贝叶斯统计分析及其应用

韦程东　著

科学出版社

贝叶斯统计分析及其应用

韦程东 著

科学出版社
北京

内 容 简 介

本书在独立随机变量条件下运用贝叶斯理论对在自然科学、工程技术、医学、环境科学、保险精算学、经济等诸多学科起重要作用的 Rayleigh 分布、Pareto 分布、Poisson 分布、Burr 分布等进行深入研究. 独立性假设在某些时候是合理的，但要验证一个样本的独立性是非常困难的，而在许多理论和实际问题中，随机变量不一定是独立的，所以，本书选择新颖的研究视角与相关的分布，把建立在样本独立同分布基础上的贝叶斯统计推断理论推广到混合随机变量上，丰富了贝叶斯统计学理论.

本书可作为统计学、经济学、金融工程及数学与应用数学本科高年级学生和研究生的教材，也可供从事统计学、经济学、金融工程及数学科研工作者学习参考.

图书在版编目(CIP)数据

贝叶斯统计分析及其应用 / 韦程东著. —北京：科学出版社，2015
ISBN 978-7-03-043498-2

Ⅰ.①贝… Ⅱ.①韦… Ⅲ.①贝叶斯统计量 Ⅳ.①O212.8

中国版本图书馆 CIP 数据核字（2015）第 033302 号

责任编辑：李静科 / 责任校对：张凤琴
责任印制：张倩 / 封面设计：陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

双青印刷厂印刷

科学出版社发行 各地新华书店经销

*

2015 年 2 月第 一 版 开本：720×1000 1/16
2015 年 2 月第一次印刷 印张：14 3/4
字数：297 000

定价：78.00 元

（如有印装质量问题，我社负责调换）

前言

贝叶斯统计推断理论源于英国学者贝叶斯(Thomas Bayes)于1763年在英国国家学会哲学学报上发表的论文*An essay towards solving a problem in the doctrine of chances*,文中提到通过参数先验分布和似然函数的合并可获得参数的后验分布,即贝叶斯公式.尽管从现在的角度看,该文内容相当简单,但是Cox和Hinckley指出:它提供了统计推断的一种新思维.该方法提出后,由于理论上的不完善和应用中出现的问题,导致了该方法长期未被接受.直到第二次世界大战后,随着贝叶斯统计在决策理论和信息论上的发展,才使其方法和理论得以推进.20世纪50年代以后,通过De Finetti、Savaga、Raiffa、Schlaifer、Jeffreys、Good等统计学家大量开拓性的研究工作,贝叶斯统计推断理论获得了迅速发展和完善.而召开有组织的会议和建立相应的协会更加推动了贝叶斯统计的传播.具有标志性的进展是从1979年开始,每隔四年召开一次的国际贝叶斯统计会议(巴伦西亚会议);另一个是在1992年创建的国际贝叶斯分析(International Society of Bayesian Analysis, ISBA).马尔可夫链蒙特卡罗(Markov Chain Monte Carlo,MCMC)方法在贝叶斯分析中的应用,使原先异常复杂的数值计算问题迎刃而解,参数后验分布的模拟也更为方便,贝叶斯统计理论在生物统计、医学统计、可靠性工程、经济预测与决策等诸多领域中均获得了广泛的应用.

目前,贝叶斯统计学派已经成为了与经典统计学派并驾齐驱的当今两大学派之一,著名统计学家Lindely认为21世纪将是贝叶斯统计的世界.贝叶斯统计学如此之重要,1763年发表的*An essay towards solving a problem in the doctrine of chances*被视为贝叶斯统计学的基础,因该论文从发表至今恰好为250周年,美国统计协会、数理统计学会、国际生物计量学学会、国际生物统计学会、伯努利学会和皇家统计学会联合宣布2013年为“国际统计年”.国际统计年得到了全世界1700多个组织的支持.作者于20世纪90年代末开始接触贝叶斯统计学,呈现在大家面前的《贝叶斯统计分析及其应用》一书是作者对贝叶斯统计学的一些点滴感悟,它恰逢在国际统计年中诞生,犹如大海中的一朵小小浪花,但愿它的出版对统计学的发展有所裨益.

本书的出版得到了广西“十二五”博士点建设经费、广西自然科学基金项目经费、混合与缺失数据统计分析广西高校重点实验室建设经费、广西高校科学研究项目经费的资助,在此表示衷心感谢.

作者的研究生陈志强、韦师、吴文俊、吕孝亮、韦莹莹、薛婷婷、程艳琴、邓立凤、苏韩、仲崇刚、邱燕、胡莎莎、王弗、魏换其、杨敏、庞夏等为本书部分章节的撰写、修改做了大量的工作，在此表示衷心的感谢.在本书的撰写过程中，作者还参考了有关的文献，吸取了其中的经验，并引用了其中的一些材料与数据，在此，谨向这些文献作者表示衷心的感谢.

由于作者水平有限，不妥之处在所难免，恳请同行与广大读者批评指正.

韦程东

2013 年 1 月

目　　录

第二篇 独立样本下的贝叶斯估计

第三篇 相依样本下的贝叶斯估计

第四篇　贝叶斯检验问题

第一篇　贝叶斯分析基础

第 1 章　Bayes 统计推断

1.1　先验分布与后验分布

近几十年以来，统计学中的 Bayes 学派有了重大的发展，如今已成为与经典学派并驾齐驱的两大统计学派之一. Bayes 统计得名于英国学者 T. Bayes，Bayes 统计的理论与方法由其论文 *An essay towards solving a problem in the doctrine of chances* 发展而成.

经典统计的出发点是根据样本在一定的统计模型下作出统计推断. 本章假设统计模型为参数统计模型. 在取得样本观测值 x 前，往往对参数统计模型中的参数 θ 有某些先验知识；在数学上，关于 θ 的先验知识的数学描述就是先验分布. Bayes 统计的主要特点就是使用先验分布，而在得到样本观测值 $x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 后，由 x 与先验分布提供的信息，得到后验分布. 这一后验分布综合了样本与先验信息，组成较完整的后验信息，是 Bayes 统计推断的基础. 经典统计对样本量较大的样本，有较好的统计推断效果. Bayes 推断由于利用了先验知识，因而对小样本一般也有较好的统计推断效果.

1.1.1　Bayes 统计模型

设事件 $A_1,A_2,\cdots,A_n$ 构成互不相容的事件完备组，概率论中的 Bayes 公式为

$$P(A_i\,|\,B)=\frac{P(B\,|\,A_i)P(A_i)}{\sum\limits_{j=1}^{n}P(B\,|\,A_j)P(A_j)} \tag{1.1.1}$$

这时，先验信息以 $\{P(A_j);j=1,2,\cdots,n\}$ 这一概率分布给出，即先验分布. 由于事件 B 的发生，可以对 $A_1,A_2,\cdots,A_n$ 发生的概率重新估计，以后验分布

$$\{P(A_i\,|\,B);i=1,2,\cdots,n\}$$

体现出来，Bayes 公式反映了先验分布向后验分布的转化. 下面通过一例说明 Bayes 统计推断的特征.

例 1.1.1　设有金、银、铜三种盒子，其中金盒 5 个，银盒 4 个，铜盒 3 个. 每个盒子里放了红、黄、蓝、白四种球，个数为金盒：红 70，黄 20，蓝 8，白 2；银盒：红 10，黄 75，蓝 3，白 12；铜盒：红 5，黄 12，蓝 80，白 3. 从这 12 个盒子中随机抽一个盒子，再从这一个盒子里随机地抽一个球. 问：如何从抽出的球的颜色推断它所来自的盒子的材料.

解　利用 Bayes 公式(1.1.1)，求得

P(金|红)=70/81,　　P(银|红)=8/81,　　P(铜|红)=3/81

P(金|黄)=25/109,　　P(银|黄)=75/109,　　P(铜|黄)=9/109

P(金|蓝)=10/73,　　P(银|蓝)=3/73,　　P(铜|蓝)=60/73

P(金|白)=10/55,　　P(银|白)=36/55,　　P(铜|白)=9/55

利用上述计算结果可对上述问题作统计推断.比如,若取到的球为黄色,从后验分布

P(金|黄)=25/109,　P(银|黄)=75/109,　P(铜|黄)=9/109

可以看出,P(银|黄)=75/109 最大.故应作下列推断:"黄→银",即取到黄球应推断盒子为银.同理,可作下列推断:

"红→金";　"蓝→铜";　"白→银"

在例 1.1.1 中,可引进参数 θ 与样本 X. θ 是盒子材料所对应的数,如令:θ(金)=1,θ(银)=2,θ(铜)=3,参数空间 $\Theta=\{1,2,3\}$.又设 X 是球的颜色所对应的数,如令:X(红)=1,X(黄)=2,X(蓝)=3,X(白)=4.样本空间 $\aleph=\{1,2,3,4\}$.上述问题归结为由样本 X 推断 θ,这符合一般统计推断问题的提法.但此例中 θ 是随机变量,且知道了 θ 的分布,即

$$P(\theta=1)=5/12,\quad P(\theta=2)=4/12,\quad P(\theta=3)=3/12$$

因为参数 θ 为随机变量,所以本例中的样本分布族为条件概率分布 $\{P(x|\theta:\theta\in\Theta)\}$.依上述,有

$$\{P(1|2),P(2|1),P(3|1),P(4|1)\}=\{70/100,20/100,8/100,2/100\}$$

$$\{P(1|2),P(2|2),P(3|2),P(4|2)\}=\{10/100,75/100,3/100,12/100\}$$

$$\{P(1|3),P(2|3),P(3|3),P(4|3)\}=\{5/100,12/100,80/100,3/100\}$$

设 $\hat{\theta}$ 为 θ 的估计量,按上述符号,有

$$\hat{\theta}(1)=1,\quad \hat{\theta}(2)=2,\quad \hat{\theta}(3)=3,\quad \hat{\theta}(4)=2$$

例 1.1.1 体现了 Bayes 统计模型的思想.现在阐述 Bayes 参数统计模型.首先,Bayes 参数统计模型中的参数 θ 是参数空间 Θ 上的随机变量,它的概率分布称为参数 θ 的先验分布.又因 θ 是随机变量,因此参数统计模型中,样本 $X=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 的分布族,即样本分布族应理解为条件分布族 $\{F(x|\theta):\theta\in\Theta\}$.在连续型或离散型两种情况,样本分布族皆可以表为条件密度函数族 $\{f(x|\theta):\theta\in\Theta\}$,其中 $f(x|\theta)$ 为样本密度函数.

定义 1.1.1　(1) 参数 θ 的参数空间 Θ 上的一个概率分布称为 θ 的先验分布,其(连续或离散)密度记为 $\{\pi(\theta):\theta\in\Theta\}$.

(2) 样本 $X=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 的条件密度函数 $\{f(x|\theta):\theta\in\Theta\}$(连续或离散)称为样本分布族.

(3) 先验分布 $\{\pi(\theta):\theta\in\Theta\}$ 与样本分布族 $\{f(x|\theta):\theta\in\Theta\}$ 构成 Bayes 参数统计模型.

Bayes 统计模型的特点是将参数 θ 视为随机变量，并具有先验分布 $\pi(\theta)$. 将 θ 视为随机变量，在很多场合是合理的. 如某厂某产品的废品率 p，在较长期间会有一些随机波动. 若有相当长的逐日废品频率记录，就可以确定 p 的先验分布. 而在有些情况下，将参数 θ 看成随机变量似乎是不自然的. 如要估计某个范围内确定的铁矿含铁百分率. 此时 θ 为一个未知常数，这时可以根据已开采的类似铁矿的经验选择一个 θ 的先验分布. Bayes 学派与经典学派的分歧主要是关于参数的认识上的分歧，经典学派视 θ 为未知常数，Bayes 学派则视 θ 为随机变量且具有先验分布，两个学派分歧的根源在于对于概率的理解. 经典学派视概率为事件大量独立重复试验频率的稳定值；而 Bayes 学派则视 θ 为主观概率，将事件的概率理解为认识主体对事件发生的相信程度，当然对于可以独立重复试验的事件，概率仍可视为频率稳定值. 显然，将 θ 视为随机变量且具有先验分布具有实际意义，能拓广统计学应用的范围.

1.1.2　后验分布

给定了 Bayes 参数统计模型，即给定了先验分布 $\{\pi(\theta):\theta\in\Theta\}$ 与样本分布族 $\{f(x|\theta):\theta\in\Theta\}$. 这里，$\pi(\theta)$ 也视为密度函数（连续或离散）. 给定了这样的 Bayes 统计模型，就可以确定 (θ,X) 的联合分布. 以 θ,X 皆为连续型分布说明. 这时 (θ,X) 的联合密度函数为

$$f(x|\theta)\pi(\theta) \tag{1.1.2}$$

又 X 的边缘密度函数为

$$q(x)=\int_{\Theta} f(x|\theta)\,\mathrm{d}\theta \tag{1.1.3}$$

由此可见，样本分布 $f(x|\theta)$ 与 θ 有关，而边缘分布(1.1.3)式是样本分布按先验分布的“平均”，与 θ 无关. 有了 (θ,X) 的联合分布与 X 的边缘分布，可以求得已知 $X=x$ 的条件下 θ 的分布. 在 $X=x$ 时，θ 的条件密度函数为

$$h(\theta|x)=\frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)\,\mathrm{d}\theta} \tag{1.1.4}$$

(1.1.4)式也称为 Bayes 公式. 当 θ,X 为各种类型的分布时，可得到各种情况下的 Bayes 公式. 例如，当 X 为连续型，θ 为离散型时，有

$$h(\theta|x)=\frac{f(x|\theta)\pi(\theta)}{\sum f(x|\theta_i)\pi(\theta_i)} \tag{1.1.5}$$

可看出(1.1.4)，(1.1.5)式的分母只与 x 有关，而与 θ 无关.

$h(\theta|x)$ 反映了得到样本观测值 x 后（即 $X=x$），θ 取各种可能值概率大小的新认识，称为 θ 的后验分布. $h(\theta|x)$ 称为后验密度函数.

定义 1.1.2　在 $X=x$ 的条件下，θ 的条件分布称为 θ 的后验分布. 后验分布由后验密度函数 $\{h(\theta|x),\theta\in\Theta\}$ 描述.

后验分布的意义在于综合了关于θ的先验信息(反映在先验分布$\pi(\theta)$中)和样本x关于θ的信息(反映在样本分布$f(x|\theta)$中).先验分布概括了在试验前对θ的认识,而得到样本观测值x后,对θ的认识有了深化,这集中反映在后验分布中.Bayes公式反映了先验分布到后验分布的转化,即Bayes自己所说的"归纳推理"的统计方法.

若随机变量X的密度函数为$f(x)=cg(x)$,其中c为与x无关的数,则可记为$f(x)\propto g(x)$,$g(x)$称为密度函数$f(x)$的核.如三个常见分布的核为

正态分布$N(\mu,\sigma^2)$:$\exp\left[-\dfrac{1}{2\sigma^2}(x-\mu)^2\right]$;

二项分布$b(n,p)$:$\dbinom{n}{x}\left(\dfrac{p}{1-p}\right)^x$;

B分布$\mathrm{B}(a,b)$:$x^{a-1}(1-x)^{b-1}$.

当给定样本,记$X=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$,则样本密度$f(x|\theta)=f(x_1,x_2,\cdots,x_n|\theta)$.此即为似然函数,现记为$L(\theta|x)$.在(1.1.2)式中,分母$\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta$与$\theta$无关(可能与$x$有关),故有

$$h(\theta|x)\propto\pi(\theta)L(\theta|x) \tag{1.1.6}$$

或

$$h(\theta|x_1,x_2,\cdots,x_n)\propto\pi(\theta)L(\theta|x_1,x_2,\cdots,x_n) \tag{1.1.7}$$

在Bayes统计中,$h(\theta|x)$采取这样的表达形式运算方便.(1.1.6)式表明:先验信息包含在$\pi(\theta)$中,样本信息包含在$L(\theta|x)$中,它们结合起来得到后验信息,包含在$h(x|\theta)$中.(1.1.7)式可以描述当多个观测值相继得到时,关于θ的信息的不断更新.设$X=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$为独立同分布(i. i. d.)样本.当得到第一个观测值时,有

$$h(\theta|x_1)\propto\pi(\theta)L(\theta|x_1)$$

当得到第二个观测值x_2时,有

$$h(\theta|x_1,x_2)\propto\pi(\theta)L(\theta|x_1,x_2)=\pi(\theta)L(\theta|x_1)L(\theta|x_2)\propto h(\theta|x_1)L(\theta|x_2)$$

一般地有,对$m=2,\cdots,n$,

$$h(\theta|x_1,x_2,\cdots,x_m)\propto h(\theta|x_1,x_2,\cdots,x_{m-1})L(\theta|x_m) \tag{1.1.8}$$

其中$h(\theta|x_1,x_2,\cdots,x_{m-1})$包含了$x_1,x_2,\cdots,x_{m-1}$的先验信息,新观测值$x_m$提供的信息包含在$L(\theta|x_m)$中,它们组合起来构成$h(\theta|x_1,x_2,\cdots,x_m)$.可以看出:所谓"先验"与"后验"是相对的.对$h(\theta|x_1,x_2,\cdots,x_m)$而言,$h(\theta|x_1,x_2,\cdots,x_m)$相当于"先验分布".

1.1.3 Bayes统计推断原则

(1.1.6)式是Bayes统计的基本公式.先验分布概括了试验前对θ的认识.而得