贝叶斯统计模型及其应用研究

BeiYeSi TongJi MoXing JiQi YingYong YanJiu

王明高／著

中国财经出版传媒集团
经济科学出版社
Economic Science Press

贝叶斯统计模型及其应用研究

BeiYeSi TongJi MoXing JiQi YingYong YanJiu

王明高／著

中国财经出版传媒集团
经济科学出版社
Economic Science Press

图书在版编目（CIP）数据

贝叶斯统计模型及其应用研究/王明高著. —北京：
经济科学出版社，2017.1
ISBN 978-7-5141-7754-1

Ⅰ.①贝… Ⅱ.①王… Ⅲ.①贝叶斯统计量-
研究 Ⅳ.①O212.8

中国版本图书馆 CIP 数据核字（2017）第 025732 号

责任编辑：李 雪 李 建
责任校对：刘 昕
责任印制：邱 天

贝叶斯统计模型及其应用研究

王明高 著

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲 28 号 邮编：100142

总编部电话：010-88191217 发行部电话：010-88191522

网址：www.esp.com.cn

电子邮件：esp@esp.com.cn

天猫网店：经济科学出版社旗舰店

网址：http://jjkxcbs.tmall.com

北京汉德鼎印刷有限公司印刷

三河市华玉装订厂装订

710×1000 16 开 19 印张 280000 字

2017 年 1 月第 1 版 2017 年 1 月第 1 次印刷

ISBN 978-7-5141-7754-1 定价：58.00 元

前　言

蒙特卡洛作为一种统计模拟方法，在统计计算中具有重要地位。后来发展起来的马尔科夫蒙特卡洛方法极大地推进了统计计算的发展，特别是有力地促进了贝叶斯理论及其应用的发展。基于先验信息和实际数据建模的贝叶斯方法，对一些复杂问题特别有效。事实上，贝叶斯原理已经成了当今技术进步和科技创新的基础，譬如，在无人驾驶、搜索引擎、语音识别和人工智能等领域都离不开贝叶斯方法的应用。在大数据时代，贝叶斯方法必将是数据分析的一把利器。

自从20世纪80年代中期以来，随着计算机技术的发展以及马尔科夫链蒙特卡洛方法（MCMC）的应用，贝叶斯统计及其建模受到越来越多的关注。在20世纪90年代后期，基于MCMC方法的BUGS软件问世，该软件能够以简洁的形式建立复杂的模型，所以在各个领域的统计分析中得到广泛应用。此外，R软件在统计建模和数据分析中的应用也不断普及，并且可以与BUGS软件有机结合，所以，本书在介绍蒙特卡洛理论和贝叶斯统计模型时，将同时给出应用R和OpenBUGS编写的程序代码。

本书内容分为3部分，共9章。第1部分为蒙特卡洛方法及其应用，其中第1章介绍了随机变量的模拟理论，这部分是统计计算的基础，也是蒙特卡洛方法和贝叶斯理论的基础；第2章介绍了蒙特卡洛积分，它对贝叶斯模型的求解具有重要意义；第3

章介绍了蒙特卡洛 EM 方法，该方法有利于一些复杂模型的求解。

第 2 部分为马尔科夫蒙特卡洛方法（MCMC），其中第 4 章介绍了贝叶斯理论的基础知识；第 5 章介绍了贝叶斯马尔科夫链蒙特卡洛方法（MCMC），主要包括 MH 方法和 Gibbs 方法；第 6 章介绍了 MCMC 方法的发展，如自适应 MH 方法和可逆跳跃 MCMC 方法等。

第 3 部分为贝叶斯模型，其中第 7 章介绍了贝叶斯混合分布模型，该模型对异常数据建模具有一定的优势，可以体现贝叶斯方法的灵活性；第 8 章介绍了贝叶斯线性回归模型，通过该模型说明了贝叶斯建模的基本方法以及 OpenBUGS 的应用；第 9 章介绍了贝叶斯分层模型，该模型又称为随机效应模型，主要用来分析比较复杂的分层数据，应用贝叶斯方法处理该模型具有一定优势。

本书后面的两个附录，其中附录 1 介绍了贝叶斯模型的检验方法。贝叶斯方法主要通过大量的抽样对模型参数进行估计，所以对抽样样本的收敛性进行检验非常重要；附录 2 主要介绍了如何应用 OpenBUGS 软件建立贝叶斯模型，给出了基本的操作说明。

本书的应用案例主要来自非寿险定价和准备金评估中的有关问题，使用 R 和 OpenBUGS 软件编写程序代码，所以阅读本书的读者需要具备一定的统计基础知识，并且对 R 和 OpenBUGS 软件具有一定了解。

本书的部分内容是作者参与教育部人文社会科学重点研究基地重大项目“基于大数据的精算统计模型及其应用”（项目编号：16JJD910001）的研究成果。

目　录

第 1 章

随机变量的生成

在统计计算中，经常需要从特定概率分布中模拟随机变量，这些模拟随机分布的方法依赖于均匀分布随机变量生成其他分布的随机变量。比如，蒙特卡洛方法中，由某分布中生成一系列随机变量，模拟主要基于区间［0，1］上的均匀分布。在 R 软件中，均匀分布伪随机数生成函数为 runif，具体应用 runif(n) 生成一个 0 与 1 之间的 n 维向量随机数。

1.1 均匀分布模拟

在 R 中均匀分布的函数是 runif，该函数需要给出变量生成个数，以及取值区间。例如：

runif(100，min = 2，max = 5)

即在［2，5］内生成 100 个均匀分布数。对于生成的数据需要进行检验，确保这些数据服从均匀分布，比较简便的方法有：生成数据直方图、相邻数据（X_i，X_{i+1}）的关系图、数据的自相关函数。具体 R 程序如下所示：

```
n = 10^4 #随机数总个数
x = runif(n)
x1 = x[-n]; x2 = x[-1] #相邻数对
par(mfrow = c(1,3)); hist(x); plot(x1,x2); acf(x) #自相关函数
```

根据生成的图形（见图 1－1）可以看出，runif 函数生成均匀分布数据是合理的。

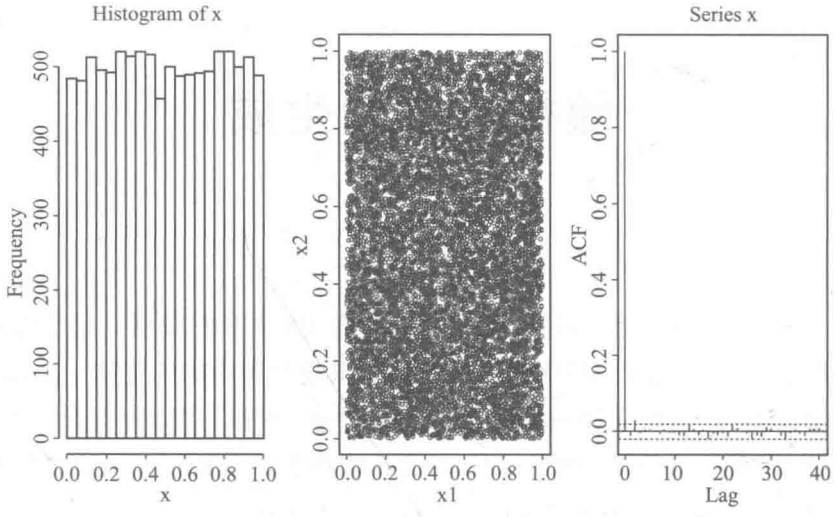

图 1－1　模拟数据比较图

1.2　逆　变　换

逆变换也可以称为概率积分变换，它可以将任意随机变量转换为均匀分布的随机变量，或者进行相反的变换。例如，如果 X 具有密度函数 f 和分布函数 F，则他们具有关系：

$$F(x) = \int_{-\infty}^{x} f(t)$$

假设 $U = F(X)$，随机变量 U 服从均匀分布 $u(0, 1)$，可以得到下面的等式：

$$P(U \leqslant u) = P(F(X) \leqslant F(x)) = P(F^{-1}(F(X)) \leqslant F^{-1}(F(x))) = P(X \leqslant x)$$

这里假设分布函数 F 具有逆函数，该等式对于大多数连续函数都成立。

例 1－1　假设 $X \sim \exp(1)$，则 $F(x) = 1 - e^{-x}$，对于等式 $u = 1 - e^{-x}$ 进行变换，得到 $x = -\log(1-u)$。如果 $U \sim u(0, 1)$，可以得到

$$X = -\log U \sim \exp(1)$$

这里 U 和 $1-U$ 都是服从均匀分布，其相应的 R 代码如下所示：

```
n=10^4#随机样本个数
U=runif(n); X= -log(U)#分布变换
Y=rexp(n) #R 中的指数分布随机数
par(mfrow=c(1,2)) #图表分布
hist(X,freq=F,main="Exp from Uniform");hist(Y,freq=
  F,main="Exp from R")
```

上面分别应用逆变换方法和 R 随机数函数生成指数随机数，根据图 1-2 看出这两种方法生成的随机数具有相同的分布形式，说明逆变换方法的有效性。

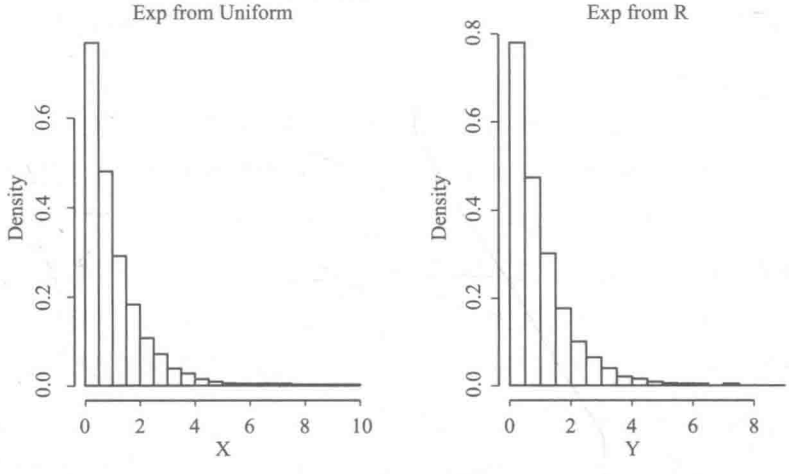

图 1-2　指数分布比较图

其他分布都可以由均匀分布转变生成，所以均匀分布随机变量的产生是模拟方法的关键因素。逆变换方法也可以用于生成离散分布随机变量，如果 X 是离散随机变量，并且各点具有如下关系，

$$\cdots < x_{i-1} < x_i < x_{i+1} < \cdots$$

它们是分布函数 $F_X(x)$ 中的点，如果 $F_X(x_{i-1}) < u \leqslant F_X(x_i)$，根据逆变换可以得出 $F_X^{-1}(u) = x_i$。

随机变量的生成步骤如下：

（1）从均匀分布 $Uniform(0, 1)$ 中生成随机变量 u。

（2）如果 $F(x_{i-1}) < u \leqslant F(x_i)$，生成随机变量 x_i。

例 1－2 生成两点分布，应用逆变换方法生成参数 $p = 0.4$ 贝努力分布，在这个例子中，分布函数取值分别为 $F_X(0) = f_X(0) = 1 - p$ 和 $F_X(1) = 1$。如果生成的随机变量 $u > 0.6$，则 $F_X^{-1}(u) = 1$；如果 $u \leqslant 0.6$，可得 $F_X^{-1}(u) = 0$。具体 R 程序如下所示：

```
set.seed(10); n <- 1000; p <- 0.4
u <- runif(n); x <- as.integer(u >0.6)
mean(x)
[1] 0.405
var(x)
[1] 0.2412
```

基于逆变换方法生成的 1000 样本，得出他们的均值和方差，并与贝努力分布的理论均值和方差（$p = 0.4$ 和 $p(1-p) = 0.24$）进行比较，可以看出模拟结果和理论结果比较接近。

例 1－3 几何分布，应用逆变换方法生成参数 $p = 0.25$ 的几何分布，其概率函数为 $f(x) = pq^x$，这里 $x = 0, 1, 2, \cdots$ 和 $q = 1 - p$，均值和方差分别为 $EX = (1-p)/p = 3$ 和 $VarX = (1-p)/p^2 = 12$，分布函数为 $F(x) = 1 - q^{x+1}$。生成均匀分布随机变量 u，求解等式：$1 - q^x < u \leqslant 1 - q^{x+1}$，为了方便求解需要进行变换 $x < \log(1-u)/\log(q) \leqslant x + 1$，得出 $x + 1 = [\log(1-u)/\log(q)]$，这里 $[t]$ 表示顶函数，其 R 函数表示为 ceiling ()。该例的 R 程序如下所示：

```
n <- 1000;p <- 0.25;u <- runif(n)
x <- ceiling(log(1 -u) / log(1 -p)) -1
mean(x)
[1] 4.034
var(x)
[1] 12.18703
```

应用逆变换方法模拟贝努力分布和几何分布比较容易，因为它的不等式 $F(x-1)<u\leqslant F(x)$ 比较容易求解。但是其他一些离散分布的该不等式难以求解。

例1-4　对数分布，应用逆变换生成对数分布随机变量，其概率函数为：

$$f(x)=P(X=x)=\frac{a\theta^x}{x},\ x=1,\ 2,\ \cdots,\ 0<\theta<1,\ a=(-\log(1-\theta))^{-1}.$$

生成特定 u 值，满足关系式 $F(x-1)<u\leqslant F(x)$，计数满足不等式 $F(x-1)<u$ 的个数确定 x 值。具体 R 程序如下所示：

```
n<-1000; N<- 30; theta<-0.5; k<- 1:N; a <- -1/log(1-theta)
fk<- exp(log(a) + k * log(theta) - log(k)) #对数分布密度函数
Fk<- cumsum(fk) #对数分布函数
x<- integer(n)
for (i in 1:n) {
  u<- runif(1)
  x[i]<- as.integer(sum(u > Fk)) +1}
Ex<-a* theta/(1-theta);mean(x) #理论均值与样本均值比较
[1] 1.442695;[2] 1.477
Varx<- -a^2* theta* (theta+log(1-theta))/(1-theta)^2;
  var(x) #理论方差与样本方差比较
[1] 0.8040211;[2] 0.8262973
```

上面程序中具体 N 值根据不同的概率函数选定，确保 $F(N)=1$。n 表示生成模拟值的个数，*Ex* 表示对数分布的理论均值，*Varx* 是其理论方差，通过与模拟值的均值与方差的比较，可以确保模拟程序的准确性。

1.3　一般变换方法

当一个具有密度函数 f 的分布跟其他易于模拟的分布具有简单关系，根据他们之间关系可以模拟得到 f 的随机变量。

（1）如果随机变量 Z 服从标准正态分布 $Z\sim N(0,\ 1)$，则 Z^2 服从参数

为 1 的卡方分布 $V=Z^2\sim\chi^2(1)$。

（2）如果随机变量 U 和 V 分别服从卡方分布，即 $U\sim\chi^2(m)$，$V\sim\chi^2(n)$，则随机变量 $F=\frac{U/m}{V/n}$服从自由度为（m，n）的 F 分布。

（3）如果随机变量 Z 服从标准正态分布 $Z\sim N(0,1)$，V 服从卡方分布 $V\sim\chi^2(n)$，并且他们相互独立，则 $T=\frac{Z}{\sqrt{V/n}}$服从参数为 n 的 t 分布。

（4）如果随机变量 U 和 V 服从相互独立的均匀分布 U，$V\sim Unif(0,1)$，则 $Z_1=\mu+\sigma\sqrt{-2\log U}\cos(2\pi V)$ 和 $Z_2=\mu+\sigma\sqrt{-2\log U}\sin(2\pi V)$ 为相互独立的正态分布 $N(\mu,\sigma^2)$，随机变量的指数服从对数正态分布 $\exp(Z)\sim lognormal(\mu,\sigma^2)$。

（5）如果随机变量 U 和 V 服从相互独立的伽玛分布，$U\sim Gamma(r,\lambda)$ 和 $V\sim Gamma(s,\lambda)$，则 $X=\frac{U}{U+V}$服从贝塔分布 $Beta(r,s)$。

（6）如果随机变量 U 和 V 服从相互独立的均匀分布 U，$V\sim Unif(0,1)$，则

$$X=\left[1+\frac{\log(V)}{\log(1-(1-\theta)^U)}\right]$$

服从对数分布 $logarithmic(\theta)$，$[x]$ 表示 x 值的整数部分。

（7）如果随机变量 U 服从均匀分布即 $U\sim Unif(0,1)$，则 $X=\alpha+\beta\tan\{\pi(U-0.5)\}$ 服从 $Cauchy(\alpha,\beta)$ 分布。

（8）如果随机变量 U 服从均匀分布即 $U\sim Unif(0,1)$，则 $X=-(\log U)/\lambda$ 服从指数分布 $Exp(\lambda)$。

例 1-5 对数分布的生成，上面介绍过对数分布的生成，这里应用一般变换的方法可以生成对数分布随机变量，如果随机变量 U 和 V 服从相互独立的均匀分布 $Uniform(0,1)$，则

$$X=\left[1+\frac{\log(V)}{\log(1-(1-\theta)^U)}\right]$$

服从对数分布 $logarithmic(\theta)$。该方法提供一个简单、有效的对数分布随机变量的方法。具体 R 程序如下所示：

```
n <- 1000; theta <- 0.5; u <- runif(n); v <- runif(n) #生
  成对数分布样本
x <- floor(1 +log(v) / log(1 - (1 - theta)^u)); k <- 1:max(x)
p <- -1 / log(1 -theta) * theta^k / k  #理论概率
[1] 0.721 0.180 0.060 0.023 0.009 0.004 0.002 0.001
p. hat <- tabulate(x)/n  #经验概率
[1] 0.723 0.191 0.060 0.018 0.004 0.002 0.001 0.001
```

上面程序中 p 表示理论概率，p. hat 表示经验概率，通过两者的比较分析，说明模拟方法的有效性。下面是更加简洁的对数分布函数随机数 R 函数。

```
rlogarithmic <- function(n, theta) {
  stopifnot(all(theta > 0 & theta < 1))
  th <- rep(theta, length =n)
  u <- runif(n); v <- runif(n)
  x <- floor(1 + log(v) / log(1 - (1 - th)^u))
  return(x)
}
```

例 1-6　如果 $X_i \sim \exp(1)$ 是独立同分布，则可以得到下面的关系式

$$Y_1 = 2\sum_{j=1}^{v} X_j \sim \chi^2_{2v}$$

$$Y_2 = \beta \sum_{j=1}^{\alpha} X_j \sim G(\alpha, \beta)$$

$$Y_3 = \frac{\sum_{j=1}^{a} X_j}{\sum_{j=1}^{a+b} X_j} \sim Beta(a, b)$$

上面三个指数变换表达式，分别表示生成卡方分布、伽玛分布和贝塔分布。假设

```
U =runif(3* 10^4)
U =matrix(data =U,nrow =3) #用于矩阵求和
X = -log(U) #由均匀分布生成指数分布
X1 =2* apply(X,2,sum) #求和得到卡方分布 χ²₆,其理论均值为 6.
```

```
mean(X1)
[1] 5.988547
X2 <-0.5* apply(X,2,sum) #求和得到伽玛分布 G(3,0.5)
mean(X2)
[1] 1.497137
U1 =runif(10^4);Y = -log(U1)
Y =Y +X1* 0.5;X3 <- (X1* 0.5)/Y #得出贝塔分布 Beta(3,1)
mean(X3)
[1] 0.7496498
```

例 1-7 生成均值参数 $\lambda=100$ 泊松分布随机数，因为 $P(X<70)+P(X>130)=0.00268$，所以主要取值在 $\lambda\pm3\sqrt{\lambda}$ 区间内，即（70，130）。所以生成的随机数可以限定在值域（70，130）中，具体 R 程序如下所示：

```
n =10^4; lambda =100;spread =3* sqrt(lambda)
t =round(seq(max(0,lambda -spread),lambda +spread,1)) #
  随机数生成值域
prob =ppois(t, lambda);X =rep(0,n)
for (i in 1:n){
   u =runif(1)
   X[i] =t[1] +sum(prob <u) }
Mean(X)
[1] 99.9546
```

1.4 混合分布

在很多情况下，一个分布可以表示成混合分布的形式，比如：

$$f(x)=\int_y g(x\mid y)p(y)dy \text{ 或 } f(x)=\sum_{i\in y}p_i f_i(x)$$

分别表示连续分布和离散分布的情况。

如果随机变量 X 是一个离散混合分布，其分布函数表示成加权之和的

形式 $F_X(x)=\sum p_iF_{X_i}(x)$，这里 X_1，X_2，…表示随机变量，$p_i>0$ 表示权重或混合概率，并且具有关系 $\sum_i p_i=1$。如果随机变量 X 是一个连续混合分布，其分布函数表示成 $F_X(x)=\int_{-\infty}^{\infty}F_{X\mid Y=y}(x)f_Y(y)dy$，其中 $f_Y(y)$ 表示权重函数，且 $\int_{-\infty}^{\infty}f_Y(y)dy=1$。

例1-8　假设两随机变量 $X_1\sim Gamma(2,2)$ 和 $X_2\sim Gamma(2,4)$ 相互独立，比较分析卷积分布 $S=X_1+X_2$ 和混合分布 $F_X=0.5F_{X_1}+0.5F_{X_2}$。通过下面R程序进行比较分析：

```
n <- 1000 #样本个数
x1 <- rgamma(n, 2, 2); x2 <- rgamma(n, 2, 4) #生成伽玛分布模拟值
s <- x1 + x2 #卷积分布
u <- runif(n); k <- as.integer(u > 0.5) #生成0与1组成的向量
x <- k * x1 + (1 - k) * x2 #生成混合分布
par(mfcol = c(1,2)); hist(s, prob = TRUE); hist(x, prob = TRUE)
```

从图1-3中的直方图可以看出卷积分布 S 和混合分布 F_X 明显不同，上面的例子只考虑了两个分布的简单混合，下面考虑更多分布的混合模型。

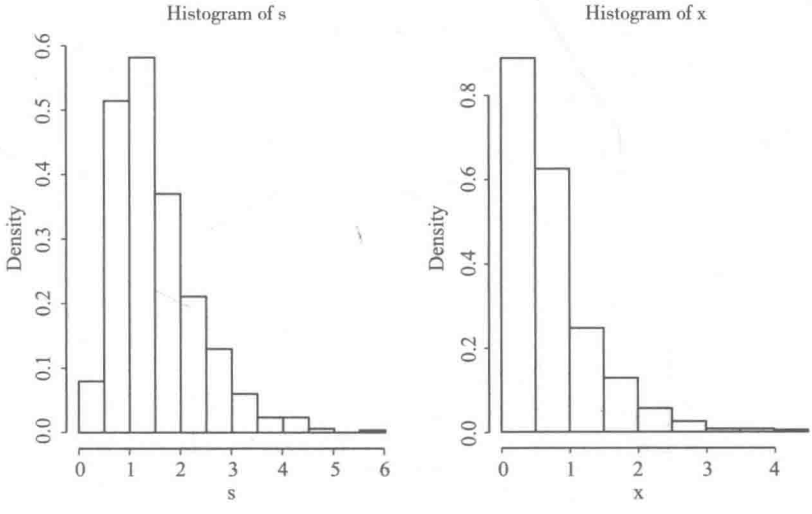

图1-3　分布比较图