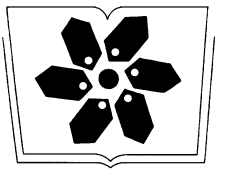

中国科学院科学出版基金资助出版

贝叶斯网引论

张连文　郭海鹏　著

科学出版社

北　京

内 容 简 介

贝叶斯网是将概率、统计应用于复杂系统的不确定性推理和数据分析的一种有效工具，它起源于20世纪80年代中期对人工智能中的不确定性问题的研究，近年来在国际上的影响不断扩大。本书是第一本系统论述贝叶斯网的基本理论、算法及其应用的中文专著。内容包括概率论及贝叶斯网基本概念、贝叶斯网推理、贝叶斯网学习，以及贝叶斯网在中医中的应用四大部分。本书从实例出发，由浅入深，直观与严谨相结合，并提供了详尽的参考文献。本书的读者对象是相关专业的高年级本科生、研究生和科研人员。

图书在版编目(CIP)数据

贝叶斯网引论/张连文，郭海鹏著．—北京：科学出版社，2006
ISBN 7-03-018170-0

Ⅰ．贝… Ⅱ．①张… ②郭… Ⅲ．贝叶斯推断 Ⅳ．O212

中国版本图书馆CIP数据核字(2006)第125753号

责任编辑：赵卫江 陈砺川/责任校对：赵 燕
责任印制：吕春珉/封面设计：耕者设计工作室

科学出版社发行 各地新华书店经销

*

2006年11月第 一 版 开本：B5(720×1000)
2006年11月第一次印刷 印张：19
印数：1—2 000 字数：360 000

定价：58.00元

(如有印装质量问题，我社负责调换〈环伟〉)

销售部电话 010-62136131 编辑部电话 010-62138017(BI01)

序

这是一本系统论述由概率、统计与图论结合而发展起来的贝叶斯网的专著. 贝叶斯网起源于 20 世纪 80 年代中期对人工智能中的不确定性问题的研究，已经成为人工智能的一个重要领域. 近年来它的影响在国际上不断扩大，成为将概率、统计应用于复杂系统的不确定性推理和数据分析的一种有效工具. 本书作者根据多年来在这方面的研究和与国际交流的体会，采取由实例出发、由浅入深、直观与严谨相结合等较易为读者理解的方式，对这一新的学科的基本理论、方法及其应用进行了系统的、较全面的介绍，列举了相当详尽的参考文献. 我认为，本书在国内的出版是一桩在学术上有意义的事件，将促进国内对这一新兴而又有潜能的领域的进一步学习、研究和应用. 作者希望我给本书写序，作为人工智能研究的一个热心观众，借这个机会谈一点有关人工智能以及它与数学、电子计算机（下面采用现在通用的称谓——电脑）的关系的粗浅看法，供有志于人工智能研究的读者参考.

人工智能的兴起大致源于 20 世纪 40 年代电脑的出现. 人们一方面兴奋地运用电脑强大的计算功能，去完成以往无法完成以及新提出的庞大计算任务；另一方面，对电脑所具有的类似于人脑的某些功能，包括一些逻辑判断功能，产生了巨大兴趣，从而对它寄予了很高的期望，期望它能够在智能方面起作用. 于是人们开始研究自动机的理论、人类智能的机理以及智能行为的计算模型，进而期望构造与人脑相类似的智能系统，这大概是人工智能的最初想法. 在当时，已经有人认为要制造完全代替人脑的电脑是不可能的. 50 年代，我从著名的数理逻辑界前辈胡世华先生那里知道了一种相当有说服力的说法（我认为现在看来也是正确的）：机械是人手的延伸，而电脑是人脑的延伸. 机械常常能够更好、更快地完成人手所能完成的机械性任务，但是不能完全代替人手；同样地，电脑能够完成人脑所能完成的机械性任务，但是不可能完全代替人脑. 当然这种看法丝毫没有贬低人工智能的意义，相反，人工智能是一个引起人们浓厚兴趣、已经有相当成就而又被人们寄予巨大希望的科学领域.

与探索和模拟人类智能这一长远目标相比，人工智能研究更多的是集中在针对实际具体问题的智能系统的研究和开发. 专家系统就是一个例子，它的目标是将专家们对某一复杂领域的知识和经验引入电脑系统，使得一般人能够借助电脑系统得到对问题的专家级解答. 60 年代，在国内由于肝炎的流行，人们研制出学习和模仿著名中医专家关幼波的诊治系统就是一例. 正如本书的 1.4 节所表述

的，人工智能的实质进展有赖于不断针对人类的某种智能行为，运用数学理论和方法，结合计算机技术来建立适当的数学计算模型．只有智能的目标和计算机技术而没有数学的深层次介入是不可能有显著进展的．不确定性是人工智能所面临的一个重要课题，因此概率论自然就成为处理人工智能问题的一个工具．对于静态的概率问题的研究和计算，理论上只需要一个联合概率分布就足够了．但是联合概率分布的复杂度相对于变量个数呈指数增长，以致当变量个数很大时，计算复杂度极高而不可行．虽有电脑的强大计算功能也会成为不可行．贝叶斯网的提出就是要解决这个问题．它用图论的语言直观揭示问题的结构，又按照概率论的原则对问题的结构加以分析，降低推理计算的复杂度，使得人们能够将概率论应用于大型问题．近年来贝叶斯网能够在众多领域得到广泛应用，原因就在于此．从贝叶斯网的发展，我们同样也看到计算机技术的广泛应用有赖于数学的深层次介入．

对于概率论与数理统计的广泛应用，人们有普遍的共识．但是在很长一段时期，人们普遍地关注和运用概率的频率解释，对于贝叶斯观点比较忽视，认为将概率看作一种主观信任程度是没有客观标准的．20 世纪中、后期以来，贝叶斯观点的应用逐步得到了认同，贝叶斯方法再度兴起，贝叶斯网就与这次潮流密切相关，这些是概率论与数理统计的应用中一个值得重视的趋势．现在看来，主观、客观观念属于哲学范畴，应用于具体的对象时，需要根据实践来判断．主观信度或在贝叶斯观点下的先验概率，实际上可以认为是在本次试验或观察以前或以外的某种经验积累的表述．所以，对于一种数学观点和方法，应该依据实践来认识、判断和应用，这样才不至于因噎废食．

本书作者之一张连文 1986 年在北京师范大学汪培庄教授的指导下获得应用数学硕士学位，接着在我们的概率统计学科点攻读博士，作为我和 Kansas 大学的 Glenn Shafer 教授联合指导的博士生，以论文《人工智能中不确定性的三个模型》通过论文答辩（1989）．我当时觉得他比较年轻，对人工智能有兴趣，就建议他再在国外攻读一个计算机博士学位，这样对以后研究人工智能会有好处．1994 年，他在加拿大不列颠哥伦比亚大学 David Poole 教授的指导下获得计算机博士学位，论文标题是《决策网络的计算理论》．这些年来，他实际上一直围绕人工智能中不确定性问题而工作，对贝叶斯网的发展做出了一系列的重要工作；而且，他还和中医学者合作，应用贝叶斯网研究中医的辨证施治．这次能在科学出版社出版《贝叶斯网引论》，是向内地对他的培养的一项汇报和宝贵的回报．我希望通过本书的出版，使他能够和内地建立更好的学术交流，帮助推动贝叶斯网及相关方面在内地的应用和发展．

严士健

于北京师范大学数学科学学院

前　言

贝叶斯网（Bayesian networks）是一种帮助人们将概率统计应用于复杂领域、进行不确定性推理和数据分析的工具．它起源于人工智能领域的研究，近年来对众多其它领域也产生了重要影响．本书系统地介绍贝叶斯网的基本理论和方法，为读者了解和进入这个新兴领域提供一条相对平坦的途径．

从技术层面上讲，贝叶斯网是一种系统地描述随机变量之间关系的语言．构造贝叶斯网的主要目的是进行概率推理，即计算一些事件发生的概率．要在一些随机变量之间进行概率推理，理论上只需要一个联合概率分布即可．但是，联合概率分布的复杂度相对于变量个数成指数增长，所以当变量众多时不可行．贝叶斯网的提出就是要解决这个问题，它把复杂的联合概率分布分解成一系列相对简单的模块，从而大大降低了知识获取的难度和概率推理的复杂度，使得人们可以把概率论应用于大型问题．

贝叶斯网是概率论与图论相结合的产物，它一方面用图论的语言直观揭示问题的结构，另一方面又按照概率论的原则对问题的结构加以利用，降低推理的计算复杂度．近年来，贝叶斯网之所以能够在众多不同领域得到广泛应用，其根本原因就在于此．另外，由于贝叶斯网直观易懂，它也是学者们喜爱的讨论和交流工具．

统计学、系统工程、信息论以及模式识别等学科中许多经典的多元概率模型都是贝叶斯网的特例，包括朴素贝叶斯模型（naïve Bayes models）（Titterington et al.，1981）、隐类模型（latent class models）（Lazarsfeld and Henry，1968；Goodman，1974）、混合模型（mixture models）（Everitt and Hand，1981；McLachlan and Basford，1988）、隐马尔可夫模型（hidden Markov models）（Baum and Petrie，1966）、卡尔曼滤波器（Kalman filters）（Kalman，1960）等等．贝叶斯网为这些模型提供了一个共同的框架，使得在一个领域获得的结果可以推广到其它领域．更重要的是，它也为发展新模型提供了一个自然的框架．例如，动态贝叶斯网（dynamic Bayesian networks）（Dean and Kanazawa，1990；Russell and Norvig，2003）就是最近几年在这个框架下发展起来的一类新模型，它主要用于对多维离散时间序列的监控和预测．另一个例子是多层隐类模型（hierarchical latent class models）（见第 9 章），它是对隐类模型的推广，能够揭示观测变量（observed variables）背后的隐结构．

构造贝叶斯网的方法因问题而定．有时，网络结构和参数由应用问题的定义

所决定. 例如，在解码问题中（见 2.6.2 节），网络结构完全取决于编码器的设计，而网络参数则由编码器的设计和信道的特征所决定. 在专家系统中，贝叶斯网的结构和参数往往通过咨询专家来获得. 近年来，贝叶斯网越来越多地被用于数据分析，成了数据分析的工具. 统计学视之为图模型的一种，而人工智能学科把从数据出发获得贝叶斯网的过程视为是机器学习的一个特例，称为贝叶斯网学习.

在国际上，贝叶斯网已经成为人工智能和机器学习教材的重要内容，同时还有多本专著已经或即将出版，这些均凸显贝叶斯网受到的重视. 我们从事贝叶斯网研究多年，觉得有责任把它系统地介绍到国内，于是决定写这本书.

本书分四个部分. 第一部分介绍贝叶斯网基础，包括第 1、2、3 章. 第 1 章回顾一些概率论及信息论中与贝叶斯网密切相关的概念和结果. 第 2 章以降低概率推理的复杂度为出发点，引入贝叶斯网的概念，介绍如何手工构造贝叶斯网，并给出诸多应用实例. 第 3 章讨论贝叶斯网的图论侧面与它的概率论侧面之间的密切关系.

第二部分介绍贝叶斯网推理，包括第 4、5、6 章. 第 4 章揭示贝叶斯网推理的原理，并给出最基本的推理算法，即变量消元算法. 第 5 章介绍另一个推理算法，即团树传播算法. 与变量消元法相比，团树传播法的主要优点是它使得两次不同推理的中间结果可以共享. 因此，当需要做多次推理时，团树传播法比变量消元法更为合适. 第 6 章介绍近似推理算法，包括随机抽样法和变分法以及最新的研究进展.

第三部分讨论贝叶斯网学习，包括第 7、8、9 章. 第 7 章假设已知网络结构，讨论参数估计方法. 第 8 章探讨在不知道网络结构的情况下，如何通过数据分析同时确定贝叶斯网的结构和参数. 第 9 章假设观测变量背后有一个隐结构，研究如何通过数据分析对这个隐结构进行推测.

第四部分只有一章，即第 10 章，它介绍的是贝叶斯网在中医研究中的一个应用实例. 贝叶斯网的应用有很多，之所以选用中医的例子是因为贝叶斯网有机会对中医现代化起重要推动作用，而中医也为贝叶斯网研究提出了一些有趣的新问题.

本书的读者对象是相关专业的高年级本科生、研究生和科研人员. 我们选择贝叶斯网最核心的内容，基于我们自己的体会将其组织起来，因此与国际上现有的专著相比，本书更适合作为贝叶斯网的入门教材. 当然，本书也可以用于自学，书中不仅详细阐述贝叶斯网的基本理论，还简要介绍它的各种应用实例，综述最新研究进展，并提供大量的参考文献以便读者进一步深入学习. 在注意理论深度和严谨性的同时，我们也注重理论背后的直观含意以及如何将理论应用于实际.

本书的筹划始于2002年秋. 那时，张连文在香港科技大学开设了一门关于贝叶斯网的研究生课程，备课时已为将来把讲稿扩充成书做了许多考虑. 具体写作开始于2003年秋，初稿完成于2005年夏. 陈弢和王燚几次仔细阅读初稿，提出了许多改进意见，并参与了一些章节的写作，我们表示衷心感谢. 严士健教授是张连文的博士导师，长期关心和支持张连文的工作，这次又拨冗为本书写序，在此特表感激之意. 本书的写作得到香港研究资助局项目622105、国家重点基础研究发展计划（973计划）No. 2003CB517101，以及香港科技大学博士后基金的资助，本书的出版得到中国科学院科学出版基金的资助，史忠植教授、刘际明教授提供了协助，陈砺川、赵卫江女士做了大量细致的编辑工作，这里一并致谢.

由于我们水平有限，书中难免有疏漏和错误，请读者不吝指正. 错误更正会在本书的网页（http://www.cs.ust.hk/bnbook/）上公布. 在此网页上，读者还可以找到一些关于贝叶斯网的信息，如相关学术组织、学术会议、学术期刊、学术团队以及常用软件等.

张连文　郭海鹏

2006年春于香港清水湾

目　录

第一部分　贝叶斯网基础

第二部分 贝叶斯网推理

第三部分 贝叶斯网学习

第四部分　贝叶斯网应用

第 一 部 分

贝 叶 斯 网 基 础

第 1 章　概率论基础

本章简要介绍概率论中与贝叶斯网密切相关的一些基本概念．与一般的概率论教材不同，我们将侧重于概念的直观含意．深入理解诸如条件独立、贝叶斯定理等概念和结果，对以后理解贝叶斯网至关重要．另外，本章还将简述概率方法在人工智能研究中的崛起过程，并介绍在第三部分中会被用到的一些信息论知识．

1.1　随机事件与随机变量

世界上许多事情都具有不确定性．例如掷硬币，其结果可能正面朝上，也可能反面朝上，在抛掷之前无法预知．又如赌马，理论上每匹马都有跑第一的可能，事先无法预料哪匹马一定会赢．再如火星上是否曾有生命存在？答案有两种可能，是或不是，但根据目前掌握的证据判断，无法给出绝对的答复．概率论是研究处理这类现象的数学理论．本节介绍概率论中几个最基本的概念，包括样本空间、事件、概率测度、随机变量以及概率函数．

1．样本空间和事件

在概率论中，随机试验指的是事先不能完全预知其结果的试验．随机试验的所有可能结果组成该试验的样本空间，通常记为 Ω．样本空间可以是离散的，也可以是连续的．如无特殊说明，本书所论及的样本空间都将是离散的．

样本空间中的点，即随机试验的可能结果，称为样本点，或原子事件，记为 ω．样本空间的子集称为事件，通常用大写字母表示：A，B，…．如果随机试验的结果包含在一个事件之中，则称该事件发生了．样本空间 Ω 本身也是一个事件，而且是一定会发生的必然事件．空集也是一个事件，是不可能事件，通常记为 ∅．事件之间可以进行交（⋂）、并（⋃）、差（\）等各种集合运算．若两事件 A 和 B 交空，即 $A \cap B = \varnothing$，则称它们为互斥事件，又称不相容事件．两互斥事件不能同时发生．若 A 和 B 互斥，且 $A \cup B = \Omega$，则称它们为互补事件．

例 1.1　考虑掷硬币试验，其结果有正反面两种可能，因此样本空间为 $\Omega = \{h, t\}$，其中 h 表示正面，t 表示反面，h 和 t 为两个互补的原子事件．□

例 1.2　考虑掷骰子试验，有 6 种可能的结果，样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$．子集合 {1，3，5} 表示的是“掷出结果为奇数”这一事件，其互补事件为 {2，4，6}，即“掷出结果为偶数”．□

例 1.3　考虑从香港科技大学的所有研究生中随机抽样的试验，其结果可能

是任一研究生，样本空间为 $\Omega=\{x \mid x$ 是科大研究生$\}$.{研究生张三} 为一原子事件[①]. {所有科大男研究生} 或 {所有年龄超过 25 岁的研究生} 是两个非原子事件. □

2. 概率测度

概率测度给样本空间中的每一事件 A 赋予一个数值 $P(A)\in[0,1]$，以度量该事件发生的可能性. 在数学上，它是一个从样本空间 Ω 的幂集 2^{Ω} 到区间 $[0,1]$ 的映射 $P:2^{\Omega}\to[0,1]$，且满足以下 3 个 Kolmogorov 公理 (Kolmogorov, 1933, 1950)：

(1) $P(\Omega)=1$；

(2) $P(A)\geqslant 0$，$\forall A\in 2^{\Omega}$；

(3) $P(A\cup B)=P(A)+P(B)$，$\forall A, B\in 2^{\Omega}$，$A\cap B=\varnothing$.

$P(A)$ 称为事件 A 的概率. 上面这 3 个公理分别被称为概率测度的规范性、非负性和有限可加性. 规范性规定必然事件的概率为 1，非负性规定概率不能为负，有限可加性规定互斥事件的并集的概率等于它们各自概率的和. 从这 3 个公理出发，可以推出概率测度的诸多基本性质和定理.

例 1.4 例 1.1 中的样本空间的所有子集是：Ω，$\varnothing$，{h}，{t}. 设所掷为均匀硬币，则有 $P(\Omega)=1$，$P(\varnothing)=0$，$P(h)=P(t)=0.5$. 这里 $P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

例 1.5 在例 1.1 中，若所掷为不均匀硬币，且已知在 1000 次重复试验中正面出现 700 次，反面出现 300 次，可因此设 $P(\Omega)=1$，$P(\varnothing)=0$，$P(h)=0.7$，$P(t)=0.3$. 这里 $P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

例 1.6 在例 1.3 中，对 Ω 的任一子集 A，定义 $P(A)=\dfrac{|A|}{|\Omega|}$，其中 $|A|$ 是 A 中元素的个数，称为 A 的势. 那么，$P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

3. 随机变量

随机变量是定义在样本空间 Ω 上的函数，通常用大写字母表示，如 X，Y，Z. 随机变量的取值随试验的结果而定，通常用小写字母表示，如 x，y，z. 随机变量 X 的所有可能取值的集合称为它的值域，也称状态空间，记为 Ω_x. 随机变量可以是离散的，也可以是连续的. 离散随机变量的状态空间是离散的，包含有限个或无穷可数个状态. 连续随机变量的状态空间是连续的，包含无穷不可数个

① 假设只有一个叫张三的研究生.

状态. 本书主要考虑离散随机变量.

例 1.7　在例 1.3 中，设 X 为“随机抽出的一个学生的性别”，则 X 是定义在“科大研究生”这个样本空间上的随机变量：$\Omega_X = \{m, f\}$，其中 m 表示男，f 表示女. □

例 1.8　同时掷两个质地均匀的硬币，其样本空间为 $\Omega = \{(h, h), (h, t), (t, h), (t, t)\}$. 对任一 $\omega \in \Omega$，$P(\{\omega\}) = 1/4$. 设 X 为“正面朝上的硬币个数”，那么 X 是定义在 Ω 上的一个随机变量：$X((h, h)) = 2$，$X((h, t)) = 1$，$X((t, h)) = 1$，$X((t, t)) = 0$，故 $\Omega_X = \{0, 1, 2\}$. □

4. 概率函数

设 X 为一随机变量，x 是它的一个取值. 在样本空间中，所有使 X 取值为 x 的原子事件组成一个事件，记之为 $\Omega_{X=x} = \{\omega \in \Omega \mid X(w) = x\}$，有时也简记为“$X = x$”. 注意，$\Omega_{X=x}$ 与 Ω_X 的含意完全不同，后者是随机变量 X 的状态空间，包括 X 的所有可能取值.

事件“$X = x$”的概率 $P(X = x) = P(\Omega_{X=x})$ 依赖于 X 的取值 x. 让 x 在 Ω_X 上变动，$P(X = x)$ 就成为 Ω_X 的一个取值于 $[0,1]$ 的函数，称之为随机变量 X 的概率质量函数，记为 $P(X)$. 根据概率测度的定义，有

$$P(X = x) \geqslant 0, \forall x \in \Omega_X; \sum_{x \in \Omega_X} P(X = x) = 1.$$

为了记号上的方便，上面两式有时简记为

$$P(X) \geqslant 0; \quad \sum_X P(X) = 1.$$

例 1.9　在例 1.7 中，设科大共有 500 名研究生，其中 400 名男生，100 名女生. 这里 $\Omega_X = \{m, f\}$，$\Omega_{X=f} = \{\text{科大所有女研究生}\}$. $X = f$ 的概率为

$$P(X = f) = P(\Omega_{X=f}) = P(\{\text{科大所有女研究生}\}) = 100/500 = 0.2.$$

X 的概率函数为

X	m	f
P(X)	0.8	0.2

□

例 1.10　在例 1.8 中，变量 X 的值域是 $\Omega_X = \{0, 1, 2\}$，$\Omega_{X=0} = \{(t, t)\}$，$\Omega_{X=1} = \{(t, h), (h, t)\}$，$\Omega_{X=2} = \{(h, h)\}$. X 的概率函数为

X	0	1	2
P(X)	0.25	0.5	0.25

□

离散随机变量有概率质量函数. 与之对应，连续随机变量有概率密度函数. 由于本书主要关心离散随机变量，所以对概率密度函数不做详细介绍. 以后，我们有时会用“概率分布”一词来泛指概率质量函数、概率密度函数，或与它们等价的其它概念.

1.2 概率的解释

概率的解释主要有 5 种：古典解释、频率解释、主观解释、特性解释（Popper，1957）以及逻辑解释（Carnap，1950）. 本节主要介绍前 3 种解释，重点放在概率的主观解释上.

1.2.1 古典解释

概率的古典解释起源于 16 世纪数学家们对掷骰子等赌博活动的研究. 对一粒质地均匀的正方体骰子，投掷后其任何一面朝上的可能性相等，因此掷出每面的概率都应为 1/6. 一般地讲，如果事件 A 包含的样本数为 m，而样本空间的总样本数为 n，则事件 A 的概率应为

$$P(A)=\frac{\text{事件 A 包含的样本数}}{\text{样本空间的总样本数}}=\frac{m}{n}. \tag{1.1}$$

用这种方法定义的概率称为古典概率. 古典概率的一个前提条件就是等可能性. 在实际应用中，这个前提一般很难满足，因此古典概率的应用范围很有限.

例 1.11 考虑如下掷 3 粒骰子的赌局：若结果之和为 {3，4，5，6，7，14，15,16，17，18} 中任一数字，则赌客胜，否则庄家胜. 3 粒骰子之和有从 3 ～ 18 共 16 种可能结果，其中赌客胜的有 10 种，所以表面看来似乎这是一个对赌客有利的赌局. 但是，由于结果不具有等可能性，因此不能简单认为赌客获胜的概率是 10/16. 实际上，这是一个对庄家有利的赌局，投掷 3 粒骰子有 216 种等可能结果：{(1，1，1)，(1，1，2)，…，(6，1，1)，…，(6，6，6)}，其中使赌客获胜的结果只有 69 种，而使庄家获胜的结果有 147 种. □

1.2.2 频率解释

给定一个质地不均匀的骰子，掷出 6 的概率为多大？古典解释无法处理这种情况，因为这时等可能性前提不成立. 为近似计算这一概率，人们通常进行多次重复试验，记下其中掷出 6 的次数，除之以总的试验次数，将结果作为掷出 6 的概率. 一般地讲，对于一个可在同样条件下重复进行的试验，如果事件 A 在所有 N 次试验中共发生了 M 次，则它的概率可以用其发生的频率来近似：$P(A)\approx M/N$. 这个近似的理论支持是大数定律：当 N 趋于无穷大时，频率几乎处处趋

于概率. 即当 N 较大时，频率经常稳定地出现在概率附近，而当 N 越大时，越是更经常地稳定于概率，而且幅度也越小. 这就是概率的频率解释.

按照频率解释，概率只有当试验可以在同等条件下无限次重复时才有意义. 然而，实际中人们往往需要研究一些不可重复的事件发生的概率，例如总统竞选或体育比赛的结果. 频率解释对这些一次性事件无法处理. 在早期的人工智能研究中，概率的频率解释曾占据主导地位，这一度为概率论的应用造成了概念上的困难.

1.2.3　主观解释

主观解释又称贝叶斯解释，它认为概率即合理信度，反映的是个体的知识状态和主观信念. 在这种意义下的概率称为主观概率.

1. 主观概率的评估

相对于频率解释，主观解释的长处是它允许对一次性事件也进行概率评估. 例如：巴西队赢得下届世界杯足球赛冠军的概率是多大？频率解释认为此问题无意义，因为“下届世界杯决赛巴西队夺冠”不是一个可以重复的事件. 但是，主观解释仍可以根据各种先验知识给出一个主观概率评估. 主观概率的评估有许多方法，其中之一是下面的概率轮方法.

例 1.12（概率轮与概率评估）　设想有一质地均匀的概率轮（图 1.1），其上仅包含黑白两个连续区域，转动后指针停在任一位置的概率相等，因而其停在黑区的概率应等于黑区角度所占的百分比. 概率轮提供了一个进行主观概率评估的客观参照. 当评估“巴西队赢得下届世界杯足球赛冠军的概率”时，首先问如下问题：巴西队夺冠的可能性大，还是指针停在黑区的可能性大？如果认为巴西队夺冠更有可能，那么就设想一个更大的黑区，反之设想一个较小的黑区，再次问同样的问题. 如此反复，直到认为巴西队夺冠和指针停在黑区具有相同的可能性，然后测量黑区角度的大小，除以 360，即得到巴西队夺冠的概率. □

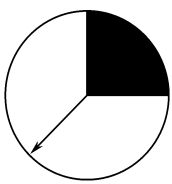

图 1.1　用于主观概率评估的概率轮

作为一种操作性强的手段，用概率轮进行主观概率评估在管理科学、心理学以及运筹学中被广泛使用，详见有关文献（Spetzler and von Holstein，1975）. 不难看出，这里存在一个评估精度的问题：概率值为 0.201 或 0.202，有什么根本的差别吗？两者差别如此之小，以至于人们往往很难断定自己的真实信度究竟是哪一个. 所幸的是，在贝叶斯网应用中，这往往不是一个大问题. 原因有三：首先，概率值的微小差别对决策的影响一般不大；其次，实际中往往会同时考虑多个事件的

概率，由于概率必须满足 Kolmogorov 公理，因此不同事件的概率之间存在一定的关系，而这些关系限制了主观概率的任意性；第三，在数据分析中，当数据量足够大时，主观概率的影响不大，这一点将在本小节最后详细论述.

2. 主观概率与 Kolmogorov 公理

为什么主观概率必须满足 Kolmogorov 公理呢？围绕这个问题，人们提出了几种理论和论证方法（Shafer and Pearl, 1990）. 其中之一与赌博有关，它的基本思想是：如果一个人的主观概率不满足 Kolmogorov 公理，那么就可以构造一个赌局，使他认为合理而接受，但又将必输无疑，这样的赌局称为 Dutch book[①]. 对于理性个体，不应该有 Dutch book 存在，因此理性个体的主观概率必须满足 Kolmogorov 公理.

如果一个人对某事件 S 出现的主观概率为 P，$0 \leqslant P \leqslant 1$，则他最多愿意付 P 元去购买一个关于 S 的单位筹码，即如果 S 为真则可拿它兑换 1 元赔金的筹码. 同时，他也愿意以 P 元或更高的价格卖掉一个单位筹码. 下面的两个例子分别说明，如果一个人的主观概率不满足 Kolmogorov 第(1)和第(3)公理，那么就有针对他的 Dutch book 存在.

例 1.13　考虑 $\Omega = \{S, \neg S\}$. 假设一赌客的主观概率为 $P(S) = 0.51$，$P(\neg S) = 0.51$，从而总和 $P(\Omega) = P(S) + P(\neg S)$ 大于 1，违反了 Kolmogorov 第 (1) 公理. 因为 $P(S) = 0.51$，所以他愿意以 0.51 元买入一个赌 S 为真的单位筹码 C_S；同时因为 $P(\neg S) = 0.51$，所以他也愿意以 0.51 元买入一个赌 $\neg S$ 为真的单位筹码 $C_{\neg S}$. 构造如下的 Dutch book：让赌客以 1.02 元的价格同时买进 C_S 和 $C_{\neg S}$ 两个筹码. 对这个赌客来讲，此 Dutch book 是公平的. 但是一旦他接受，无论结果是 S 还是 $\neg S$，他都只能拿回 1 元，从而损失 0.02 元. □

例 1.14　考虑一个由 3 匹马 H_1，H_2，H_3 参加的跑马比赛. 假设有 8 种彩票 T_0，T_1，T_2，T_3，T_{12}，T_{13}，T_{23}，T_{123}，它们的赌金如下：

T_1：100 元 如果 H_1 赢	T_{12}：100 元 如果 H_1 或 H_2 赢
T_2：100 元 如果 H_2 赢	T_{13}：100 元 如果 H_1 或 H_3 赢
T_3：100 元 如果 H_3 赢	T_{23}：100 元 如果 H_2 或 H_3 赢
T_0：100 元 如果没有任何一匹马赢	T_{123}：100 元 如果任何一匹马赢

分别以 $P(H_1)$，$P(H_2)$，$P(H_1 \cup H_2)$ 记某人对 H_1 赢、H_2 赢、H_1 或 H_2 赢

① 在英式赌马比赛中，一个“book”是指为某人所接受的一套下注组合.

的主观概率估计，那么对他来说，彩票 T_1、T_2 和 T_{12} 的公平价格分别为 $P(H_1)\times 100$，$P(H_2)\times 100$ 和 $P(H_1 \cup H_2)\times 100$.

假设一赌客的概率评估如下：$P(H_1)=0.3$，$P(H_2)=0.4$，$P(H_1 \cup H_2)=0.5$. 这里 $P(H_1)+P(H_2)\neq P(H_1 \cup H_2)$，违反了 Kolmogorov 第(3)公理. 一个针对此赌客的 Dutch book 如下：假设他手中开始时有彩票 T_{12}，从他手中以 50 元买走 T_{12}，再分别以 30 元和 40 元的价格将 T_1 和 T_2 卖给他. 对他来说，这个交易是公平的，但是他却蒙受了损失. 交易前后赌客手中彩票的价值变化如下：

	交易前	交易后
如果 H_1 赢	100 元	100＋50－30－40＝80(元)
如果 H_2 赢	100 元	100＋50－30－40＝80(元)
如果 H_3 赢	0 元	50－30－40＝－20(元)

无论哪匹马赢，该赌客都损失 20 元. □

3. 主观概率与贝叶斯网

贝叶斯网早期主要应用于专家系统. 在专家系统应用中，贝叶斯网的结构和参数是通过咨询专家而获得的，因此需要用类似于概率轮的方法进行概率评估，主观概率占有重要地位.

随着时间的推移，贝叶斯网越来越多地被用于分析数据，也就是要基于数据建立贝叶斯网模型. 这有两种情形：一是已知网络结构，对网络参数进行估计，称为参数学习；二是不知道网络结构，要通过分析数据，同时获得网络结构和网络参数，称为结构学习. 参数学习有两种方法——最大似然估计和贝叶斯估计. 最大似然估计完全基于数据，不需要先验概率. 贝叶斯估计则假定在考虑数据以前，网络参数服从某个先验分布. 这是先验的主观概率，它的影响随着数据量的增大而减小. 结构学习情形假设在考虑数据以前，不同结构的可能性相等，这也是先验的主观概率，它的影响也随着数据量的增大而减小. 所以，当有足够多的数据时，主观概率对数据分析的影响不大.

尽管概率的主观解释在贝叶斯网的实际应用中并不扮演非常重要的角色，但是在概念上，它对贝叶斯网却是至关重要的. 贝叶斯网所依赖的一个核心概念是条件独立，而概率的主观解释为直观理解条件概率和条件独立提供了一个自然的角度. 这一点的论证将在第 1.3.3 和 1.3.4 节中看到.

1.2.4　特性解释与逻辑解释

在特性解释中，均匀硬币“正面朝上”的概率为 1/2 是这个硬币的固有物理属性，与其是否投掷或投掷次数无关. 特性解释没有为概率提供可操作的运算方

法，因此很难应用于实际之中.

逻辑解释则认为概率是对知识状态的总结，是由从证据到假设的逻辑关系所决定的. 一旦相关的知识得到确定，则事件的可能性就已经被客观地确定下来，并应该能够通过逻辑分析来得到. 古典解释可以看做是逻辑解释的一个特例，它从等可能性的前提条件出发来计算概率. 同特性解释一样，逻辑解释的缺点在于它没能为概率提供一个可操作的运算方法.

1.3 多元概率分布

随机现象往往涉及多个随机因素，因而需要用多个随机变量来描述. 本节介绍多元概率的一些基本概念.

1.3.1 联合概率分布

我们知道，对单个随机变量 X，可以用概率函数 $P(X)$ 来描述它的各个状态的概率. 而对于多个随机变量 $X_1, \cdots, X_n$，则可以用联合概率分布 $P(X_1, \cdots, X_n)$，简称联合分布来描述各变量所有可能的状态组合的概率. 它是一个定义在所有变量状态空间的笛卡儿乘积之上的函数：

$$P(X_1, \cdots, X_n): \bigotimes_{i=1}^{n} \Omega_{X_i} \rightarrow [0, 1],$$

其中所有函数值之和为 1，即

$$\sum_{X_1, \cdots, X_n} P(X_1, \cdots, X_n) = 1.$$

联合分布经常被表示为一张表，其中包含了 $\prod_{i=1}^{n} |\Omega_{X_i}|$ 个状态组合及其概率值. 如果所有变量都只取两个状态，则联合分布表共有 2^n 个项，刻画了变量之间的各种关系.

例 1.15 考虑香港市场上所有出租房屋.从中随机抽取一间,考查其月租(R)和类型(T)这两个随机变量.月租分为 4 等:{low(低于 2000 元), medium (2000 ～ 6000 元), upper medium (6000 ～ 12000 元), high(高于 12000 元)}. 类型有 3 种:{public(公屋), private(私家屋), others(其它)}. 联合分布 P(R, T) 如下：

R \ T	public	private	others
low	0.17	0.01	0.02
medium	0.44	0.03	0.01
upper medium	0.09	0.07	0.01
high	0	0.14	0.01

从表中可知，随机抽到中价公屋的可能性最大，为 44%. □

1.3.2　边缘概率分布

在例 1.15 中，由于有了联合分布 P(R,T)，所以可以回答这样的问题：随机抽取一间出租房屋为公屋的概率 P(T＝public) 是多少？根据概率的有限可加性

$$\begin{aligned}P(T=\text{public}) &= P(T=\text{public}, R=\text{low})\\ &\quad + P(T=\text{public}, R=\text{medium})\\ &\quad + P(T=\text{public}, R=\text{upper medium})\\ &\quad + P(T=\text{public}, R=\text{high})\\ &=0.7.\end{aligned}$$

同样地，可以计算 P(T＝private) 和 P(T＝others)：

$$\begin{aligned}P(T=\text{private}) &= P(T=\text{private}, R=\text{low})\\ &\quad + P(T=\text{private}, R=\text{medium})\\ &\quad + P(T=\text{private}, R=\text{upper medium})\\ &\quad + P(T=\text{private}, R=\text{high})\\ &=0.25,\\ P(T=\text{others}) &= P(T=\text{others}, R=\text{low})\\ &\quad + P(T=\text{others}, R=\text{medium})\\ &\quad + P(T=\text{others}, R=\text{upper medium})\\ &\quad + P(T=\text{others}, R=\text{high})\\ &=0.05.\end{aligned}$$

为了简化记号，上面三式可分别缩写为

$$P(T=\text{public})=\sum_R P(T=\text{public}, R),$$

$$P(T=\text{private})=\sum_R P(T=\text{private}, R),$$

$$P(T=\text{others})=\sum_R P(T=\text{others}, R).$$

这三个式子还可以进一步合并为下面一式：

$$P(T)=\sum_R P(T, R).$$

相对于联合分布 P(R, T)，P(T) 称为边缘分布[①]. 下表同时给出了联合分布 P(R, T) 和边缘分布 P(T)，P(R)：

① 这一术语来源于保险统计行业. 保险统计师通常把观察到的频率数据相加，并把结果写在保险统计报表的边缘，所以叫“边缘概率”.

R \ T	public	private	others	P(R)
low	0.17	0.01	0.02	0.20
medium	0.44	0.03	0.01	0.48
upper medium	0.09	0.07	0.01	0.17
high	0	0.14	0.01	0.15
P(T)	0.70	0.25	0.05	

记 $X=\{X_1,\cdots,X_n\}$，Y 是 X 的真子集，即 $Y\subset X$，$Z=X\setminus Y$. 则相对于 P(X)，Y 的边缘分布 P(Y) 定义为

$$P(Y)=\sum_{Z}P(X_1,\cdots,X_n). \tag{1.2}$$

从联合分布 P(X) 到边缘分布 P(Y) 的过程称为边缘化.

例 1.16　设有 3 个装有黑白两色球的口袋，第 1 个口袋黑白球各半，第 2 个口袋黑白球比例为 4∶1，第 3 个则全是黑球. 设随机变量 X，Y，Z 分别代表从这 3 个口袋随机抽出的球的颜色，其状态空间为 $\Omega_X=\Omega_Y=\Omega_Z=\{w,b\}$，其中 w 表示白，b 表示黑. 联合概率分布 P(X,Y,Z) 如下：

X	Y	Z	P(X,Y,Z)
w	w	w	0
w	w	b	0.1
w	b	w	0
w	b	b	0.4
b	w	w	0
b	w	b	0.1
b	b	w	0
b	b	b	0.4

表中给出了 X，Y，Z 的所有 8 个可能的状态组合及其概率. 从表中可知抽球结果为 (w，b，b) 和 (b，b，b) 的概率一样大，都是 0.4；结果为 (w，w，b) 和 (b，w，b) 的概率也一样，都是 0.1；而其余所有满足 Z＝w 的状态组合的概率都为零，因为第 3 个袋子里没有白球. 这里边缘分布 P(X) 为 (0.5，0.5)，P(Y) 为 (0.2,0.8)，P(Z) 为 (0,1). □

1.3.3　条件概率分布

条件概率与条件分布是用来刻画事件之间及变量之间关系的基本工具.

1. 条件概率

设 A，B 为两随机事件且 $P(B)>0$，事件 A 在给定事件 B 发生时的条件概率定义为

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)}. \tag{1.3}$$

直观上，$P(A\mid B)$ 是在已知 B 发生时，对 A 发生的信度；而 $P(A)$ 则是在不知道 B 是否发生时，对 A 发生的信度[①]. 由式 (1.3) 可得

$$P(A\cap B)=P(B)P(A\mid B). \tag{1.4}$$

这称为概率的乘法定律，其含意非常直观：对 A 和 B 同时发生的信度，等于对 B 发生的信度乘以已知 B 发生时对 A 发生的信度. 当然乘法定律也可以写为

$$P(A\cap B)=P(A)P(B\mid A). \tag{1.5}$$

例 1.17　在例 1.2 的掷骰子试验中，掷出 6 的概率为 1/6. 假定投掷后被告知“掷出的结果是偶数”，问此时对结果为 6 的信度是多少？设掷出 6 为事件 A，掷出结果为偶数为事件 B，则 $P(A)=1/6$，$P(B)=1/2$，$P(A\cap B)=1/6$. 所问的问题即是要计算如下条件概率：

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)}=\frac{1/6}{1/2}=\frac{1}{3}.$$

□

2. 条件分布

设 X 和 Y 是两随机变量，x 和 y 分别是它们的一个取值. 考虑事件 $X=x$ 在给定 $Y=y$ 时的条件概率为

$$P(X=x\mid Y=y)=\frac{P(X=x,\ Y=y)}{P(Y=y)}. \tag{1.6}$$

在式(1.6)中，固定 y，让 x 在 Ω_X 上变动，则得到一个在 Ω_X 上的函数. 这个函数称为在给定 $Y=y$ 时变量 X 的条件概率分布，记为 $P(X\mid Y=y)$. 用 $P(X\mid Y)$ 记 $\{P(X\mid Y=y)\mid y\in\Omega_Y\}$，即在 Y 取不同值时 X 的条件概率分布的集合. $P(X\mid Y)$ 称为给定 Y 时变量 X 的条件概率分布. 在式(1.6)中，让 x 和 y 在 Ω_X

① 这里用到的是概率的主观解释.

和 Ω_Y 上变动，则得到一组等式．与 1.3.2 节类似，把这些等式缩写为

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}. \tag{1.7}$$

式 (1.7) 可视为是 P(X | Y) 的直接定义．

更一般地，设 $X = \{X_1, \cdots, X_n\}$ 和 $Y = \{Y_1, \cdots, Y_m\}$ 为两个变量集合，P(X,Y) 为 $X \cup Y$ 的联合概率分布，P(Y) 为 Y 的边缘概率分布．则给定 Y 时 X 的条件概率分布定义为

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}. \tag{1.8}$$

例 1.18　在例 1.15 中问：随机抽取一间私家屋，其租金为 low 的概率多大？这即是问给定 T = private 时 R = low 的条件概率 P(R = low | T = private)．按定义，有

$$P(R = \text{low} \mid T = \text{private}) = \frac{P(R = \text{low}, T = \text{private})}{P(T = \text{private})} = \frac{0.01}{0.25} = 0.04.$$

给定 T 时，变量 R 的条件分布 P(R | T) 如下：

T \ R	low	medium	upper medium	high
public	$\frac{0.17}{0.7}$	$\frac{0.44}{0.7}$	$\frac{0.09}{0.7}$	$\frac{0}{0.7}$
private	$\frac{0.01}{0.25}$	$\frac{0.03}{0.25}$	$\frac{0.07}{0.25}$	$\frac{0.14}{0.25}$
others	$\frac{0.02}{0.05}$	$\frac{0.01}{0.05}$	$\frac{0.01}{0.05}$	$\frac{0.01}{0.05}$

表中第 1 行显示的是在给定 T = public 时，R 的条件概率分布，第 2 行是在给定 T = private 时，R 的条件概率分布，等等．这里每行的数字之和为 1，即 $\sum_R P(R \mid T) = 1$．这与例 1.15 中所列的联合分布 P(R, T) 不同，那里表中所有数字之和为 1，即 $\sum_{R,T} P(R, T) = 1$．□

下面的例子涉及本节前面所介绍的 3 个主要概念，即联合分布、边缘分布和条件分布．其目的是为读者建立这样的直观印象：几个随机变量的联合分布对应的是一张表，其中一些变量的边缘分布以及条件分布对应的也是表．

例 1.19　设有一袋积木，每块积木有 3 个属性：颜色、材料和形状．设积木的颜色只能是红(r)或蓝(b)两种，材料只能是金属(m)或木头(w)，形状只能是正方体(6)或正四面体(4)．设 C，M，S 为 3 个随机变量，分别代表从袋中随机取出一块积木的颜色、材料和形状，则 $\Omega_C = \{r, b\}$，$\Omega_M = \{m, w\}$，$\Omega_S =$

{6, 4}. 设联合概率分布 P(C, M, S) 为

C	M	S	P(C, M, S)
r	m	6	0.10
r	m	4	0.10
r	w	6	0.25
r	w	4	0.05
b	m	6	0.15
b	m	4	0.10
b	w	6	0.20
b	w	4	0.05

那么，变量 C 和 M 的边缘分布 P(C, M) 为

C	M	P(C, M)
r	m	0.20
r	w	0.30
b	m	0.25
b	w	0.25

条件分布 P(C | M) 为

M \ C	r	b
m	$\frac{4}{9}$	$\frac{5}{9}$
w	$\frac{6}{11}$	$\frac{5}{11}$

□

3. 链规则

对两变量 X、Y 的联合分布 P(X,Y)，按照条件分布的定义，可得

$$P(X,Y) = P(X)P(Y \mid X). \tag{1.9}$$

将其推广到 n 个变量的联合分布 $P(X_1, X_2, \cdots, X_n)$，有

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2 \mid X_1) \cdots P(X_n \mid X_1, \cdots, X_{n-1}). \tag{1.10}$$

式(1.10)将一个联合分布分解为一系列条件分布的乘积，它称为链规则.

1.3.4 边缘独立与条件独立

1. 事件独立

设 A，B 为同一随机试验的两个不同事件. 我们称事件 A 与 B 相互独立，如果下式成立：

$$P(A \cap B) = P(A)P(B), \tag{1.11}$$

当 $P(B) > 0$ 时，由式 (1.11) 可得 $P(A) = P(A \mid B)$. $P(A \mid B)$ 是已知事件 B 发生时对 A 发生的信度，而 $P(A)$ 是在未知事件 B 是否发生时对 A 发生的信度. 所以 A 与 B 相互独立的直观含义是：对于事件 B 是否发生的了解不影响对事件 A 发生的信度. 当 $P(A) > 0$ 时，由式 (1.11) 可得 $P(B) = P(B \mid A)$，所以 A 与 B 相互独立意味着：对于事件 A 是否发生的了解也不影响对事件 B 发生的信度.

考虑 3 个事件 A，B 和 C，假定 $P(C) > 0$. 我们称事件 A 与 B 在给定 C 时相互条件独立，如果下式成立：

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C), \tag{1.12}$$

当 $P(B \cap C) > 0$ 时，由式 (1.12) 可得 $P(A \mid C) = P(A \mid B \cap C)$. $P(A \mid C)$ 是已知事件 C 发生时对事件 A 发生的信度，而 $P(A \mid B \cap C)$ 是已知 事件 B 和 C 都已发生时对事件 A 发生的信度. 所以，事件 A 与 B 在给定 C 时相互条件独立的直观意义是：在已知事件 C 发生的前提下，对事件 B 是否发生的了解不会改变对事件 A 发生的信度；同样，对事件 A 是否发生的了解也不影响对事件 B 发生的信度.

2. 变量独立

两个随机变量 X 和 Y 称为相互 (边缘) 独立，记为 $X \perp Y$，如果下式成立：

$$P(X, Y) = P(X)P(Y), \tag{1.13}$$

考虑变量 Y 的某个取值 y，如果 $P(Y = y) > 0$，则由式 (1.13) 可得

$$P(X) = P(X \mid Y = y).$$

$P(X \mid Y = y)$ 是已知 $Y = y$ 时，变量 X 的概率(信度)分布，而 $P(X)$ 是未知 Y 的取值时 X 的概率(信度)分布. 所以，变量 X 与 Y 相互独立意味着：对变量 Y 的取值的了解不会改变变量 X 的概率(信度)分布；同样，对变量 X 的取值的了解也不会改变变量 Y 的概率(信度)分布.

更一般地，我们称随机变量 X_1，X_2，…，X_n 相互 (边缘) 独立，如果

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2)\cdots P(X_n).$$

例 1.20 在例 1.16 中，从 3 个袋子中抽球，所得球的颜色的联合概率分布

P(X,Y,Z) 如下：

X	Y	Z	P(X,Y,Z)
w	w	w	0
w	w	b	0.1
w	b	w	0
w	b	b	0.4
b	w	w	0
b	w	b	0.1
b	b	w	0
b	b	b	0.4

而边缘分布 P(X)，P(Y) 和 P(Z) 则分别如下：

X	w	b
P(X)	0.5	0.5

Y	w	b
P(Y)	0.2	0.8

Z	w	b
P(Z)	0	1

容易验证 $P(X,Y,Z)=P(X)P(Y)P(Z)$，即 X,Y,Z 相互边缘独立. □

考虑 3 个随机变量 X，Y 和 Z，设 $P(Z=z)>0,\forall z\in\Omega_Z$. 我们说 X 和 Y 在给定 Z 时相互条件独立，记为 $X\perp Y\mid Z$. 如果下式成立：

$$P(X,Y\mid Z)=P(X\mid Z)P(Y\mid Z). \tag{1.14}$$

设 y 和 z 分别是 Y 和 Z 的任意取值，且 $P(Y=y,Z=z)>0$，由式 (1.14) 可得

$$P(X\mid Y=y,Z=z)=P(X\mid Z=z).$$

$P(X\mid Z=z)$ 是在已知 $Z=z$ 时，X 的概率（信度）分布，而 $P(X\mid Y=y,Z=z)$ 是在已知 $Y=y$ 以及 $Z=z$ 时，X 的概率（信度）分布. 因此，$X\perp Y\mid Z$ 的直观含意是：在已知 Z 的前提下，对 Y 的取值的了解不影响 X 的概率（信度）分布. 注意，这并不意味着在未知 Z 的取值时，X 和 Y 相互独立. $Y=y$ 有可能含有关于 X 的信息，只是所有这样的信息也都包含于 $Z=z$ 中，所以当已知 $Z=z$ 时，进一步了解到 $Y=y$ 并不增加关于 X 的信息. 当然，$X\perp Y\mid Z$ 也意味着，在已知 Z 的取值时，对 X 的取值的了解不影响 Y 的概率（信度）分布.

例 1.21　设有一装有两种硬币的口袋，其中一些是均匀硬币，掷出正面朝上的概率为 0.5；另一些为非均匀硬币，掷出正面朝上的概率为 0.8. 现从袋中随机取出一个硬币，抛掷若干次. 令 X_i 表示第 i 次抛掷硬币的结果，Y 表示该硬币

是否均匀. 这里，X_i 与 X_j $(i \neq j)$ 之间不是相互（边缘）独立的，因为如果掷了 10 次硬币，其中 9 次都是正面朝上，那么有理由相信这枚硬币是不均匀的，从而增大了下一次掷出正面朝上的信度. 所以 X_i 的值给了我们关于这枚硬币的一些信息，它有助于我们继续判断 X_j 的值.

另一方面，如果已经知道了 Y 的值，例如该硬币是不均匀的，那么不管前面的结果如何，以后每次掷硬币的结果为正面的概率都是 0.8，我们将不能从前面的试验得到什么信息. 所以给定 Y 的值后，X_i 与 X_j 之间就是相互条件独立的. 本例中变量间的依赖关系可以用图 1.2 来表示：变量 Y 切断了变量 X_i 与变量 X_j 之间的“信息通道”. □

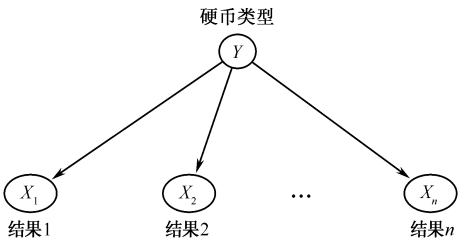

图 1.2 条件独立示意：给定硬币类型，各投掷结果相互独立

命题 1.1 考虑 3 个随机变量 X，Y 和 Z，设 $P(Z)>0$，下列条件相互等价：

(1) $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$；

(2) $P(X \mid Y, Z) = P(X \mid Z)$，当 $P(Y, Z)>0$；

(3) $P(X, Y \mid Z) = f(X, Z)g(Y, Z)$，f 和 g 均为函数；

(4) $P(X \mid Y, Z) = f(X, Z)$，f 为一函数，当 $P(Y, Z)>0$；

(5) $P(X, Y, Z) = P(X \mid Z)P(Y \mid Z)P(Z)$；

(6) $P(X, Y, Z) = \dfrac{P(X, Z)P(Y, Z)}{P(Z)}$；

(7) $P(X, Y, Z) = f(X, Z)g(Y, Z)$，f 和 g 均为函数.

1.3.5 贝叶斯定理

先验概率和后验概率这两个概念是相对于某组证据而言的. 设 H 和 E 为两个随机变量，$H = h$ 为某一假设，$E = e$ 为一组证据. 在考虑证据 $E = e$ 之前，对事件 $H = h$ 的概率估计 $P(H = h)$ 称为先验概率. 而在考虑证据之后，对 $H = h$ 的概率估计 $P(H = h \mid E = e)$ 称为后验概率. 贝叶斯定理描述了先验概率和后验概率之间的关系：

$$P(H = h \mid E = e) = \frac{P(H = h)\ P(E = e \mid H = h)}{P(E = e)}, \qquad (1.15)$$