贝叶斯统计学
Bayesian Statistics

主讲教师: 黄长全 **(Ph.D., CUHK)**

办公室：经济楼 **D111**

Email: **cqhuang@xmu.edu.cn**

辅导教师: 王智博

手机：**18805929103**

贝叶斯统计开课的话

- 历史悠久：R. T. Bayes(1702-1761，见下页)
  P. C. Laplace(1749-1827，见下下页)
- 争论不休：经典学派**VS**贝叶斯学派
- 困难所在：在应用上模型复杂，计算量巨大
- 应用广泛：不但在统计本身而且在许多其它学科上都有重要应用（大数据，人工智能等等）
- 欣欣向荣：电子计算机发展；优良算法发现；近三十年来蓬勃发展

没有学习过贝叶斯统计，就不能说了解现代统计学！

祖师爷：R. T. Bayes(1702-1761)

Reverend Thomas Bayes

P. C. Laplace(1749-1827)

课程要求

- **课堂纪律**：有病有事一律向系里请假，而不是向我请假（没批假的权力）。有系里批准的假条给我，我将没异议，但上课内容你要补上。每次上课都点名，出勤率关系到你的成绩。如果三次无假条而未到堂上课，则本课程按规定不给学分。在课堂上手机一律静音或关机，不然，响一次总分扣五分！
- **学习态度**：强烈的求知欲望（就如同你们在高考时期），否则学不好贝叶斯统计。
- **作业**：每次作业都有登记评分，作为平时成绩的一部分。
- **课堂随机提问**：每个同学以一样的概率被提问。

教材与参考书

- 教材：黄长全(2017) 贝叶斯统计及其R实现，清华大学出版社
- 参考书：1.茆诗松等(2012) 贝叶斯统计(第二版), 中国统计出版社
  2. Karl-Rudolf Koch (2007) Introduction to Bayesian Statistics, 2nd ed, Springer

注： **每位同学写一份贝叶斯统计课程学习计划，两周后（三月四日）提交你的课程学习计划电子版。**

第1章贝叶斯统计基本概念

§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，当然也让我们统计人大感骄傲。西尔弗的预测模型有什么神秘之处呢？那就是利用了大数据和我们将要学习的贝叶斯统计理论和方法。

Nate Silver （生于1978年1月13日）

补充：

读者可能知道对于2016年美国大选Nate Silver也预测不对。之所以如此是因为民调的数据有极大的问题。(1)对于偏远地区，民调的电话访问往往没有涉及，而该次大选这些地区投票非常积极而且投给特朗普居多。(2)不少人口是心非，接受访问时声称会投给希拉里，但在正式投票时其实是投给特朗普。所以得到的样本要么是有偏的（代表性差），要么根本是假样本。如此，用这些所谓的样本，任何预测方法（模型）都不可能做出正确的预测！

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，却热衷于概率统计研究，还是英国皇家学会会员。但是，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才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，于1763年发表在《皇家学会哲学会刊》（1763，Vol. 53, pp370-418）上。在该论文中，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。

Q: 这里Chances是何意？下次上课提问，要说明理由。

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，在应用上又计算复杂且计算量巨大，贝叶斯统计理论和方法长期未被普遍接受,甚至于被看作是一种旁门左道。直到二十世纪中叶，有一小批统计学家例如杰弗里斯（H. Jeffreys，1939）、萨维奇(L. J. Savage，1954)等才对贝叶斯统计做了更加深入的研究，特别是罗马尼亚裔美国统计学家阿布拉汉·瓦尔德(Abraham Wald，(1902-10-31)1902 —1950(1950-12-13))提出了统计决策函数理论，而后才又引起许多人对贝叶斯统计的兴趣，因为该理论把经典统计学与贝叶斯统计学有机地联系到了一起，得到了很有意义的理论结果。这样，从二十世纪中叶开始（特别是二十世纪九十年代以来），在一批学者的努力下，人们先是对贝叶斯统计在观点、方法和理论上不断完善，认识不断加深，而后，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，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，是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。我们回忆一下概率统计课程中概率的定义，就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。其次，它又认为样本是随机变量，从而样本统计量也是随机的，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和假设检验都不是容易的事，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）都看作是一个随机变量，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它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等统计推断都基于后验分布来进行，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，但是在现实世界要进行重复试验要么需要花费大量的人力物力要么根本无法重复，例如，我们无法重复昨天的天气和去年的经济活动，我们也不可能人为的重复2008年汶川大地震。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。其次，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，有时甚至于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!

§1.2 概率空间与随机事件贝叶斯公式

1.2.1 概率空间与事件贝叶斯公式

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$P(A) \geq 0$

（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$A_i A_j = \Phi(i \neq j)$（称为两两互不相容），则

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

另外，对于任意两个事件 A,B 且 $P(A)>0$，则定义在 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

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

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 Ω 的一个划分（也称为完全事件组，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

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

利用可列可加性及乘法公式就得

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

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

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

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？与“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(怀孕)=0.15， P(检测阳性|怀孕)=0.90， P(检测阳性|未怀孕)=0.10

由已知得 P(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

$$P(\text{怀孕}|\text{检验阳性})=\frac{P(\text{检验阳性}|\text{怀孕})P(\text{怀孕})}{P(\text{检验阳性}|\text{怀孕})P(\text{怀孕})+P(\text{检验阳性}|\text{未怀孕})P(\text{未怀孕})}$$

$$=\frac{0.90\times0.15}{0.90\times0.15+0.10\times0.85}=\frac{0.135}{0.135+0.085}=0.614$$

这里 P(怀孕)=0.15 就是怀孕的先验概率， P(怀孕|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？）（2003 年）“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 25000 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 1000 人，试求：（1）该区中某人患“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

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

易知 A，B，C，D 构成了一个划分。根据对该区 25000 人进行检查的结果，有

$$P(A)=\frac{250}{25000},\ P(B)=\frac{500}{25000},\ P(C)=\frac{1000}{25000},$$
$$P(D)=\frac{25000-(250+500+1000)}{25000}=\frac{23250}{25000}$$

注：想不到此时此刻（2019 末~2020 初）我们又遇到了影响全球的新冠肺炎瘟疫，以至于同学无法返校上课！而其主要病症仍然为发热、干咳。

由全概率公式得人患“非典”的概率

$$P(E)=P(A)P(E|A)+P(B)P(E|B)+P(C)P(E|C)+P(D)P(E|D)$$
$$=\frac{250}{25000}\cdot 5\%+\frac{500}{25000}\cdot 3\%+\frac{1000}{25000}\cdot 1\%+\frac{23250}{25000}\cdot 0.01\%=0.001593$$

由贝叶斯公式知，“非典”患者是仅发热的病人的概率

$$P(B|E)=\frac{P(B)P(E|B)}{P(E)}=\frac{\frac{500}{25000}\times 3\%}{0.001593}=0.3766478$$

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

$$P(A|E)=\frac{P(A)P(E|A)}{P(E)}=\frac{\frac{250}{25000}\times 5\%}{0.001593}=0.3138732$$

$$P(C|E)=\frac{P(C)P(E|C)}{P(E)}=\frac{\frac{1000}{25000}\times 1\%}{0.001593}=0.2510986$$

$$P(D|E)=\frac{P(D)P(E|D)}{P(E)}=\frac{\frac{23250}{25000}\times 0.01\%}{0.001593}=0.05838041$$

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”时期动不动就要测量病人的体温的原因。

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它语音输入法了吗？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)。从上个世纪六十年代开始，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

1.3.1 总体与总体信息

我们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，我们要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。再比如说，我们要考察一个经济指标 X（可以把它设想为某一支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，X 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。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，做大量的寿命实验，获得大量数据后才确认其寿命概率分布是什么。简而言之，总体信息非常重要，要获得它虽然不容易但又是必须做的，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 $\mathbf{x}=(x_1,x_2,\cdots,x_n)$，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）等等。样本是统计学（无论频率学派或贝叶斯学派）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，但大发展却是从十九世纪末到二十世纪上半叶。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空前的发展，成为当时统计学的主流。在二十世纪下半叶，经典统计学在工业、农业、医学、经济、金融、管理、军事等等领域里获得广泛的应用，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之处也逐渐暴露出来了。



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，理应善加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)曾考察如下两个统计实验：

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 $2^{-10} = 0.0009766$，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
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：

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

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

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它是“免检产品”，或者每月抽检一、二次就足够了，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，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

$$f_X(x)=\int_R f(x,y)dy\,,\quad f_Y(y)=\int_R f(x,y)dx$$

其中，R表示实数集，而Y在X已知的条件密度是

$$f(y\mid x)=\frac{f(x,y)}{f_X(x)}=\frac{f(x,y)}{\int_R f(x,y)dy}$$

从而又有

$$f(x,y)=f(y\mid x)f_X(x)=f(x\mid y)f_Y(y)$$

其次引入高等数学中的两个重要函数：贝塔函数和伽玛函数。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

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

它们有两个重要性质

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

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

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

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

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和体验研读贝叶斯统计的乐趣，从而激发起对贝叶斯统计的兴趣。

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风行不止呢？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 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（无论你在中国何处）

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，我们不但可以进行各种算术计算也可以进行通常的统计计算（建模）和绘图了。为了方便初学者的学习和实践，本书制作了一个专用R包BayesianStat，把书中所有案例数据和主要程序都放入了此包中，读者可免费下载此包，然后把它放入文件夹library中即可应用，此文件夹的一个路径示例是

C:\Program Files\R\R-3.3.1\library

从现在开始，我们就要充分利用R软件来进行贝叶斯统计的学习了。

1.4.3 一般形式的贝叶斯公式

现在我们要对一个总体 X 进行统计推断，假设其分布密度为 $p(x|\theta)$，其中 θ 是未知参数，之所以写成条件密度的形式是因为在贝叶斯统计中未知参数 θ 被看成是随机变量。进一步，假设参数 θ 已经有了先验分布 $\pi(\theta)$ 而且从总体 X 那里得到了新样本 $\mathbf{x}=(x_1,x_2,\cdots,x_n)$。现在的问题是怎样利用样本对先验分布 $\pi(\theta)$ 进行更新，以期得到更适当的分布。我们知道样本信息综合体现在其联合分布密度 $p(\mathbf{x}|\theta)$ 中，而且如果样本是简单随机样本，则

$$p(\mathbf{x}|\theta)=\prod_{i=1}^{n}p(x_i|\theta)$$

现在假设更新后的分布是 $\pi(\theta|\mathbf{x})$，即 θ 的以样本 $\mathbf{x}=(x_1,x_2,\cdots,x_n)$ 为条件的分布。根据条件密度的公式，$\pi(\theta|\mathbf{x})$ 可以写成

$$\pi(\theta|\mathbf{x})=\frac{h(\mathbf{x},\theta)}{m(\mathbf{x})}$$

其中，$h(\mathbf{x},\theta)$ 是样本 $\mathbf{x}$ 和参数 θ 的联合密度，$m(\mathbf{x})$ 是 $\mathbf{x}$ 的边际密度而且

$$m(\mathbf{x})=\int_{\Theta} h(\mathbf{x},\theta)d\theta$$ （Θ 是参数空间）

另一方面，利用先验分布 $\pi(\theta)$ 和样本的分布密度 $p(\mathbf{x}|\theta)$，我们可得样本 $\mathbf{x}$ 和参数 θ 的联合密度

$$h(\mathbf{x},\theta)=p(\mathbf{x}|\theta)\pi(\theta)$$

于是

$$\pi(\theta\,|\,\mathbf{x})=\frac{h(\mathbf{x},\theta)}{m(\mathbf{x})}=\frac{p(\mathbf{x}|\theta)\pi(\theta)}{\int_{\Theta} p(\mathbf{x}|\theta)\pi(\theta)d\theta}$$

显而易见这个公式把总体信息、样本信息和先验信息都综合进去了。这就是密度函数形式的贝叶斯公式，其中 $\pi(\theta\,|\,\mathbf{x})$ 被称为 θ 的后验分布，它是集中了总体、样本和先验三种信息后对于先验分布 $\pi(\theta)$ 的更新，以期得到参数 θ 的更符合实际的分布。

如果θ是离散参数，其先验分布可用先验分布列$\{\pi(\theta_j)\mid j=1,2,\cdots\}$来表示。则后验分布也是离散形式而且容易得到

$$\pi(\theta_j\mid\mathbf{x})=\frac{p(\mathbf{x}\mid\theta_j)\pi(\theta_j)}{\sum_i p(\mathbf{x}\mid\theta_i)\pi(\theta_i)},\ j=1,2,\cdots$$

这个公式与事件形式的贝叶斯公式是何其相似！

注：

1. 从贝叶斯公式显而易见无论是样本分布$p(\mathbf{x}\mid\theta)$还是先验分布$\pi(\theta)$乘以一个常数都不会改变后验分布$\pi(\theta\mid\mathbf{x})$。

2. 当样本观察值$\mathbf{x}$得到后，样本分布密度$p(\mathbf{x}\mid\theta)$也就是似然函数，并常常记为$l(\theta)=l(\theta\mid\mathbf{x})=p(\mathbf{x}\mid\theta)$。

3. 先验分布$\pi(\theta)$当然也有参数（如λ），但是在这里假定它已知了，所以没有写出来。如果它未知或为了强调而写出来，那就是$\pi(\theta)=\pi(\theta\mid\lambda)$，并且我们称先验分布中的参数为超参数。

1.4.4 计算后验分布的例

例 1.5（例 1.4 续）该工厂为了进一步改善产品质量，采用了更先进可行的技术，不合格品率θ因此有可能发生变化。为了对θ的先验分布进行更新，我们来计算θ的后验分布。为此，我们对n件产品进行独立检测，不合格品出现的个数记为 X，显然，X 服从二项分布$Bin(n,\theta)$，即

$$P(X=x|\theta)=p(x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},x=0,1,...,n.$$

再根据贝叶斯公式和θ的先验分布（表 1.2），我们就可以把θ的后验分布算出来，其一般表达式是

$$\pi(\theta_j|x)=\frac{p(x|\theta_j)\pi(\theta_j)}{\sum_i p(x|\theta_i)\pi(\theta_i)},x=0,1,2,\cdots,n\ ;\ j=1,2,\cdots,6$$

在 R 平台中利用如下命令就可以把以二项分布$Bin(n,\theta)$为总体，参数θ为离散情形的后验概率分布具体算出来，例如，若$n=10,x=0$，则可以算得相应的后验概率分布表 1.3。从该表可以看出通过采用新技术，产品质量有了很大的提高。

以下就是所用的 R 命令

```
library(BayesianStat)       #计算后验概率的命令 bindiscrete 在此包中
theta<-c(0,0.2,0.4,0.6,0.8,1)
prior<-c(0.94,0.03,0.02,0.01,0.00,0.00)
bindiscrete(x=0, n=10,pi=theta,pi.prior=prior,n.pi=6)
```

这里参变量 x 是样本值；n 是样本量；pi 是不合格品率θ的取值向量；pi.prior 是θ的先验概率向量；n.pi 是θ的取值个数。另外，最后这个命令可以同时得到先验概率与后验概率的比较图（图 1.1）。该图形象地把后验概率相对于先验概率的变化显示出来，从该图可以看出不合格品率$\theta=0$的后验概率比先验概率大，而其它情形的后验概率都不大于先验概率，这就更生动形象地说明了产品质量有了很大的提高。

表 1.3　不合格品率后验概率分布表

不合格品率θ	0.0	0.2	0.4	0.6	0.8	1.0
后验概率	0.9965	0.0034	0.0001	0.0000	0.0000	0.0000

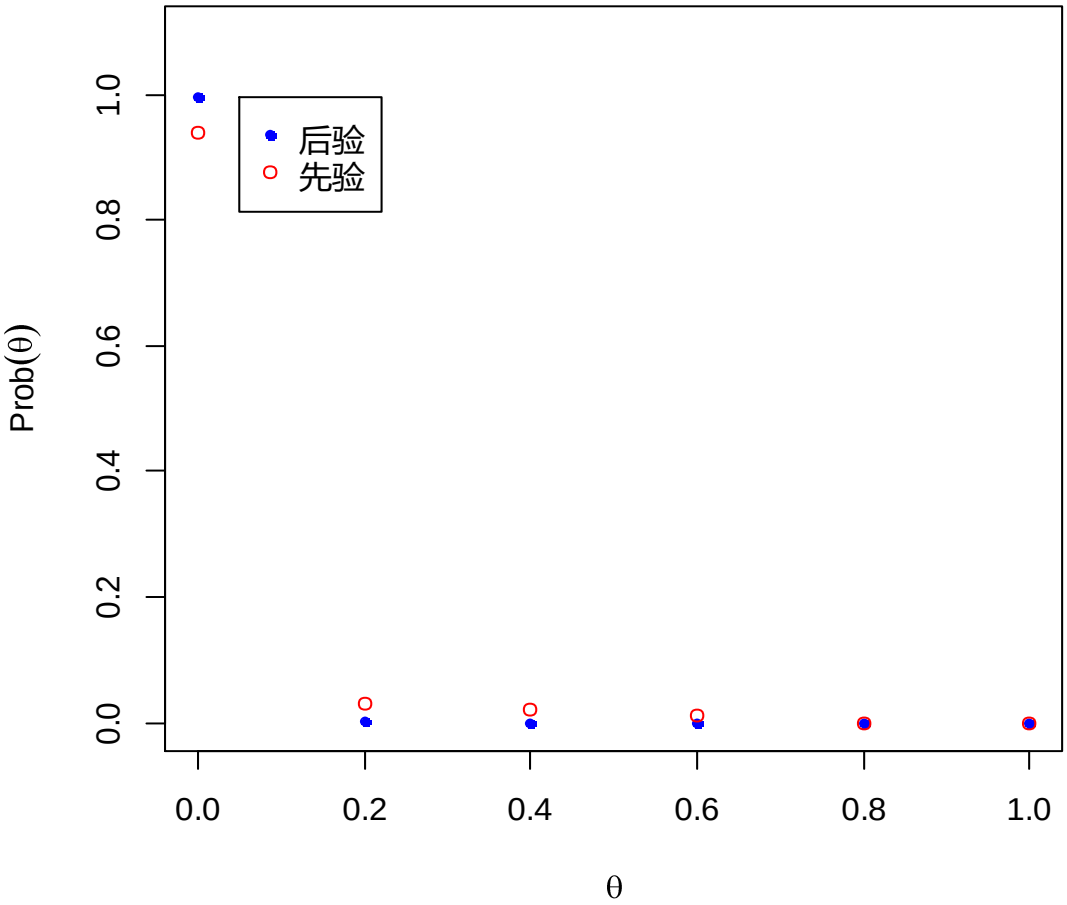

图 1.1 不合格品率先验与后验概率比较

现在假设在检测该产品之前我们对不合格品率θ没有任何先验信息（比如说，这是新产品）。在这种情况下，贝叶斯建议用区间(0，1)上的均匀分布$U(0,1)$作为θ的先验分布，因为该分布在区间（0，1）上机会均等地取到每一点。贝叶斯的这个建议被后人称为贝叶斯假设。这时θ的先验分布密度为

$$\pi(\theta)=\begin{cases}1, & 0<\theta<1\\ 0, & \text{其它场合}\end{cases}$$

于是，样本 X 与参数θ的联合分布

$$h(x,\theta)=p(x|\theta)\pi(\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x},\ x=0,1,...,n;0<\theta<1.$$

而 X 的边缘分布

$$m(x)=\binom{n}{x}\int_0^1\theta^{x}(1-\theta)^{n-x}d\theta=\binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}=\frac{1}{n+1},x=0,1,\cdots,n.$$

利用贝叶斯公式，最后可得θ的后验分布

$$\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1},0<\theta<1$$

这正是参数为 x+1 和 n-x+1 的贝塔分布$Beta(x+1,n-x+1)$。

在 R 平台中利用如下命令就可以把以二项分布 $Bin(n,\theta)$ 为总体，参数 θ 服从贝塔分布的先验密度和后验分布密度图形画出来（如图 1.2 所示）。

```
library(BayesianStat)
binbeta(x, n, a = 1, b = 1, pi = seq(0.01, 0.999, by = 0.001), plot = TRUE)
```

在函数 binbeta 中，参变量 x 是样本值；n 是样本量；a 和 b 是贝塔分布的两个参数（在本例中，因为先验是（0，1）区间上的均匀分布，所以 a = b = 1）；pi 是不合格品率 θ 的取值向量；plot 是逻辑变量（取“TRUE”表示要作图；取“FASLE”表示不要作图）。注意：这里样本量（产品抽取个数）$n=10$，按照从左到右从上到下的顺序各图对应的样本值分别是 $x=0, x=2, x=5, x=8$。从图 1.2 可以看出随着样本值的变化，后验密度曲线也发生了重大变化，换句话说，样本对先验分布产生了重大影响，先验被实质性更新了。

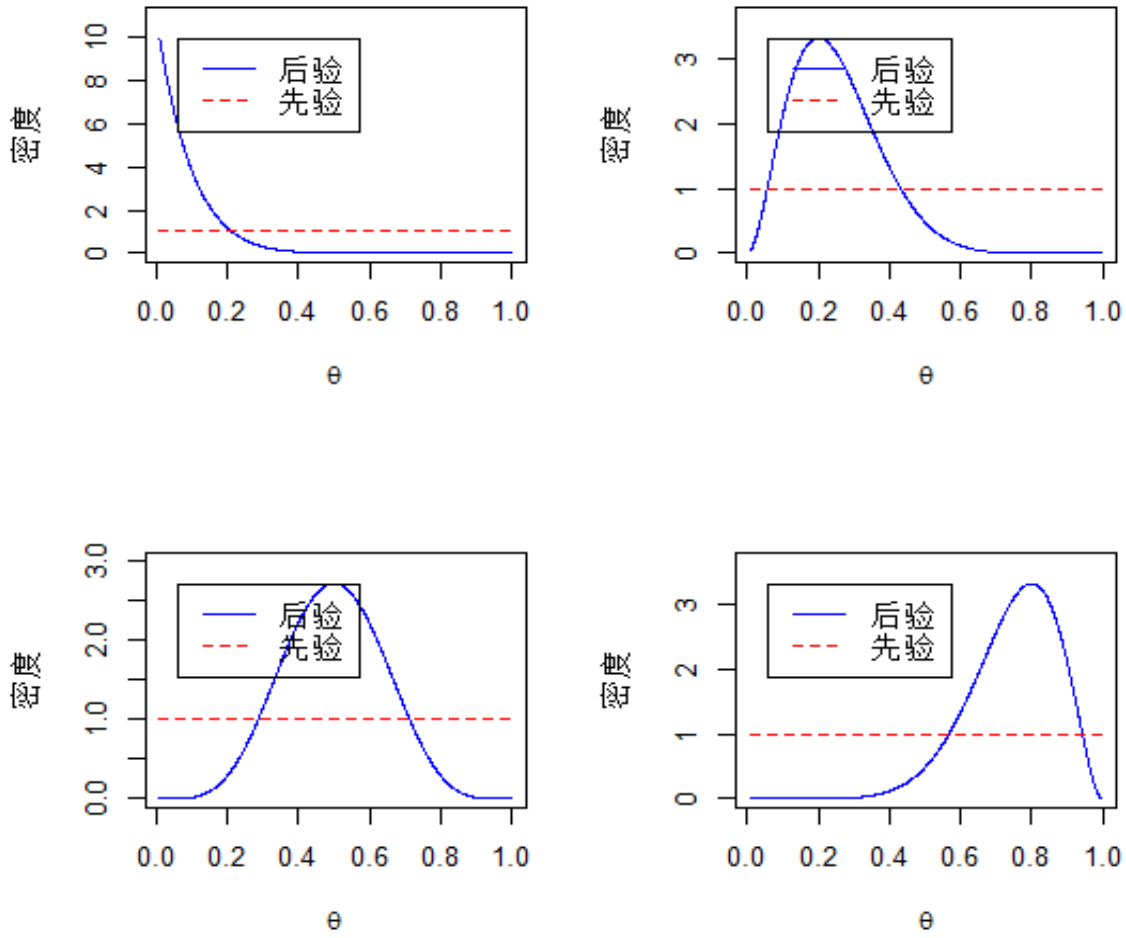

图 1.2 均匀分布先验与二项分布形成的后验分布密度图

Homework：

1. 在你的电脑上安装R软件和包BayesianStat.
2. 熟悉R的基本操作和描述统计计算。
3. 书面作业

PP14-15，1，2，3，4，5，6，7，8，9

非常时期，作业交电子版即可，但用软件得到的结果不能原封不动地拷贝到作业上，要归纳整理并翻译成中文。作业设计成A4纸的规格。