2021年数理统计上机课

–随机数发生

黄启岳

北京师范大学统计学院

2021 年 4 月 28 日

目录

引入

生活是由大量的随机事件构成的。随机事件即不确定结果的事件，通常统计学中研究的随机事件在大量重复试验中具有某种规律性。

生活中的例子

- 例1：盲盒。相同的盒子中放置不同的商品，消费者事先不知道盒子里装的是哪一款，但有一定概率能够抽到自己心仪的商品。
- 例2：部分游戏中的抽卡抽奖玩法。使用现实中的货币购买游戏中虚拟货币，再将虚拟货币转化为抽奖机会，一定概率获得自己需要的道具或角色。

盲盒抽取满足结果不固定以及重复次数上升后具备一定规律的特点。但网络上的随机抽奖则是应用随机数发生的知识，并不是完全意义上的“随机事件”。以下讲述随机数发生的原理以及部分随机数的生成。

常见分布随机数

分布	R中函数
二项分布	binom(size, prob)
泊松分布	pois(lambda)
均匀分布	unif(min, max)
指数分布	exp(rate)
正态分布	norm(mean, sd)
χ^2分布	chisq(df, ncp)
t分布	t(df)
F分布	f(df1, df2, ncp)

随机数发生方法

除了R语言中已经给出的函数可以用于随机数生成，还可以通过一些数学变换基于均匀分布随机数生成指定分布随机数，常用于对未知分布或不常用分布的模拟。这里主要介绍逆变换法，即通过概率论中的单调逆定理导出服从特定分布的随机变量与均匀分布的关系，进而生成符合分布要求的随机数。

均匀分布随机数的生成机制

现在随机模拟采用的随机数通常是由计算机按确定的递推公式实时地产生的伪随机数，其在一定程度上体现随机性。好的伪随机数序列与真实随机数序列在统计检验上表现几乎相同，很难区分，因此伪随机数通常也被称为随机数。

计算机最容易产生的随机数是均匀分布随机数，产生这些随机数的发生器主要有线性同余发生器、反馈位移寄存器发生器以及组合发生器。这里介绍最简单的线性同余发生器。

线性同余发生器(选读)

线性同余发生器于1951年由Lehmer提出，使用同余运算产生随机数，即：

$$x_n = (ax_{n-1} + c) \mod M, a > 0, c \geq 0, M > 0 \qquad (1)$$

其中x_0是小于M的整数，当c与M互质；M所有质因子的乘积能够整除(a-1)；(a-1)是4的倍数，若M是4的倍数。

C语言rand()常取$M = 2^{31}, a = 1103515245, c = 12345$，每次递归过程，只截取X最高的16位作为最终的随机数输出。这种方法周期最高为2^{32}。一般直接使用runif函数生成随机数即可。

逆变换法生成随机数

根据概率论中的单调逆定理，分布函数的逆存在，这是逆变换法生成随机数的基础。

THEOREM (单调逆定理)

若F为连续的严格单增分布函数，反函数F^{-1}存在。即：

$$F^{-1}(x) = \inf\{y : F(y) \geq x\}$$

则：

(1)$\xi \sim F(x)$，则$F(\xi) \sim U(0,1)$;

(2)$R \sim U(0,1)$，则$F^{-1}(R) \sim F(x)$.

两点分布随机数

离散分布两点分布B(1,p)即伯努利试验结果的分布，具有概率分布为：

$$f(x) = p^x(1-p)^{(1-x)} \tag{2}$$

若针对单次试验，结果以0与1记，即取1的概率为p，取0概率为1-p。则分布函数可以记为：

$$F(x) = \chi_{\{x>1-p\}}$$

初步算法构建：生成$U \sim U(0,1)$，若$U < p$则$\xi = 1$，反之，$\xi = 0$，则$\xi \sim B(1,p)$。本质上是将[0,1]区间分段。类似可以生成多项分布的随机数，即：

$$P(X = x_j) = p_j, \text{其中} \sum_j p_j = 1$$

思考：使用逆变换法如何生成离散均匀分布,即$P(x = j) = \frac{1}{n}$?

几何分布随机数

回忆：几何分布是怎么构造的？

几何分布概率分布函数：

$$P(X = x) = p(1-p)^{x-1} = pq^{x-1} \tag{3}$$

均值与方差分别为$E(X) = \frac{1}{p}, Var(X) = \frac{q}{p^2}$

考虑几何分布的分布函数：

$$\begin{aligned} P(X \leq j) &= 1 - P(x > j) \\ &= 1 - q^j \end{aligned} \tag{4}$$

则将分布函数值域n等分，令$U \sim U(0,1)$：

$$1 - q^{j-1} < U < 1 - q^j \tag{5}$$

(5)式成立后$X = j$。反解出j，即：

$$j - 1 < \frac{\ln(1-U)}{\ln(q)} < j \tag{6}$$

即算法为：

- 生成$U \sim U(0,1)$;
- 令$X = \left[\frac{\ln(1-U)}{\ln(q)}\right] + 1$;
- X服从几何分布$G(p)$.

注意：$1-U$与U同分布。R语言代码为rgeom(n,p)+1。

二项分布随机数

考虑二项分布$B(n,p)$在x处密度函数为：

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{(n-x)} \tag{7}$$

思考：生成二项分布随机数可以有几种方法？

思路

构思：

- 变换法：根据两点分布构造，哪一个统计量服从二项分布？
- 逆变换法

逆变换法

考虑：若记$P(X=i)=P_i$，则

$$\frac{P_{i+1}}{P_i}=\frac{\binom{n}{i+1}p^{i+1}(1-p)^{(n-i-1)}}{\binom{n}{i}p^{i}(1-p)^{(n-i)}}=\frac{n-i}{i+1}\frac{p}{1-p} \tag{8}$$

即得到：

$$P_{i+1}=\frac{n-i}{i+1}\frac{p}{1-p}P_i \tag{9}$$

算法

- (1)生成$U \sim U(0,1)$;
- (2)设定初值$i = 0$，从一次没有成功开始模拟，
  $P = (1-p)^n$，$F = F + P = P(x \leq i)$;
- (3)若$U < F$则循环停止，令$x = i$;
- (4)若$U > F$则：$P = \frac{n-i}{i+1}\frac{p}{1-p}P$，$F = F + P, i = i + 1$;
- (5)回到(3).

这样生成的$x \sim B(n,p)$。R语言中的命令为$rbinom(m,n,p)$。

指数分布随机数生成

服从指数分布的随机变量为连续型随机变量。强度为λ的指数分布$e(\lambda)$概率密度与概率分布为：

$$f(x) = \lambda e^{-\lambda x}, F(x) = 1 - e^{-\lambda x} \tag{10}$$

若令$1 - e^{-\lambda x} = U \sim U(0,1)$，则：

$$x = -\frac{1}{\lambda} \ln(1 - U) \tag{11}$$

算法

根据(11)式生成算法为：

- 生成$U \sim U(0,1)$;
- 令$x = -\frac{1}{\lambda}\ln(U)$;
- $x \sim e(\lambda)$

统计模拟

统计模拟即使用计算机生成大量随机数模拟真实事件发生并计算感兴趣的统计量的过程，即蒙特卡洛方法（Monte Carlo method）。尤其像一些复杂系统，理论计算非常困难，但使用随机模拟的方式即可估计出需要的参数。

常见估计策略

针对被估计参数，常常通过模拟生成一组被估参数的估计值，利用这些模拟出的估计值近似被估计参数的分布，进而通过模拟出的分布信息估计参数本身。

练习

某手游设定有抽奖机制，当期SSR平均抽取概率为0.8%。请使用统计模拟的方法完成：

- (1)：当玩家存有足够326次抽奖的道具时，假定每次抽取服从两点分布，模拟其抽取到SSR的总次数并搜索哪几次抽取到了SSR。
- (2)：模拟在(1)的条件下，不同总抽奖次数没有抽取到SSR的概率，并使用统计模拟判断：总抽奖次数不少于多少次时可以保证没有抽取到SSR的概率低于5%？

抽样与BOOTSTRAP方法

R语言中抽取数据中样本的函数为sample(size = 抽取量, replace = T(放回)/F(不放回))。

以该抽样方法为基础，简单介绍Bootstrap方法的思想。

BOOTSTRAP思想

Bootstrap于1979年由Efron提出，英文原意为“靴带”，引申义为“To improve your situation yourself without help from other people”。总体思想是通过对原始数据进行放回重抽样模拟生成很多个Bootstrap样本，利用抽取的样本进行统计推断。

算法

最一般的Bootstrap算法可以总结如下：

- 获得原始数据，用于Bootstrap重抽样;
- 使用放回简单随机抽样抽取Bootstrap样本;
- 使用抽取的Bootstrap样本进行统计推断。

BOOTSTRAP方法估计具体例子

例1：自助法估计均值的方差：

设统计量$T_n = g(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$，$x_i \sim^{i.i.d} (\mu, \sigma^2)$，$T_n$的方差表示为：

$$Var_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 \mathbf{d}F - \left(\int x \mathbf{d}F\right)^2}{n}$$

理论上知道了F即可计算出均值的方差。实际情形则是观测到了数据$x_1, \ldots, x_n$，希望通过数据获知F的信息。

算法设计：

- (1)获取数据为$x_1, \ldots, x_n$
- (2)对数据进行放回简单随机抽样，获得Bootstrap数据为$x_{11}, \ldots, x_{1n}$
- (3)计算统计量在Bootstrap数据下的取值，即

$$T_1^* = \frac{1}{n}\sum_{i=1}^{n} x_{1i}$$

- (4)将(2)和(3)重复B次，得到：$T_1^*, \ldots, T_B^*$，

$$v_{boot} = \frac{1}{B}\sum_{j=1}^{B}\left(T_j^* - \frac{1}{B}\sum_{i=1}^{B} T_i^*\right)^2$$

例2：估计一组数据的中位数标准差的Bootstrap估计。算法应该如何设计？