第1章 贝叶斯统计基本概念

俗话说，万事开头难。为了提高读者的学习兴趣，本章从一个贝叶斯统计的真实应用开始，介绍贝叶斯统计的基本概念和公式，概述贝叶斯统计学的历史和发展趋势以及与经典统计学的比较。

1.1 引言

1.1.1 一个美国书呆子的故事

在2012年美国总统大选期间，一个一直都被人称作“书呆子”的美国人纳特·西尔弗(Nate Silver，生于1978年1月13日)用以统计为主要工具的模型准确预测了美国全部50个州的选举结果。在大选日当天早晨，他的模型最新预测到时任总统巴拉克·奥巴马(Barack Obama)将有90.9%的可能获得多数选举人票从而连任，而选举结果确确实实就是奥巴马总统赢得了这次美国总统大选。于是，他凭借自己的模型及其准确的预测打败了所有时事政治记者、政党媒体顾问和政治评论员。“你们知道谁是今晚(大选日当夜)的赢家吗?”美国全国广播公司新闻节目主播自问自答，“是纳特·西尔弗”。其实，早在2008年的美国总统大选期间，西尔弗就准确预测了整个美国50个州中49个州的选举结果。两次极为准确的预测，让这个“书呆子”扬眉吐气、名声大震，各种荣誉接踵而来，甚至于被四所大学授予了四个荣誉博士学位，当然这也让我们从事统计领域的人士大感骄傲。西尔弗的预测模型有什么神秘之处呢？答案就是其利用了大数据和我们将要学习的贝叶斯统计理论和方法。

1.1.2 贝叶斯统计简史

贝叶斯统计学是以英国人托马斯·贝叶斯(Thomas Bayes，1702—1761)的名字命名

的。贝叶斯是一位英国牧师，但他却热衷于概率统计等科学研究，还是英国皇家学会会员。遗憾的是，现在人们对他的生平却知之甚少，甚至没有人知道贝叶斯的相貌如何，现存所有他的画像都是传说，并不能证实是他的真容。贝叶斯统计学起源于贝叶斯逝世后公开发表的一篇论文——《论一个概率理论问题的求解》(*An Essay Towards Solving a Problem in the Doctrine of Chances*)。在贝叶斯去世两年之后，这篇论文由他的朋友理查德·普莱斯(Richard Price)介绍到英国皇家学会，引起了该学会的注意和讨论，并于1763年发表在《皇家学会哲学会刊》上。在该篇论文中，贝叶斯首次提出了贝叶斯统计的基本思想和归纳推理方法。

五十一年后，法国数学、统计学、天文学和物理学家拉普拉斯(P. S. Laplace，1749—1827)在1814年出版了著作《关于概率的哲学评述》(*A Philosophical Essay on Probabilities*)，在该著作中他将贝叶斯提出的公式进行了推广并导出了一些很有意义的新结果。然而，之后相当长的一段时间里虽然有一些理论和应用研究，但由于其理论与经典统计学相比显得另类，而且人们对它的理解还不够深刻，在应用上其计算复杂且计算量巨大，因此贝叶斯统计理论和方法长期未被普遍接受，甚至被一些学者看作一种旁门左道。直到20世纪中叶开始，有一批统计学家，例如杰弗里斯(H. Jeffreys，1939)、萨维奇(L. J. Savage，1954)、雷法和施莱弗(H. Raiffa and R. Schlaifer，1961)以及伯杰(J. O. Berger，1985)等，才对贝叶斯统计做了更加深入的研究，特别是罗马尼亚(匈牙利)裔美国统计学家阿布拉汉·瓦尔德(Abraham Wald，1939，1950)通过将损失函数引入统计学并利用决策概念和思想把经典统计推断纳入决策理论框架中而形成了统计决策理论，这样经典统计学和贝叶斯统计学通过决策理论有机地联系到了一起，才得到了很有意义的理论结果。从20世纪中叶开始，在一批学者的努力下，人们对贝叶斯统计在观点、方法和理论上的认识不断加深。从20世纪90年代以来，伴随着计算机科学技术的发展和有效的贝叶斯统计计算方法的发现和应用，贝叶斯统计解决了相当一批经典统计难以解决的实际问题，从而得到了人们极大的重视。现在，贝叶斯理论和方法获得了人们的普遍接受，贝叶斯统计不仅在统计学本身而且在众多学科中都得到了广泛的应用，解决了各个不同学科中大量的复杂统计问题。贝叶斯统计表现出了勃勃生机和欣欣向荣的景象，在统计学领域牢牢地站稳了一席之地，也成为现代统计学的重要分支，可以这么说，没有学习过贝叶斯统计，就不能说了解过现代统计学。

1.1.3 经典统计方法

我们先来回顾一下经典统计学的思想方法，以便与下一小节的贝叶斯统计思想方法进行比较。回顾一下概率统计课程中概率的定义，便容易明白经典统计学思想方法也就是“频率方法”，它把概率定义为频率的极限，也就是说如果随着随机试验重复次数的增多，随机事件发生的频率会稳定在一个常数附近，这个常数就是该随机事件发生的概率。同时，它认为总体的数字特征(如均值、方差)和别的参数仅仅是未知的常数，可以用样本统计量来估计。而且，它又认为样本是随机变量，从而样本统计量也是随机变量，因此具有概率分布，即它的抽样分布。如果统计量的分布可以求出，利用该分布，就可以进行区

间估计和假设检验等统计推断。然而，我们知道寻求统计量的概率分布和进行区间估计以及假设检验等都不是容易的事，而且参数的区间估计既不容易理解也不容易解释。

1.1.4　贝叶斯统计方法

贝叶斯统计学虽然也认可经典统计学的概率定义，但它同时把概率理解为人对随机事件发生可能性的一种信念（有时被称为“可信度”），当然，这种信念不是信口开河，而是基于学识和经验之上的审慎度量。其次，贝叶斯统计把任意一个未知量（参数）都看作一个随机变量，可用一个概率分布去描述它。我们说这种观点是合理的，因为即使是一个确定性的未知量，也可以把它看成随机变量的特殊情形，即服从 0—1 分布的随机变量。所以说，任一个未知量都可用一个适当的概率分布去描述它。这个概率分布利用历史数据或其他历史信息或研究人员的经验和学识而确定，称为该未知量（参数）的**先验分布**。而后利用新样本信息（即抽样信息）对先验分布进行更新，更新之后的这个新概率分布称为该未知量的**后验分布**。由此，未知参数的点估计、区间估计和假设检验等统计推断都是基于后验分布来进行的，而且参数的区间估计既容易理解也容易解释，假设检验则简单明了。

经典统计学把概率定义为频率的极限，初看起来似乎客观、严谨，但是在现实世界中要进行重复试验需要花费大量的人力、物力，而且有时根本无法重复，例如，我们无法重复昨天的天气和去年的经济活动。因此，用频率的极限来定义概率在实际应用中受到了极大的限制。相反，贝叶斯统计把概率理解为人对随机事件发生可能性的信念，则在实际应用中没有任何限制，因为它不需要重复，事件甚至可以一次都没有发生。而且，在贝叶斯统计中一旦后验分布建立起来了，所有的统计推断都是基于后验分布来进行的，因此，至少从理论上而言，贝叶斯统计推断比经典统计推断要简单明了得多。当然，现代统计学的发展趋势是，根据实际问题的条件和需要挑选经典统计方法或贝叶斯统计方法，有时甚至是综合利用这两种统计理论和方法进行统计推断。所以，不管是经典统计还是贝叶斯统计，能够解决问题的就是“好统计”！

对于经典统计学与贝叶斯统计学的比较，有待学完本书的内容后才能有更深刻的体会，因此希望读者在研读完本书后，再好好对它们做一个详细的比较分析。

1.2　概率空间与随机事件贝叶斯公式

1.2.1　概率空间与随机事件贝叶斯公式

我们从概率论知道概率空间是三位一体的一个研究对象(Ω, F, P)，其中 Ω 是样本点全体，也称为样本空间；F 是事件域（简单说就是所要研究的随机事件全体，包含必然事件 Ω 和不可能事件 Φ）；P 是定义在事件域 F 上的概率（测度），满足以下三条公理：

（1）非负性：对于任意事件 A，其概率 $P(A)\geqslant 0$；

（2）规范性：必然事件 Ω 的概率等于 1，即 $P(\Omega)=1$；

(3) 可列可加性：如$\{A_i\}_{i=1}^{\infty}$是一列事件，满足$A_iA_j=\Phi(i\neq j)$(称为两两互不相容)，则

$$P(\bigcup_{i=1}^{\infty}A_i)=P(\sum_{i=1}^{\infty}A_i)=\sum_{i=1}^{\infty}P(A_i)$$

这一公理体系称为**柯尔莫哥洛夫概率论公理体系**，是苏联著名数学家柯尔莫哥洛夫于1933年建立的，得到了概率统计学者们的广泛认可，从而为概率论建立了坚实的理论基础。

另外，对于任意两个事件A,B且$P(A)>0$，定义在A发生的条件下，B发生的条件概率为

$$P(B|A)=\frac{P(AB)}{P(A)}$$

从而，$P(AB)=P(A)P(B|A)$，这就是乘法公式。推而广之，设$\{A_k\}_{k=1}^{n}$是任意n个随机事件，则有更一般的乘法公式

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$$

现设$\{A_i\}_{i=1}^{\infty}$是事件域F中的一列事件，若$\bigcup_{i=1}^{\infty}A_i=\Omega$，且$A_iA_j=\Phi(i\neq j)$，则称$\{A_i\}_{i=1}^{\infty}$为$\Omega$的一个划分(也称为$\Omega$的完全事件组，这里事件的个数也可以是有限多个，比如说n个，这相当于$k>n$时都有$A_k=\Phi$)。显然，任一个事件A与其补$\overline{A}$就是Ω的一个划分。现在设$\{A_i\}_{i=1}^{\infty}$为Ω的一个划分且$P(A_i)>0$，则对任一个事件$B\in F$有全概率公式

$$P(B)=\sum_{i=1}^{\infty}P(A_i)P(B|A_i)$$

事实上，由

$$B=B(\bigcup_{i=1}^{\infty}A_i)=\bigcup_{i=1}^{\infty}(A_iB)\ \text{且}(A_iB)\cap(A_jB)=(A_iA_j)B=\Phi,i\neq j$$

利用可列可加性及乘法公式就得

$$P(B)=P(\bigcup_{i=1}^{\infty}A_iB)=\sum_{i=1}^{\infty}P(A_iB)=\sum_{i=1}^{\infty}P(A_i)P(B|A_i)$$

现在将全概率公式以及乘法公式应用到条件概率$P(A_j|B)$的公式上就有

$$P(A_j|B)=\frac{P(A_jB)}{P(B)}=\frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{\infty}P(A_i)P(B|A_i)}\quad j=1,2,\cdots,n,\cdots$$

这就是著名的**随机事件形式的贝叶斯公式(定理或法则)**，也称为逆概率公式，这里$\{A_j\}$可以认为是事件B发生的所有可能的原因，而贝叶斯公式就是计算在已知事件B发生的条件下每个原因的可能性大小(概率)，也就是说由结果去推测原因，因此叫逆概率公式。在贝叶斯公式中，$P(A_j)$称为A_j的先验概率，因为这是事先已知的，而$P(A_j|B)$自然称为A_j的后验概率。

1.2.2 两例：她怀孕了吗？“非典”时期病人为何要测量体温？

贝叶斯公式与全概率公式都是概率论中的著名公式，在许多学科中都有重要应用，下

面我们来看两个例子。

例 1.1　(她怀孕了吗?)根据历史资料知道:女性一次性交后怀孕的概率为 15%。假如一个女性某次性交后怀疑自己怀孕了,但又不能确定。于是,她做了个准确率为 90%的验孕测试,即 90%的怀孕案例会给出阳性反应的检验结果,同时知道该测试当未怀孕时阳性反应占 10%。她当然想知道在检验结果为阳性的条件下的怀孕概率。然而,她不懂贝叶斯统计,所以请你帮助她算出该概率。

解　已知

$$P(\text{怀孕})=0.15,P(\text{检测阳性}|\text{怀孕})=0.90,P(\text{检测阳性}|\text{未怀孕})=0.10$$

由已知得,$P(\text{未怀孕})=0.85$。由贝叶斯公式知在检验结果为阳性的条件下的怀孕概率:

$$\begin{aligned}P(\text{怀孕}|\text{检验阳性})&=\frac{P(\text{检验阳性}|\text{怀孕})P(\text{怀孕})}{P(\text{检验阳性}|\text{怀孕})P(\text{怀孕})+P(\text{检验阳性}|\text{未怀孕})P(\text{未怀孕})}\\&=\frac{0.90\times0.15}{0.90\times0.15+0.10\times0.85}=\frac{0.135}{0.135+0.085}=0.614\end{aligned}$$

这里 $P(\text{怀孕})=0.15$ 就是怀孕的先验概率,$P(\text{怀孕}|\text{检验阳性})=0.614$ 就是怀孕的后验概率,它是在观察数据(阳性测试)后怀孕概率的更新,表明如果测验呈阳性,则怀孕的可能性大大提高。

例 1.2　(非典时期病人为何要测量体温?)“非典(SARS)”患者的主要病症表现为发热、干咳。根据某地区历史资料,已知人群中既发热又干咳的病人患“非典”的概率为 5%;仅发热的病人患“非典”的概率为 3%;仅干咳的病人患“非典”的概率为 1%;无上述病症而患“非典”的概率为 0.01%;现对该区 25 000 人进行检查,发现其中既发热又干咳的病人为 250 人,仅发热的病人为 500 人,仅干咳的病人为 1 000 人,试求:

(1) 该地区中某人患“非典”的概率;

(2) “非典”患者是仅发热的病人的概率。

解　引入记号

$$\begin{aligned}&A=\{\text{既发热又干咳的病人}\},B=\{\text{仅发热的病人}\},\\&C=\{\text{仅干咳的病人}\},D=\{\text{无明显症状的人}\},\\&E=\{\text{“非典”患者}\}\end{aligned}$$

易知 A,B,C,D 构成了一个划分。根据对该区 25 000 人进行检查的结果,有

$$P(A)=\frac{250}{25\,000},P(B)=\frac{500}{25\,000},P(C)=\frac{1\,000}{25\,000},$$

$$P(D)=\frac{25\,000-(250+500+1\,000)}{25\,000}=\frac{23\,250}{25\,000}$$

由全概率公式得患“非典”的概率:

$$\begin{aligned}P(E)&=P(A)P(E|A)+P(B)P(E|B)+P(C)P(E|C)+P(D)P(E|D)\\&=\frac{250}{25\,000}\times5\%+\frac{500}{25\,000}\times3\%+\frac{1\,000}{25\,000}\times1\%+\frac{23\,250}{25\,000}\times0.01\%=0.001\,593\end{aligned}$$

由贝叶斯公式知，“非典”患者是仅发热的病人的概率：

$$P(B|E)=\frac{P(B)P(E|B)}{P(E)}=\frac{\frac{500}{25\ 000}\times 3\%}{0.001\ 593}=0.376\ 647\ 8$$

同理，可以算出“非典”患者是既发热又干咳、仅干咳、无明显症状的病人的概率分别为

$$P(A|E)=\frac{P(A)P(E|A)}{P(E)}=\frac{\frac{250}{25\ 000}\times 5\%}{0.001\ 593}=0.313\ 873\ 2$$

$$P(C|E)=\frac{P(C)P(E|C)}{P(E)}=\frac{\frac{1\ 000}{25\ 000}\times 1\%}{0.001\ 593}=0.251\ 098\ 6$$

$$P(D|E)=\frac{P(D)P(E|D)}{P(E)}=\frac{\frac{23\ 250}{25\ 000}\times 0.01\%}{0.001\ 593}=0.058\ 380\ 41$$

不难看出

$$P(A|E)+P(B|E)+P(C|E)+P(D|E)=1$$

而一个人患“非典”时最可能的症状是发热。这就是为什么在非典时期要测量病人体温的原因。

1.2.3 案例：自动语音识别——神奇的语音输入法

你的手机里安装了讯飞语音输入法或其他语音输入法了吗？是不是觉得它很神奇呢？想不想知道它为什么能够把你说的话转换为文字呢？这个转换过程其实就是自动语音识别。简单地说，自动语音识别是指由机器自动将语音信号转换为文字的方法和过程。人类的语言可以说是各种信息里最复杂和最动态的一种，著名语言学家乔姆斯基(A. N. Chomsky)和信息论的祖师爷香农(C. Shannon)等学者都关注过自动语音识别问题，然而那时自动语音识别并没有获得很大进展。在这个领域率先取得突破的是捷克裔美国语音和语言处理大师贾里尼克(F. Jelinek)。从20世纪60年代开始，贾里尼克开创性地将语音识别问题看成一个通信问题，认为语音识别就是根据接收到的信号序列推测说话人实际发出的信号序列(即说的话)和要表达的意思，并且用贝叶斯公式和两个隐含马尔可夫模型建立起统计语音识别系统，把对应的一套模型称为声学模型和语言模型，从而极大地改变了这一领域的研究方向。此外，他还与其他合作者提出了数字通信领域最重要的算法之一——BCJR(L. R. Bahl, J. Cocke, F. Jelinek, J. Raviv, 1974)算法。难能可贵的是，这种统计语音识别系统不但能够识别静态的词库里的语音，而且对动态变化的词库语音具有很好的适应性，即对新出现的词汇，只要这个词已经被高频使用，可用于训练的数据量足够多，系统就能通过训练而正确地识别之。这实际上表明贝叶斯公式对新词汇语音信息有非常好的适应能力。由于本书的性质，这里我们不可能对问题展开详细的讨论，有兴趣者可以去研读有关文献资料。但我们从已经开发出来的语音输入法知道这种统计语音识别系统是非常成功的！

1.3 三种信息与先验分布

在 1.1 节中，我们初步了解到统计学中有两个主要学派：经典统计学派与贝叶斯统计学派。在本节我们将从这两个学派使用的信息种类来讨论它们之间的异同。首先我们来了解统计推断问题中存在的三种信息。

1.3.1 总体与总体信息

我们从已学课程知道统计学中总体就是根据一定的目的和要求所确定的研究对象的全体。例如，如果要统计调查全国大学男生的身高，那么，我们就可以把全国大学男生的集合作为总体，而大学男生身高这个指标就是关于该总体的一个数量，可以用一个符号 X 来标记它。由于在对随机抽出的一个大学男生具体测量之前，并不知道该大学男生的确切身高，而且人的身高是受遗传、营养等随机因素影响而确定的，所以 X 是一个随机变量并且服从某种概率分布。再如，我们要考察一个经济指标 Y（可以把它设想为某一只股票的收益率或一个国家的 GDP），由于受各种各样的随机因素的影响，Y 是一个随机变量，它的所有可能取值就构成了一个总体并且也服从某一种概率分布。由于一个随机变量的概率分布完全刻画了该随机变量的统计规律性，因此，我们实际上甚至可以抽象地把这个随机变量的概率分布看作总体。总体信息就是我们对总体概率分布的了解或知识，一般而言，对总体信息最大的了解是知道总体概率分布所属的分布族，例如，若我们知道总体服从正态分布族 $N(\mu,\sigma^2)$，虽然这时两个参数还是未知的，我们也知道它的密度函数是一条关于总体均值对称的钟形曲线并且它的各阶矩都存在，同时也知道第一个参数 μ 是分布的均值，第二个参数 σ^2 是分布的方差。当然，总体到底服从怎样的概率分布族对一个新研究问题而言通常不得而知，这正是统计学的一个分支——非参数统计所要研究的。显而易见，要获得总体信息往往必须投入大量的人力、物力，例如，美国军队为了获得某种新的电子元件的寿命分布，购买了上万个此种电子元件，做了大量的寿命实验，在获得大量数据后才确认其寿命概率分布是什么。简言之，总体信息非常重要，要获得它虽然不容易但又是必须要做的，因为它是统计推断的基础。

1.3.2 样本信息

为了对所研究的总体有更多的了解，我们必须从总体抽取（观察或收集）一定的样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$，这些样本给我们提供的信息就是样本信息，也称为抽样信息。样本信息两种最重要的表现形式是样本的联合分布与样本统计量的抽样分布，其次是样本对总体特征的各种估计，例如，样本均值、样本方差（标准差）等。样本是统计学（无论是频率学派还是贝叶斯学派）的粮食，没有样本就如同巧妇难为无米之炊一样，做不成统计学上的任何事情，也就没有统计学了。

仅仅基于总体信息和样本信息进行统计推断的统计学理论和方法称为**经典统计学**。它的历史悠久，但大发展却是从 19 世纪末到 20 世纪上半叶。由于统计学家皮尔逊

(K. Pearson, 1857—1936)、费雪(R. A. Fisher, 1890—1962)和奈曼(J. Neyman, 1894—1981)等人的杰出工作,经典统计学理论得到了空前的发展,成为当时统计学的主流。在20世纪下半叶,经典统计学在工业、农业、医学、经济、金融、管理、军事等领域里获得了广泛的应用,并取得了巨大的成功,同时,在这些领域又不断提出新的统计问题,于是又反过来促进了经典统计学的进一步发展。但是,伴随着经典统计学的持续发展与广泛应用,它本身的缺陷与某些方面的矛盾也逐渐暴露出来了。

1.3.3 先验信息与先验分布

所谓先验信息是指在抽样之前对所研究的统计问题的了解或知识,一般说来,先验信息主要来源于研究者的知识和经验以及历史资料(数据),而且常常是零散的,需要进行提炼加工才可以应用。

先验信息是人们对所研究的统计问题长期观察或研究积累起来的重要历史信息,理应加以利用到统计推断中来,以提高统计推断的质量。从后面的章节我们可以看到经典统计学由于忽视了先验信息的使用,有时会导致不合理的结论。关于先验信息在帮助人们进行推断的作用,请看下面有趣的例子。

例 1.3 统计学家萨维奇(L. J. Savage, 1962)曾考察过两个统计实验:

1. 一位常饮奶茶的妇女声称,对于一杯奶茶,她能辨别先倒进杯子里的是茶还是奶。对此做了十次试验,她都正确地说出了。

2. 一位音乐家声称,他能从一页乐谱中辨别出是海顿(Haydn)还是莫扎特(Mozart)的作品。在十次这样的试验中,他都正确辨别了。

现在的问题是被实验者是完全在猜测吗?假如被实验者完全是在猜测,则每次成功的概率为0.5,那么十次都猜中的概率为$2^{-10}=0.000\ 976\ 6$,这是一个很小的概率,是几乎不可能发生的,所以假设"被实验者完全是在猜测"是不对的,被实验者每次成功的概率要比0.5大得多。换句话说,这不是纯粹的猜测了,而是由于这两位被实验者都有丰富的经验,是经验帮助他们做出了正确的判断。由此可见,经验(也就是一种先验信息)在推断中不可忽视,应被善加利用才是正确之举。

例 1.4 (产品质量管理问题)有一句话说得好,"产品质量是企业的生命线"。企业能否生存下去,其产品质量是关键因素之一。我们可以用一个指标来衡量产品质量的高低,那就是不合格品率。为了了解产品的质量,某厂每天都要抽检5件产品,以获得不合格品率θ的估计。经过100个工作日后就积累了大量的数据,通过整理得到表1.1。

表 1.1 产品抽查数据表

不合格品	出现次数	频率
0	94	0.94
1	3	0.03
2	2	0.02

续表

不合格品	出现次数	频率
3	1	0.01
4	0	0.00
5	0	0.00

根据这些历史资料(就是一种先验信息),对过去产品的不合格率就可以构造一个分布,如表 1.2 所示。

表 1.2　不合格品率先验概率分布表

不合格品率 θ	0.0	0.2	0.4	0.6	0.8	1.0
先验概率	0.94	0.03	0.02	0.01	0.00	0.00

从这个分布列表可以看出,不合格品率 θ 大于等于 0.2 的概率:

$$P(\theta \geqslant 0.2)=0.03+0.02+0.01=0.06$$

是一个相当小的数。

对先验信息进行提炼加工获得的分布称为先验分布。在这个例子中,先验分布(表 1.2)综合了该厂过去产品的质量情况。我们看到这个分布的概率绝大部分集中在 $\theta=0$ 附近。因此,该产品可认为是“信得过产品”。如果以后的多次抽检结果与历史资料提供的先验分布是一致或更好的,质检单位就可以按照要求授予其是“免检产品”,或者每月抽检一两次就足够了,这样,就省去了大量的人力、物力。可见先验信息在统计推断及统计应用中是大有用武之地的。当然,如果以后的多次抽检结果与先验分布有较大的区别,那么我们就应该考虑利用新样本对先验分布进行更新,以期获得更符合实际的新分布,这正是贝叶斯统计所要做的重要工作。

基于总体信息、样本信息和先验信息进行统计推断的理论和方法被称为**贝叶斯统计学**。从使用信息的角度来看,它与经典统计学的差别在于是否利用先验信息。贝叶斯学派重视先验信息的收集、挖掘和提炼,并综合先验信息形成先验分布,将其应用到统计推断中来,以提高统计推断的质量。

1.4　一般形式的贝叶斯公式与后验分布

1.4.1　知识准备

首先回忆一下在概率论中有关随机向量和条件分布的几个概念。我们以二维情形为例,设 (X,Y) 是二维随机向量且分布密度为 $f(x,y)$,则 X 和 Y 的边际密度分别是

$$f_X(x)=\int_{\mathbf{R}} f(x,y)\mathrm{d}y, f_Y(y)=\int_{\mathbf{R}} f(x,y)\mathrm{d}x$$

其中,$\mathbf{R}$ 表示实数集,而 Y 在 X 已知的条件密度是

$$f(y \mid x)=\frac{f(x,y)}{f_X(x)}=\frac{f(x,y)}{\int_{\mathbf{R}} f(x,y)\mathrm{d}y}$$

从而又有

$$f(x,y)=f(y|x)f_X(x)=f(x|y)f_Y(y)$$

其次，引入高等数学中的两个重要函数：贝塔函数(β)和伽马函数(Γ)。它们在贝叶斯统计中经常出现，值得记住。它们分别定义如下：

$$\beta(z,w)=\int_0^1 t^{z-1}(1-t)^{w-1}\mathrm{d}t,\Gamma(z)=\int_0^\infty \mathrm{e}^{-t}t^{z-1}\mathrm{d}t$$

它们有两个重要性质

$$\Gamma(z+1)=z\Gamma(z),\beta(z,w)=\frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)}$$

第一个性质表明伽马函数是阶乘 $n!=n\cdot(n-1)!$ 的推广，第二个性质说明贝塔函数和伽马函数密切关联。

最后，引入一个在贝叶斯统计中常用的分布族，即贝塔分布族 Beta(a,b)，其中 $a>0$，$b>0$ 是两个参数。贝塔分布的密度函数如下

$$\beta(x|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},x\in[0,1]$$

并且具有性质

$$\mathrm{Mode}(X)=\frac{a-1}{a+b-2},\mathrm{E}(X)=\frac{a}{a+b},\mathrm{Var}(X)=\frac{ab}{(a+b)^2(a+b+1)}$$

当 $a=b=1$ 时，贝塔分布的密度函数变成

$$\beta(x|a=1,b=1)=1,x\in(0,1)$$

这正是均匀分布 $U(0,1)$的密度，所以均匀分布 $U(0,1)$是一个特殊的贝塔分布。

1.4.2 R 语言与 R 软件包

本书从下一小节开始就要求读者用软件进行统计计算和作图，并把这一要求贯穿全书，目的是通过动手使用软件让读者培养起自己的数据感并体验研读贝叶斯统计的乐趣，从而激发起对贝叶斯统计学的兴趣。

“R you ready for R?”这是国外高校校园里一句时髦的问句，它表明了 R 在国外高校盛行的程度。那么 R 到底是何方神圣而在校园里如此盛行呢？R 是著名的贝尔实验室(Bell Laboratory)的编程语言 S 的实现版，最初的两位设计者是当时任教于新西兰奥克兰大学的 Ross Ihaka 教授和 Robert Gentleman 教授。由他们的名字拼写大家可以看出这套软件系统叫 R 的原因了。现在 R 由其核心团队负责维护和发展，每半年左右会更新一次。R 是用于统计计算和绘图的编程语言和软件环境；R 是一个自由、免费、源代码开放的软件包；R 是一套完整的用于数据处理、统计计算和制图的软件系统。R 的功能还包括：数据的输入、输出以及存储；数组运算(其数组种类丰富，向量、矩阵运算功能尤其强大)。由于全球学者的贡献，R 有成千上万用于不同领域的软件包，但它的基本包为 base，我们可从其官网“镜像”(http://mirrors.xmu.edu.cn/CRAN/)中下载并安装，本书安装的版本是 R-3.3.1-win。由于基本包 base 实际上还包括了 stats 和 graphics 等诸多包，所以安装好 base 后，我们不仅可以进行各种算术计算，而且可以进行通常的统计计算