of 2 random numbers

(3)原随机数及预留随机数与新随机数的经验分布图和箱线图。

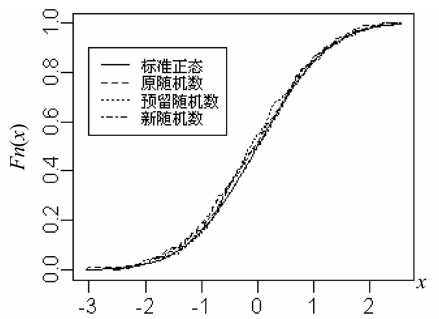

图 2　三组随机数经验分布图的比较

Fig. 2　Comparison among the experience distribution curves of 3 sets of random numbers

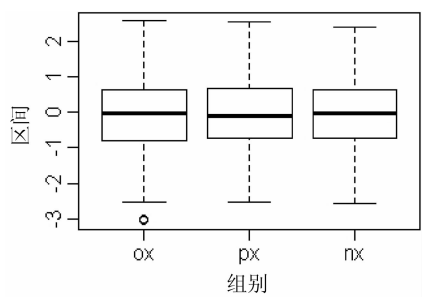

图 3　三组数据箱线图的比较

Fig. 3　Comparison of box line figures among 3 sets of data

从图 2 可看出,3 组数据的经验分布函数图交织在一起,几乎无法区分,直观上说明 3 组随机数是同分布的。图 3 的箱线图可看出,3 组随机数分布形状非常接近。

(4)独立性的直观检验,作 1 到 6 阶自相关三点图, 从数据分布的趋势上作独立性的判断。

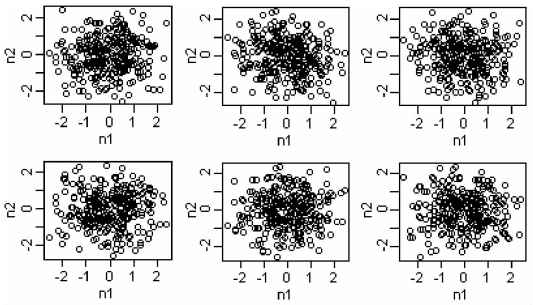

图 4 1 –6 阶自相关图形

Fig. 4 Self-correlated figures of order 1 to 6

从图 4 的自相关散点图看出，新随机数之间并不存在某种一致性的趋势，直观上可认定随机数之间是相互独立的。这点从步骤 2 和步骤 3 是可以推出来的。模拟实验表明，核估计方法生成的随机数同原总体产生的随机数并无统计意义上的偏差，且在统计性能上均可满足随机数发生器的要求，因此可将其作为一种可行的随机数发生器。

在实际问题中经常处理小样本数据，而有些经典方法在处理小样本数据的时候往往效果较差。因此如何将小样本数据变成“大样本”数据，并将其应用于随机模拟中，还是有意义的。下面从小样本角度对本文方法进行检验。

3.2 模拟实验 2

对 N(0,1)，t(5)，$\chi^2(5)$，F(5,5) 分别抽样 20 个随机数，然后用本文方法产生 100 个新随机数。反复做 10 个流程，并分别与 20 个随机数作同分布检验。计算均值、方差和检验 p 值的平均值。

表 2 计算结果

Tab. 2 Calculation results of mean value, square root value and *p* value

分布		N(0,1)	t(5)	$\chi^2(5)$	F(5,5)
均值	原	0.107 867	0.146 007	4.422 129	1.195 808
平均	新	0.136 535	0.182 869	4.513 236	1.254 544
方差	原	0.516 335	0.846 102	4.263 615	1.085 520
平均	新	0.554 457	0.861 503	5.109 133	2.072 740
检验 p 值		0.788 071	0.986 328	0.883 265	0.349 183

从表 2 计算结果看出由小样本出发采用核估计生成的”大样本”随机数，其统计指标还是能满足随机数模拟要求的。

4 结论

通过对非参数核密度估计的分析，利用密度函数和分布函数的积分关系，导出分布函数的非参数核估计进而得到该分布的近似随机数，并通过模拟数据的计算和检验，验证了非参数核估计随机数发生器是一种实用有效的方法。非参数核估计的计算量大，模型中还涉及到窗宽及核函数的选取，如何有效地减少计算量并提高效率，如何选取合适的窗宽和核函数一直是非参数领域的课题，这些都有待于进一步研究。

参考文献：

[1] 卢振泰，陈武凡. Monte Carlo 随机数在图像加密中的应用[J]. 计算机工程与应用，2009，45(7)：7 –9.
[2] 石琴，李友文，郑与波. 随机数在汽车行驶工况构建中的应用[J]. 西南交通大学学报，2010，45(6)：938 –945.
[3] 肖化昆. 系统仿真中任意概率分布的伪随机数的研究[J]. 计算机工程与设计，2005，26(1)：168 –171.
[4] 庄光明，夏建伟，彭作祥，等. 基于 Matlab 的 Poisson 分布随机数的 Monte carlo 模拟[J]. 数学的实践与认识，2012，42(5)：87 –92.
[5] 汪龙，马海强，李申，等. 基于光子间隙随机分布的真随机数源[J]. 物理学报，2013，62(10)：1 –5.
[6] 杨振海，程维虎. 非均匀随机数产生[J]. 数理统计与管理，2006，25(6)：750 –757
[7] 杨振海，张国志. 随机数生成[J]. 数理统计与管理，2006，25(2)：244 –252.
[8] 沈华韵，张鹏，王侃. 改进线性同余法随机数发生器[J]. 清华大学学报：自然科学版，2009，49(2)：191 –193.
[9] Parzen E. On estimation of a probability density function and mode[J]. The Annals of Mathematical Statistics，1962，33(3)：1065 –1076.
[10] 陈希孺，方兆本，李国英，等. 非参数统计[M]. 上海：科学技术出版社，1989.

（责任编辑：肖锡湘）